OTTO VON GUERICKE UNIVERSITÄT MAGDEBURG

FACULTY OF MATHEMATICS

Selected Aspects of Complex Flow Problems

Modelling - Analysis - Numerics

Habilitation

zur Erlangung des akademischen Grades

doctor rerum naturalium habilitatus
(Dr. rer. nat. habil.)

von Dr. Piotr Minakowski
geb. am 16.04.1986 in Breslau, Polen

genehmigt durch die Fakultät für Mathematik
der Otto-von-Guericke-Universität Magdeburg

Gugachter: Prof. Dr. Thomas Richter
Prof. Dr. Stefan Turek
Prof. Christian Vergara

eingereicht am: 01.12.2021
Verteidigung am: 10.11.2022

Abstract

The research presented in this thesis focuses on the mathematical modelling, analysis, and development of numerical schemes to capture complex flows. We present results on selected partial and ordinary differential equation problems arising in continuum mechanics and collective dynamics.

As a result of the research of this thesis, the author worked on several research projects together with collaborators. The applications concern four main topics. We first focus on the Navier-Stokes system and fluid-structure interactions that describe blood flow and blood vessel wall mechanics. The second direction in fluid mechanics is a complex flow of the Euler system with variable congestion in an application to crowd dynamics. The research in collective dynamics revolves around a density induced consensus protocol and its relevance to data segmentation. Finally, we consider novel neural network approximations of PDEs, for which we develop dual weight residual error estimator.

The investigation of theoretical and practical aspects of blood flow simulations focuses on error estimates for the finite element approximation of elliptic partial differential equations on perturbed domains. Later, we study the impact of using a full Fluid Structure Interaction model on important hemodynamical factors by developing a stenotic coronary artery benchmark. Finally, we present patient specific simulations of blood flow in cephalic arch, and discuss their implications is clinical decision support.

The two-phase compressible/incompressible fluid model with variable congestion constraint is a model of the macroscopic motion of a crowd with individual congestion preferences that exhibits typical features of crowd dynamics. The results includes proof of the existence of weak solutions, the development of asymptotic preserving finite volume schemes, and numerical simulations.

The research in collective dynamics introduces a novel density induced consensus protocol. The communication rule is based on the following general paradigm: to influence the behavior of an individual, communication with a sufficiently dense nearby crowd is required. The proposed protocol is applied to both first-order and second-order systems. We investigate mathematical properties, as well as presenting applications to data clustering and image processing.

Finally, we present an error estimator based on the dual weighted residual method for neural network approximations of partial differential equations. It is destined to serve as an accurate and simple stopping criterion that guarantees the accuracy of the solution independently of the design of the neural network training.

All of the topics motioned above are supplemented with numerous computational examples. By means of numerical simulations, we present the characteristics of the models under consideration that validate theoretical findings.

Zusammenfassung

Diese Habilitationsschrift befasst sich mit der mathematischen Modellierung, Analyse und Entwicklung von numerischen Verfahren zur Erfassung komplexer Strömungen. Wir präsentieren Ergebnisse zu ausgewählten gewöhnlichen und partiellen Differentialgleichungen (PDG), die in der Kontinuumsmechanik und kollektiven Dynamik auftreten.

Im Rahmen der Forschung dieser Arbeit arbeitete der Autor an mehreren Forschungsprojekten mit. Die Anwendungen betreffen vier Hauptthemen. Wir konzentrieren uns zunächst auf das Navier-Stokes-System und die Fluid-Struktur-Interaktion, die den Blutfluss und die Mechanik der Blutgefäßwände beschreiben. Letzteres ist eine komplexe Strömung des Euler-Systems mit variabler Stauung, die auf die Dynamik von Menschenmengen angewandt wird. Die Forschung im Bereich der kollektiven Dynamik dreht sich um ein dichteinduziertes Konsensprotokoll und seine Bedeutung für die Datensegmentierung. Schließlich betrachten wir neuartige Approximationen neuronaler Netze von PDG, für die wir einen dual gewichteten Restfehlerschätzer entwickeln.

Wir untersuchen die theoretischen und praktischen Aspekte von Blutflusssimulationen, wobei wir uns auf Fehlerschätzungen für die Finite-Elemente-Approximation von elliptischen PDG auf gestörten Gebieten konzentrieren. Später untersuchen wir, wie sich die Verwendung eines vollständigen Fluid-Struktur-Interaktionsmodells auf wichtige hämodynamische Faktoren auswirkt, indem wir eine stenotischen Koronararterie-Benchmark entwickeln. Schließlich präsentieren wir patientenspezifische Simulationen des Blutflusses in der Kopfader und erörtern deren Bedeutung für die klinische Entscheidungsfindung.

Das zweiphasige kompressible/inkompressible Fluidmodell mit variabler Stauung ist ein Modell der makroskopischen Bewegung einer Menschenmenge mit individuellen Staupräferenzen. Die Ergebnisse umfassen den Nachweis über die Existenz schwacher Lösungen, die Entwicklung von Finite-Volumen-Verfahren und numerische Simulationen.

Die Forschung im Bereich der kollektiven Dynamik stellt ein neuartiges dichteinduziertes Konsens-Protokoll dar. Die Kommunikationsregel entspricht dem folgenden Prinzip: Um das Verhalten eines Individuums zu beeinflussen, ist die Kommunikation mit einer ausreichend großen Menschenmenge in der Nähe erforderlich. Das vorgeschlagene Protokoll wird sowohl auf Systeme erster als auch zweiter Ordnung angewendet. Wir untersuchen die mathematischen Eigenschaften und stellen Anwendungen für Datenclustering und Bildverarbeitung vor.

Schließlich präsentieren wir einen Fehlerschätzer auf Grundlage der Dual Weighted Residual Methode für die Approximation von PDG durch neuronale Netze. Er soll als genaues und einfaches Abbruchkriterium dienen, das die Genauigkeit der Lösung unabhängig von der Gestaltung des Trainings des neuronalen Netzes garantiert.

Alle oben behandelten Themen werden durch zahlreiche Rechenbeispiele ergänzt. Anhand von numerischen Simulationen werden die Eigenschaften der betrachteten Modelle aufgezeigt und die theoretischen Erkenntnisse validiert.

Contents

CHAPTER 1

Introduction and Outline of Thesis

The research presented in this thesis focuses on the mathematical modelling, analysis, and development of numerical schemes to capture complex flows. We present results on selected partial and ordinary differential equation problems arising in continuum mechanics and collective dynamics. As a result of the research of this thesis, the author worked on several research projects together with collaborators. The applications concern four main topics.

We first focus on the Navier-Stokes system and fluid-structure interactions that describe blood flow and blood vessel wall mechanics. The second direction in fluid mechanics is a complex flow of the Euler system with variable congestion in application to crowd dynamics. The research in collective dynamics revolves around a density induced consensus protocol and its relevance to data segmentation. Finally, we consider novel neural network approximations of PDEs, for which we develop dual weight residual error estimator.

The investigation of theoretical and practical aspects of blood flow simulations focuses on the study of the impact of modelling assumptions on the computational results. First, we consider how a domain related error affects the discretization error and develop error estimates for the finite element approximation of elliptic partial differential equations on perturbed domains. Moreover, we develop a stenotic coronary artery benchmark to compare typical flow configurations using Navier-Stokes in a rigid geometry with a fully coupled FSI model. The relevance of vascular elasticity is investigated and its impact on the computation of important hemodynamical factors.

The second example of a complex fluid flow is the two-phase compressible/incompressible fluid model with a variable congestion constraint. It describes the macroscopic motion of a crowd with individual congestion preferences. We prove that this system can be approximated by the fully compressible Navier-Stokes system with a singular pressure supplemented with transport equation for the congestion density. Furthermore, we develop a finite volume asymptotic preserving scheme based on a conservative system formulation, carry out two-dimensional numerical simulations, and show that the model exhibits typical crowd dynamics.

The research in collective dynamics introduces a novel density induced consensus protocol that operates within the following general paradigm: to influence the behavior of an individual, communication with a sufficiently dense nearby crowd is required. The system is an asymmetric density-induced version of the Cucker-Smale model with short-range interactions. We examine the basic mathematical properties of the system and concentrate on the presentation of interesting behaviors of the solutions. Moreover, the first-order version of the protocol is considered, for which we present applications to data clustering and image segmentation.

Finally, we explore the emerging field of (deep) neural networks applied to the numerical approximation of PDEs. In this context, we develop an error estimator based on the dual weighted residual method for neural network approximations of partial differential equations. It is destined to serve as an accurate and simple stopping criterion that guarantees the accuracy of the solution independently of the design of the neural network training.

All of these models share common features and mathematical challenges and are of great interest from the applications point of view. This is a strong motivation to investigate their mathematical structure, qualitative properties and design effective numerical solving techniques.

Below an overview of the results is given. We briefly introduce mathematical description of considered phenomena, summarise the content and refer to the publications and corresponding chapters.

Furthermore, we would like to state equal contributions made by the co-authors to the presented results and corresponding publications.

1.1 Theoretical and Practical Aspects of Blood Flow Simulations

The application of computational fluid dynamics (CFD) to blood flow is a rapidly growing field of biomedical and mathematical research. Currently the development of numerical methods in hemodynamics is evolving from a purely academic tool to aiding in clinical decision making. However, the complexity of living organisms require special mathematical approaches that provide accurate description and deal with its intrinsic variability. Hemodynamical simulations face numerous challenges that are connected with measurements, medical image segmentation, mathematical modelling and development of numerical methods.

In this direction of research we approached the topic from theoretical and application perspectives. First, we developed finite element error estimates for the approximation of elliptic problems under geometric uncertainty and showed that the uncertainty related to the domain can be a dominating factor in the finite element discretization. Second, we studied the impact of fluid-structure interactions (FSI) and its practical consequences in the blood flow simulation regime. We solve the FSI problem in the Arbitrary Lagrangian Eulerian formulation and compare the solutions to that of a Navier-Stokes flow on a prestressed domain. The solver is based on a monolithic parallel Newton multigrid framework. The result shows that the compliance of the vessel has a significant impact.

Collaborators and Publications The simulation of blood flow is a joint work with Thomas Richter (Otto von Guericke University Magdeburg) and Lukas Failer (Siemens AG). The impact of geometry perturbation and the impact of FSI in blood flow simulations is the topic of Chapter 2 and Chapter 3, respectively. Chapter 4 contains the description of a research project on cephalic arch stenosis.

Chapter 2 is published in the paper

P. Minakowski and T. Richter. 'Finite Element Error Estimates on Geometrically Perturbed Domains'. In: *Journal of Scientific Computing* 84.2 (July 2020). DOI: 10.1007/s10915-020-01285-y

We develop error estimates for the finite element approximation of elliptic partial differential equations on perturbed domains, i.e. when the computational domain does not match the real geometry. The result shows that the error related to the domain can be a dominating factor in the finite element discretization error. The main result consists of H^1- and L_2- error estimates for the Laplace problem. Theoretical considerations are validated by a computational example.

Chapter 3 is published in the paper

L. Failer, P. Minakowski and T. Richter. 'On the Impact of Fluid Structure Interaction in Blood Flow Simulations'. In: *Vietnam Journal of Mathematics* (Jan. 2021). DOI: 10.1007/s10013-020-00456-6

We study the impact of using fluid-structure interactions (FSI) to simulate the blood flow in a stenosed artery. We compare typical flow configurations using Navier-Stokes in a rigid geometry setting to a fully coupled FSI model. The relevance of vascular elasticity is investigated with respect to several questions of clinical importance. Namely, we study the effect of using FSI on the wall shear stress distribution, on the Fractional Flow Reserve and on the damping effect of a stenosis on the pressure amplitude during the pulsatile cycle. The coupled problem is described in a monolithic variational formulation based on Arbitrary Lagrangian Eulerian (ALE) coordinates. For comparison, we perform pure Navier–Stokes simulations on a pre-stressed geometry to give a good matching of both configurations. A series of numerical simulations that cover important hemodynamical factors are presented and discussed.

Chapter 4 In Chapter 4 we present research results of the project "Simulation and Optimization of Blood Flows in damaged Vessels" founded by the Federal Ministry of Education and Research of Germany, grant number 05M16NMA.

We develop a prototypical computational tool that supports clinical decision making for patients with stenosis. Blood flow is simulated as Newtonian fluid and Fluid-structure interactions on patient-specific geometry that is reconstructed from medical images.

Focus is put on the cephalic arch in patients with end-stage renal disease. These patients require a surgical arteriovenous connection for dialysis treatment. As a result, the pressure and the flow rate in the cephalic vein exceed psychological values by order of magnitude, leading to cephalic arch stenosis.

We perform virtual stenting and evaluate the effect of placing stents on flow characteristics and hemodynamical factors. Finally, we discuss the application of adapted patient-specific simulation to the evaluation of surgical treatment.

1.2 Variable Congestion in Two-Phase Compressible/Incompressible Flows

Our aim is to analyse the free-boundary two-phase fluid system that can be used to model the congestions in the large group of individuals in a bounded area. Individuals have their own preferences for how close they let the closest neighbour to approach and they carry this information with them in the course of motion. They do not follow any neighbour trying to align their velocities, nor are they trying to reach a certain target, as for example, the evacuation point. We simply prescribe their initial velocity that determines their direction of motion and check how the individual preferences as well as the initial distribution of the agents determines creation of congestions.

We study two phase compressible/incompressible Euler/Navier Stokes systems with variable congestion:

$$\begin{aligned}
\partial_t \varrho + \operatorname{div}(\varrho \mathbf{u}) = 0,\\
\partial_t(\varrho \mathbf{u}) + \operatorname{div}(\varrho \mathbf{u}\otimes \mathbf{u}) + \nabla \pi + \nabla p\left(\frac{\varrho}{\varrho^*}\right) - \operatorname{div}\mathbf{S} = \mathbf{0},\\
\partial_t \varrho^* + \mathbf{u}\cdot\nabla\varrho^* = 0,\\
0 \le \varrho \le \varrho^*,\\
\pi(\varrho^* - \varrho) = 0, \quad \pi \ge 0,
\end{aligned}$$

with the initial data

$$\varrho(0,x) = \varrho_0(x) \ge 0, \quad \mathbf{u}(0,x) = \mathbf{u}_0(x), \quad \varrho^*(0,x) = \varrho^*_0(x), \quad \varrho_0 < \varrho^*_0.$$

The unknowns are, $\varrho = \varrho(t,x)$ – the mass density, $\mathbf{u} = \mathbf{u}(t,x)$ – the velocity, $\varrho^* = \varrho^*(t,x)$ – the congestion density, and π – the congestion pressure. The barotropic pressure p is an explicit function of the density fraction $p\left(\frac{\varrho}{\varrho^*}\right) = \left(\frac{\varrho}{\varrho^*}\right)^\gamma, \quad \gamma > 1$, and plays the role of the background pressure. The stress tensor $\mathbf{S}$ is a known function of $\mathbf{u}$, characteristic for the Newtonian fluid, namely

$$\mathbf{S} = \mathbf{S}(\mathbf{u}) = 2\mu\,\boldsymbol{D}(\mathbf{u}) + \lambda \operatorname{div}\mathbf{u}\,\boldsymbol{I}, \quad \mu > 0,\ 2\mu + \lambda > 0,$$

where $\boldsymbol{D}(\mathbf{u}) = (\nabla\mathbf{u} + (\nabla\mathbf{u})^T)/2$ denotes the symmetric part of the gradient of $\mathbf{u}$, and $\boldsymbol{I} = \boldsymbol{I}_3$ is the identity matrix.

The congestion pressure π appears only when the density ϱ achieves its maximal value, the congestion density ϱ^*. Therefore ϱ^* can be referred to as the barrier or the threshold density. It was observed that the restriction on the density is equivalent to the condition

$$\operatorname{div}\mathbf{u} = 0 \text{ in } \{\varrho = \varrho^*\},$$

if $\varrho, \mathbf{u}, \varrho^*$ are sufficiently regular solutions of the continuity equation and the transport equation. For that reason, the system can be seen as a free boundary problem for the interface between the compressible (uncongested) regime $\{\varrho < \varrho^*\}$ and the incompressible (congested) regime $\{\varrho = \varrho^*\}$.

Moreover, in the inviscid case, $\mathbf{S}(\mathbf{u}) = 0$, we analyze the system numerically, i.e. we propose the numerical scheme capturing the phase transition. To this end, we use the fact that incompressibility can be obtained as a limit, when $\varepsilon \to 0$, of the compressible Euler system with the singular approximation π_ε of the congestion pressure:

$$\pi_\varepsilon\left(\frac{\varrho}{\varrho^*}\right) = \varepsilon\left(\frac{\varrho}{\varrho^* - \varrho}\right)^\alpha, \quad \alpha > 0.$$

Collaborators and Publications The results concerning congestion in two-phase compressible/incompressible flows are joint work with Pierre Degond (Imperial College London), Ewelina Zatorska (Imperial College London), and Laurent Navoret (University of Strasbourg).

The existence of weak solutions and finite volume approximation to the two-phase compressible/incompressible flows are addressed in Chapter 5 and Chapter 6, respectively.

Chapter 5 is published in the paper

P. Degond, P. Minakowski and E. Zatorska. ‘Transport of Congestion in Two-phase Compressible/Incompressible Flows’. In: *Nonlinear Analysis: Real World Applications* 42 (2018), pp. 485–510. doi: 10.1016/j.nonrwa.2018.02.001

The existence of weak solutions to the two-phase fluid model with congestion constraint was studied. The model encompasses the flow in the uncongested regime (compressible) and the congested one (incompressible) with the free boundary separating the two phases. The congested regime appears when the density in the uncongested regime achieves a threshold value that describes the comfort zone of individuals. This quantity is prescribed initially and transported along with the flow. We prove that this system can be approximated by the fully compressible Navier–Stokes system with a singular pressure, supplemented with a transport equation for the congestion density. We also present the application of this approximation for the purposes of numerical simulations in the one-dimensional domain.

Chapter 6 is published in the paper

P. Degond, P. Minakowski, L. Navoret and E. Zatorska. ‘Finite Volume Approximations of the Euler System with Variable Congestion’. In: *Computers & Fluids* 169 (2018). Recent progress in nonlinear numerical methods for time-dependent flow & transport problems, pp. 23–39. doi: 10.1016/j.compfluid.2017.09.007

We are interested in the numerical simulations of the Euler system with variable congestion encoded by a singular pressure. This model describes for instance the macroscopic motion of a crowd with individual congestion preferences. We propose an asymptotic preserving scheme based on a conservative formulation of the system in terms of density, momentum and density fraction. A second order accuracy version of the scheme is also presented. We validate the scheme on one-dimensional test-cases. We finally carry out two dimensional numerical simulations and show that the model exhibits typical crowd dynamics.

1.3 Density Induced Consensus Protocol

Many phenomena in daily life can be viewed as an interplay between ensembles of objects that communicate and interact in a non-local manner. A basic example of such a phenomenon is the flocking of birds where the objects – individual birds – interact with each other forming a larger ensemble – a flock. Such systems of interacting individuals/agents/particles are often referred to as multi-agent systems. Furthermore, when agents interact with their environment, i.e. fluids, we naturally arrive at multi-scale agents-environment framework. Such systems have received much attention due to broad range of engineering, physical and biological applications.

We introduced density-induced consensus protocol (DI protocol) for agent-based collective dynamics. The communication rule of the DI protocol reads as follows: the ith agent is influenced by jth agent if

- the density of agents in close proximity to the ith agent is substantial enough (otherwise the ith agent is an outlier),
- the jth agent is in close proximity to the ith agent.

The DI protocol operates within the following general paradigm:

To influence the behavior of an individual,
communication with a sufficiently dense nearby crowd is required.

The proposed protocol was applied to both first-order and second-order systems. The presentation starts with the former.

Consider an ensemble of N agents with $(x_i(t), v_i(t)) \in \mathbb{R}^{2d}$ denoting the position and the velocity of ith agent at the time $t \geq 0$. The agents follow the density-induced consensus protocol

$$\begin{cases} \dot{x}_i = v_i, & x_i(0) = x_{i0} \in \mathbb{R}^d, \\ \dot{v}_i = \sum\limits_{k \in \mathcal{N}_i} \kappa(v_k - v_i), & v_i(0) = v_{i0} \in \mathbb{R}^d. \end{cases}$$

Here $\kappa > 0$ is a fixed coupling strength. The neighbor set $\mathcal{N}_i$ of ith agent is defined through the following relation: given positive parameters δ and m, for $t \geq 0$ we define,

$$\begin{aligned} k \in \mathcal{N}_i(t) \;\; \Leftrightarrow \;\; & x_k(t) \in B(x_i(t), \delta) \text{ and} \\ & \#\Big\{ k \in \{1, ..., N\} : x_k(t) \in B(x_i(t), \delta) \Big\} > m, \end{aligned}$$

where $B(x_i(t), \delta)$ is an open ball centered at $x_i(t)$ with radius δ and $\#A$ denotes the cardinal number of A.

Based on the above notation, we introduce first-order variant of the DI model. Consider N agents with $x_i(t) \in \mathbb{R}^d$ denoting the position of ith agent in a d-dimensional space at the time $t \geq 0$. The agents follow the protocol

$$\dot{x}_i = \kappa \sum_{k \in \mathcal{N}_i} (x_k - x_i), \quad x_i(0) = x_{i0} \in \mathbb{R}^d.$$

The protocol is specific to real-life phenomena related to societal dynamics, where individuals do not interact with separate agents but are highly susceptible to the influence of crowds. It is inspired by such phenomena as emergence of trends in decision-making and viral videos in social media. In such a setting we view x_i and v_i as the opinion and the tendency of the individual, respectively.

It is noteworthy that the considered communication rule is equivalent to the density-based spatial clustering algorithm (DBSCAN) widely used in data classification.

Collaborators and Publications The results concerning density-induced consensus protocol are joint work with Piotr Mucha (University of Warsaw) and Jan Peszek (University of Warsaw). The first- and second-order systems are the topic of Chapter 8 and Chapter 7, respectively.

Chapter 7 is published in the paper

- P. Minakowski, P. B. Mucha and J. Peszek. ‘Density-Induced Consensus Protocol’. In: *Mathematical Models and Methods in Applied Sciences* 30.12 (July 2020), pp. 2389–2415. doi: 10.1142/s0218202520500451

We study basic qualitative and quantitative analysis of second-order density induced consensus protocol. That includes: the existence and uniqueness of classical solutions, conditional flocking estimates, and conditional cluster stability in terms of cluster density. Moreover, we provide analytical examples signifying the rich dynamics of the system. Finally, we illustrate the variability of possible behaviors by a number of numerical simulations.

Chapter 8 is based on the preprint

- P. Minakowski and J. Peszek. ‘Data Clustering as an Emergent Consensus of Autonomous Agents’. In: (2022). doi: 10.48550/ARXIV.2204.10585

We develop a data segmentation method based on a first-order density-induced consensus protocol. We provide a mathematically rigorous analysis of the consensus model leading to the stopping criteria of the data segmentation algorithm. To illustrate our method the algorithm is applied on selected images from Berkeley Segmentation Dataset.

1.4 Dual Weight Residual Error Estimates for Neural Network Solutions of Partial Differential Equations

Scientific machine learning is a rapidly evolving field of research that combines and further develops techniques of scientific computing and machine learning. This relatively new research field aims to employ machine learning techniques in order to improve current numerical solvers, e.g. significantly decrease runtime.

In recent years, the emerging field of (deep) neural networks has reached the numerical approximation of partial differential equations (PDE). Several approaches have been proposed that aim at directly representing the solution to the PDE by a deep neural network. For this purpose the network is considered as (differentiable) function $\mathscr{N} : \Omega \to \mathbb{R}^c$, where $\Omega \subset \mathbb{R}^d$ is the computational domain of dimension $d \in \mathbb{N}$ and $c \in \mathbb{N}$ is the size of the differential system.

The network is trained by integrating the differential equation (and the boundary conditions) into the loss function. Several different approaches based on different realizations have been presented:

The *Deep Ritz* method aims at minimizing the energy functional and it can be applied to symmetric problems. For the Laplace equation, $-\Delta u = f$ in Ω with $u = g$ on $\partial\Omega$, this means to minimize

$$E(\mathscr{N}) = \frac{1}{2}\int_\Omega |\nabla \mathscr{N}(x)|^2 \,\mathrm{d}x - \int_\Omega \mathscr{N}(x) \cdot f(x)\,\mathrm{d}x + \lambda \int_{\partial\Omega} |\mathscr{N}(x) - g(x)|^2 \,\mathrm{d}x,$$

with a parameter $\lambda > 0$, where the integrals are approximated by Monte-Carlo integration. *Training data* is generated by picking random integration points.

The common rationale for the such an approaches discussed above is the excellent approximation property of neural networks, in particular the capability of deep neural networks to uniformly and simultaneously approximate differential functions and their derivatives. On the other hand, it must be noted that the previous approaches, applied to common, low-dimensional ($d = 1, 2, 3$) problems, cannot compete with established methods in terms of efficiency. While algorithms of $O(N)$ complexity exist for finite element or finite difference approximations of elliptic problems, the training of the deep neural network is a by far more challenging task. Deep learning techniques can be applied in the context of numerical simulation, as an extension of existing CFD codes to increase their efficiency. One can generalize existing numerical methods as artificial neural networks with a set of trainable parameters.

Estimator We aim at estimating the functional error of the neural network solution $u_{\mathscr{N}} \in V_{\mathscr{N}}$ obtained with the *Deep Ritz* approach. Let $J : \mathscr{V} \to \mathbb{R}$ be a linear functional and let $z \in \mathscr{V}$ be the solution to the adjoint problem

$$-\Delta z = J \text{ in } \Omega, \quad z = 0 \text{ on } \partial\Omega.$$

Since the network minimizer $u_{\mathscr{N}} \notin \mathscr{V}$ does not satisfy the Dirichlet condition $u = 0$, a consistency term remains when multiplying the error $u - u_{\mathscr{N}}$ with the adjoint problem

$$J(u - u_{\mathscr{N}}) = (\nabla(u - u_{\mathscr{N}}), \nabla z) + \langle \partial_n z, u_{\mathscr{N}} \rangle.$$

From primal problem we derive the error identity

$$J(u-u_{\mathscr{N}}) = (f,z) - (\nabla u_{\mathscr{N}}, \nabla z) + \langle \partial_n z, u_{\mathscr{N}} \rangle.$$

We must approximate $z \in \mathscr{V}$ by a discrete solution which is accurate, efficiently achievable and which does not fall into the vicinity of Galerkin orthogonality, which, in the case of the neural network error $u - u_{\mathscr{N}}$ imposes the condition $z_h \notin V_{\mathscr{N}}$. The neural network spaces have no similarity to local finite element spaces. Hence, we will approximate the adjoint solution in as coarse as possible finite element space $z_h \in V_h \subset \mathscr{V}$ and approximate

$$J(u-u_{\mathscr{N}}) \approx \eta(u_{\mathscr{N}}, z_h) = (f, z_h)_\Omega - (\nabla u_{\mathscr{N}}, \nabla z_h) + \langle \partial_n z_h, u_{\mathscr{N}} \rangle.$$

This error estimator is efficiently evaluated on the finite element mesh using a numerical quadrature rule within the domain and along the boundaries.

Collaborators and Publications The application of dual weighted residual method to the error estimator of neural network approximations of PDEs is a joint work with Thomas Richter (Otto von Guericke University Magdeburg) and is the topic of Chapter 9.

Chapter 9 is published in the preprint

- P. Minakowski and T. Richter. *Error Estimates for Neural Network Solutions of Partial Differential Equations.* 2021. arXiv: 2107.11035 [math.NA]

Extended version is published in the paper

- P. Minakowski and T. Richter. ‘A priori and a posteriori error estimates for the Deep Ritz method applied to the Laplace and Stokes problem’. In: *Journal of Computational and Applied Mathematics* 421 (2023), p. 114845. issn: 0377-0427. doi: 10.1016/j.cam.2022.114845

We derive a dual weight residual estimator for a functional of interest of neural network approximations of PDEs and demonstrate its performance for the Laplace and Stokes equation. The method is independent of the design of the neural network and the training procedure. The evaluation on very coarse meshes already shows very good accuracy, such that little computational overhead is brought along. The estimator can be used as accurate and simple stopping criterion during the training process. Hereby, we gain a first validation of the neural network approximation and the error controlled training also helps to reduce the computational effort by avoiding excessive training epochs.

1.5 Additional Contributions

In addition to the previously presented works of the thesis', we would like to mention additional scientific results of the author together with collaborators. These do not represent a major contributions to the thesis. However, are complementary to conducted research.

1.5.1 Numerical Benchmarking of the Fluid-Rigid Body Interactions

The research on fluid-structure interactions presented in Chapter 3 is complementary to the benchmark studies

- H. von Wahl, T. Richter, C. Lehrenfeld, J. Heiland and P. Minakowski. 'Numerical Benchmarking of Fluid-rigid Body Interactions'. In: *Computers & Fluids* 193 (Oct. 2019), p. 104290. doi: 10.1016/j.compfluid.2019.104290
- H. von Wahl, T. Richter, C. Lehrenfeld, J. Heiland and P. Minakowski. *Numerical Benchmarking of Fluid-rigid Body Interactions.* Version v1. June 2019. doi: 10.5281/zenodo.3253455

We propose a fluid-rigid body interaction benchmark problem, consisting of a solid spherical obstacle in a Newtonian fluid, whose centre of mass is fixed but is free to rotate. A number of different problems are defined for both two and three spatial dimensions. The geometry is chosen specifically, such that the fluid-solid partition does not change over time and classical fluid solvers are able to solve the fluid-structure interaction problem. We summarise the different approaches used to handle the fluid-solid coupling and numerical methods used to solve the arising problems. The results obtained by the described methods are presented and we give reference intervals for the relevant quantities of interest.

1.5.2 Binary Mixtures

The research on the two-phase model presented in Chapter 5 and Chapter 6 is complementary to the binary mixtures studies

- M. Sykora, M. Pavelka, I. Peshkov, P. Minakowski, V. Klika and E. Romenski. 'Comparison of the Symmetric Hyperbolic Thermodynamically Compatible framework with Hamiltonian mechanics of binary mixtures'. In: (2022). doi: 10.48550/ARXIV.2201.04460

We ask the question, how to adequately describe continuum thermodynamics of binary mixtures where each of the constituents has its momentum? The Symmetric Hyperbolic Thermodynamically Consistent (SHTC) framework and Hamiltonian mechanics in the form of the General Equation for Non-Equilibrium Reversible-Irreversible Coupling (GENERIC) provides two answers, which are similar, but not identical, and are compared in this article. The comparison is made on several levels of description, varying by the amount of detail involved, both analytically and numerically. The GENERIC equations, stemming from the Liouville equation, contain terms expressing self-advection of the relative velocity by itself, which lead to a vorticity-dependent diffusion matrix after a reduction. The SHTC equations, on the other hand, do not contain such terms. We also show how to formulate a theory of mixtures with two momenta and only one temperature that is compatible with the Liouville equation and possesses the Hamiltonian structure, including Jacobi identity.

1.5.3 Singular Cucker-Smale Dynamics

The research on collective dynamics presented in Chapter 7 and Chapter 8, is complementary to the review on the singular Cucker-Smale model

P. Minakowski, P. B. Mucha, J. Peszek and E. Zatorska. ‘Singular Cucker–Smale Dynamics’. In: *Active Particles, Volume 2: Advances in Theory, Models, and Applications*. Ed. by N. Bellomo, P. Degond and E. Tadmor. Cham: Springer International Publishing, 2019, pp. 201–243. doi: 10.1007/978-3-030-20297-2_7

This chapter is dedicated to singular models of flocking. We give an overview of the existing literature starting from microscopic Cucker-Smale model with singular communication weight, through its mesoscopic mean-field limit, up to the corresponding macroscopic regime. For the microscopic CS model and its selected variants, the collision-avoidance phenomenon is discussed. For the kinetic mean-field model we sketch the existence of global-in-time measure-valued solutions, paying special attention to weak-atomic uniqueness of solutions. Ultimately, for the macroscopic singular model, we provide a summary of existence results for the Euler-type alignment system. This includes the existence of strong solutions on a one-dimensional torus, and the extension of this result to higher dimensions by restricting the size of the initial data. Additionally, we present the pressureless Navier–Stokes-type system corresponding to a particular choice of alignment kernel. This system is then compared—analytically and numerically—to the porous medium equation.

1.6 Numerical Software

The results of numerical simulations presented in this thesis are implemented using several numerical software packages.

Blood flow simulation results presented in Chapter 2 and Chapter 3 are computed in the finite element toolkit Gascoigne3D

- M. Braack, R. Becker, D. Meidner, T. Richter and B. Vexler. *The Finite Element Toolkit Gascoigne.* Version v1.01. Oct. 2021. doi: 10.5281/zenodo.5574969

Chapter 3 resulted in documented example,

- L. Failer, P. Minakowski and T. Richter. *Gascoigne Documented Example: Blood Flow and Virtual Stenting.* url: https://gascoigne.math.uni-magdeburg.de/index.php?-show=documentedexamples_05bloodflowsandvirtualstenting (visited on 01/10/2021)

The source code of the finite volume simulations in Chapter 5 and Chapter 6 is available on demand. In case of interest please send a request via email at *piotr.minakowski@ovgu.de*.

The source code for the simulations of the density induced consensus protocol has been performed in Python and is available at

- P. Minakowski. *Density-induced Consensus Protocol (source code).* Nov. 2019. doi: 10.5281/zenodo.3551842

The implementation of the dual weighed residual error estimator for neural network solutions of partial differential equations is based on GascoignePytorch, a library implementing bindings between Gascoigne3D and PyTorch [193].

- N. Margenberg, P. Minakowski and T. Richter. *GascoignePytorch.* url: https://kosinus.math.uni-magdeburg.de/gascoigne/gascoignepytorch (visited on 01/10/2021)

Simulation Videos We extend the presentation of numerical results by providing videos. They are available under hyperlinks listed in the figures caption or by scanning QR codes placed on the margins.

CHAPTER 2

Finite Element Error Estimates on Geometrically Perturbed Domains

The content of this chapter is joint work with Thomas Richter, and is published in the paper

- P. Minakowski and T. Richter. ‘Finite Element Error Estimates on Geometrically Perturbed Domains’. In: *Journal of Scientific Computing* 84.2 (July 2020). doi: 10.1007/s10915-020-01285-y

Chapter Summary. We develop error estimates for the finite element approximation of elliptic partial differential equations on perturbed domains, i.e. when the computational domain does not match the real geometry. The result shows that the error related to the domain can be a dominating factor in the finite element discretization error. The main result consists of H^1- and L_2- error estimates for the Laplace problem. Theoretical considerations are validated by a computational example.

Chapter Organisation. After introductory Section 2.1, in Section 2.2 we provide the mathematical setting and some required auxiliary results. Section 2.3 covers finite element discretization and proves the main results of this work. We illustrate our result with computational examples in Section 2.4.

Contents of Chapter

2.1 Introduction

The main aim of this work is to develop finite element (FE) error estimates in the case when there is uncertainty with respect to the computational domain. We consider the question of how a domain related error affects the finite element discretization error. We use the conforming finite element method (FEM) which is well established in the scientific computing community and allows for a rigorous analysis of the approximation error [92].

Our motivation is as follows. The steps to obtain a mesh for FE computations often come with some uncertainty, for example related to empirical measurements or image processing techniques, e.g. medical image segmentation [188, 189]. Therefore, we often perform computations on a domain which is an approximation of the real geometry, i.e., the computational domain is close to but does not match the real domain. In this work we do not specify the source of the error, but we take the error into account by explicitly using the error laden reconstructed domains.

This theoretical result is of great importance for scientific computations. Vast numbers of engineering branches rely on the results of computational fluid dynamics simulations, where there is often uncertainty connected to the computational domain. A prime example of this is computational based medical diagnostics, where shapes are reconstructed from inverse problems, such as computer tomography. The assessment of error attributed to the limited spatial resolution of magnetic resonance techniques has been discussed in [176, 175]. For a survey on computational vascular fluid dynamics, where modelling and reconstruction related issues are discussed, we refer to [207]. Error analysis of computational models is a key factor for assessing the reliability for virtual predictions.

Uncertainties in the computational domain have been studied from the numerical perspective. Rigorous bounds for elliptic problems on random domains have been derived, for approximate problems defined on a sequence of domains that is supposed to converge in the set sense to a limit domain, for both Dirichlet [15] and Neumann [14] boundary conditions. Although our techniques are similar, we consider a case where the geometrical error is not small, but where it might dominate the discretization error.

When measurement data is available the accuracy of numerical predictions can be improved by data assimilation techniques. Applications of variational data assimilation in computational hemodynamics have been reviewed in [69]. For recent developments we refer to [108] and [186]. On the other hand, the treatment of boundary uncertainty can be cast into a probabilistic framework. The domain mapping method is based entirely on stochastic mappings to transform the original deterministic/stochastic problem in a random domain into a stochastic problem in a deterministic domain, see [241, 227, 119]. The perturbation method starts with a prescribed perturbation field at the boundary of a reference configuration and uses a shape Taylor expansion with respect to this perturbation field to represent the solution [120]. In [7] and [70] a similar technique was used to incorporate random perturbations of a given domain in the context of shape optimization. Moreover, the fictitious domain approach and a polynomial chaos expansion have been applied in [49]. We note, that the probabilistic approach is beyond the scope of this work and the introduction of the boundary uncertainty as random variable increases the complexity of the problem.

The above approaches incorporate additional information on the domain reconstruction, such as measurement data or a probabilistic distribution of the approximation error. In comparison to these approaches our result can be seen as the worst case scenario. We only require that the distance between the two domains is bounded.

The analysis presented in this paper starts with well-known results regarding the finite element approximation on domains with curved boundaries. But in contrast to these estimates we cannot expect the error coming from the approximation of the geometry is small or even converging to zero. Instead we split the error into a geometric approximation error between real domain and perturbed domain and into an error coming from the finite element discretization of the problem on the perturbed domain. A central step is Lemma 2.3 which estimates the geometry perturbation. Having in mind that this error is not small and cannot be reduced by means of tuning the discretization, the typical application case is to balance both error contributions to efficiently reach the barrier of the geometry error. Theorem 2.6 gives such optimally balanced estimates that include both error contributions.

2.2 Mathematical Setting and Auxiliary Result

2.2.1 Notation

Let $\Omega \subset \mathbb{R}^d$ be a domain with dimension $d \in \{2,3\}$. By $L^2(\Omega)$ we denote the Lebesgue space of square integrable functions equipped with the norm $\|\cdot\|_\Omega$. By $H^1(\Omega)$ we denote the space of $L^2(\Omega)$ functions with first weak derivative in $L^2(\Omega)$ and by $H^m(\Omega)$ for $m \in \mathbb{N}_0$ we denote the corresponding generalizations with weak derivatives up to degree $m \in \mathbb{N}_0$. The norms in $H^m(\Omega)$ are denoted by $\|\cdot\|_{H^m(\Omega)}$. For convenience we use the notation $H^0(\Omega) := L^2(\Omega)$. By $H^1_0(\Omega)$ we denote the space of those $H^1(\Omega)$ functions that have vanishing trace on the domain's boundary $\partial\Omega$ and we use the notation $H^1_0(\Omega;\Gamma)$ if the trace only vanishes on a part of the boundary, $\Gamma \subset \partial\Omega$. Further, by $(\cdot,\cdot)_\Omega$ we denote the $L^2(\Omega)$-scalar product and $\langle\cdot,\cdot\rangle_\Gamma$ the L^2-scalar product on a $d-1$ dimensional manifold Γ, e.g. $\Gamma = \partial\Omega$. Moreover, $[\partial_n\psi]$ is the jump of the normal derivative of ψ, i.e. for $x \in \Gamma$ with normal $\vec{n}$ (that is normal to Γ) $[\partial_n\psi](x) := \lim_{h\searrow 0}\partial_n\psi(x+h\vec{n}) - \lim_{h\searrow 0}\partial_n\psi(x-h\vec{n})$.

2.2.2 Laplace Equation and Domain Perturbation

On $\Omega \subset \mathbb{R}^d$ let $f \in L^2(\Omega)$ be the given right hand side. We consider the Laplace problem with homogeneous Dirichlet boundary conditions,

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega. \tag{2.2.1}$$

The variational formulation of this problem is given by: find $u \in H^1_0(\Omega)$, such that

$$(\nabla u, \nabla\phi)_\Omega = (f,\phi)_\Omega \quad \forall\phi \in H^1_0(\Omega). \tag{2.2.2}$$

The boundary $\partial\Omega$ is supposed to have a parametrization in C^{m+2}, where $m \in \mathbb{N}_0$. Given the additional regularity $f \in H^m(\Omega)$, $H^0(\Omega) := L^2(\Omega)$, there exists a unique solution satisfying the a-priori estimate

$$\|u\|_{H^{m+2}(\Omega)} \le c\|f\|_{H^m(\Omega)}, \tag{2.2.3}$$

see e.g. [94].

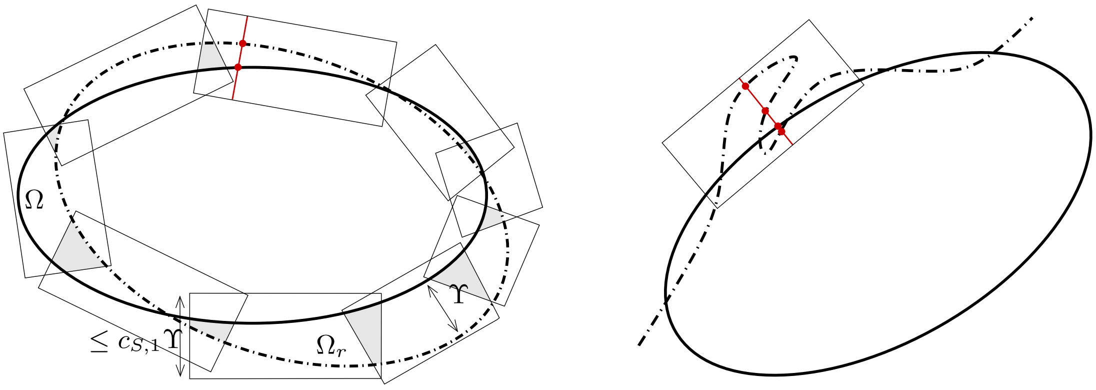


Figure 2.1: The domain Ω (bold line) and its reconstruction Ω_r (dashed). The cover of the domain remainders $S = (\Omega \setminus \Omega_r) \cup (\Omega_r \setminus \Omega)$ by a set of rectangles. The left configuration fulfils Assumption A1. The height h_i of each rectangle R_i is bounded by $c_{S,1}\Upsilon$ and the intersections of each rectangle with the domain remainder $R_i \cap S$ do not overlap excessively. The shaded areas show the overlap. The right configuration is excluded by Assumption A1.

In the following we assume that the *real domain* Ω is not exactly known but only given up to an uncertainty. We hence define a second domain, the *reconstructed domain* Ω_r with a boundary that allows for C^{m+2} parametrization. The Hausdorff distance between both domains is then denoted by $\Upsilon \in \mathbb{R}$,

$$\Upsilon := \operatorname{dist}(\partial\Omega, \partial\Omega_r) := \max\{ \sup_{x\in\partial\Omega} \inf_{y\in\partial\Omega_r} |x-y|, \sup_{y\in\partial\Omega_r} \inf_{x\in\partial\Omega} |x-y| \}.$$

This distance Υ is not necessarily small. When it comes to spatial discretization we will be interested in both cases, $h \ll \Upsilon$ as well as $\Upsilon \ll h$, where $h > 0$ is the mesh size. The two domains do not match and either domain can protrude from the other, see Figure 2.1. In order to prove our error estimates we require the following technical assumption on the relation between the two domains Ω and Ω_r.

Assumption A1 (*Domains*). Let Ω and Ω_r be two domains with $\Omega \cap \Omega_r \neq \emptyset$ and with Hausdorff distance $\Upsilon \in \mathbb{R}$. Both boundaries allow for a local C^{m+2} parametrization, $m \in \mathbb{N}_0$. Let

$$S := (\Omega_r \setminus \Omega) \cup (\Omega \setminus \Omega_r).$$

We assume that there exists a cover of S by a finite number of open rectangles (or rectangular cuboids) $\{R_1, \dots, R_{n(S)}\}$. Each rectangle R_i is given as translation and rotation of $(0, h_i) \times (0, t_i)$ for $d = 2$, or $(0, h_i) \times (0, t_i^1) \times (0, t_i^2)$ for $d = 3$, where the height h_i is bounded by $h_i \leq c_{S,1}\,\Upsilon$ with a constant $c_{S,1} \geq 0$. Following conditions hold:

A1.a) On each rectangle R, the boundary lines $\partial\Omega \cap R$ and $\partial\Omega_r \cap R$ allow for unique parametrizations $g_\Omega^R(t)$ and $g_{\Omega_r}^R(t)$ over the base t, or (t^1, t^2) for $d = 3$, respectively.

A1.b) The area of the cover is bounded by the area of the remainder S, i.e.

$$\Big| \bigcup_{i=1}^{n(S)} R_i \cap S \Big| \leq c_{S,2} |S|,$$

where $c_{S,2} > 0$ is a constant.

For the following we set $c_S := \max\{c_{S,1}, c_{S,2}\}$.

Figure 2.1 shows such a cover for different domain remainders. From Assumption A1 we deduce that each line through the height of the rectangle (marked in red in the figure) cuts each of the two boundaries exactly one time. The second assumption limits the overlap of the rectangles. These are shown as the shaded in the left sketch in Figure 2.1. Both assumptions on the domain are required for the proof of Lemma 2.2 that is based on Fubini's integral theorem. A more flexible framework that allows for a wider variety of domains, e.g. with boundaries that feature hooks, could be based on the construction of a map between two boundary segments on $\partial\Omega_r$ and $\partial\Omega$. Such approaches play an important role in isogeometric analysis. We refer to [243] and [244] for examples on the construction of such maps.

To formulate the Laplace equation on the reconstructed domain Ω_r we must face the technical difficulty that the right hand side $f \in H^m(\Omega)$ is not necessarily defined on Ω_r. We therefore weaken the assumptions on the right hand side.

Assumption A2 (*Right hand side*). Let $f \in H^m_{\text{loc}}(\mathbb{R}^d)$, i.e. $f \in H^m(G)$ for each compact subset $G \subset \mathbb{R}^d$. In addition we assume that the right hand side on Ω_r can be bounded by the right hand side on Ω, i.e.

$$\|f\|_{H^m(\Omega_r)} \le c\|f\|_{H^m(\Omega)}. \tag{2.2.4}$$

An alternative would be to use Sobolev extension theorems to extend functions $f \in H^m(\Omega)$ from Ω to Ω_r, see [45].

On Ω_r we define the solution $u_r \in H^1_0(\Omega_r)$ to the perturbed Laplace problem

$$(\nabla u_r, \nabla\phi_r)_{\Omega_r} = (f, \phi_r)_{\Omega_r} \quad \forall \phi_r \in H^1_0(\Omega_r), \tag{2.2.5}$$

The unique solution to (2.2.5) satisfies the bound

$$\|u_r\|_{H^{m+2}(\Omega_r)} \le c\|f\|_{H^m(\Omega_r)} \le c\|f\|_{H^m(\Omega)}. \tag{2.2.6}$$

Remark 2.1 (Extension of the solutions). A difficulty for deriving error estimates is that u is defined on Ω and u_r on $\Omega_r \neq \Omega$. Since the domains do not match, u may not be defined on all of Ω_r and vice versa. To give the expression $u - u_r$ a meaning on all domains we extend both solutions by zero outside their defining domains, i.e. $u := 0$ on $\Omega_r \setminus \Omega$ and $u_r := 0$ on $\Omega \setminus \Omega_r$. Globally, both functions still have the regularity $u, u_r \in H^1(\Omega \cup \Omega_r)$. We will use the same notation for discrete functions $u_h \in V_h$ defined on a mesh Ω_h and extend them by zero to $\mathbb{R}^d$. □

The following preliminary results are necessary in the proof of the main estimates. They can be considered as variants of the trace inequality and of Poincaré's estimate, respectively.

Lemma 2.2. *Let $\gamma \in \mathbb{R}, \gamma > 0$, $V \subset \mathbb{R}^d$ and $W \subset \mathbb{R}^d$ for $d \in \{2,3\}$ be two domains with boundaries ∂V and ∂W that satisfy Assumption A1 with distance*

$$\gamma := \operatorname{dist}(\partial V, \partial W).$$

For $\psi \in C^1(V) \cap C(\bar{V})$ it holds

$$\begin{aligned} \|\psi\|_{\partial W \cap V} &\le c\Big(\|\psi\|_{\partial V} + \gamma^{\frac{1}{2}}\|\nabla\psi\|_{V\setminus W}\Big), \\ \|\psi\|_{V\setminus W} &\le c\gamma^{\frac{1}{2}}\Big(\|\psi\|_{\partial V} + \gamma^{\frac{1}{2}}\|\nabla\psi\|_{V\setminus W}\Big), \end{aligned} \tag{2.2.7}$$

where the constants $c > 0$ depend on c_S from Assumption A1 and the curvature of the domain boundaries.

Proof. Let R be one rectangle of the cover and let $x_{\partial W} = g_W^R(t) \in \partial(W \cap V) \cap R$, see Figure 2.2. By $x_{\partial V} = g_V^R(t) \in \partial(W \cap V) \cap R$ we denote the corresponding unique point on $\partial W \cap R$. The connecting line segment $\overline{x_{\partial V} x_{\partial W}}$ completely runs through $V \setminus W$, as, if the line would leave this remainder, it would cut each line more than once which opposes Assumption A1.b). Integrating the function ψ along this line gives

$$\Big|\psi(x_{\partial W})\Big|^2 \leq 2\Big|\psi(x_{\partial V})\Big|^2 + 2\Big|\int_{x_{\partial V}}^{x_{\partial W}} \psi'(s)\,\mathrm{d}s\Big|^2.$$

Applying Hölder's inequality to the second term on the right hand side, with the length of the line bounded by $c_S\gamma$, we obtain

$$|\psi(x_{\partial W})|^2 \leq 2|\psi(x_{\partial V})|^2 + 2c_S\gamma \int_{x_{\partial V}}^{x_{\partial W}} |\nabla\psi(s)|^2\,\mathrm{d}s.$$

Using the parametrizations $x_{\partial W} = g_W^R(t)$ and $x_{\partial V} = g_V^R(t)$ we integrate over t which gives

$$\int |\psi(g_W^R(t))|^2\,\mathrm{d}t \leq \int |\psi(g_V^R(t))|^2\,\mathrm{d}t + 2c_S\gamma \int \int_{g_V^R(t)}^{g_W^R(t)} |\nabla\psi(s)|^2\,\mathrm{d}s\,\mathrm{d}t. \tag{2.2.8}$$

The volume integral on the right hand side is exactly the integral over $R\cap(V\setminus W)$. The boundary integrals can be interpreted as path integrals and therefore be estimated by

$$\begin{aligned}\frac{1}{\max_t\{1+|\nabla g_W^R(t)|^2\}} &\int_{(\partial W\cap V)\cap R} |\psi|^2\,\mathrm{d}s \\ &\leq \frac{1}{\min_t\{1+|\nabla g_V^R(t)|^2\}} \int_{R\cap\partial V} |\psi|^2\,\mathrm{d}s + 2c_S\gamma\int_{R\cap(V\setminus W)} |\nabla\psi|^2\,\mathrm{d}x.\end{aligned} \tag{2.2.9}$$

As the boundaries allow for a C^2 parametrization, we estimate

$$\|\psi\|^2_{(\partial W\cap V)\cap R} \leq c(\partial V, \partial W)c_S\Big(\|\psi\|^2_{R\cap\partial V} + \gamma\|\nabla\psi\|^2_{R\cap S}\Big). \tag{2.2.10}$$

Summation over all rectangles and estimation of all overlaps by means of Assumption A1 gives

$$\|\psi\|^2_{\partial W\cap V} \leq c(\partial V, \partial W)c_S\Big(\|\psi\|^2_{\partial V} + \gamma\|\nabla\psi\|^2_{V\setminus W}\Big).$$

For $\psi \in H_0^1(V)$, the term on ∂V vanishes.

To show the second estimate on $V \setminus W$ we again pick one rectangle R and consider a point $x \in V \setminus W$ on the line connecting $x_{\partial V} = g_V^R(t)$ and $x_{\partial W} = g_W^R(t)$ such that we introduce the notation $x = x(t, s)$. By the same arguments as above it holds

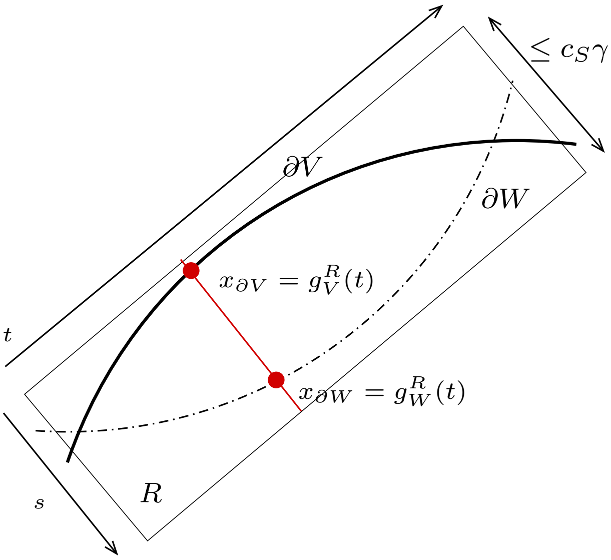
$$|\psi(x(t,s))|^2 \leq 2|\psi(g_V^R(t))|^2 + 2c_S\gamma \int \int_{x(t)}^{g_V^R(t)} |\nabla\psi|^2\,\mathrm{d}s\mathrm{d}t.$$

We integrate over s and t to obtain

$$\|\psi(x)\|^2_{R\cap(V\setminus W)} \leq \frac{2}{\min_t\{1+|\nabla g_V^R(t)|^2\}}\gamma\|\psi\|^2_{R\cap\partial V} + 2c_S\gamma^2\|\nabla\psi\|^2_{R\cap(V\setminus W)}.$$

Summing over all rectangles gives the desired result.

Figure 2.2: Illustration of the proof of Lemma 2.2. Points in each rectangle R can be represented by local coordinates (t, s), where $0 \leq s \leq c_S\gamma$ and the range of t depends on the size of the rectangle. The two boundary segments are given by the (smooth) parametrizations $g_V^R(t)$ and $g_W^R(t)$. Each line in the direction of s cuts both boundaries exactly once. In the 3d setting, the base is represented by two coordinates $\mathbf{t} = (t_1, t_2)$.



The above lemma is later used in such a way that V and W can be substituted as both Ω and Ω_r, specifically to the case of use.

We continue by estimating the difference between the solutions of the Laplace equations on Ω and on Ω_r.

Lemma 2.3. *Let $\Omega, \Omega_r \in \mathbb{R}^d$ with $\partial\Omega, \partial\Omega_r \in C^{m+2}$ satisfying $\operatorname{dist}(\partial\Omega, \partial\Omega_r) < \Upsilon$ as well as Assumption A1. Furthermore, let $f \in L^2_{loc}(\mathbb{R}^d)$ satisfy Assumption A2 and let $f_r := f|_{\Omega_r}$. For the solutions $u \in H_0^1(\Omega) \cap H^2(\Omega)$ and $u_r \in H_0^1(\Omega_r) \cap H^2(\Omega_r)$ to (2.2.2) and (2.2.5) respectively, it holds that*

$$\|u - u_r\|_\Omega + \Upsilon^{\frac{1}{2}}\|\nabla(u - u_r)\|_\Omega \leq c\Upsilon\|f\|_{\Omega\cup\Omega_r}.$$

Proof. (i) We continuously extend u and u_r by zero to $\mathbb{R}^d$, c.f. Remark 2.1, such that $u - u_r \in H^1(\Omega \cup \Omega_r)$ is well defined. We separate the domains of integration and integrate by parts

$$\begin{aligned} \|\nabla(u - u_r)\|_\Omega^2 =& \Big(\nabla(u - u_r), \nabla(u - u_r)\Big)_{\Omega\cap\Omega_r} + \Big(\nabla(u - u_r), \nabla(u - u_r)\Big)_{\Omega\setminus\Omega_r} \\ =& - \Big(\Delta(u - u_r), u - u_r\Big)_\Omega + \langle\partial_n(u - u_r), u - u_r\rangle_{\partial(\Omega\cap\Omega_r)} \\ &+ \langle\partial_n(u - u_r), u - u_r\rangle_{\partial(\Omega\setminus\Omega_r)}. \end{aligned} \tag{2.2.11}$$

Combining the boundary terms on the right-hand side of (2.2.11), into an integral over $\partial\Omega$ and a jump term over $\partial\Omega_r \cap \Omega$, we obtain

$$\begin{aligned} \|\nabla(u - u_r)\|_\Omega^2 = & - \Big(\Delta(u - u_r), u - u_r\Big)_\Omega + \langle\partial_n(u - u_r), u - u_r\rangle_{\partial\Omega} \\ & + \langle[\partial_n(u - u_r)], u - u_r\rangle_{\partial\Omega_r\cap\Omega}. \end{aligned} \tag{2.2.12}$$

In $\Omega \cap \Omega_r$ it holds $f = f_r$ and hence (weakly) $-\Delta(u - u_r) = 0$, such that

$$-(\Delta(u - u_r), u - u_r)_\Omega = -(\Delta(u - u_r), u - u_r)_{\Omega\cap\Omega_r} - (\Delta u, u)_{\Omega\setminus\Omega_r} = (f, u)_{\Omega\setminus\Omega_r}. \tag{2.2.13}$$

On $\partial\Omega$ it holds $u = 0$ and on $\partial\Omega_r \cap \Omega$ it holds $u_r = 0$. Further, since $u \in H^2(\Omega)$ it holds that $[\partial_n u] = 0$ on $\partial\Omega_r \cap \Omega$. Finally, $u_r = 0$ on $\Omega \setminus \Omega_r$, such that the boundary terms reduce to

$$\begin{aligned}\langle \partial_n(u - u_r), u - u_r\rangle_{\partial\Omega} &+ \langle [\partial_n(u - u_r)], u - u_r\rangle_{\partial\Omega_r\cap\Omega} \\ &= -\langle \partial_n(u - u_r), u_r\rangle_{\partial\Omega\cap\Omega_r} - \langle \partial_n u_r, u\rangle_{\partial\Omega_r\cap\Omega}.\end{aligned} \tag{2.2.14}$$

Combining (2.2.12)-(2.2.14) and using the Cauchy-Schwarz inequality, we estimate

$$\begin{aligned}\|\nabla(u - u_r)\|_\Omega^2 \le \|f\|_{\Omega\setminus\Omega_r}\|u\|_{\Omega\setminus\Omega_r} \\ + \|\partial_n(u - u_r)\|_{\partial\Omega\cap\Omega_r}\|u_r\|_{\partial\Omega\cap\Omega_r} + \|\partial_n u\|_{\partial\Omega_r\cap\Omega}\|u\|_{\partial\Omega_r\cap\Omega}.\end{aligned} \tag{2.2.15}$$

Since $u, u_r \in H^2(\Omega \cap \Omega_r)$, the trace inequality gives

$$\begin{aligned}\|\nabla(u - u_r)\|_\Omega^2 \le \|f\|_{\Omega\setminus\Omega_r}\|u\|_{\Omega\setminus\Omega_r} \\ + c\Big(\|u\|_{H^2(\Omega)} + \|u_r\|_{H^2(\Omega_r)}\Big)\Big(\|u_r\|_{\partial\Omega\cap\Omega_r} + \|u\|_{\partial\Omega_r\cap\Omega}\Big).\end{aligned} \tag{2.2.16}$$

Applying Lemma 2.2 twice: to $\psi = u$ and to $\psi = \nabla u$ (same for u_r), and extending the norms from $\Omega \setminus \Omega_r$ to Ω and from $\Omega_r \setminus \Omega$ to Ω_r give the bounds

$$\begin{aligned}\|u\|_{\partial\Omega_r\cap\Omega} &\le c\Upsilon^{\frac12}\|\nabla u\|_{\Omega\setminus\Omega_r} \le c\Upsilon\Big(\|\nabla u\|_{\partial\Omega} + \Upsilon^{\frac12}\|u\|_{H^2(\Omega)}\Big), \\ \|u_r\|_{\partial\Omega\cap\Omega_r} &\le c\Upsilon^{\frac12}\|\nabla u_r\|_{\Omega_r\setminus\Omega} \le c\Upsilon\Big(\|\nabla u_r\|_{\partial\Omega_r} + \Upsilon^{\frac12}\|u_r\|_{H^2(\Omega_r)}\Big).\end{aligned} \tag{2.2.17}$$

With the trace inequality and the a priori estimates $\|u\|_{H^2(\Omega)} \le c\|f\|_\Omega$ and $\|u_r\|_{H^2(\Omega_r)} \le c\|f_r\|_{\Omega_r} \le c\|f\|_\Omega$ we obtain the bounds

$$\|u\|_{\partial\Omega_r\cap\Omega} \le c\Upsilon\|f\|_\Omega, \quad \|u_r\|_{\partial\Omega\cap\Omega_r} \le c\Upsilon\|f\|_\Omega. \tag{2.2.18}$$

Using the fact that $u = 0$ on $\partial\Omega$ we apply (2.2.7) twice and use the trace inequality to get the estimate

$$\|u\|_{\Omega\setminus\Omega_r} \le c\Upsilon\|\nabla u\|_{\Omega\setminus\Omega_r} \le c\Upsilon^{\frac32}\Big(\|u\|_{H^2(\Omega)} + \Upsilon^{\frac12}\|u\|_{H^2(\Omega)}\Big) \le c\Upsilon^{\frac32}\|f\|_\Omega. \tag{2.2.19}$$

We can then estimate $\|f\|_{\Omega\setminus\Omega_r} \le \|f\|_\Omega$ by extending to the complete domain. Combining (2.2.16) with (2.2.18) and (2.2.19) we obtain the estimate

$$\|\nabla(u - u_r)\|_\Omega^2 \le c\Big(\Upsilon^{\frac32} + \Upsilon\Big)\|f\|_\Omega^2,$$

which concludes the H^1-norm bound.

(ii) For the L^2-estimate we introduce the adjoint problem

$$z \in H_0^1(\Omega): \quad -\Delta z = \frac{u - u_r}{\|u - u_r\|_\Omega} \text{ in } \Omega,$$

which allows for a unique solution satisfying the a-priori bound $\|z\|_{H^2(\Omega)} \leq c_s$ with the stability constant $c_s < \infty$. Testing with $u - u_r$ and integrating by parts twice gives

$$\begin{aligned}\|u-u_r\|_\Omega =& -(z,\Delta(u-u_r))_\Omega + \langle z, \partial_n(u-u_r)\rangle_{\partial\Omega} \\ &+ \langle z, [\partial_n(u-u_r)]\rangle_{\partial\Omega_r\cap\Omega} - \langle \partial_n z, u-u_r\rangle_{\partial\Omega}.\end{aligned}$$

It holds $z = 0$ and $u = 0$ on $\partial\Omega$, $[\partial_n u] = 0$ on $\partial\Omega_r \cap \Omega$ and $-\Delta(u-u_r) = 0$ in $\Omega\cap\Omega_r$ such that we get

$$\begin{aligned}\|u-u_r\|_\Omega &= (z,f)_{\Omega\setminus\Omega_r} - \langle z, \partial_n u_r\rangle_{\partial\Omega_r\cap\Omega} + \langle \partial_n z, u_r\rangle_{\partial\Omega} \\ &\leq \|z\|_{\Omega\setminus\Omega_r}\|f\|_{\Omega\setminus\Omega_r} + \|z\|_{\partial\Omega_r\cap\Omega}\|\partial_n u_r\|_{\partial\Omega_r\cap\Omega} + \|\partial_n z\|_{\partial\Omega}\|u_r\|_{\partial\Omega}.\end{aligned}$$

The boundary terms $\|z\|_{\partial\Omega_r\cap\Omega}$ and $\|u_r\|_{\partial\Omega}$ are estimated with Lemma 2.2, the normal derivatives by the trace inequality and the terms on $\Omega\setminus\Omega_r$ by (2.2.7)

$$\|u-u_r\|_\Omega \leq c\Upsilon^{\frac{3}{2}}\|z\|_{H^2(\Omega)}\|f\|_\Omega + c\Upsilon\|z\|_{H^2(\Omega)}\|u_r\|_{H^2(\Omega_r)} + c\Upsilon\|z\|_{H^2(\Omega)}\|u_r\|_{H^2(\Omega_r)}.$$

The L^2-norm estimate follows by using the bounds $\|u\|_{H^2(\Omega)} \leq c\|f\|_\Omega$, $\|u_r\|_{H^2(\Omega_r)} \leq c\|f\|_\Omega$ and $\|z\|_{H^2(\Omega)} \leq c$. □

Remark 2.4. The estimate $\|f\|_{\Omega\setminus\Omega_r} \leq c\|f\|_\Omega$ is not optimal. Further powers of Υ are easily generated at the cost of a higher right hand side regularity. Also, the estimate $\|\partial_n(u-u_r)\| \leq c(\|u\|_{H^2(\Omega)} + \|u_r\|_{H^2(\Omega_r)})$ by Cauchy Schwarz and the trace inequality could be enhanced to produce powers of Υ. The limiting term in (2.2.12) however is the boundary integral $|\langle\partial_n u_r, u\rangle_{\partial\Omega_r\cap\Omega}| = \mathscr{O}(\Upsilon^{\frac{1}{2}})$ which is optimal in the H^1-estimate. In Remark 2.8 and Corollary 2.9 we present an estimate that focuses on the intersection $\Omega\cap\Omega_r$ only and that allows us to improve the order to $\mathscr{O}(\Upsilon)$ in the H^1-case by avoiding exactly this boundary integral. □

2.3 Discretization

The starting point of a finite element discretization is the mesh of the domain Ω. In our setting we do not mesh Ω directly, because the domain Ω is not exactly known. Instead, we consider a mesh of the reconstructed domain Ω_r.

We partition Ω_r into a parametric triangulation Ω_h, consisting of open elements $T \subset \mathbb{R}^d$. Each element $T \in \Omega_h$ stems from a unique reference element $\widehat{T}$ which is a simple geometric structure such as a triangle, quadrilateral or tetrahedron. The numerical examples in Section 2.4 are based on quadrilateral meshes. The map $T_T : \widehat{T} \to T$ is a polynomial of degree $r \in \mathbb{N}$. We will consider iso-parametric finite element spaces, that are based on polynomials of the same degree $r \geq 1$. In the following we assume structural and shape regularity of the mesh such that standard interpolation estimates

$$\begin{aligned}\|\nabla^k(u_r - I_h u_r)\|_T &\leq ch^{r+1-k}\|u_r\|_{H^{r+1}(T)}, \quad k = 0,\dots,r \leq m, \\ \|\nabla^k(u_r - I_h u_r)\|_{\partial T} &\leq ch^{r+\frac{1}{2}-k}\|u_r\|_{H^{r+1}(T)}, \quad k = 0,\dots,r \leq m,\end{aligned} \tag{2.3.1}$$

will hold for all elements $T \in \Omega_h$, c.f. [30]. The discretization parameter h represents the size of the largest element in the mesh. See [209, Section 4.2.2] for a detailed description.

On the reference element $\widehat{T}$ let $\widehat{P}$ be a polynomial space of degree r, e.g.

$$\widehat{P} \mathrel{\widehat{=}} Q^r := \operatorname{span}\{x_1^{\alpha_1} \cdots x_d^{\alpha_d} \;:\; 0 \leq \alpha_1, \dots, \alpha_d \leq r\}$$

on quadrilateral and hexahedral meshes. Then, the finite element space V_h^r on the mesh Ω_h is defined as

$$V_h^r = \{\phi_h \in C(\bar{\Omega}_h) \;:\; \phi_h \circ T_T \in \widehat{P} \text{ on every } T \in \Omega_h\}.$$

This parametric finite element space does not exactly match the domain Ω_r. Given an iso-parametric mapping of degree r it holds $\operatorname{dist}(\partial\Omega_r, \partial\Omega_h) = \mathscr{O}(h^{r+1})$ and finite element approximation error and geometry approximation error are balanced. Iso-parametric finite elements for the approximation on domains with curved boundaries are well established [91], optimal interpolation and finite element error estimates have been presented in [61, Section 4.4]. The case of higher order elements with optimal order energy norm estimates is covered in [148]. From [209, Theorem 4.37] we cite the following approximation result for the iso-parametric approximation of the Laplace equation that also covers the L^2-error and which is formulated in a similar notation.

Theorem 2.5. *Let $m \in \mathbb{N}_0$ and let Ω_r be a domain with a boundary that allows for a parametrization of degree $m+2$. Let $f_r \in H^m(\Omega_r)$ and $u_h \in V_h^r \cap H_0^1(\Omega_h)$ be the iso-parametric finite element discretization of degree $1 \leq r \leq m+1$*

$$(\nabla u_h, \nabla \phi_h)_{\Omega_h} = (f_r, \phi_h)_{\Omega_h} \quad \forall \phi_h \in V_h^r.$$

It holds

$$\|u_r - u_h\|_{H^1(\Omega_r)} \leq c h^r \|f_r\|_{H^{r-1}(\Omega_r)}, \quad \|u_r - u_h\|_{\Omega_r} \leq c h^{r+1} \|f_r\|_{H^{r-1}(\Omega_r)}.$$

We formulated the error estimate on the domain Ω_r although the finite element functions are given on Ω_h only. To give Theorem 2.5 meaning, we consider all functions extended by zero as described in Remark 2.1. Combining these preliminary results directly yields the a priori error estimates.

Theorem 2.6. *Let $m \in \mathbb{N}_0$, Ω and Ω_r be domains with C^{m+2} boundary, distance Υ and that satisfy Assumption A1. Let Ω_h be the iso-parametric mesh of Ω_r with degree $1 \leq r \leq m+1$ and let $f \in H_{loc}^{r-1}(\mathbb{R}^d)$ satisfy Assumption A2. For the finite element error between the fully discrete solution $u_h \in V_h^r$*

$$(\nabla u_h, \nabla \phi_h)_{\Omega_h} = (f, \phi_h)_{\Omega_h} \quad \forall \phi_h \in V_h^r$$

and the true solution *$u \in H_0^1(\Omega) \cap H^{m+2}(\Omega)$ it holds*

$$\|u - u_h\|_{H^1(\Omega)} \leq c\Big(\Upsilon^{\frac{1}{2}} + h^r\Big)\|f\|_{H^{r-1}(\Omega)},$$

as well as

$$\|u - u_h\|_{\Omega} \leq c(\Upsilon + h^{r+1})\|f\|_{H^{r-1}(\Omega)}.$$

Proof. *(i)* We start with the H^1 error. Inserting $\pm u_r$ and extending the finite element error $u_r - u_h$ from Ω to Ω_r, where a small remainder appears, we have

$$\|\nabla(u - u_h)\|_{\Omega}^2 \leq 2\Big(\|\nabla(u - u_r)\|_{\Omega}^2 + \|\nabla(u_r - u_h)\|_{\Omega_r}^2 + \|\nabla(u_r - u_h)\|_{\Omega\setminus\Omega_r}^2\Big). \tag{2.3.2}$$

The first and the second term on the right hand side are estimated by Lemma 2.3 and Theorem 2.5 and, since $u_r = 0$ on $\Omega \setminus \Omega_r$, we obtain

$$\|\nabla(u-u_h)\|_\Omega^2 \le c\Big(\Upsilon + h^{2r}\Big)\|f\|_{H^{r-1}(\Omega)}^2 + 2\|\nabla u_h\|_{\Omega\setminus\Omega_r}^2. \tag{2.3.3}$$

We continue with the remainder ∇u_h on $\Omega \setminus \Omega_r$, which is non-zero on Ω_h only

$$\|\nabla u_h\|_{\Omega\setminus\Omega_r}^2 = \|\nabla u_h\|_{(\Omega_h\setminus\Omega_r)\cap(\Omega\setminus\Omega_r)}^2.$$

This remaining stripe has the width

$$\gamma_{h,\Upsilon} \coloneqq \mathscr{O}(\min\{h^{r+1}, \Upsilon\}),$$

and we apply Lemma 2.2 to get

$$\|\nabla u_h\|_{(\Omega\setminus\Omega_r)\cap(\Omega_h\setminus\Omega_r)}^2 \le c\gamma_{h,\Upsilon}\|\nabla u_h\|_{\partial\Omega_r}^2 + c\gamma_{h,\Upsilon}^2\|\nabla^2 u_h\|_{\Omega_h\setminus\Omega_r}^2, \tag{2.3.4}$$

where the second derivative $\nabla^2 u_h$ is understood element wise. This term is extended to Ω_h and with the inverse estimate and the a priori estimate for the discrete solution we obtain with $\gamma_{h,\Upsilon}^2 = \mathscr{O}(h^{2r+2})$ that

$$\gamma_{h,\Upsilon}^2\|\nabla^2 u_h\|_{\Omega_h\setminus\Omega_r}^2 \le c_{inv}\gamma_{h,\Upsilon}^2\, h^{-2}\|\nabla u_h\|_{\Omega_h}^2 \le c_{inv}h^{2r}\|f\|_{\Omega_h}^2 \le ch^{2r}\|f\|_\Omega^2. \tag{2.3.5}$$

To the first term on the right hand side of (2.3.4) we add $\pm u_r$ and $\pm I_h u_r$, the nodal interpolation of u_r into the finite element space

$$\gamma_{h,\Upsilon}\|\nabla u_h\|_{\partial\Omega_r}^2 \le c\gamma_{h,\Upsilon}\Big(\|\nabla u_r\|_{\partial\Omega_r}^2 + \|\nabla(u_r - I_h u_r)\|_{\partial\Omega_r}^2 + \|\nabla(u_h - I_h u_r)\|_{\partial\Omega_r}^2\Big). \tag{2.3.6}$$

Here, the first and last terms are estimated with the trace inequalities and, in the case of the discrete term with the inverse inequality[1], followed by adding $\pm u_r$ we get

$$\begin{aligned}\gamma_{h,\Upsilon}\|\nabla u_h\|_{\partial\Omega_r}^2 \le c\Upsilon\|f\|_{L^2(\Omega)}^2 &+ ch^{r+1}\|\nabla(u_r - I_h u_r)\|_{\partial\Omega_r}^2 \\ &+ ch^r\|\nabla(u_r - I_h u_r)\|_{\Omega_r}^2 + h^r\|\nabla(u_h - u_r)\|_{\Omega_r}^2.\end{aligned} \tag{2.3.7}$$

We used both $\gamma_{h,\Upsilon} = \mathscr{O}(h^{r+1})$ and $\gamma_{h,\Upsilon} = \mathscr{O}(\Upsilon)$. Then, collecting all terms in (2.3.3)-(2.3.7) and using the interpolation estimates as well as Theorem 2.5 we finally get

$$\|\nabla(u-u_h)\|_\Omega^2 \le c\Big(\Upsilon + h^{2r}\Big)\|f\|_{H^{r-1}(\Omega)}^2 + ch^{3r-1}\|f\|_{H^{r-1}(\Omega)}^2, \tag{2.3.8}$$

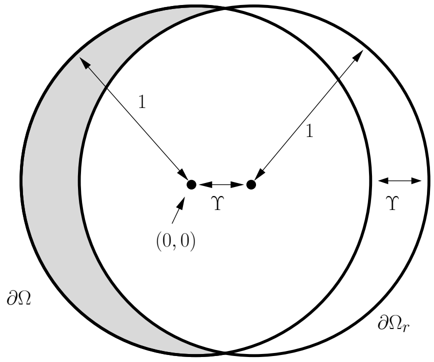
which shows the a priori estimate since $3r - 1 \le 2r$ for all $r \ge 1$.

(ii) For the L^2-error we proceed in the same way, but the remainder appearing in (2.3.2) does not carry any derivative, such that, instead of (2.3.4) the optimal order variant of Lemma 2.2 with integration to the boundary $\partial\Omega_h$, where $u_h = 0$, can be applied, i.e.

$$\|u-u_h\|_\Omega^2 \le c\Big(\|u-u_r\|_\Omega^2 + \|u_r - u_h\|_{\Omega_r}^2 + \Upsilon^2\|\nabla u_h\|_{\Omega\setminus\Omega_r}^2\Big).$$

The L^2-estimate directly follows with Lemma 2.3, Theorem 2.5 and by the a priori estimate $\|\nabla u_h\|_{\Omega\setminus\Omega_r}^2 \le c\|\nabla u_h\|_{\Omega_h}^2 \le c\|f\|_{\Omega_r}^2$.

□



On $\Omega = B_1(0)$ and $\Omega_r = B_1(\Upsilon)$ consider $-\Delta u = 4$ and $-\Delta u_r = 4$, respectively with homogeneous Dirichlet conditions and the solutions

$$u(x,y) = 1 - x^2 - y^2,\ u_r(x,y) = 1 - (x-\Upsilon)^2 - y^2$$

and the errors

$$\|\nabla(u-u_r)\|_\Omega = \sqrt{8}\Upsilon^{\frac{1}{2}} + \mathscr{O}(\Upsilon),\ \|u-u_r\|_\Omega = \sqrt{\pi}\Upsilon + \mathscr{O}(\Upsilon^3).$$

Figure 2.3: Illustration concerning Remark 2.8. The error estimates for $u - u_h$ are optimal, if the error is evaluated on Ω. The lowest order terms $\mathscr{O}(\Upsilon^{\frac{1}{2}})$ appear in the shaded area $\Omega \setminus \Omega_r$ where u_r and (most of) u_h are zero.

Remark 2.7 (Polygonal domains). In two dimensions, the extension of the error estimates to the case of convex polygonal domains, where $u \in H^2(\Omega)$ and $u_r \in H^2(\Omega_r)$, is relatively straightforward. In this case, Ω_h fits Ω_r such that the finite element error $u_r - u_h$ can be estimated with the standard a priori result $\|u_r - u_h\| + h\|\nabla(u_r - u_h)\| \le c\|f\|$. The extension of Lemma 2.2, which locally requires smoothness of the parametrizations $g_W^R(\cdot)$ and $g_V^R(\cdot)$, see steps (2.2.8)-(2.2.10), can be accomplished by refining the cover of the domain which is described in Assumption A1, see also Figure 2.1: All rectangles are split in such a way that the corners of $\partial\Omega$ and Ω_r are cut by the edges of rectangles. This allows to derive the optimal error estimates $\|u - u_h\|_{H^1(\Omega)} = \mathscr{O}(\Upsilon^{\frac{1}{2}} + h)$ and $\|u - u_h\|_{H^1(\Omega)} = \mathscr{O}(\Upsilon + h^2)$. In three dimensions, such a simple refinement of the cover is not possible and the extension to polygonal domains is more involved. □

Remark 2.8 (Optimality of the estimates). Two ingredients govern the error estimates:

1. A geometrical error of order $\mathscr{O}(\Upsilon^{\frac{1}{2}})$ and $\mathscr{O}(\Upsilon)$, that describes the discrepancy between Ω and Ω_r, in the H^1 and L^2 norms respectively. This term is optimal which is easily understood by considering a simple example illustrated in Figure 2.3, namely $-\Delta u = 4$ on the unit disc $\Omega = B_1(0)$ and $-\Delta u_r = 4$ on the shifted domain $\Omega_r = B_1(\Upsilon)$. The errors in H^1 norm and L^2 norms expressed on the complete domain Ω are estimated by

$$\|u - u_r\|_\Omega = \sqrt{\pi}\Upsilon + \mathscr{O}(\Upsilon^3), \quad \|\nabla(u - u_r)\|_\Omega = \sqrt{8\Upsilon} + \mathscr{O}(\Upsilon).$$

   A closer analysis shows that the main error – in the H^1-case – occurs on the small shaded stripe $\Omega \setminus \Omega_r$ such that

$$\|\nabla(u - u_r)\|_{\Omega\setminus\Omega_r} = \mathscr{O}(\Upsilon^{\frac{1}{2}}), \quad \|\nabla(u - u_r)\|_{\Omega\cap\Omega_r} = \mathscr{O}(\Upsilon),$$

   while the L^2-error in $\Omega \cap \Omega_r$ is optimal

$$\|u - u_r\|_{\Omega\setminus\Omega_r} = \mathscr{O}(\Upsilon^{\frac{3}{2}}), \quad \|u - u_r\|_{\Omega\cap\Omega_r} = \mathscr{O}(\Upsilon).$$

2. The usual Galerkin error $\|u_r - u_h\|_{\Omega_r} + h\|\nabla(u - u_r)\|_{\Omega_r} = \mathscr{O}(h^{r+1})$ of iso-parametric finite element approximations contributes to the overall error. For $\Omega = \Omega_r$, i.e. $\Upsilon = 0$, this

[1] We refer to [204, Chapter 1.4.3] or [20, 46] for recent developments on the local trace inequality and the inverse estimate on meshes with curved boundaries.

would be the complete error. This estimate is optimal, as it shows the same order as usual finite element bounds on meshes that resolve the geometry.

□

In Section 2.4 we discuss the difficulty of measuring errors on an unknown domain Ω. The optimality of the error estimates is difficult to verify which is mainly due to the technical problems in evaluating norms on the domain remainders $\Omega \setminus \Omega_r$, where no finite element mesh is given. These remainders contribute the lowest order parts $\Upsilon^{\frac{1}{2}}$ in the overall error. The following corollary is closer to the setting of the numerical examples and it yields the approximation of order Υ in the H^1-norm error. In addition to the previous setting we require a regular map $T_r : \Omega \to \Omega_r$ between the two domains. By pulling back Ω_r to Ω via this map a Jacobian arises that controls the geometrical error and that hence has to be controllable by Υ.

Corollary 2.9. *In addition to the assumptions of Theorem 2.6 let there be a C^1-diffeomorphism*

$$T_r : \Omega \to \Omega_r$$

satisfying

$$\|I - \det(\nabla T_r)\nabla T_r^{-1}\nabla T_r^{-T}\|_{L^\infty(\Omega)} = \mathscr{O}(\Upsilon). \tag{2.3.9}$$

Further, let the following regularity of problem data hold in addition to Assumption A2

$$f \in W^{1,\infty}_{loc}(\mathbb{R}^d) \cap H^{r-1}_{loc}(\mathbb{R}^d) \tag{2.3.10}$$

and let the solution satisfy

$$\|u\|_{W^{2,\infty}(\Omega)} + \|u_r\|_{W^{2,\infty}(\Omega_r)} \leq c. \tag{2.3.11}$$

Then, it holds

$$\|\nabla(u - u_h)\|_{\Omega\cap\Omega_r\cap\Omega_h} \leq c\Big(\Upsilon + h^r\Big).$$

Proof. We start by splitting the error into domain approximation and finite element approximation errors

$$\|\nabla(u - u_h)\|_{\Omega\cap\Omega_r\cap\Omega_h} \leq \|\nabla(u - u_r)\|_{\Omega\cap\Omega_r} + \|\nabla(u_r - u_h)\|_{\Omega_r\cap\Omega_h}. \tag{2.3.12}$$

An optimal order estimate of the finite element error

$$\|\nabla(u_r - u_h)\|_{\Omega_r\cap\Omega_h} \leq \|\nabla(u_r - u_h)\|_{\Omega_r} = \mathscr{O}(h^r) \tag{2.3.13}$$

is given in Theorem 2.5. To estimate the first term of the right hand side of (2.3.12) we introduce the function

$$\widehat{u}_r(x) := u_r(T_r(x)),$$

which satisfies $\widehat{u}_r \in H^1_0(\Omega)$ and solves the problem

$$(J_r F_r^{-1} F_r^{-T} \nabla\widehat{u}_r, \nabla\widehat{\phi}_r)_\Omega = (\widehat{f}_r, \widehat{\phi}_r) \quad \forall\widehat{\phi}_r \in H^1_0(\Omega),$$

where $\widehat{f}_r(x) := f(T_r(x))$ and where $F_r := \nabla T_r$ and $J_r := \det(F_r)$. See [209, Section 2.1.2] for details of this transformation of the variational formulation. To estimate the domain approximation error in (2.3.12) we introduce $\pm\widehat{u}_r$ to obtain

$$\|\nabla(u - u_r)\|_{\Omega\cap\Omega_r} \leq \|\nabla(u - \widehat{u}_r)\|_{\Omega\cap\Omega_r} + \|\nabla(\widehat{u}_r - u_r)\|_{\Omega\cap\Omega_r}. \tag{2.3.14}$$

We introduce the notation $e_r := u - \widehat{u}_r$, extend the first term from $\Omega \cap \Omega_r$ to Ω and insert $\pm J_r F_r^{-1} F_r^{-T} \nabla \widehat{u}_r$ which gives

$$\begin{aligned}\|\nabla(u - \widehat{u}_r)\|^2_{\Omega\cap\Omega_r} &\le \|\nabla(u - \widehat{u}_r)\|^2_{\Omega}\\ &= (\nabla u, \nabla e_r)_\Omega - (J_r F_r^{-1} F_r^{-T} \nabla \widehat{u}_r, \nabla e_r)_\Omega + (J_r F_r^{-1} F_r^{-T} \nabla \widehat{u}_r, \nabla e_r)_\Omega - (\nabla \widehat{u}_r, \nabla e_r)_\Omega\\ &= (f - \widehat{f}_r, e_r)_\Omega + ([J_r F_r^{-1} F_r^{-T} - I]\nabla \widehat{u}_r, \nabla e_r)_\Omega\\ &\le \|f - \widehat{f}_r\|_\Omega c \|\nabla e_r\|_\Omega + \|[J_r F_r^{-1} F_r^{-T} - I]\|_{L^\infty(\Omega)} \|\nabla e_r\|_\Omega, \end{aligned} \tag{2.3.15}$$

where we also used Poincaré's estimate. For bounding $f - \widehat{f}_r$ we consider a point $x \in \Omega \cap \Omega_r$, use the higher regularity of the right hand side (2.3.10) to estimate by a Taylor expansion

$$|f(x) - \widehat{f}_r(x)| = |f(x) - f(T_r(x))| = |\nabla f(\xi) \cdot (T_r(x) - x)| \le \Upsilon |\nabla f(\xi)|, \tag{2.3.16}$$

where $\xi \in \Omega$ is some point on the line from x to $T_r(x)$. We take the square and integrate over Ω to get the estimate

$$\|f - \widehat{f}_r\|_\Omega \le c\Upsilon \|f\|_{W^{1,\infty}(\Omega_\Upsilon)}, \tag{2.3.17}$$

where Ω_Υ is a enlargement of Ω by at most $\mathscr{O}(\Upsilon)$, since intermediate values ξ used in (2.3.16) are not necessarily part of $\Omega \cup \Omega_r$. This argument is also applicable to the second term on the right hand side of (2.3.14) such that it holds

$$\|\nabla(\widehat{u}_r - u_r)\|_M \le c\Upsilon \|u_r\|_{W^{2,\infty}(\Omega\cap\Omega_r)} \le c\Upsilon.$$

Combining this with (2.3.12), (2.3.13), (2.3.14) and (2.3.15) finishes the proof.

Unfortunately this corollary can not be applied universally as the existence of a suitable map $T_r : \Omega \to \Omega_r$ depends on the given application. Here a construction, corresponding to the ALE map, can be realised by means of a *domain deformation* $\widehat{d} : \Omega \to \mathbb{R}^2$

$$T_r(x) = x + \widehat{d}(x), \quad F_r(x) = I + \nabla \widehat{d}(x).$$

Such a construction is common in fluid-structure interactions, see [209, Section 2.5.2]. Given that $|\widehat{d}|, |\nabla \widehat{d}| = \mathscr{O}(\Upsilon)$ it holds

$$\|J_r\|_{L^\infty(\Omega)} = 1 + \mathscr{O}(\Upsilon), \quad \|I - J_r F_r^{-1} F_r^{-T}\|_{L^\infty(\Omega)} = \mathscr{O}(\Upsilon).$$

While the assumption $|\widehat{d}| = \mathscr{O}(\Upsilon)$ is easy to satisfy since $\operatorname{dist}(\partial\Omega, \partial\Omega_r) \le \Upsilon$, the condition $|\nabla \widehat{d}| = \mathscr{O}(\Upsilon)$ will strongly depend on the shape and regularity of the boundary.

We conclude by discussing a simple application of this corollary. Figure 2.4 illustrates the setting. Let Ω be the unit sphere, Ω_r be an ellipse

$$\Omega = \{x \in \mathbb{R}^2 \,:\, x_1^2 + x_2^2 < 1\}, \quad \Omega_r = \{x \in \mathbb{R}^2 \,:\, (1+\Upsilon)^2 x_1^2 + (1+\Upsilon)^{-2} x_2^2 < 1\}.$$

It holds $\operatorname{dist}(\partial\Omega, \partial\Omega_r) \le \Upsilon$ and we define the map $T_r : \Omega \to \Omega_r$ by

$$T_r(x) = \begin{pmatrix} (1+\Upsilon)^{-1} x_1 \\ (1+\Upsilon) x_2, \end{pmatrix}, \quad F_r = \nabla T_r = \begin{pmatrix} (1+\Upsilon)^{-1} & 0 \\ 0 & (1+\Upsilon) \end{pmatrix}, \quad J_r = 1.$$

This map satisfies the assumptions of the corollary

$$I - J_r F_r^{-1} F_r^{-T} = \Upsilon(\Upsilon + 2) \begin{pmatrix} -1 & 0 \\ 0 & (1+\Upsilon)^{-2} \end{pmatrix}, \quad \|I - J_r F_r^{-1} F_r^{-T}\|_\infty = 2\Upsilon + \Upsilon^2.$$

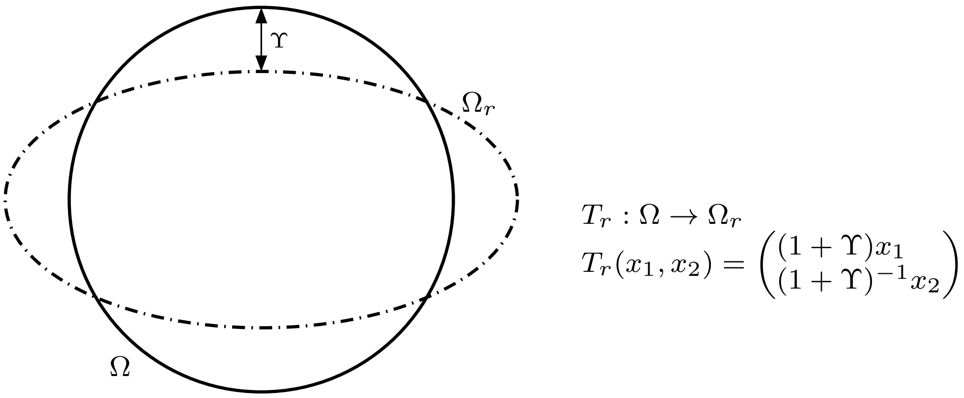

Υ
Ω_r
Ω
$T_r : \Omega \to \Omega_r$
$T_r(x_1, x_2) = \begin{pmatrix} (1+\Upsilon)x_1 \\ (1+\Upsilon)^{-1}x_2 \end{pmatrix}$

Figure 2.4: Illustration of an example for the application of Corollary 2.9.

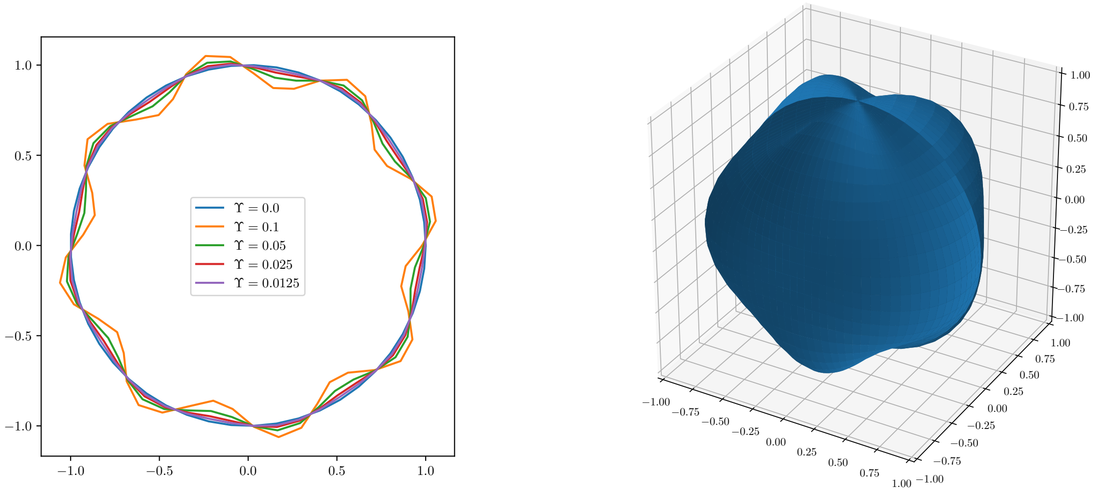


Figure 2.5: Sketch of the computational domains w.r.t. the parameter Υ in two dimensions (left) and for $\Upsilon = 0.1$ in three dimensions (right).

2.4 Numerical Illustration

In this section we illustrate the theoretical results from the previous section. We compute the Laplace problem on a family of domains representing different values of Υ. Moreover, we numerically extend the analytical predictions and show that a similar behavior holds for the Stokes system.

We consider Ω to be a unit ball in two and three dimensions and define a family of perturbed domains Ω_Υ, with the amplitude of the perturbation being dependent on the coefficient Υ, cf. Figure 2.5.

In two dimensions, the boundary of the domain Ω_Υ is given in polar coordinates (ϱ, φ) by

$$\partial\Omega_\Upsilon = \{(1 - \Upsilon/5 + \Upsilon \sin(8\varphi), \varphi) \text{ for } \varphi \in [0, 2\pi)\},$$

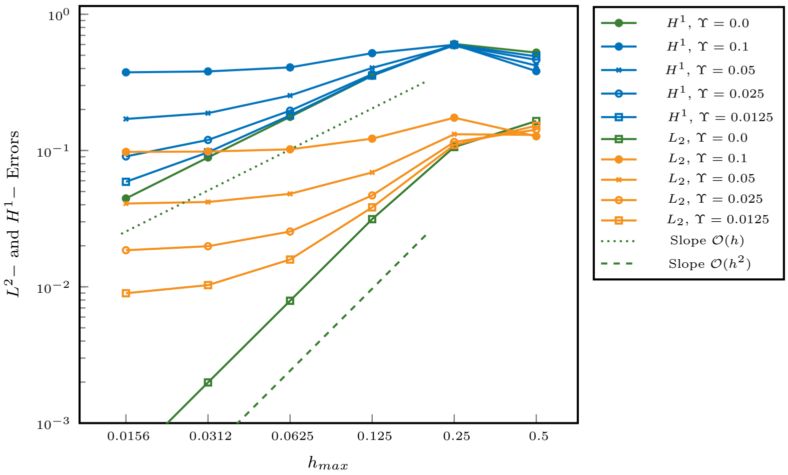


Figure 2.6: L^2- and H^1-errors w.r.t. mesh-size h_{max} for varying parameter Υ computed for the Laplace problem in three-dimensions with linear finite elements.

and in three dimensions in spherical coordinates $(\varrho, \theta, \varphi)$ by

$$\partial\Omega_\Upsilon = \{(1 - \Upsilon/5 + \Upsilon \sin(3\varphi)\sin(3\theta), \theta, \varphi) \text{ for } \theta \in [0, \pi), \varphi \in [0, 2\pi)\}.$$

For computations we take

$$\Upsilon \in \{0, 0.0125, 0.025, 0.05, 0.1\}.$$

In order to illustrate the convergence result from Theorem 2.6, we compute the model problem on a series of uniformly refined meshes. The dependence between the mesh size h and the refinement level L reads $h = 2^{-L}$. We denote the mesh approximating Ω_Υ, with a mesh size h, by $\Omega_{h,\Upsilon}$.

The numerical implementation is realized in the software library Gascoigne 3D [39], using isoparametric finite elements of degree 1 and 2. A detailed description of the underlying numerical methods is given in [209].

2.4.1 Laplace Equation in Two and Three Dimensions

We consider the following problem

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{2.4.1}$$

where Ω is the unit ball in two dimensions and the unit sphere in three dimensions.

To compute errors we choose a rotationally symmetric analytical solution to (2.4.1) as

$$u(r) = -\cos\left(\frac{\pi}{2} r\right)$$

with $r = \sqrt{x^2 + y^2}$ in two and $r = \sqrt{x^2 + y^2 + z^2}$ in three dimensions, respectively, which results in the right hand sides

$$f_{2d}(r) = \frac{\pi}{2r} \sin\left(\frac{\pi}{2} r\right) + \frac{\pi^2}{4} \cos\left(\frac{\pi}{2} r\right), \quad f_{3d}(r) = \frac{\pi}{r} \sin\left(\frac{\pi}{2} r\right) + \frac{\pi^2}{4} \cos\left(\frac{\pi}{2} r\right).$$

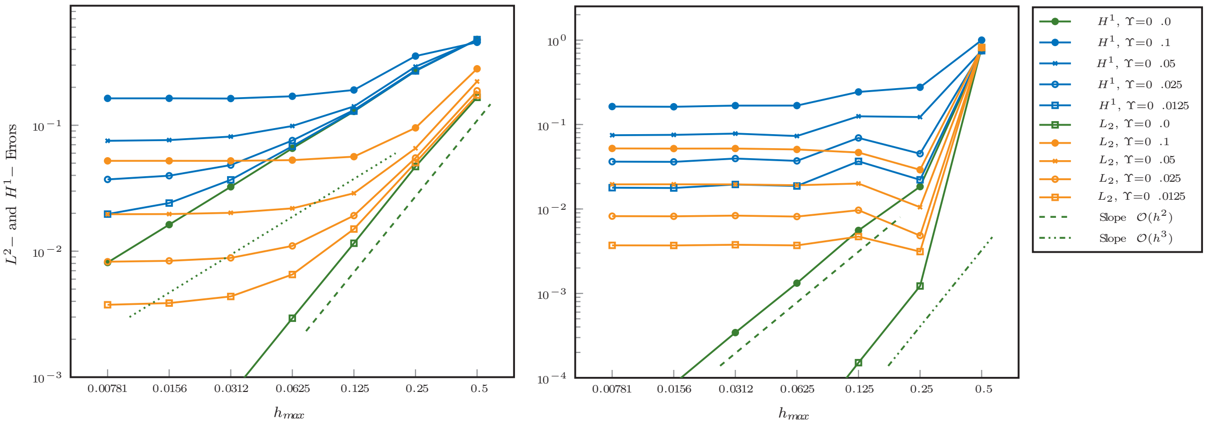


Figure 2.7: L^2- and H^1-errors w.r.t. mesh-size h_{max} for varying parameter Υ computed for the Laplace problem in two-dimensions with FE. Left: linear finite elements. Right: quadratic finite elements.

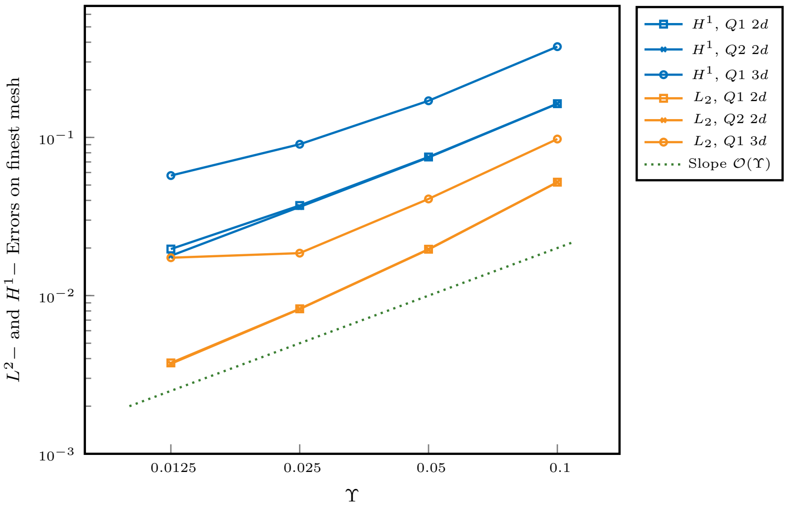


Figure 2.8: L^2- and H^1-errors w.r.t. parameter Υ computed for the Laplace problem in two and three-dimensions with linear and quadratic finite elements.

For the ease of evaluations the errors, the H^1- and L^2-norms will be computed on the truncated domains

$$\Omega'_{2d} = \{(\varphi, \varrho) \text{ for } \varphi \in [0, 2\pi) \text{ and } \varrho \in (0, 0.88)\},$$
$$\Omega'_{3d} = \{(\varphi, \theta, \varrho) \text{ for } \theta \in [0, \pi), \varphi \in [0, 2\pi) \text{ and } \varrho \in (0, 0.88)\},$$

see also Remark 2.8. We hence do not compute the errors $\|\nabla(u - u_h)\|$ and $\|u - u_h\|$ on the remainders $\Omega \setminus \Omega_r$. Therefore we expect optimal order convergence in the spirit of Corollary 2.9. The restriction of the domain to an area within Ω_h is also by technical reasons, as the evaluation of integrals outside of the meshed area is not easily possible.

In Figures 2.6 and 2.7 we see the resulting L^2- and H^1-errors. We observe that for finer meshes, Υ becomes the dominating factor of the error. In particular the use of quadratic finite elements shows a strong imbalance between FE error and geometric error, which quickly dominates as seen in the left part of Fig. 2.7. The result is consistent with Corollary 2.9. As soon as the FE error is smaller than the geometry perturbation Υ, we do not observe any further improvement of the error. In Fig. 2.8 we show the convergence in both norms in terms of the geometry parameter Υ. Linear convergence is clearly observed. The apparent decay of convergence rate in case of the L^2-error in three dimensions is due to the still dominating FE error in this case.

2.4.2 Stokes System in Two Dimensions

To go beyond the Laplace problem, we investigate the behavior of the solution to the Stokes system with respect to the domain variation in two spatial dimensions. The problem is to find the velocity $\mathbf{u}$ and the pressure p such that

$$\operatorname{div}\mathbf{u} = 0, \quad -\Delta\mathbf{u} + \nabla p = \mathbf{f} \text{ in } \Omega, \tag{2.4.2}$$

with homogeneous Dirichlet condition $\mathbf{u} = 0$ on the boundary $\partial\Omega$ and a right hand side vector $\mathbf{f}$. System (2.4.2) is solved with equal-order iso-parametric finite elements using pressure stabilization by local projections, see [24].

We prescribe an analytical solution for comparison with the finite element approximation

$$\mathbf{u}(x,y) = \cos\left(\frac{\pi}{2}(x^2+y^2)\right)\begin{pmatrix} y \\ -x \end{pmatrix},$$

where the corresponding forcing term reads

$$\mathbf{f}(x,y) = \pi\cos\left(\frac{\pi}{2}(x^2+y^2)\right)\begin{pmatrix} yr^2\pi + 4(y-x)\tan\left(\frac{\pi}{2}(x^2+y^2)\right) \\ -xr^2\pi - 4(x+y)\tan\left(\frac{\pi}{2}(x^2+y^2)\right) \end{pmatrix}.$$

In Figure 2.9 we see the resulting L^2- and H^1-errors. Again we observe that Υ becomes the dominant factor for finer meshes. This result is not covered by the theoretical findings, however it shows that geometric uncertainty should be taken into account for the simulations of flow models.

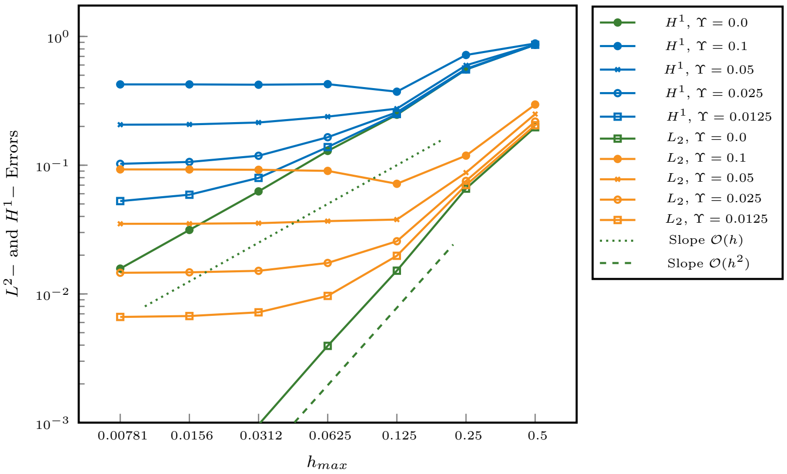


Figure 2.9: L^2- and H^1-errors w.r.t. mesh-size h_{max} for varying parameter Υ computed for the Stokes problem in two dimensions with linear finite elements.

CHAPTER 3

On the Impact of Fluid-structure Interaction in Blood Flow Simulations

The content of this chapter is joint work with Lukas Failer and Thomas Richter, and is published in the paper

- L. Failer, P. Minakowski and T. Richter. 'On the Impact of Fluid Structure Interaction in Blood Flow Simulations'. In: *Vietnam Journal of Mathematics* (Jan. 2021). doi: 10.1007/s10013-020-00456-6

Chapter Summary. We study the impact of using fluid-structure interactions (FSI) to simulate blood flow in a stenosed artery. We compare typical flow configurations using Navier-Stokes in a rigid geometry setting to a fully coupled FSI model. The relevance of vascular elasticity is investigated with respect to several questions of clinical importance. Namely, we study the effect of using FSI on the wall shear stress distribution, on the Fractional Flow Reserve and on the damping effect of a stenosis on the pressure amplitude during the pulsatile cycle. The coupled problem is described in a monolithic variational formulation based on Arbitrary Lagrangian Eulerian (ALE) coordinates. For comparison, we perform pure Navier-Stokes simulations on a pre-stressed geometry to give a good matching of both configurations. A series of numerical simulations that cover important hemodynamical factors are presented and discussed.

Chapter Organisation. After the brief introduction in Section 3.1 we present the monolithic formulation of the FSI problem, see Section 3.2. The setting of the simulations is the topic of Section 3.3. The numerical method is described in Section 3.4. Section 3.5 is dedicated to the presentation of investigated hemodynamical factors and the discussion of numerical results obtained on relevant test cases.

Contents of Chapter

3.1 Introduction

The application of computational fluid dynamics (CFD) to blood flow is a rapidly growing field of biomedical and mathematical research. Currently the development of numerical methods in hemodynamics is evolving from a purely academic tool to aiding in clinical decision making [228, 37]. In particular the investigation of blood flow in stenosed arteries can help to shape medical treatment. For example virtual/computed Fractional Flow Reserve (cFFR) can evaluate the physiological significance of sclerotic plaque [229, 178, 36]. Moreover, the correct reconstruction of wall shear stress (WSS) is of crucial importance for the cell signaling and as a consequence for the stenosis development [151] or for the assessment of rupture of cerebral aneurysms [55]. These two factors are studied in detail.

Hemodynamical simulations face numerous challenges, that are connected with measurements, medical image segmentation, mathematical modelling and development of numerical methods [207]. In this work we confine to an idealized geometry and a Newtonian fluid and focus on the comparison between a compliant wall vs. a rigid vessel wall. As a model example we have chosen a curved channel that resembles a large human artery. Stenotic and non-stenotic configurations are considered. The channel is preloaded by first considering a steady inflow to reach a certain physiological pressure and diameter before starting the pulsatile heart cycle. We investigate the aforementioned important clinical hemodynamical factors cFFR and WSS.

In the physiology of vessel macrocirculation the compliance plays a crucial role. However for individual arteries the problem is still open and there is no gold standard that can be adapted into clinical practice. For a discussion on assumptions in hemodynamic modelling of large arteries we refer to [206, 225].

Numerous numerical studies were performed to compare rigid with compliant vessel simulations. The results report reduction of WSS for compliant vessels compared to rigid walls. In the case of a carotid bifurcation the authors of [194] reported significant WSS reduction, however have not observed significant changes in the flow patterns. Furthermore, computational studies of flow in cerebral artery aneurysms indicated that rigid models tended to over estimate the WSS magnitude [220]. Moreover, it is worth to mention that even vessels like the aorta, that are often treated as rigid, can undergo substantial radial wall motion. The motion consists of bulk deformation and wall compliance that results in notable changes of flow characteristics [136, 232].

3.2 Model Description

We consider a 3-dimensional domain $\Omega \subset \mathbb{R}^3$, that represents a part of a vessel. The domain is partitioned in the reference configuration $\Omega = \mathscr{F} \cup \mathscr{I} \cup \mathscr{S}$, where $\mathscr{F}$ is the fluid domain, $\mathscr{S}$ the solid domain and $\mathscr{I} = \partial\mathscr{F} \cap \partial\mathscr{S}$ is the fluid structure interface.

The velocity field $\mathbf{v}$ and the deformation field $\mathbf{u}$ are split into fluid $\mathbf{v}_f := \mathbf{v}|_{\mathscr{F}}$, $\mathbf{u}_f := \mathbf{u}|_{\mathscr{F}}$ and solid $\mathbf{v}_s := \mathbf{v}|_{\mathscr{S}}$, $\mathbf{u}_s := \mathbf{u}|_{\mathscr{S}}$ counterparts respectively. The pressure variable p_f only exists on the fluid domain.

The boundary of the fluid domain $\Gamma_f := \partial\mathscr{F} \setminus \mathscr{I}$ is split into the inflow boundary Γ_f^{in} and the outflow boundary Γ_f^{out}. Similarly the solid boundary $\Gamma_s = \partial\mathscr{S} \setminus \mathscr{I}$ is split into inflow Γ_s^{in} and outflow Γ_s^{out} boundaries. Boundary conditions are described in Section 3.3.

3.2.1 Fluid Material Model

Although, blood exhibits many unique properties, i.e. in certain regimes blood shows a non-Newtonian behaviour, we confine this work to considering an incompressible Newtonian fluid with the viscosity of $\mu_f = 0.033\,\mathrm{g}\cdot\mathrm{cm}^{-1}\mathrm{s}^{-1}$ and the density $\varrho_f = 1\,\mathrm{g}\cdot\mathrm{cm}^{-3}$. The assumption of a Newtonian model for blood rheology is widely accepted for large and medium vessels, see e.g. [207]. The flow is governed by the Navier-Stokes equations

$$\varrho_f \left(\partial_t \mathbf{v}_f + (\mathbf{v}_f \cdot \nabla)\mathbf{v}_f\right) - \mu_f \operatorname{div}\left(\nabla \mathbf{v}_f + \nabla \mathbf{v}_f^T\right) + \nabla p_f = 0 \quad \text{in } \mathscr{F}, \tag{3.2.1a}$$

$$\operatorname{div} \mathbf{v}_f = 0 \quad \text{in } \mathscr{F}. \tag{3.2.1b}$$

3.2.2 Solid Material Model

Arterial walls consist of heterogeneous layers with significant difference in physical properties. The schematic layer construction of an arterial wall consists of intima (inner layer), media (middle layer), and adventitia (outer layer). For detail account we refer to [107] and [130]. We briefly describe how the elastic constitutional law used in this work is derived.

Since arteries hardly change their volume within the physiological range of deformation [50], they can be regarded as incompressible or nearly incompressible materials. This motivates the application of a multiplicative decomposition of the deformation tensor ($\mathbf{F}$) into its volumetric part $J^{\frac{1}{3}}$ and the deviatoric part $\bar{\mathbf{F}}$:

$$\mathbf{F} = \boldsymbol{I} + \nabla \mathbf{u}_s, \quad \mathbf{F} = J^{\frac{1}{3}}\bar{\mathbf{F}}, \text{ where } J = \det \mathbf{F} \text{ and } \bar{\mathbf{F}} = J^{-\frac{1}{3}}\mathbf{F}. \tag{3.2.2}$$

Its associated modified deviatoric Cauchy Green tensor $\bar{\mathbb{C}}$ then has the structure

$$\bar{\mathbb{C}} = \bar{\mathbf{F}}^T \bar{\mathbf{F}}. \tag{3.2.3}$$

Thereby the free-energy function Ψ can be split in a volumetric and deviatoric part as described in [130] or [115]:

$$\Psi = \Psi_{VOL}(J) + \Psi_{DEV}(\bar{\mathbb{C}}).$$

Artery and vein walls consist of elastin and colagen fibres. The measurements of stress-strain curve exhibit stiffening effects at higher pressures due to the collagen fibres, c.f. [107]. Whereas under low loading of the artery the properties of elastin dominates. This motivates to model the artery as pseudo-elastic material. Following [82] and [130] we employ an exponential deviatoric energy functional

$$\Psi_{DEV}(\bar{\mathbb{C}}) = \frac{\mu}{2\gamma}\Big(\exp^{\gamma(\mathrm{tr}\,(\bar{\mathbb{C}})-3)} - 1 \Big). \tag{3.2.4}$$

For the volumetric part of the energy functional the simple convex energy function

$$\Psi_{VOL}(J) = \frac{\kappa}{2}\left(\frac{1}{2}(J^2-1) - \ln(J)\right) \tag{3.2.5}$$

is used as stated in [22, 23]. By assuming hyperelastic stress response we obtain the second Piola–Kirchhoff stress tensor $\mathbf{\Sigma}_s$:

$$\begin{aligned}\mathbf{\Sigma}_s(J,\mathbf{F}) =& \frac{\partial \Psi_{VOL}(J)}{\partial \mathbb{C}} + \frac{\partial \Psi_{DEV}(\bar{\mathbb{C}})}{\partial \mathbb{C}} \\ =& \mu_s J^{-2/3}(\boldsymbol{I} - \frac{1}{3}\,\mathrm{tr}\,(\mathbf{F}^T\mathbf{F})(\mathbf{F}^T\mathbf{F})^{-1}) e^{\gamma(J^{-2/3}\,\mathrm{tr}\,(\mathbf{F}^T\mathbf{F})-3)} \\ &+ \frac{\kappa_s}{2}(J^2-1)J(\mathbf{F}^T\mathbf{F})^{-1},\end{aligned} \tag{3.2.6}$$

with the material parameters $\mu_s = 44.2\,\mathrm{kPa}$ and $\gamma = 20$ as well as $\kappa_s = 4998\,\mathrm{kPa}$. Similar values have been used in [19]. Moreover the solid density equals $\widehat{\varrho}_s = 1.2\,\mathrm{g}\cdot\mathrm{cm}^{-3}$.

The equation for the conservation of momentum in the elastic vessel wall is then given by

$$\widehat{\varrho}_s d_t \mathbf{v}_s - \mathrm{div}\Big(\mathbf{F}\mathbf{\Sigma}_s\Big) = \widehat{\varrho}_s \mathbf{f}_s \quad \text{in } \mathscr{S}, \tag{3.2.7a}$$

$$d_t \mathbf{u}_s = \mathbf{v}_s \quad \text{in } \mathscr{S}, \tag{3.2.7b}$$

where we formulated the hyperbolic problem as a first order system in time by introducing the solid velocity $\mathbf{v}_s$ and the deformation field $\mathbf{u}_s$ as separate variables. The solid problem is formulated in the Lagrangian coordinates on the reference domain $\mathscr{S}$, which is not moving with time. Hence, the density $\widehat{\varrho}_s = 1.2\,\mathrm{g}\cdot\mathrm{cm}^{-3}$ is the reference density which is unaffected by any compression or extension.

On the interior boundary of the solid domain, which is the interface to the fluid domain, the typical coupling conditions of fluid-structure interaction problems are given. The proper handling of the outer boundary, where the vessel is embedded in tissue, is a delicate task. For our study, this difficulty is neglected and for further reading we refer to [174].

3.2.3 Fluid-Structure Interactions

One aim of this work is to study the impact of coupled fluid-structure interactions on typical blood flow configurations found in medical applications. We therefore couple the Navier-Stokes equations (3.2.1) with the elastic material law (3.2.7) via coupling conditions on the common interface $\mathscr{I}$. The coupled dynamics of a fluid-structure interaction problem leads to a free boundary problem with moving domains and a moving interface.

In the following, we briefly sketch the derivation of the coupled fluid-structure interaction problem. For details we refer to the literature [209, Chapter 5]. To overcome the mismatch of a Navier-Stokes equations (3.2.1) defined in Eulerian coordinates, e.g. on the evolving fluid domain $\mathscr{F}(t)$, and the solid problem (3.2.7) derived on the fixed reference domain $\mathscr{S}$ we use the well established concept of Arbitrary Lagrangian Eulerian (ALE) coordinates. See [87] or [209, Chapter 5] for a detailed derivation. We denote by $\mathscr{F}$ the fixed fluid reference domain and by $T_f(t) : \mathscr{F} \to \mathscr{F}(t)$ the ALE map. Then, the velocity and the pressure can be mapped onto the fixed domain by defining $\hat{\mathbf{v}}_f := \mathbf{v}_f \circ T_f^{-1}$ and $\hat{p}_f := p_f \circ T_f^{-1}$. This allows to transform the Navier-Stokes problem in its variational formulation onto the fluid reference domain

$$\begin{aligned}\Big(J_f\Big(\partial_t\hat{\mathbf{v}}_f + (\mathbf{F}_f^{-1}(\hat{\mathbf{v}}_f - \partial_t T_f)\cdot\hat{\nabla})\hat{\mathbf{v}}_f\Big), \boldsymbol{\phi}\Big)_{\mathscr{F}} + \Big(J_f\hat{\boldsymbol{\sigma}}_f\mathbf{F}_f^{-T}, \hat{\nabla}\boldsymbol{\phi}\Big)_{\mathscr{F}}& \\ + \Big(J_f\mathbf{F}_f^{-1} : \hat{\nabla}\hat{\mathbf{v}}_f^T, \xi\Big)_{\mathscr{F}} = \Big(J_f\varrho_f\hat{\mathbf{f}}, \boldsymbol{\phi}\Big)_{\mathscr{F}}&,\end{aligned} \tag{3.2.8}$$

where $\boldsymbol{\phi}$ and ξ are test functions. We denote by $\mathbf{F}_f := \hat{\nabla}T_f$ the gradient of the deformation variable and by $J_f := \det(\mathbf{F}_f)$ its determinant. Most characteristic feature of the ALE formulation is the appearance of the domain convection term $-(\mathbf{F}_f^{-1}\partial_t T_f \cdot \nabla)\hat{\mathbf{v}}$ that takes care of the implicit motion of the fluid domain. Furthermore, the Cauchy stress tensor is mapped to the reference domain which gives rise to the Piola transform $J_f\hat{\boldsymbol{\sigma}}_f\mathbf{F}_f^{-T}$. The reference stress is given by

$$\hat{\boldsymbol{\sigma}}_f(\mathbf{v}, p) = -\hat{p}_f\boldsymbol{I} + \varrho_f\nu_f(\hat{\nabla}\hat{\mathbf{v}}\mathbf{F}^{-1} + \mathbf{F}^{-T}\hat{\nabla}\hat{\mathbf{v}}^T).$$

It remains to describe the construction of the ALE map. Typically the ALE map is defined by means of an artificial fluid domain deformation $\mathbf{u}_f$ via

$$T_f(\mathbf{x}, t) := \mathbf{x} + \hat{\mathbf{u}}_f(\mathbf{x}, t),$$

where $\hat{\mathbf{u}}_f$ is an extension of the solid deformation $\mathbf{u}_s$ from $\mathscr{S}$ to the fluid reference domain $\mathscr{F}$. The most simple choice for the extension operator is to use a harmonic extension by implicitly solving the vector Laplacian

$$-\hat{\Delta}\mathbf{u}_f = 0 \text{ in } \mathscr{F}, \quad \mathbf{u}_f = \mathbf{u}_s \text{ on } \mathscr{I}, \quad \mathbf{u}_f = 0 \text{ on } \partial\mathscr{F} \setminus \mathscr{I}. \tag{3.2.9}$$

For a discussion of this extension operator we refer to the literature [209, Sections 3.5.1 and 5.3.5]. Hereby, T_f can be considered as a natural extension of the Lagrange-Euler map $T_s(\mathbf{x}, t) := \mathbf{x}+\mathbf{u}_s(\mathbf{x}, t)$ such that we will skip the subscripts f and s when denoting the deformation $T(\mathbf{x}, t) := \mathbf{x} + \mathbf{u}(\mathbf{x}, t)$, its gradient $\mathbf{F} = \nabla T$ and its determinant $J = \det(\mathbf{F})$. In the following we skip all hats referring to the use of ALE coordinates.

As the fluid reference domain $\mathscr{F}$ and the Lagrangian solid domain $\mathscr{S}$ do not move, they always share the well defined common interface $\mathscr{I}$. Here, we require the continuity of velocities, which is denoted by the *kinematic coupling condition*

$$\mathbf{v}_f = \mathbf{v}_s \text{ on } \mathscr{I},$$

continuity of normal stresses, denoted as *dynamic coupling condition*

$$\mathbf{F}\boldsymbol{\Sigma}_s \vec{n} = J\boldsymbol{\sigma}_f \mathbf{F}^{-T}\vec{n} \text{ on } \mathscr{I},$$

and finally, the *geometric coupling condition* which says that the evolving domains $\mathscr{F}(t)$ and $\mathscr{S}(t)$ may not overlap and may not separate at the interface.

Since the velocity field and the deformation field are continuous across the interface, we formulate the coupled fluid-structure interaction problem using global solution fields $\mathbf{v} \in H^1(\Omega)^3$ and $\mathbf{u} \in H^1(\Omega)^3$. Hereby, the kinematic coupling condition and the extension condition $\mathbf{u}_f = \mathbf{u}_s$ in (3.2.9) are strongly realized as parts of the function spaces. The dynamic coupling condition is realized by testing the variational formulations of the Navier-Stokes equations and the solid problem by one common continuous test function $\boldsymbol{\phi} \in H^1(\Omega)^3$. The resulting variational system of equations is given by

$$\begin{aligned}\Big(J(\partial_t \mathbf{v} + (\mathbf{F}^{-1}(\mathbf{v} - \partial_t \mathbf{u}) \cdot \nabla)\mathbf{v}, \boldsymbol{\phi}\Big)_{\mathscr{F}} + \Big(J\widehat{\boldsymbol{\sigma}}_f \mathbf{F}^{-T}, \nabla\boldsymbol{\phi}\Big)_{\mathscr{F}} \\ +(\widehat{\varrho}_s \partial_t \mathbf{v}, \boldsymbol{\phi})_{\mathscr{S}} + (\mathbf{F}\boldsymbol{\Sigma}_s, \nabla\boldsymbol{\phi})_{\mathscr{S}} = (J\varrho_f \mathbf{f}, \boldsymbol{\phi})_{\mathscr{F}} + (\widehat{\varrho}_s \mathbf{f}, \boldsymbol{\phi})_{\mathscr{S}}\end{aligned}, \tag{3.2.10a}$$

$$\Big(J\mathbf{F}^{-1} : \nabla\mathbf{v}^T, \xi\Big)_{\mathscr{F}} = 0, \tag{3.2.10b}$$

$$(\partial_t \mathbf{u} - \mathbf{v}, \psi_s)_{\mathscr{S}} = 0, \tag{3.2.10c}$$

$$(\nabla\mathbf{u}, \nabla\psi_f)_{\mathscr{F}} = 0. \tag{3.2.10d}$$

For detailed account of monolithic formulations for fluid-structure interactions we refer to [209].

3.3 Simulation Setup

We perform pulsatile blood-flow simulations by numerically solving system (3.2.10). The simulation setup that includes geometry and material parameters has been inspired by the Benchmark paper [19]. Note however that in some of the cases the assumption of rigid vessel walls is employed such that $\mathbf{u} = 0$ and $\mathbf{F} = \boldsymbol{I}$ on Ω. Then system (3.2.10) reduces to the standard incompressible Navier-Stokes equations. In what follows we distinguish between FSI- and NS-cases, respectively.

3.3.1 Geometry

The geometry of the computational domain reflects an idealized coronary artery. We show a sketch of the geometry in Figure 3.1. It consists of three parts, two straight and one curved tube. The straight parts are aligned with the $x-$ and $y-$ axes for inflow and outflow, respectively. The centerline of the curved section is a part of a circle with center in $(1, 0, 0)$ and radius $R = 1$. It can be indicated as a parametrization $\psi_C : [0, 3] \to \mathbb{R}^3$

$$\psi_C(s) := \begin{cases} (0, -1 + s, 0)^T & 0 \le s \le 1 \\ (1 + \cos(\pi s/2), \sin(\pi s/2), 0)^T & 1 \le s \le 2 \\ (s - 1, 1, 0)^T & 2 \le s \le 3. \end{cases}$$

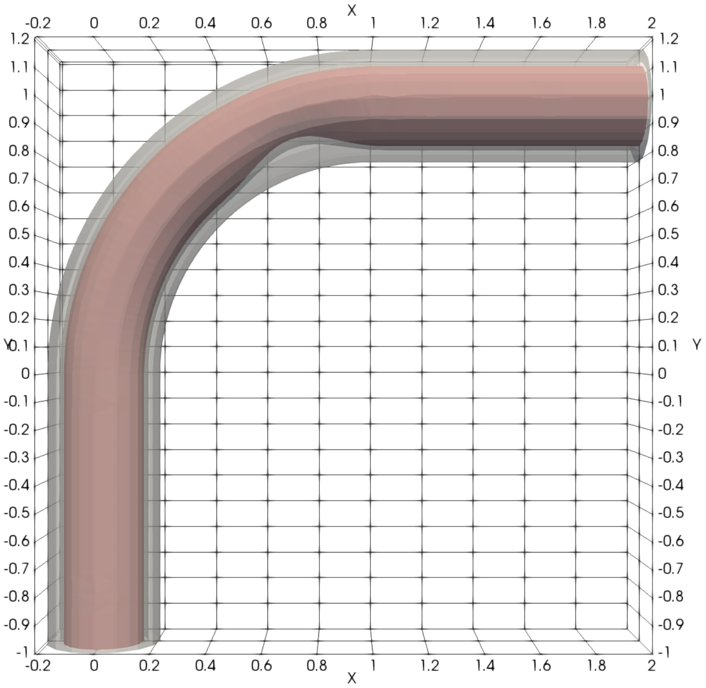


Figure 3.1: Computational domain showing Geometry 3 with a non-symmetric stenosis.

The fluid domain $\mathscr{F}$ is a cylinder around the centerline $\psi_C(s)$ with radius $r_{\mathscr{F}} = 0.15\,\text{cm}$. Furthermore, the solid domain (i.e. the vessel wall) is a cylindrical outer layer of width $0.06\,\text{cm}$, i.e. the outer radius is given by $r_{\mathscr{S}} = 0.21\,\text{cm}$. These dimensions correspond to the dimensions of realistic arteries.

We consider two stenotic and one non-stenotic configuration. The stenosis is modelled by a reduction of the inner radius $r_{\mathscr{F}}$ on the second part of the curved boundary. This can be described by the parametrization

$$r_{\mathscr{F}}^{sten}(s) := \begin{cases} 0.15\,\text{cm} & 0 \le s \le 1.5 \\ 0.15\,\text{cm} - 0.02\,\text{cm}\Big(\cos(4\pi s) - 1\Big) & 1.5 \le s \le 2 \\ 0.15\,\text{cm} & 2 \le s \le 3. \end{cases}$$

Moreover we proceed twofold. First, we consider a symmetric stenosis such that the fluid domain is a cylinder of radius $r_{\mathscr{F}}^{sten}(s)$ around the centerline $\psi_C(s)$. Second, we modify the centerline in such a way that the stenosis is only on the inner side of the curve. This is achieved by shifting the centerline to match the radius variation, i.e.

$$\psi_C^{shift}(s) := \psi_C(s) - r_{\mathscr{F}}^{sten}(s)\vec{\omega}(s),$$

where $\vec{\omega}(s)$ is the direction of the shift

$$\vec{\omega}(s) = \begin{cases} (0, -0.02\Big(\cos(4\pi s) - 1\Big), 0) & 1.5 \le s \le 2 \\ (0, 0, 0)^T & \text{elsewhere} \end{cases}.$$

	description	centerline	inner radius	outer radius
Geometry 1	without stenosis	$\psi_C(s)$	0.15 cm	0.21 cm
Geometry 2	symmetric stenosis	$\psi_C(s)$	$r_{\mathscr{F}}^{sten}(s)$	0.21 cm
Geometry 3	non-symmetric stenosis	$\psi_C^{shift}(s)$	$r_{\mathscr{F}}^{sten}(s)$	0.21 cm

Table 3.1: Summary of geometry configurations. Figure 3.1 shows Geometry 3.

In both stenotic configurations, the resulting reduction of the radius equals 0.04 cm and the lumen area shrinks by 46% from $0.15^2\pi\,\mathrm{cm}^2 \approx 0.07\,\mathrm{cm}^2$ to $0.11^2\pi\,\mathrm{cm}^2 \approx 0.038\,\mathrm{cm}^2$. Computational geometries are summarized in Table 3.1. Figure 3.1 shows Geometry 3.

3.3.2 Boundary Conditions

The blood flow is enforced by a Dirichlet inflow condition on the inflow boundary Γ_f^{in} given by

$$\Gamma_f^{in} = \{(x,-1,z) \in \mathbb{R}^3 \mid \sqrt{x^2+z^2} < 0.15\,\mathrm{cm}\}.$$

We prescribe a parabolic inflow profile

$$\mathbf{v}^{in}(x,y,z,t) = v_{\max}(t)\left(1 - \frac{x^2}{0.15^2\,\mathrm{cm}^2} - \frac{z^2}{0.15^2\,\mathrm{cm}^2}\right)(0,1,0)^T,$$

where the maximum value $v_{\max}(t)$ is presented in Figure 3.2. The temporal inflow profile (3.3.1) consists of three stages:

- Ramp phase ($0\,\mathrm{s} \le t \le 0.1\,\mathrm{s}$) with increasing inflow rate.
- Steady phase ($0.1\,\mathrm{s} < t \le 0.3\,\mathrm{s}$) with constant inflow rate.
- Pulsatile phase ($0.3\,\mathrm{s} < t$) with pulsatile inflow corresponding to heartbeats.

$$v_{\max}(t) = \frac{1}{28.3}\begin{cases} 0.5(1.0 - \cos(\pi t/0.1))3.0 & 0 \le t \le 0.1 \\ 3.0 & 0.1 \le t \le 0.3 \\ 5.931 & \\ \;-1.3933\cos(2\pi 1t) + 1.3532\sin(2\pi 1t) & \\ \;-0.9409\cos(2\pi 2t) + 0.2332\sin(2\pi 2t) & \\ \;-0.3026\cos(2\pi 3t) - 0.1190\sin(2\pi 3t) & \\ \;-0.2264\cos(2\pi 4t) - 0.0631\sin(2\pi 4t) & 0.3 \le t \le 3.3 \\ \;-0.1064\cos(2\pi 5t) - 0.2137\sin(2\pi 5t) & \\ \;+0.0402\cos(2\pi 6t) - 0.0691\sin(2\pi 6t) & \\ \;-0.0307\cos(2\pi 7t) - 0.0451\sin(2\pi 7t) & \\ \;+0.0271\cos(2\pi 8t) - 0.0735\sin(2\pi 8t) & \end{cases} \tag{3.3.1}$$

The third pulsatile part is an approximation, by means of a Fourier series of order 8, of a typical coronary velocity profile provided by [1].

For the outflow of the fluid on Γ_f^{out} in the FSI case we apply absorbing boundary conditions:

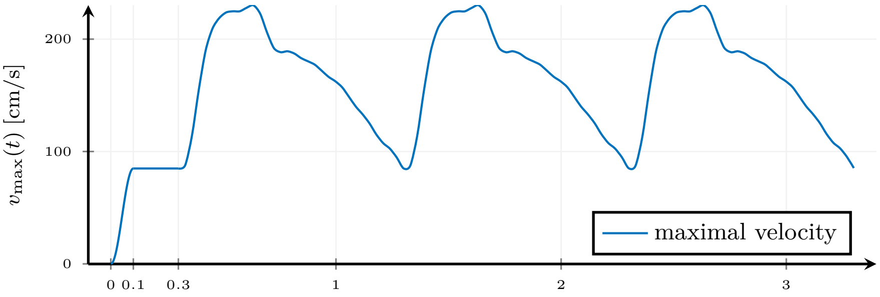


Figure 3.2: Inflow profile. The exact flow rate is given in (3.3.1).

$$(\boldsymbol{\sigma}_f^{n+1} \cdot \boldsymbol{n})|_{\Gamma_{out}} = \left(\left(\frac{\sqrt{\varrho_f}}{2\sqrt{2}} \frac{Q^n}{A^n} + \sqrt{p^*} \right)^2 + p_e - p^* \right) \boldsymbol{n}, \tag{3.3.2}$$

where $p* = \frac{E}{1-\nu^2}$. The condition is equipped with the reference pressure p_e which is chosen such that a vascular pressure of 80 mmHg at $t = 0.3$ s is achieved. The absorbing condition explicitly disallows pressure waves to reenter the domain, for details see [184]. The structure is fixed at the inflow boundary and is allowed to move in y-z direction at the outflow boundary.

In order to compare both rigid and elastic computations, we run the FSI simulation first and then prescribe the resulting reference pressure profile $P_{ref}(t)$ at the outflow boundary in the rigid wall case such that

$$\left\langle \Big(\varrho_f \nu_f \nabla \mathbf{v}_f - pI \Big) \vec{n}, \boldsymbol{\phi} \right\rangle_{\Gamma_f^{out}} = \left\langle P_{ref}(t) \vec{n}, \boldsymbol{\phi} \right\rangle_{\Gamma_f^{out}}.$$

This is the usual do-nothing outflow condition including a pressure offset, see [129].

The first two phases of (3.3.1) are designed to reach the intravascular pressure of 80 mmHg. This procedure resembles pre-stressed of the vessel wall and it is required to apply the nonlinear solid law described in Section 3.2.2 in the right tract of the stress-strain curve. We first start the FSI simulation on the undeformed stress free reference domain and pre-stress the geometry by the ramp and steady phases. The resulting geometry at $t = 0.3s$ is then extracted to compute the FSI as well as the Navier-Stokes simulation on this deformed geometry. This geometry corresponds to a reconstructed artery geometry from MRI or CT images, as during the measurements the blood flows with the pressure of at least 80 mmHg through the blood vessel. Due to the stiffening effects of the material only small deformations will occur if the pressure is increased further. If only the geometry from MRI or CT images is available one has to recompute the stress free geometry to be able to observe these minor deformations due to the stiffening of the elastin fibres. As the stress free reference domain is known here a priori, pre-stressing procedures as discussed in [111] do not need to be applied.

3.3.3 Initial Conditions

The system is initially at rest as we start the simulation with zero velocity, zero displacement and zero pressure at the boundaries. The FSI simulation is computed on the undeformed reference domain, see Section 3.3.1. In the NS case, the geometry is rigid and corresponds to the deformed pipe which we obtain in the FSI simulation at time 0.3. Thus the NS geometry already takes the deformation by the ramp and steady phases into account.

3.4 Numerical Approximation, Solution and Implementation

The realization of a numerical framework for monolithic fluid-structure interactions is very challenging and a detailed description is not possible in one manuscript. We therefore give a brief description and refer to the relevant and detailed literature. See [209] for a comprehensive overview.

We employ a very strict form of the ALE formulation, where all equations are solved on the reference domain. No mesh update is used. This prevents the necessity of projections between moving meshes and also it is the only straightforward approach for obtaining higher order discretization in time [211]. In principle this approach allows for direct Galerkin discretizations of the monolithic variational formulation (3.2.10) and the choice of finite element spaces should be based on the following considerations.

- The finite element mesh should resolve the fluid-structure interface $\mathscr{I}$ in reference framework.
- In the fluid domain, the velocity-pressure pair $V_h \times Q_h$ should fulfil the inf-sup condition to cope with the incompressibility constraint. Or, if a non-stable finite element pair is used, stabilization terms must be added. Our realization is based on equal order tri-quadratic elements for pressure and velocity, enriched with the local projection stabilization method [24].
- To reach a balanced approximation of the velocity-deformation pairing the same function space is used for the global deformation field.

To sum up, the discrete solution $U_h := (\mathbf{v}_h, \mathbf{u}_h, p_h)$ is found in the space $X_h = [V_h]^3 \times [V_h]^3 \times V_h^f$, where V_h extends over the complete domain and V_h^f over the fluid domain only.

In time we use a variant of the Crank-Nicolson time discretization scheme that gives better stability properties by an implicit shifting. We refer to [211] for details. By restricting (3.2.10) to the fully discrete setting, a system on nonlinear algebraic equations arises in each time step. As nonlinear solver we employ a Newton scheme with an analytical evaluation of the Jacobian, see [213] or [209, Section 5.2.2] for details on the derivation.

The resulting linear systems of equations are very large and extremely ill-conditioned with condition numbers that are by far larger than those of fluid and solid equation on their own, see [214, 12] for numerical studies. The approximation of these systems is still a great challenge, in particular if it comes to 3d applications. Only few fast solvers for the nonlinear setting are available [214, 12, 137]. Our approach is based on a multigrid solution that appears to be superior in 3d. In [96] we present the solution approach, which is based on a partitioning of the Jacobian based on two simple strategies:

1. Within the Navier-Stokes equations we neglect those parts of the Jacobian that come from the derivative with respect to the domain extension $\mathbf{u}_f$. The resulting nonlinear solver is an approximated Newton method that however still solves the original problem since the residual is exact. In [209, Section 5.2.3] and [96] we have found that such an approximated

Newton solver is even more efficient, despite slightly increased iteration counts. This is due to the very costly evaluation of the full Jacobian that is not required in our approach.

2. We exploit the discretization of the equation $d_t\mathbf{u}_s = \mathbf{v}_s$. Precisely, if we consider the Crank-Nicolson scheme

$$\mathbf{u}_s^n = \mathbf{u}_s^{n-1} + \frac{k}{2}\Big(\mathbf{v}_s^{n-1} + \mathbf{v}_s^n\Big)$$

we replace the dependency of the solid stress tensor on the deformation by the velocity, i.e. we replace the term $\mathbf{\Sigma}_s(\mathbf{u}_n^n)$ by

$$\mathbf{\Sigma}_s(\mathbf{u}_s) = \mathbf{\Sigma}_s\Big(\mathbf{u}_s^{n-1} + \frac{k}{2}\Big(\mathbf{v}_s^{n-1} + \mathbf{v}_s^n\Big)\Big).$$

This equivalent transformation allows us to completely remove the new solid deformation $\mathbf{u}_s^n$ from the discretized momentum equation, see [96].

The combination of these two modifications allows for a natural splitting of the Jacobian within the multigrid smoother. Further, it allows to apply very simple iterations of Vanka-type, as smoother in the geometric multigrid preconditioner, that are easy to parallelize. In [96, 95] we have demonstrated the efficiency of the approach in different 3d configurations. The implementation is based on the finite element toolkit Gascoigne3D [39].

In the context of hemodynamical fluid-structure interactions, a family of exact and inexact Newton methods based on a separate treatment of the physical coupling conditions, the interface location and the fluid- and solid-fields was previously studied in [185].

3.5 Simulations

We finally use the presented finite element framework for a computational analysis of several hemodynamic parameters that are relevant to answer clinical questions. We focus on the dependence of these parameters on the elasticity of the vessel walls. The additional effort of fully coupled fluid-structure interactions over pure Navier-Stokes simulations is immense and should be justified by corresponding effects.

The behavior of three specific parameters is investigated. The wall shear stress (WSS) plays an eminent role in several applications. It is used as indicator to model the growth of atherosclerotic plaques [245, 210] but also when it comes to deriving measures to evaluate the risk of plaque rupture [152] or the rupture of aneurysms [55]. Both the minimum wall shear stress and the distribution of the wall shear stress on the vessel walls are important measures. In Section 3.5.1 we analyze the effect of the different complexities, Navier-Stokes and fluid-structure interactions, on the WSS distribution in all three different geometries.

Furthermore we analyze the *computational fractional flow reserve* (cFFR). The *fractional flow reserve* (FFR) is a technique that measures pressure differences across a stenosis. In medical practice, the FFR is determined by inserting a catheter in the artery and measuring flow parameters at maximum blood flow. During the procedure, the tip of the catheter, where the sensor is located, is retrieved, such that measurements are available along the affected section of the vessel. Healthy vessels should give a pressure ratio close to 1. If the ratio drops below 0.8, i.e. a 20% drop in pressure, the stenosis is considered to be severe [230]. Aim of the *computational* *fractional flow reserve* is to replace the risky intervention by computer simulations based on medical imaging.

Finally, we discuss the amplitude of the pressure oscillation during one heart cycle. In clinical observation it is usually observed that the amplitude significantly drops after a stenotic region in a blood vessel. By comparing Navier-Stokes simulations with coupled fluid-structure interactions, we will show that this effect can only be described by considering the fully coupled model including elastic vessel walls.

3.5.1 Wall Shear Stress (WSS)

The tangential component of the surface force at the vessel wall is denoted as wall shear stress (WSS). By means of the Cauchy's theorem we have

$$\mathrm{WSS} = (\boldsymbol{\sigma}_f(\mathbf{v}, p)\vec{n} \cdot \boldsymbol{\tau})\boldsymbol{\tau} = [I - \vec{n}\vec{n}^T]\boldsymbol{\sigma}_f(\mathbf{v}, p)\vec{n}, \tag{3.5.1}$$

where $\boldsymbol{n}$ is a unit outward normal vector to the vessel wall and $\boldsymbol{\tau}$ i corresponding tangential vector. Note that WSS is a vector and its often confused with its magnitude $|\mathrm{WSS}|_2$ which is a scalar quantity denoted by wss.

The plots of the wall shear stress are presented on the boundary of the fluid domain, see Figure 3.3, Figure 3.4 and Figure 3.5. The surrounding solid is removed from these plots such that only the fluid domain is given.

We use the same scale ranging from 0 Pa to 20 Pa in all figures. The distribution of wss is always presented for 3 different points in time. First, when the pulsatile flow reaches its minimum inflow pressure, then, at maximum pressure and finally for an intermediate pressure value. Each specific situation is shown from two different angles, such that the inner and outer surfaces are well visible.

Comparing the coupled fluid-structure interaction model (left) with the pure Navier-Stokes flow (right) always shows much higher wss values in the Navier-Stokes case. Regions of very high wss are only found in the outside of the curve in the case of Navier-Stokes. In the case of the FSI model, the values are smoothly spread. In general, the vessels are widened and show a larger diameter of the lumen in the case of FSI. Due to computation of the Navier-Stokes simulations on the deformed domain from the FSI simulation after the ramp phase (3.3.1) with pressure of 80mmHg and the stiffening effects of the material law the diameter between NS and FSI only varies slightly.

Figures 3.4 and 3.5 show a shift of the wall shear stress distribution. High values are found all around the stenosed parts, on the inside and the outside of the curved areas. Again, the Navier-Stokes model is not able to yield an equal distribution of wss around the cylinder surface and it only concentrates on the outside of the curved section.

To sum up, it appears to be important to consider elastic fluid-structure interactions, if the spatial distribution of the wall shear stress is of interest. The Navier-Stokes model is nearly unaware of the stenosis and always concentrates the WSS on the outside of the curved region. If rupture locations [55] or localized growth processes are of interest [245, 210], the use of FSI is essential.

	description	*FSI*	*NS*
Geometry 1	without stenosis	0.99	0.96
Geometry 2	non-symmetric stenosis	0.96	0.94
Geometry 3	symmetric stenosis	0.96	0.95

Table 3.2: Values of cFFR analysis in all three geometries, considering the fully coupled fsi model and the pure Navier-Stokes case.

3.5.2 Computational Fractional Flow Reserve (cFFR)

The *computational fractional flow reserve* (cFFR) is the ratio of the maximum blood pressure after the stenosis (distal) and before the stenosis (proximal). We denote by p_d the distal pressure and by p_a the proximal pressure such that the computational fractional flow reserve is defined by

$$\text{cFFR} = \frac{p_d}{p_a}.$$

In our configuration, see Figure 3.1, we evaluate the distal and proximal values in

$$p_a = p(0,0,0) \text{ and } p_d = p(1.7,1,0). \tag{3.5.2}$$

Since the cFFR varies slightly with time the maximum value over the cardiac cycle is computed. Further computational aspects of cFFR are covered by the recent benchmarking paper [54]. In healthy vessels we expect cFFR ≈ 1 and in the medical practice, a stenosis with cFFR > 0.8 is considered to be functionally non significant [230]. No medical intervention, e.g. by placing a stent would be required.

The results of cFFR computations are presented in Table 3.2. The first striking observation is that the Navier-Stokes model is not able to reflect the presence of the stenosis at all. A pressure drop of about 5% is observed for all configurations and is only due to the curvature of the domain. The FSI model is well able to yield cFFR ≈ 1 in the case of the healthy configuration and shows a loss of about 5% in both stenotic geometries. Based on these simulations, the stenosis would not be considered to be severe and in the need of an intervention. These results clearly show that a pure Navier-Stokes simulation is not able to serve as computational basis for replacing the medical FFR procedure by simulations.

3.5.3 Pressure Amplitude

The third quantity of interest is the dynamic pressure amplitude before and after the stenosis, i.e., a temporally resolved analysis of the proximal and distal pressures p_a and p_d evaluated in the coordinates as mentioned in (3.5.2). We study the progress of the pressure to investigate possible damping effects of a stenosis. The change of the pressure profile after a strongly stenotic region, is reported as an assessment tool for cardiovascular risk [203].

We show the evolution of the pressures $p_a(t)$ and $p_d(t)$ over the third heart beat in Figure 3.6. On the left, we show results for the fully coupled FSI models, on the right we give the corresponding pressure lines for the Navier-Stokes case. Comparable to the cFFR study given in the previous section, the most striking observation is the invariance of the Navier-Stokes solution to the kind of vessel geometry. For all cases, healthy vessel, centered stenosis and non symmetric stenosis,

	Navier-Stokes			FSI		
	proximal	distal	drop	proximal	distal	drop
Geometry 1	39.2	32.7	7.5	32.1	31.4	0.7
Geometry 2	41.0	33.4	7.6	35.8	30.5	5.3
Geometry 3	41.3	33.6	7.7	35.8	30.4	5.4

Table 3.3: Pressure Amplitude in mmHg. We compare the drop of the amplitude over the stenotic region and compare the results for pure Navier-Stokes flow (left) with fully coupled fluid-structure interactions (right).

the Navier-Stokes results are nearly identical and always indicate a loss in pressure amplitude of about 7.5%. In contrast, the FSI solution is able to better preserve the oscillation in the case of the healthy vessel and shows a drop of about 5% for the two stenotic cases. A closer look reveals that the distal pressure lines are very similar in all cases. The slight oscillations in the distal pressure in case of the FSI simulation are not numerical instabilities. Instead they stem from the oscillatory behavior of the elastic vessel wall.

Finally, Table 3.3 collects the amplitudes before and after the stenosis. This third study also shows enormous qualitative and quantitative discrepancies between the simulation results depending on the model under consideration.

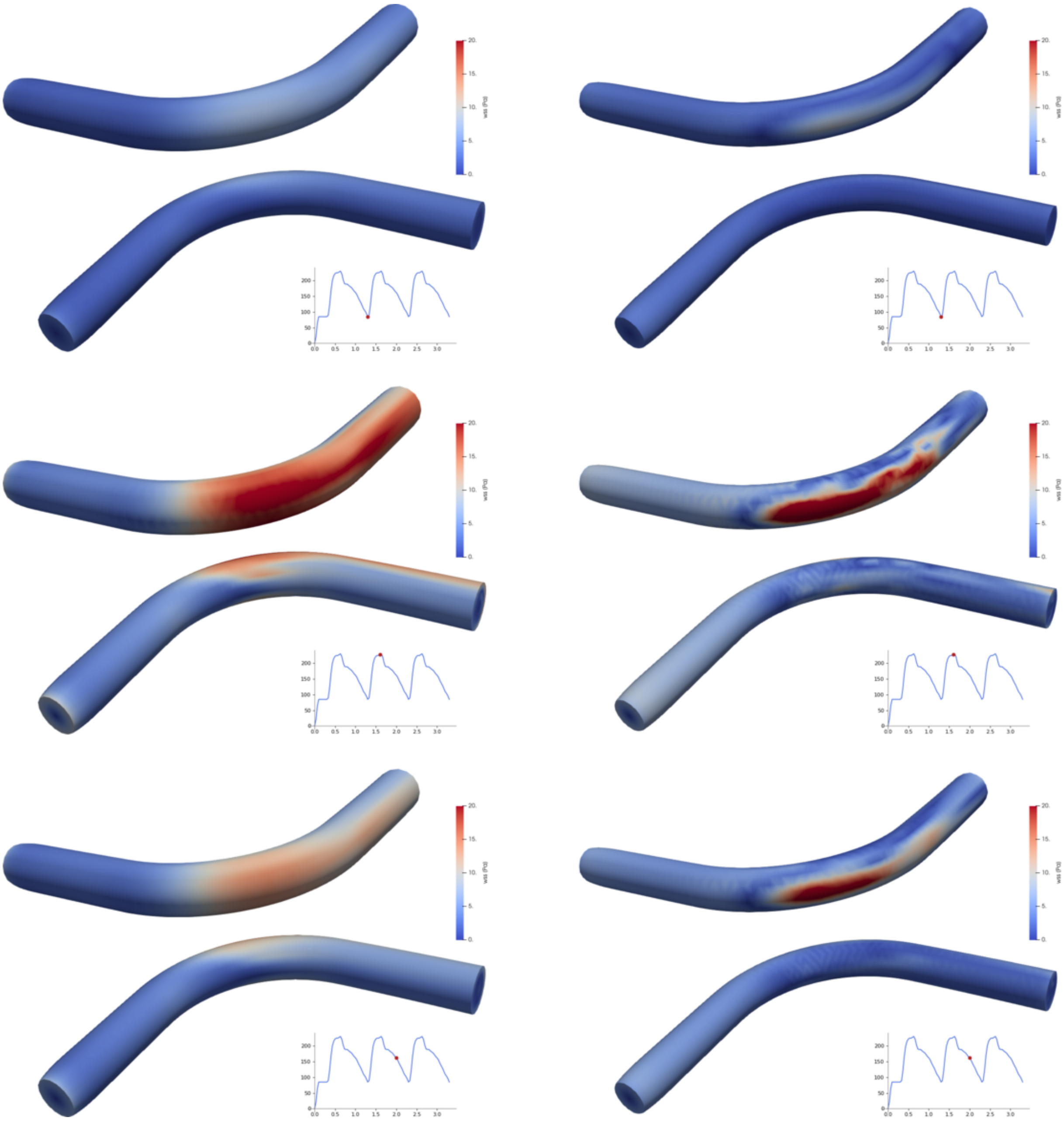


Figure 3.3: Wall Shear stress for Geometry 1 (without stenosis). On the left FSI, on the right NS. We show each configuration from two different perspectives, such that inner and outer region of the curved arch are visible. Small oscillations in the NS case are visual effects due to projections of the wall shear stress onto the curved boundary only.

https://youtu.be/YBufeDCtPro https://youtu.be/b7ac-hhMPJA

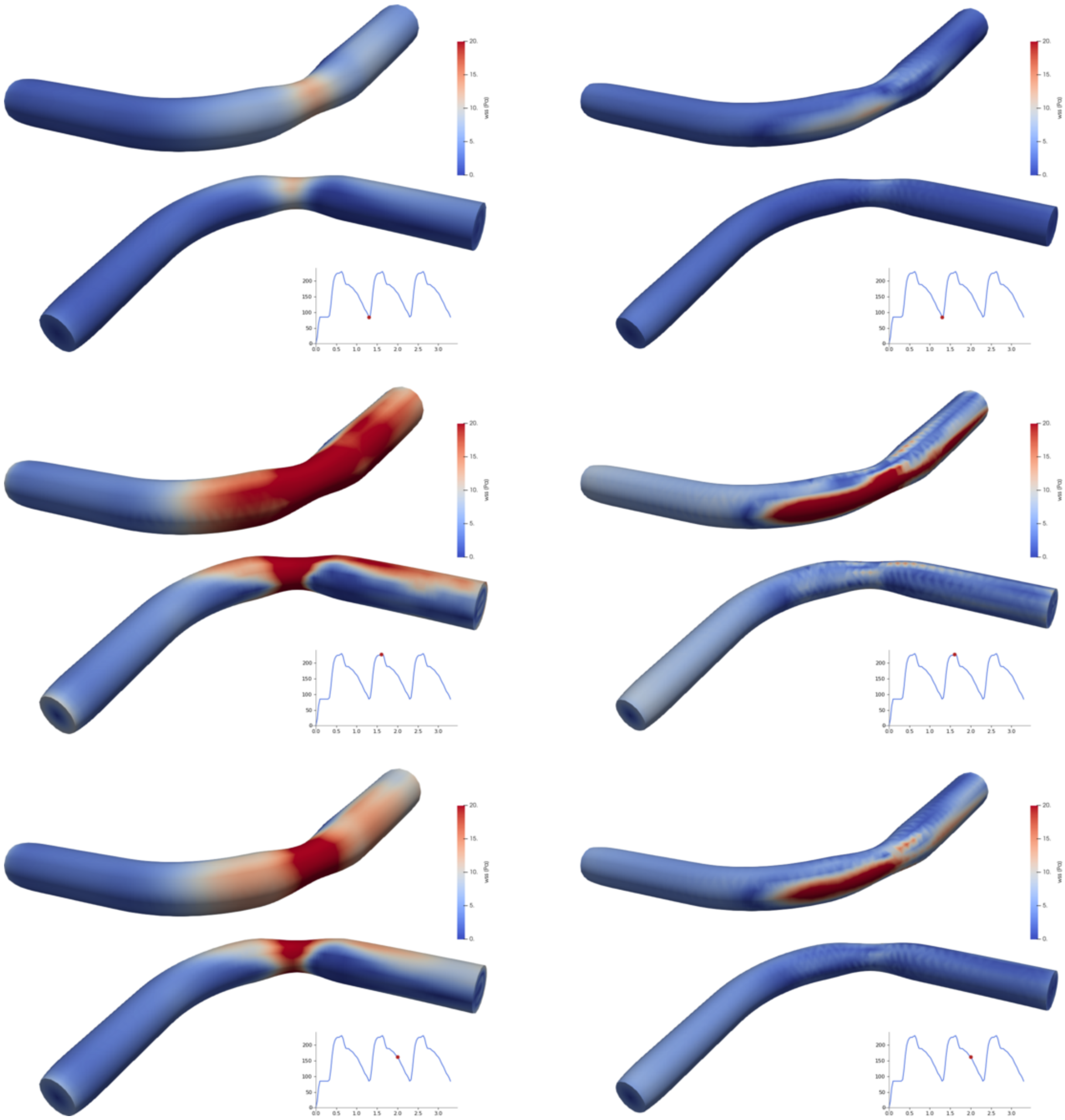


Figure 3.4: Wall Shear stress for Geometry 2 (symmetric stenosis). On the left FSI, on the right NS. We show each configuration from two different perspectives, such that inner and outer region of the curved arch are visible.

https://youtu.be/dskynFT6iqQ https://youtu.be/IIiyHBq7UmE

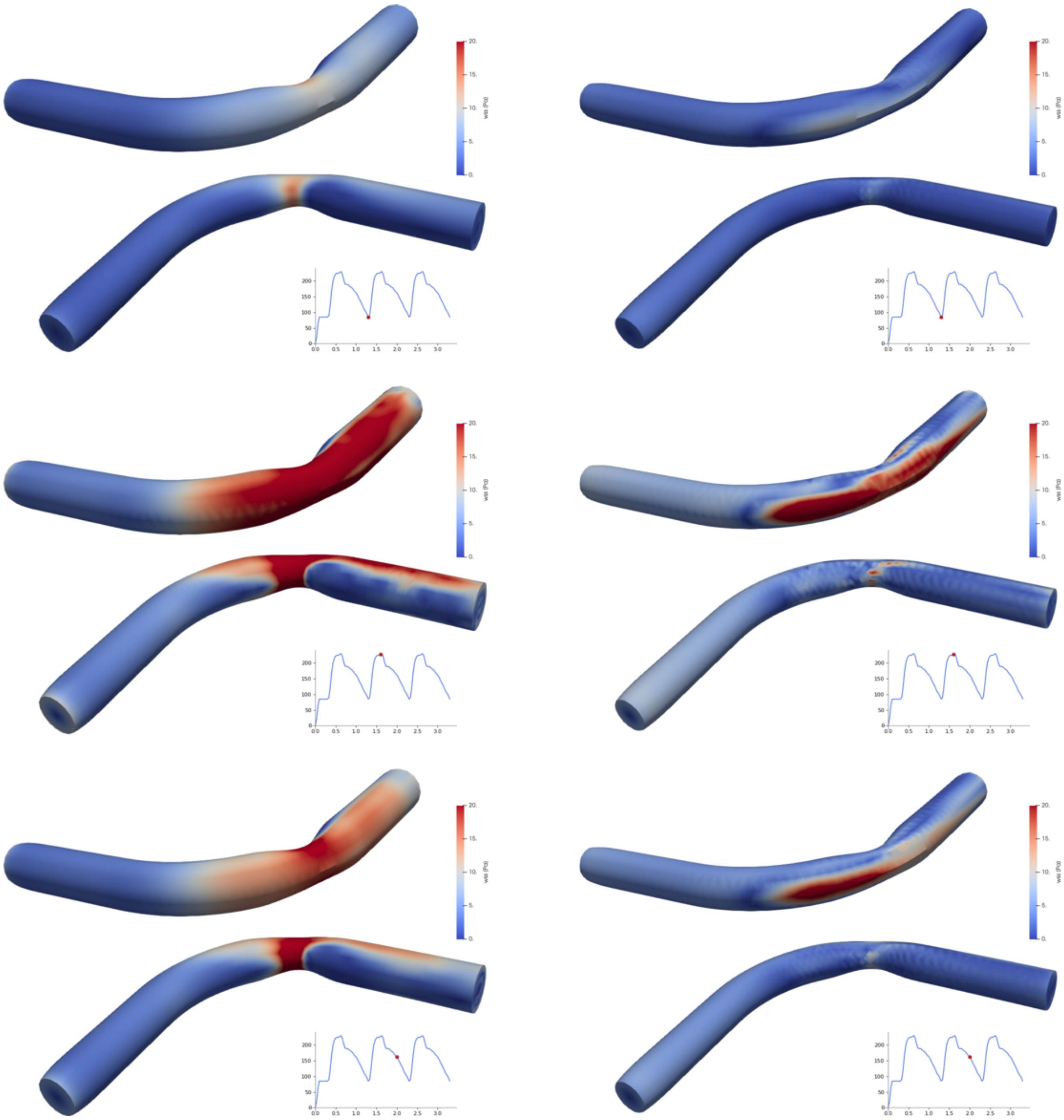


Figure 3.5: Wall Shear stress for Geometry 3 (non-symmetric stenosis). On the left FSI, on the right NS. We show each configuration from two different perspectives, such that inner and outer region of the curved arch are visible.

https://youtu.be/eGqmM7iYMwY https://youtu.be/rkQmzF-dFIo

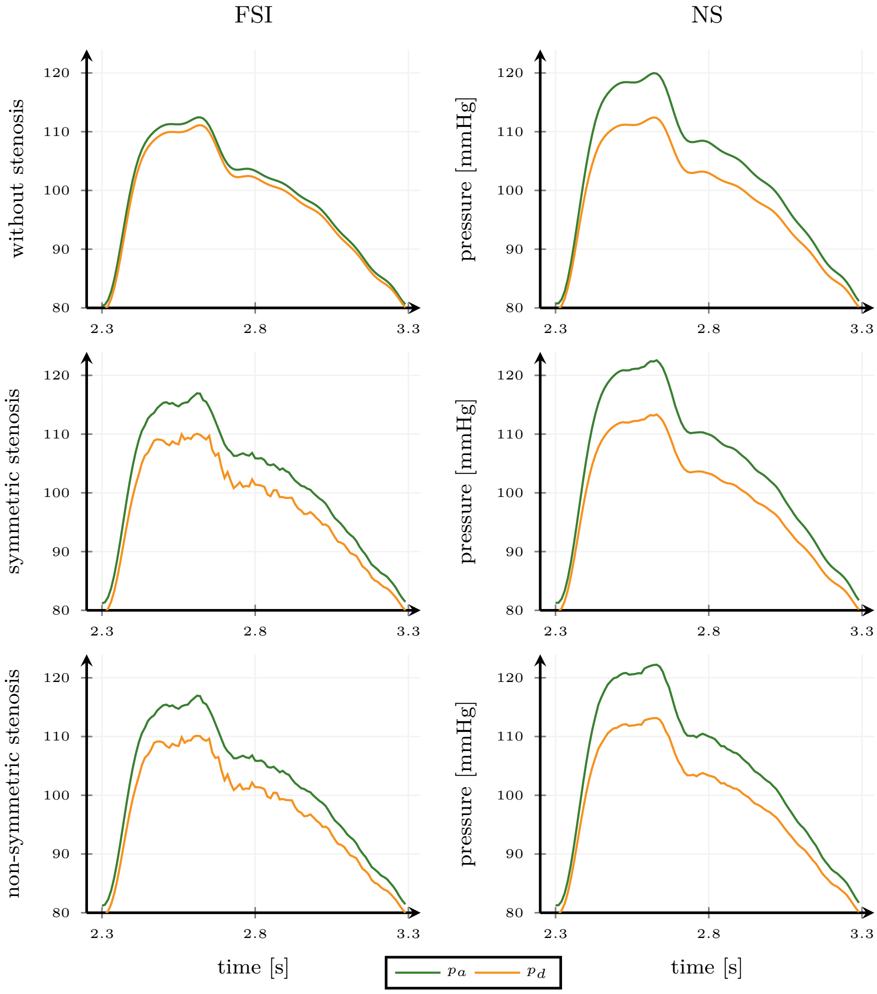


Figure 3.6: Distal and proximal pressure for one cardiac cycle

CHAPTER 4

Patient-specifc Simulation of Blood Flow in Cephalic Arch

The content of this chapter is joint work with Lukas Failer and Thomas Richter. We present research results of the project “Simulation and Optimization of Blood Flows in damaged Vessels” founded by the Federal Ministry of Education and Research of Germany, grant number 05M16NMA. In what follows, we refer to this grant as a project.

Chapter Summary. We develop a prototypical computational tool that supports clinical decision making for patients with stenosis. Blood flow is simulated as Newtonian fluid and Fluid-structure interactions on patient-specific geometry that is reconstructed from medical images.

Focus is put on the cephalic arch in patients with end-stage renal disease. These patients require a surgical arteriovenous connection for dialysis treatment. As a result, it changes flow characteristics in the cephalic vein that causes abnormal blood flow and, in consequence, cephalic arch stenosis (CAS).

We perform virtual stenting and evaluate the effect of placing stents on flow characteristics and hemodynamical factors. Finally, we discuss the application of adapted patient-specific simulation to the evaluation of surgical treatment.

Contents of Chapter

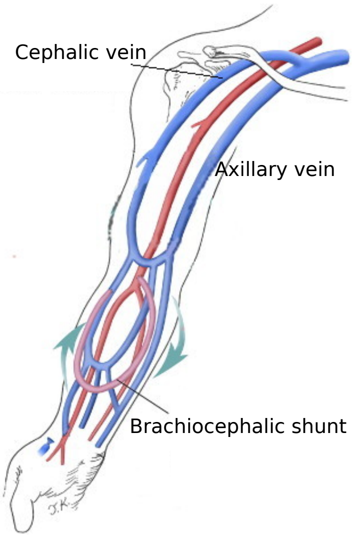


Figure 4.1: Anatomy of the cephalic arch and brachiocephalic shunt.

4.1 Introduction

We study patient-specific blood flow patterns in the cephalic arch for patients with an arteriovenous shunt. The cephalic arch is the final segment of the cephalic vein before its' confluence to the axillary vein to form the subclavian vein. A brachiocephalic shunt or fistula is an artificial connection constructed between brachial artery and cephalic vein usually at the level of the distal forearm, see Figure 4.1 for an anatomic sketch.

The construction of efficient and long-lasting artificial vascular accesses is one of the crucial points in long term treatment for patients suffering from end-stage renal disease (ESRD). When an arteriovenous shunt is placed, the cephalic vein's pressure and flow rates exceed the local physiological range by orders of magnitude. Alteration of the flow conditions is often associated with complications caused by the abnormal flow conditions In consequence, vascular access may fail to provide adequate conditions for efficient hemodialysis. Brachiocephalic shut falls prey to thrombotic incidents, mainly due to excessive neointimal hyperplasia and, consequently, to stenosis localized in the cephalic arch. For a detailed medical description of cephalic arch stenosis CAS we refer to [208, Chapter 19].

CAS is commonly treated with percutaneous transluminal angioplasty (PTA) and stenting. This procedure involves using a catheter with a balloon to widen narrowed or obstructed arteries or veins. Later if necessary, the stent is placed to widen the narrow vessel permanently. See Figure 4.2 for three intravascular images, before and after PTA and after stenting.

Patient-specific blood flow simulations in the cephalic arch is a challenging problem. The vein wall geometry and its' mechanical properties vary between patients, and they change in time due to damage and overgrowth inflicted by the altered flow conditions. In addition, the cephalic vein is artificially embedded in the arterial part of the circulatory system when an arteriovenous shunt is placed. This means that appropriate inflow and outflow conditions must be found. We address these challenges in the following sections.

Our study aims to develop a prototype of a computational tool that supports clinical decision making. A precise knowledge of the local flow conditions in the context of physiological alterations enables the appropriate adjustment of surgical treatment, e.g. the choice of stents that are used to cure the stenosis. Since cephalic arch stenosis occurs only in patients treated with hemodialysis, no specific stents have been designed deliberately for this anatomic configuration. Furthermore, the manufacturers have no recommendations on which stent could be appropriate for that particular application. The knowledge gained in the project enables the clinical partner to understand better the natural history of the progressive disorder and better assess the chances of success of different therapy options. In addition, the industrial project partners are now in a position to adapt their product range to diseases such as cephalic arch stenosis in the context of individualised medicine.

The application of computational fluid dynamics (CFD) to blood flow in cephalic arch has been studied in [3], where the authors showed that a well-defined physiological WSS target range can accurately predict the durability of patient-specific access. Similarly in [118] it has been demonstrated that fistula failure is associated with the changes in wall shear stress resulting from post-fistula alteration of hemodynamic characteristics that occur in the cephalic arch.

Commercial software solutions are currently not available to simulate the mathematical models used here. Therefore, an essential part of the project was the extension of the finite element toolkit Gascoigne [39] to compute patient-specific blood flow conditions with the necessary accuracy in adequate time. In this chapter, we describe the prototype tool that may be further developed to support clinical decision-making process.

4.2 Medical Data and Domain Reconstruction

An essential part of the project was the creation of a geometry of the cephalic arch. There is currently no commercially available software that can automatically perform this task.

In this section we describe the reconstruction of the vein geometry from medical images. The position of the centerline of the vein is determined based on two X-ray images, see Figure 4.3. The diameter of the vein is reconstructed from intravascular ultrasound IVUS images, see Figure 4.2.

Data Source Anonymised medical data presented in this chapter have been obtained by project partner OA Dr. Vladimir Matoussevitch from University Hospital of Cologne.

4.2.1 Intravascular Ultrasound Images

During the surgery IVUS images are used to determine the exact diameter of the stenosis. The IVUS device is pulled back along a wire. In the middle of the IVUS picture one can see the tip of the device. Around it, there is the black lumen of the vein surrounded by walls that we see as white. The reconstruction is based on the representation of an image in polar coordinates. We detect the vessel wall based on the steep gradient of the color value. Irregularities in the detected points (red) are smooth out by fitting an ellipse with the least squares method (blue). The thickness of the vessel wall is fixed, see the vessel outer boundary (orange, colors refer to Figure 4.2). Moreover, note that the clearly visible white circle on the right picture of Figure 4.2 is a stent.

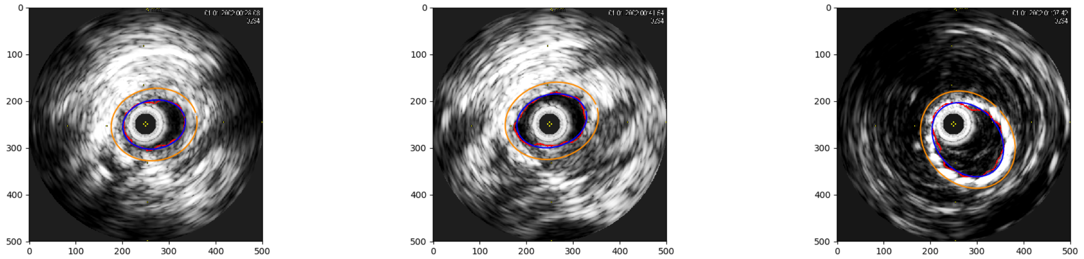


Figure 4.2: Intravascular ultrasound images for three cases: before PTA (left), after PTA (center), and with stent (right).

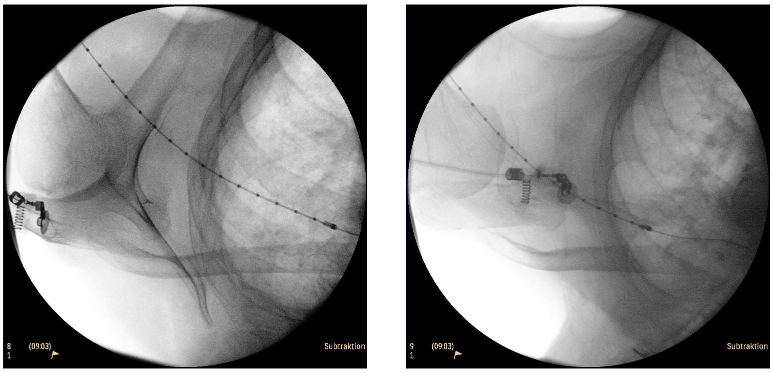


Figure 4.3: Position of the wire for IVUS from top(left) and side(right).

4.2.2 X-Ray Images

When the IVUS wire is placed into the patient X-ray images from the top and the side are available, see Figure 4.3. The IVUS pictures contain only the information about the diameter of the vein. However, we don not know a three-dimensional position of the wire. Two X-ray images enables us to reconstruct the vein centerline.

The position of the second image from the side is rotated around the x-axes by 80°. For every marker we obtained the projected position in two planes. This enables us to reconstruct the position in three-dimensions (v_x, v_y, v_z). We can directly determine v_x and v_z from one image and recompute v_z via transformation of the values in the second image in the coordinate system of the first image

$$v_z = sin(\theta) v_{y'} + \frac{cos(\theta)}{sin(\theta)} v_{y'} - \frac{1}{sin(\theta)} v_y. \tag{4.2.1}$$

In the original images a value of 60 px corresponds to 1 cm. Thereby we get a distance between the markers of 1 cm. Moreover, we assume that the IVUS device is pulled back smoothly with a constant velocity and we determine the final position of the IVUS via an additional x-ray at final time. Finally, we place the IVUS images equidistantly distributed orthogonal to the wire and perform additional smoothing in the direction of the wire to obtain final reconstruction of the vein surface.

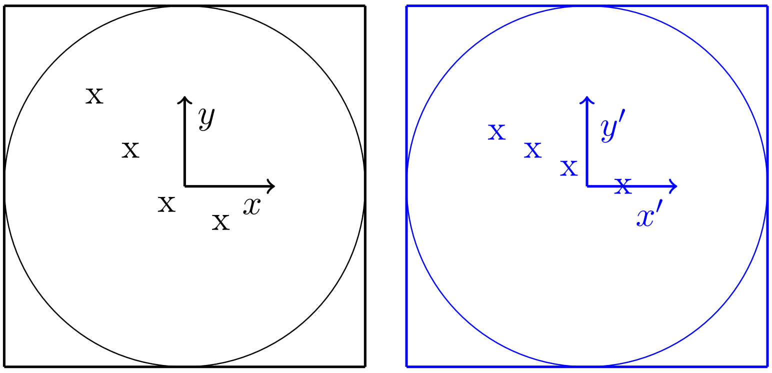


Figure 4.4: Position of the markers in images from Figure 4.3 in two coordinate systems (center at the center of the image).

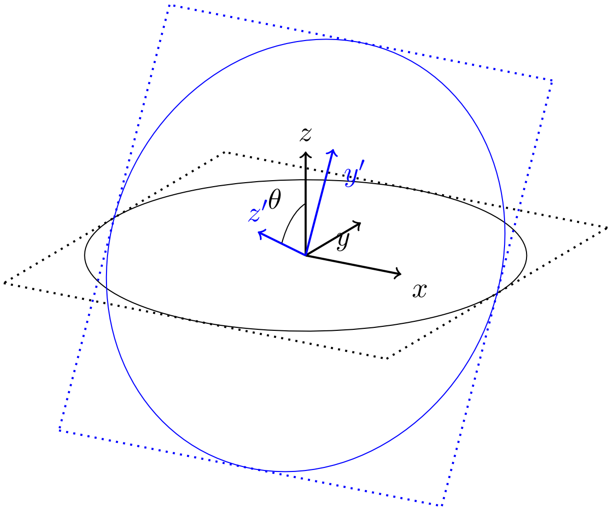


Figure 4.5: Image 2 is rotated around the x-axis via the angle θ.

4.2.3 Meshing

The next step of the workflow is a creation of computational mesh. We construct a coarse finite element grid of hexahedra, based on the centerline of IVUS catheter. The wire is placed in the cephalic vein and the outflow part of the axillary vein. The inflow part of the axillary vein is added artificially. During refinement the coarse grid is adapted to the surface geometry.

Reconstruction Error The work on domain reconstruction has led to the conclusion that there exist significant geometry uncertainty in patient-specific simulations. We addressed this issue in Chapter 2 and [169].

4.3 Model of the Vein

In Chapter 3 we have developed a stenotic coronary artery benchmark. For a simulation of the cephalic vein, previous models that described arteries had to be adapted to the special configuration of the vein. We refer to Section 3.2 for a detailed discussion on the artery modelling.

The derived hyperelastic stress-strain relation (3.2.6) models an approximately incompressible material and a stiffening effect due to collagen fibres. Since veins have a similar construction as arteries the model is still valid. However, one has to choose appropriate material parameters.

To determine the material values for the vein in the subclavian and cephalic vein we adapt the parameters μ and γ such that the behavior matches the tube law presented in [183]

$$p(x,t) = p_e(x,t) + K(x)\phi(A, A_0), \tag{4.3.1}$$

where

$$\phi(A, A_0) = \left(\frac{A}{A_0}\right)^m - \left(\frac{A}{A_0}\right)^n,$$

with practical choices

$$K(x) = K_v = \frac{c_0^2 \varrho_s}{m-n}, \quad m \approx 10, \quad n = -\frac{3}{2}.$$

Here $p_e(x,t)$ prescribed the reference pressure and $A_0(x)$ the vessel cross-sectional area at equilibrium. The wave speed for the subclavian vein is assumed to be 1.437 m/s. This corresponds to $K_v = 197.5\,\text{Pa}$. For the parameter study we use a tube of length 10 cm with radius 0.5 cm and wall thickness 0.1 cm. These values are in the same range as the values used for subclavian vein. At the in and outlet we smoothly increase the pressure in the time interval $I = [0, 1\,\text{s}]$ and observe the change in the domain in the middle of the cylinder. We set the reference pressure to 0 mmHg and assume the cylinder to be at rest.

According to the parameter study, see Figure 4.6, we use $\mu = 20\,\text{kPa}$, $\gamma = 30$ and $\kappa = 200\,\text{kPa}$. These values are in the same range as the parameters determined in [82]. According to [107] veins contain a relatively high amount of collagen; the elastin/collagen ratio is about 1 : 3. Hence the exponential stiffening effects are much higher in veins then in arteries. Therefore, the larger value of γ is appropriate.

For the remaining parameters we recall values from Chapter 3, i.e. the fluid viscosity of $\mu_f = 0.033\,\text{g}\cdot\text{cm}^{-1}\text{s}^{-1}$, the density $\varrho_f = 1\,\text{g}\cdot\text{cm}^{-3}$, and the solid density equals $\widehat{\varrho}_s = 1.2\,\text{g}\cdot\text{cm}^{-3}$.

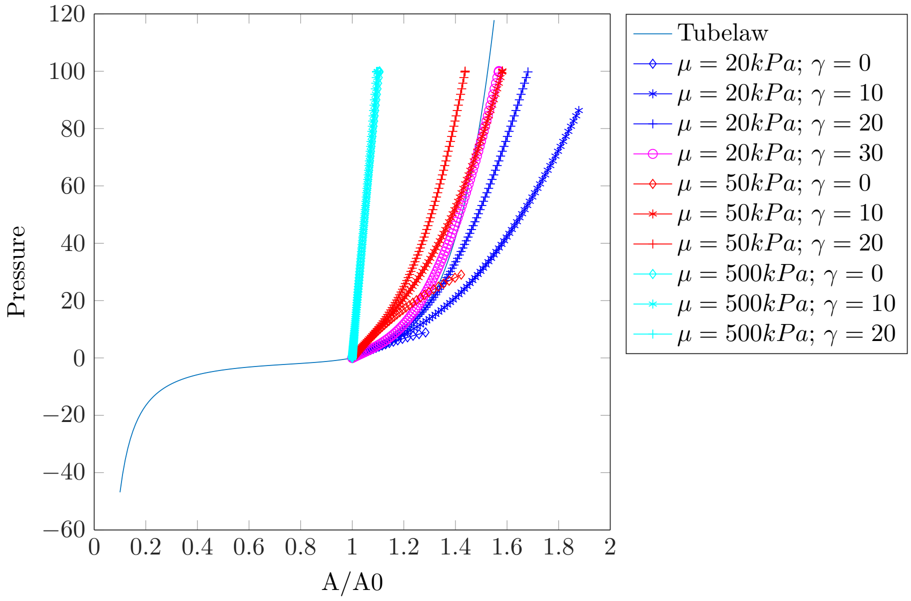


Figure 4.6: The relation of the change in area with respect to pressure.

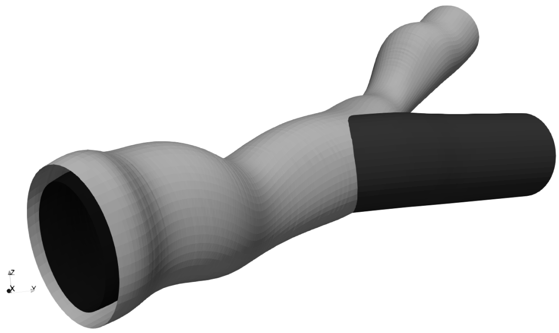


Figure 4.7: Vein surface reconstructed from IVUS and X-ray images.

4.4 Simulations Setting

We perform pulsatile blood-flow simulations, on the patient-specific computational mesh. Similarly to Chapter 3 we perform comparison between rigid and elastic vessel walls. Additionally the mesh is modified to reflect geometry after the stent is placed.

Boundary and Initial Conditions

The computational effort of three-dimensional flow models only allows the calculation of flow in a small area around the cephalic arch stenosis. Suitable pressure or flow rates must then be prescribed at the edge of the simulation area. However, these are not directly available from measurement data. Usually only average values are known or the measurements are taken in areas far from the inflow/outflow boundaries. In addition, the measurement data are usually highly error-prone and cannot be used directly for simulations.

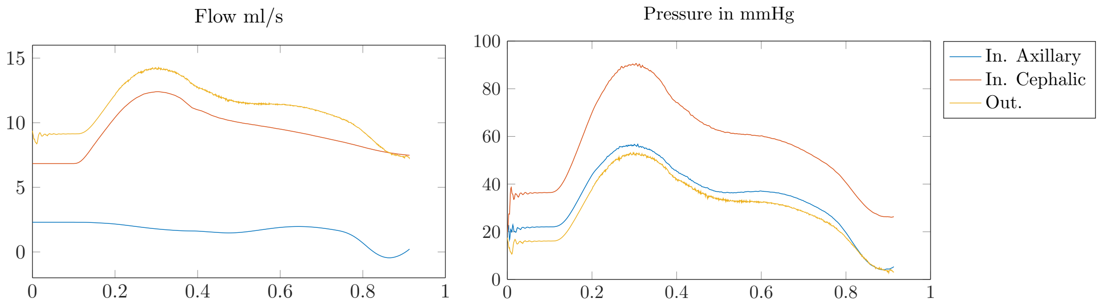


Figure 4.8: Boundary values obtained from one-dimentioal whole body circulation model.

	Navier-Stokes	FSI
WSS_{min}	$0.042\,\text{Pa} = 3.17 \cdot 10^{-4}\,\text{mmHg}$	$0.068\,\text{Pa} = 5.09 \cdot 10^{-4}\,\text{mmHg}$
WSS_{max}	$1.176 \cdot 10^{2}\,\text{Pa} = 0.88\,\text{mmHg}$	$1.018 \cdot 10^{2}\,\text{Pa} = 0.76\,\text{mmHg}$

Table 4.1: Values of WSS for stenotic cephalic arch geometry. We compare the results for Navier-Stokes flow (left) with fully coupled fluid-structure interactions (right).

Boundary values were obtained from a customized one-dimensional whole body circulation model developed by Computational Life [64]. The considered scenario is non-standard, since the vein is directly connected via the shunt with the artery. Thus, the customization was made to reflect this abnormal connection. In Figure 4.8 we present the values of the flow rate and the pressure for one heart cycle.

Implementation The realization of a numerical framework for monolithic fluid-structure interactions has been briefly discussed in Section 3.4, see also [209] for a comprehensive overview. The implementation is based on the finite element toolkit Gascoigne3D [39].

4.5 Wall Shear Stress

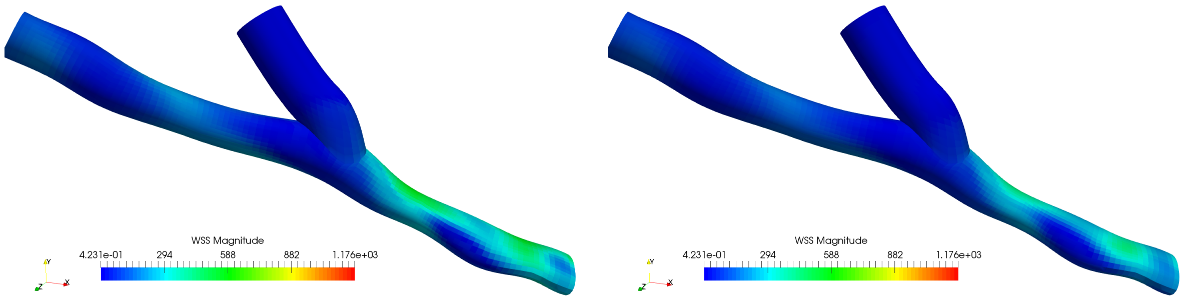


Figure 4.9: WSS for patient-specific geometry with stent at time $t = 0.3$ (left) and $t = 0.6$ (right).

Let us recall, that the wall shear stress (WSS) plays an important role in several applications, c.f. Section 3.5.1. It is used as an indicator to model the development of stenosis. We recall that

WSS is defined as the tangential component of the surface force at the vessel wall, c.f. 3.5.1.

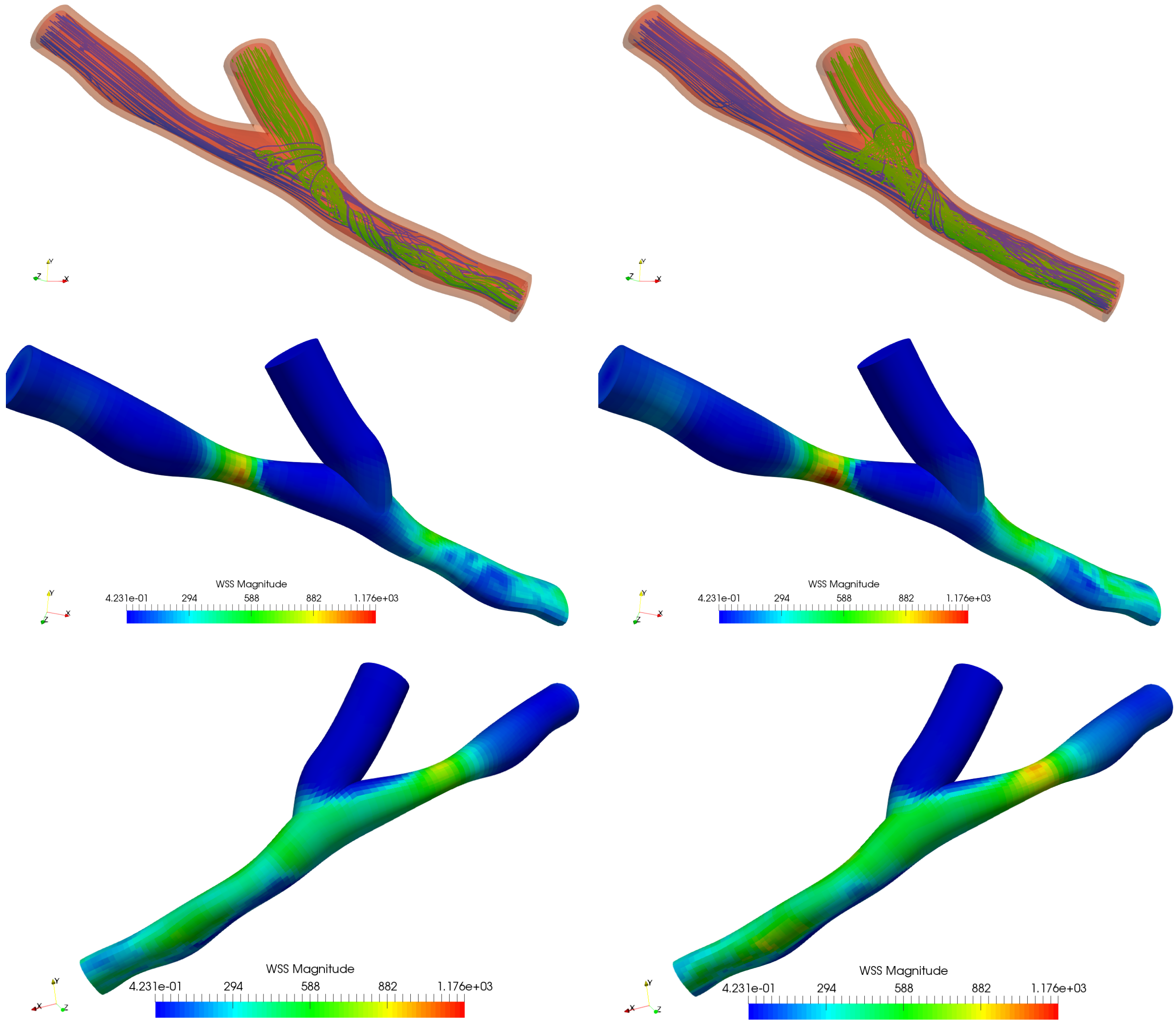


Figure 4.10: Wall shear stress for patient-specific geometry with stenosis, at time $t = 0.3$. On the left FSI, on the right NS. For both models we show streamlines (top) and WSS from two different perspectives (middle and bottom). A video of the FSI simulation of the blood flow in the cephalic arch is available at `https://youtu.be/6JspqmnnCQc`.

We compare fully coupled fluid-structure interactions with Navier-Stokes simulations in patient-specific geometry. The plots of the wall shear stress are presented on the boundary of the fluid domain, see Figure 4.10. The surrounding solid is removed from these plots such that only the fluid domain is given.

We observe the distribution of the WSS the vessel walls from two different angles with large values in the vicinity of stenosis. Note the localisation of WSS at the outer boundary. Similar bahaviour has already been observed for coronary artery benchmark.

In Table 4.1 minimal and maximal values of WSS are presented. The use of FSI model reduces WSS by 14%.

We conclude that it is important to consider elastic fluid-structure interactions, if the spatial distribution of the WSS is of interest. Computational tools based on the Navier-Stokes model tend to overestimate the WSS values.

4.6 Virtual Stenting

Based on the reconstructed surface of the vein geometry, we develop tools that enable simulation of an intervention, such as a virtual stenting. The geometry is altered in order to represent the vein with a stent.

The computations are performed for rigid walls, see Figure 4.11. We observe that stented geometry reduces maximal flow rate. Comparing Figure 4.9 with Figure 4.10 shows significant reduction of the WSS in the formerly stenosed arena of the cephalic vein.

Virtual stenting procedures can be performed for the range of stent geometries. In such a way the results of simulation could be introduced for the support of individualised therapy.

4.7 Support of Clinical Decision Making

Computer simulations have become an integral part of product development and testing in mechanical engineering. This allows modifications to be implemented virtually quickly and cost-effectively and potential weak points to be found. In medicine, simulations are still very rarely used in investigating diseases and developing possible therapies. However, they could contribute to a great gain in knowledge. For example, surgical interventions could be adapted virtually and information from CT and MRI could be processed further.

The central object of the project work was the development of new numerical approximation methods as well as the implementation of powerful software components on high-performance computers so that realistic configurations can be simulated using complex and detailed mathematical models. One of the main mathematical difficulties encountered in the simulations of such complex flow processes are the grid generation, which must lead to hierarchically nested computational grids in order to be coupled with the fast multigrid solvers within the CFD tools.

The project aimed to evaluate precisely these possibilities in the field of vascular diseases. As an example, a cephalic arch stenosis in dialysis patients was examined more closely.

Accurate knowledge of the flow allows for the adjustment of stents used to treat the stenosis. Because cephalic arch stenosis is a complication that only occurs in dialysis patients, no specific stents have been developed for this configuration. Furthermore, there are no recommendations from the manufacturers as to which stent developed so far can be optimally used for this stenosis.

The knowledge gained in the project enables the clinical partner to better understand the course of the disease and to better assess the chances of success of different therapy options. In addition, the industrial project partners are now in a position to adapt their product ranges to diseases such as cephalic arch stenosis in the sense of individualised medicine.

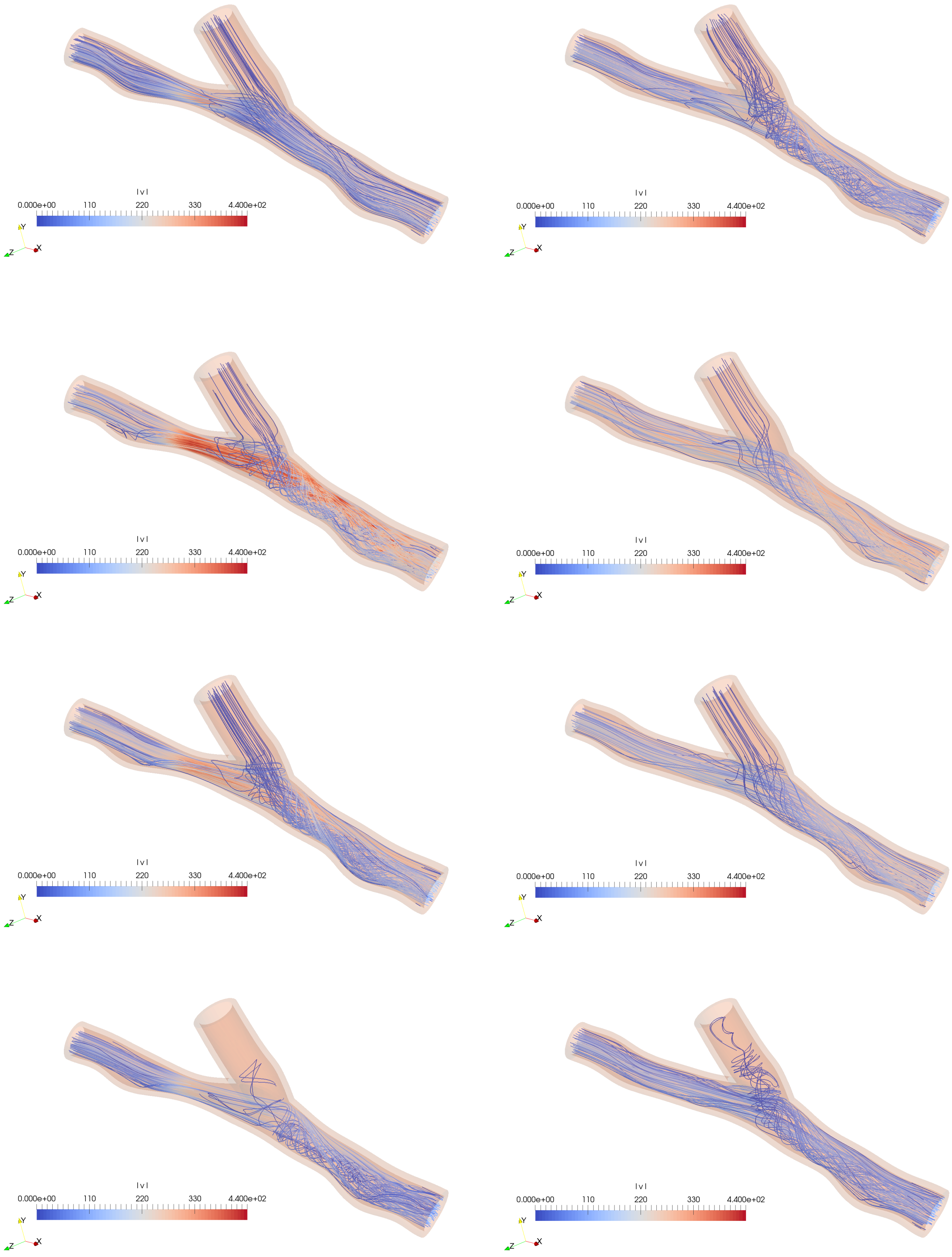


Figure 4.11: Wall Shear stress for patient-specific geometry with stenosis, times $t = \{0.0, 0.3, 0.6, 0.9\}$. On the left FSI, on the right NS. For both models we show streamlines (top) and WSS from two different perspectives (middle and bottom).

https://youtu.be/fYX1BiZkYtE https://youtu.be/ea3LSbbl63g

CHAPTER 5

Transport of Congestion in Two-Phase Compressible/Incompressible Flows

The content of this chapter is joint work with Pierre Degond and Ewelina Zatorska, and is published in the paper

P. Degond, P. Minakowski and E. Zatorska. 'Transport of Congestion in Two-phase Compressible/Incompressible Flows'. In: *Nonlinear Analysis: Real World Applications* 42 (2018), pp. 485–510. doi: 10.1016/j.nonrwa.2018.02.001

Chapter Summary. We study the existence of weak solutions to the two-phase fluid model with congestion constraint. The model encompasses the flow in the uncongested regime (compressible) and the congested one (incompressible) with the free boundary separating the two phases. The congested regime appears when the density in the uncongested regime $\varrho(t, x)$ achieves a threshold value $\varrho^*(t, x)$ that describes the comfort zone of individuals. This quantity is prescribed initially and transported along with the flow. We prove that this system can be approximated by the fully compressible Navier-Stokes system with a singular pressure, supplemented with transport equation for the congestion density. We also present the application of this approximation for the purposes of numerical simulations in the one-dimensional domain.

Chapter Organization. In Section 5.2, we present details of approximation and prove the existence of solutions to the system (5.1.10) for ε fixed. Then, in Section 5.3, we recover the two-phase system (5.1.15) by letting $\varepsilon \to 0$. After this, in Section 5.4 we recover the solution to the original two-phase system (5.1.1). Finally, in Section 5.5 we briefly describe the numerical scheme and present computational examples that illustrate the behaviour of approximate solutions to the system (5.1.1).

Contents of Chapter

5.1 Introduction

Our aim is to analyze the free-boundary two-phase fluid system that could be used to model the congestions in the large group of individuals in a bounded area. Individuals are just the agents that have their own preferences for how close they let the closest neighbour to approach and they carry this information with them in the course of motion. They do not follow any neighbour trying to align their velocities, nor they are trying to reach a certain target, as for example, the evacuation point. We simply prescribe their initial velocity that determines their direction of motion and check how the individual preferences as well as the initial distribution of the agents determines creation of congestions. Such model could be used as a building block of more involved crowds modelling [161], some progress in this direction has been made in our recent numerical work [79].

Crowd modelling is a problem of strategic importance for safety reasons. It has been studied in many parallel approaches. We can distinguish, for example, the mean-field game models [86, 146], in which the individuals behave as the players following some strategy, or optimizing certain cost; the microscopic models which describe precise position and velocity of an individual (Individual-Based-Models) using Newtonian framework [114, 124, 233, 123]; or the macroscopic models formulated in the language developed for the fluids [52, 125, 76, 132, 202]. The behaviour of the crowd in the later is characterised by some averaged quantities such as the number density or mean velocity. The macroscopic models, although less precise than the microscopic ones, are computationally more affordable. Moreover, they allow for asymptotic studies that proved to be useful for understanding various aspects like: swarming or pattern formation observed in the experiments.

It is an extensive filed of research to develop continuum models that are able to exhibit features of the kinetic approach. Although it would be desirable to use computationally cheap fluid approach to describe the crowd dynamics, the nowadays models are not developed enough to recreate behaviour observed in real world. In this paper we present, as far as we know, the first mathematical result for the fluid model that incorporates various sizes of the individuals/particles and their inhomogeneities. Similar approach has been recently applied in [195] in the context of granular media flow with memory effects. In order to use this model for more specific applications, its current version needs to be supplemented with agents specific features and we postpone this topic to further research.

Our system writes as follows:

$$\partial_t\varrho + \operatorname{div}(\varrho\mathbf{u}) = 0, \tag{5.1.1a}$$

$$\partial_t(\varrho\mathbf{u}) + \operatorname{div}(\varrho\mathbf{u}\otimes\mathbf{u}) + \nabla\pi + \nabla p\left(\frac{\varrho}{\varrho^*}\right) - \operatorname{div}\mathbf{S}(\mathbf{u}) = \mathbf{0}, \tag{5.1.1b}$$

$$\partial_t\varrho^* + \mathbf{u}\cdot\nabla\varrho^* = 0, \tag{5.1.1c}$$

$$0 \le \varrho \le \varrho^*, \tag{5.1.1d}$$

$$\operatorname{div}\mathbf{u} = 0 \text{ in } \{\varrho = \varrho^*\}, \tag{5.1.1e}$$

$$\pi \ge 0 \text{ in } \{\varrho = \varrho^*\}, \quad \pi = 0 \text{ in } \{\varrho < \varrho^*\}. \tag{5.1.1f}$$

with the unknowns: $\varrho = \varrho(t,x)$ – the mass density, $\mathbf{u} = \mathbf{u}(t,x)$ – the velocity vector field, $\varrho^* = \varrho^*(t,x)$ – the congestion density, also referred to as the barrier or the threshold density, and π – the congestion pressure, that appears only when $\varrho = \varrho^*$.

The barotropic pressure is an explicit function of $\frac{\varrho}{\varrho^*}$

$$p\left(\frac{\varrho}{\varrho^*}\right) = \left(\frac{\varrho}{\varrho^*}\right)^\gamma, \quad \gamma > 1, \tag{5.1.2}$$

and plays the role of the background pressure.

The stress tensor $\mathbf{S}$ is a known function of $\mathbf{u}$, characteristic for the Newtonian fluid, namely

$$\mathbf{S} = \mathbf{S}(\mathbf{u}) = 2\mu\,\boldsymbol{D}(\mathbf{u}) + \lambda\operatorname{div}\mathbf{u}\,\boldsymbol{I}, \quad \mu > 0,\ 2\mu+\lambda > 0, \tag{5.1.3}$$

where $\boldsymbol{D}(\mathbf{u}) = (\nabla\mathbf{u} + \nabla^T\mathbf{u})/2$ denotes the symmetric part of the gradient of $\mathbf{u}$, and $\boldsymbol{I} = \boldsymbol{I}_3$ is the identity matrix.

In the system (5.1.1) variable ϱ^* models preferences of the individuals, it is given initially and then transported with the flow. Therefore, ϱ^* depends on time and position, but more importantly it depends on initial configuration ϱ_0^*. The form of ϱ^* relaxes the restrictions from the models studied in [43, 197], where the threshold density ϱ^* was either assumed to be constant or independent of time. This allows to cover more physical applications. Including the transport of the congestion density ϱ^* allows also to study the system (5.1.8) with the contribution from the pressure in the form of the pure gradient, without factor ϱ^* as it was done in [197].

It is justified to call the above system the two-phase system because for $\varrho(t,x) < \varrho^*(t,x)$ it behaves as the compressible Navier-Stokeas system with the barotropic pressure, while when the congestion is achieved, i.e. for $\varrho(t,x) = \varrho^*(t,x)$, the system behaves like the incompressible Navier-Stokes equations. We thus observe a switching between two phases: compressible and incompressible depending on the size of the density ratio $\frac{\varrho}{\varrho^*}$. The fluid systems with congestion constraints have been recently intensively studied, especially in the hyperbolic regime [32, 33, 38]. The first analytical result for system (5.1.1) with $\varrho^* = 1$ is due to P.-L. Lions and N. Masmoudi [153], who showed that it can be obtained as a limit of compressible Navier-Stokes equations with barotropic pressure ϱ^γ with $\gamma \to \infty$, similar studies were performed recently for the model of tumour growth [198, 237].

We will consider the system (5.1.1) in the 3-dimensional domain Ω with the smooth boundary $\partial\Omega$, and the Dirichlet boundary conditions for the velocity vector field

$$\mathbf{u}|_{\partial\Omega} = \mathbf{0}. \tag{5.1.4}$$

The initial conditions are given by:

$$\varrho(0,x) = \varrho_0(x), \quad (\varrho\mathbf{u})(0,x) = \mathbf{m}_0(x), \quad \varrho^*(0,x) = \varrho_0^*(x), \tag{5.1.5}$$

and we assume that they satisfy:

$$\begin{gathered} \varrho_0 \geq 0, \quad \int_\Omega \varrho_0 \ \mathrm{d}x > 0, \\ \mathbf{m}_0 = \mathbf{0} \text{ a.e. in } \{\varrho_0 = 0\}, \quad \frac{\mathbf{m}_0}{\varrho_0}\mathbf{1}_{\{\varrho_0>0\}} \in L^2(\Omega), \\ \varrho_0 \leq \varrho_0^*, \text{ a.e. in } \Omega, \quad \varrho_0 \not\equiv \varrho_0^*, \quad \varrho_0^* \in L^\infty(\Omega). \end{gathered} \tag{5.1.6}$$

Moreover, we assume that in the region of the absence of the individuals $\varrho_0(x) = 0$, the congestion density is equal to a constant value, being the characteristic mean preference of the group:

$$\varrho_0^*\Big|_{\{\varrho_0=0\}} = \widetilde{\varrho^*} > 0. \tag{5.1.7}$$

The main result of this paper is the existence of solutions to the system (5.1.1) under the aforementioned assumptions on the constitutive relations and the initial condition, in the sense of the following definition.

Definition 5.1 (*Weak solution*). A quadruple $(\varrho, \mathbf{u}, \varrho^*, \pi)$ is called a global finite energy weak solution to (5.1.1), (5.1.4), with the initial data (5.1.5), (5.1.6), (5.1.7) if for any $T > 0$:

(i) There holds:

$$\begin{gathered} 0 \leq \varrho \leq \varrho^* \quad \textit{a.e. in } (0,T)\times\Omega, \qquad \mathbf{u}|_{(0,T)\times\partial\Omega} = \mathbf{0}, \\ \operatorname{div}\mathbf{u} = 0 \quad \textit{a.e. in } \{\varrho = \varrho^*\}, \qquad (\varrho^* - \varrho)\pi = 0, \end{gathered}$$

and

$$\begin{aligned} &\varrho \in C_w([0,T]; L^\infty(\Omega)), \\ &\varrho^* \in C_w([0,T]; L^\infty(\Omega)), \\ &\mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega,\mathbb{R}^3)), \quad \varrho|\mathbf{u}|^2 \in L^\infty(0,T; L^1(\Omega)), \\ &\pi \in \mathscr{M}^+((0,T)\times\Omega). \end{aligned}$$

(ii) For any $0 \leq \tau \leq T$, equations (5.1.1a), (5.1.1b), (5.1.1c) are satisfied in the weak sense, more precisely:

$$\int_\Omega \varrho(\tau,\cdot)\varphi(\tau,\cdot) \ \mathrm{d}x - \int_\Omega \varrho_0\varphi(0,\cdot) \ \mathrm{d}x = \int_0^\tau\!\!\int_\Omega \Big(\varrho\partial_t\varphi + \varrho\mathbf{u}\cdot\nabla\varphi\Big) \ \mathrm{d}x \ \mathrm{d}t,$$

holds for all $\varphi \in C^1([0,T] \times \overline{\Omega})$,

$$-\int_\Omega \mathbf{m}_0 \cdot \boldsymbol{\psi}(0,\cdot)\ \mathrm{d}x = \int_0^\tau\!\!\int_\Omega (\varrho\mathbf{u}\cdot\partial_t\boldsymbol{\psi} + \varrho\mathbf{u}\otimes\mathbf{u} : \nabla\boldsymbol{\psi})\ \mathrm{d}x\ \mathrm{d}t$$
$$+ \int_0^\tau\!\!\int_\Omega \left(\pi \mathrm{div}\boldsymbol{\psi} + p\left(\frac{\varrho}{\varrho^*}\right)\mathrm{div}\boldsymbol{\psi} - \mathbf{S}(\mathbf{u}) : \nabla\boldsymbol{\psi}\right)\ \mathrm{d}x\ \mathrm{d}t,$$

holds for all $\boldsymbol{\psi} \in C_c^1([0,\tau) \times \Omega, \mathbb{R}^3)$,

$$\int_\Omega \varrho^*(\tau,\cdot)\varphi(\tau,\cdot)\ \mathrm{d}x - \int_\Omega \varrho_0^*\varphi(0,\cdot)\ \mathrm{d}x = \int_0^\tau\!\!\int_\Omega \left(\varrho^*\partial_t\varphi + \varrho^*\mathrm{div}(\mathbf{u}\varphi)\right)\ \mathrm{d}x\ \mathrm{d}t,$$

holds for all $\varphi \in C^1([0,T] \times \overline{\Omega})$.

(iii) For *a.e.* $\tau \in (0,T)$, there holds the energy inequality

$$\mathscr{E}(\varrho,\mathbf{u},\varrho^*)(\tau) + \int_0^\tau\!\!\int_\Omega \left(\mu|\nabla\mathbf{u}|^2 + (\mu+\lambda)(\mathrm{div}\mathbf{u})^2\right)\ \mathrm{d}x\ \mathrm{d}t \leq \mathscr{E}(\varrho_0, \frac{\mathbf{m}_0}{\varrho_0}, \varrho_0^*),$$

where

$$\mathscr{E}(\varrho,\mathbf{u},\varrho^*)(\tau) = \int_\Omega \left(\frac{1}{2}\varrho|\mathbf{u}|^2 + \left(\frac{\varrho}{\varrho^*}\right)\Gamma\left(\frac{\varrho}{\varrho^*}\right)\right)(\tau)\ \mathrm{d}x,$$

$$\Gamma\left(\frac{\varrho}{\varrho^*}\right) = \int_0^{\frac{\varrho}{\varrho^*}} \frac{p(s)}{s^2}\ \mathrm{d}S.$$

Remark 5.2. The condition $(\varrho^*-\varrho)\pi = 0$ is not satisfied in the pointwise sense, its validity is justified using further information about the regularity of weak solution ala Lions-Masmoudi [153], see (5.3.11) below.

Remark 5.3. Introduction of the value of the weak solution at the final time for ϱ and ϱ^* implies that the initial condition for both of them is fulfilled, i.e. $\varrho(0,\cdot) = \varrho_0(\cdot)$, and $\varrho^*(0,\cdot) = \varrho_0^*(\cdot)$. The initial condition for the momentum is fulfilled only if the test function is divergence-free, i.e. $\mathrm{div}\psi = 0$, due to the fact that π is only a measure.

The main theorem of the paper states as follows.

Theorem 5.4. *Let the initial conditions ϱ_0, $\mathbf{m_0}$, ϱ_0^* satisfy the conditions above. Then the system (5.1.1) with p and $\mathbf{S}$ given by (5.1.2), (5.1.3) respectively, has a weak solution in the sense of Definition 5.1.*

Remark 5.5. The same result holds for the lower dimensions, $d = 1, 2$.

Theorem 5.4 is the first mathematical result on congested fluid system with time and space variable congestion barrier ϱ^*. In the literature devoted to fluid models with constraint ϱ^* is almost always equal to a constant. This is very severe mathematical simplification tacitly assumed in most of hyperbolic as well as parabolic models. Actually, the only other earlier result in which this assumption is not imposed is the previous work [197], where, however, only special dependence on the space variable was covered.

Including the space and time dependent congestion barrier ϱ^* leads to a sequence of mathematical difficulties already at the level of derivation of a-priori estimates, and definition of a weak solution. If, in addition, ϱ^* satisfies transport equation, it may lose its initial regularity due to low regularity of the velocity vector field ($\mathbf{u} \in L^2(0,T;H^1(\Omega))$). Moreover, the fact that the most nonlinear term in the approximate compressible Navier-Stokes system – the pressure– depends on two quantities that satisfy only hyperbolic PDEs causes further difficulties with identification of the limit. In fact, the area of more general than isentropic pressure laws, depending for example on more than one variables or in non-monotone way, is currently intensively investigated [41, 42, 149, 236], but so far this field is still at its infancy.

The core of the proof of Theorem 5.4 is to show that the system (5.1.1) can be obtained as a limit when $\varepsilon \to 0$ of the following approximation

$$\partial_t \varrho + \operatorname{div}(\varrho \mathbf{u}) = 0, \tag{5.1.8a}$$

$$\partial_t(\varrho \mathbf{u}) + \operatorname{div}(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla \pi_\varepsilon \left(\frac{\varrho}{\varrho^*}\right) + \nabla p \left(\frac{\varrho}{\varrho^*}\right) - \operatorname{div} \mathbf{S}(\mathbf{u}) = \mathbf{0}, \tag{5.1.8b}$$

$$\partial_t \varrho^* + \mathbf{u} \cdot \nabla \varrho^* = 0, \tag{5.1.8c}$$

where the π_ε stands for the singular pressure of the form

$$\pi_\varepsilon \left(\frac{\varrho}{\varrho^*}\right) = \varepsilon \frac{\left(\frac{\varrho}{\varrho^*}\right)^\alpha}{\left(1 - \frac{\varrho}{\varrho^*}\right)^\beta}, \quad \alpha \geq 0, \ \beta > 0. \tag{5.1.9}$$

A similar form of the pressure

$$\varepsilon \nabla \frac{1}{\left(\frac{1}{\varrho} - \frac{1}{\varrho^*}\right)^\beta}$$

was proposed in [35] or [78]. Singularities of the pressure of this type were also previously studied in the context of traffic models [35, 31, 34], collective dynamics [78, 77], or granular flow [159, 196, 195].

Note that for $\varepsilon > 0$ fixed and $\varrho^* = const.$ system (5.1.8) is purely compressible Navier-Stokes system with singular pressure term. A similar system was studied in [103, 104]. From this perspective, the new ingredients covered by this paper are the constitutive relation depending on more than one transported quantities, and the singular limit $\varepsilon \to 0$ leading to the two-phase free boundary problem. In the limit ε, uncongested (compressible) flow changes to incompressible when the density ϱ hits the value ϱ^*. When this happens the dynamics of the system is modified abruptly, meaning that the transition from the uncongested motion ($\varrho < \varrho^*$) to the congested motion ($\varrho = \varrho^*$) is very sudden.

The first goal is to prove the existence of solutions to a certain reformulation of the system (5.1.8). Our choice of approximation of the singular pressure (5.1.9) allows us to use some of ideas developed in the previous work [155], see also the stability result from [162], and the low Mach number analysis [102], in the context of geophysical flow model. In particular, we essentially use the formulation involving a new unknown Z, that for ϱ, ϱ^* smooth enough can be identified with the density fraction $Z = \frac{\varrho}{\varrho^*}$. We can formally check, that dividing (5.1.8a) by

ϱ^*, multiplying (5.1.8c) by $-\frac{\varrho}{\varrho^{*2}}$, and summing the resulting expressions, the system (5.1.8) can be transformed to the following one

$$\partial_t\varrho + \mathrm{div}(\varrho\mathbf{u}) = 0, \tag{5.1.10a}$$

$$\partial_t(\varrho\mathbf{u}) + \mathrm{div}(\varrho\mathbf{u}\otimes\mathbf{u}) + \nabla\pi_\varepsilon(Z) + \nabla p(Z) - \mathrm{div}\mathbf{S}(\mathbf{u}) = \mathbf{0}, \tag{5.1.10b}$$

$$\partial_t Z + \mathrm{div}(Z\mathbf{u}) = 0, \tag{5.1.10c}$$

with the initial data

$$\varrho(0,x) = \varrho_0(x), \quad (\varrho\mathbf{u})(0,x) = \mathbf{m}_0(x), \quad Z(0,x) = Z_0(x). \tag{5.1.11}$$

In consistency with the assumptions on $\varrho_0,\ \mathbf{m}_0,\ \varrho_0^*$ from the previous section, we postulate that $\varrho_0,\ \mathbf{m}_0,\ Z_0$ satisfy

$$\begin{gathered} 0 \le c_\star\varrho_0 \le Z_0 \le c^\star\varrho_0 \text{ a.e. in } \Omega, \quad \text{for} \quad 0 < c_\star \le c^\star < \infty, \\ 0 < \textstyle\int_\Omega \varrho_0 \ \mathrm{d}x, \quad Z_0 \le 1, \quad \textstyle\int_\Omega Z_0 \ \mathrm{d}x < |\Omega|, \quad \left.\frac{\varrho_0}{Z_0}\right|_{\{\varrho_0=0\}} = \widetilde{\varrho^*}, \\ \mathbf{m}_0 = \mathbf{0} \text{ a.e. in } \{\varrho_0 = 0\}, \quad \frac{\mathbf{m}_0}{\varrho_0}\mathbf{1}_{\{\varrho_0>0\}} \in L^2(\Omega). \end{gathered} \tag{5.1.12}$$

The first condition from (5.1.12) should be understood as the restriction of the initial congestion density, having in mind our notation $Z = \frac{\varrho}{\varrho^*}$ we see that the condition above means that ϱ_0^* cannot be zero on the regions where the density ϱ_0 is positive. The left condition means that the congestion density is initially bounded. The restriction on the integral $\int_\Omega Z \ \mathrm{d}x < |\Omega|$ means that the assumption $\varrho_0 \neq \varrho_0^*$ holds on a set of non-zero Lebesgue measure.

Existence of solutions to system (5.1.10) with singular pressure replaced by the barotropic one i.e. $p(Z) = Z^\gamma$ was studied in the recent paper [155]. In the present work we show that a similar result holds even with a very low a-priori integrability of the pressure. Indeed, the classical energy estimate does not provide a bound for the pressure, as it is for the barotropic case. This can be somehow compensated by better estimate of the pressure argument Z. As a result, the definition of the weak solution to (5.1.10) is very similar to the one from [155], namely:

Definition 5.6 (*Weak solution of the approximate system with Z*). A triplet $(\varrho, \mathbf{u}, Z)$ is called a global finite energy weak solution to (5.1.10), (5.1.4), with the initial data (5.1.11), (5.1.12), if for any $T > 0$:

(i) There holds:

$$0 \le c_\star\varrho \le Z \le c^\star\varrho \text{ a.e. in } (0,T)\times\Omega, \quad \text{for} \quad 0 < c_\star \le c^\star < \infty,$$

$$Z \le 1 \text{ a.e. in } (0,T)\times\Omega, \quad \mathbf{u}|_{(0,T)\times\Omega} = \mathbf{0},$$

and

$$\begin{aligned} &\varrho \in C_w([0,T]; L^\infty(\Omega)), \\ &\varrho\mathbf{u} \in C_w([0,T]; L^2(\Omega,\mathbb{R}^3)),\ \mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega,\mathbb{R}^3)),\ \varrho|\mathbf{u}|^2 \in L^\infty(0,T; L^1(\Omega)), \\ &Z \in C_w([0,T]; L^\infty(\Omega)). \end{aligned}$$

(ii) For any $0 \le \tau \le T$, equations (5.1.10a), (5.1.10b), (5.1.10c) are satisfied in the weak sense, more precisely:

$$\int_\Omega \varrho(\tau,\cdot)\varphi(\tau,\cdot)\ \mathrm{d}x - \int_\Omega \varrho_0\varphi(0,\cdot)\ \mathrm{d}x = \int_0^\tau\!\!\int_\Omega \Big(\varrho\partial_t\varphi + \varrho\mathbf{u}\cdot\nabla\varphi\Big)\ \mathrm{d}x\ \mathrm{d}t,$$

holds for all $\varphi \in C^1([0,T]\times\overline{\Omega})$,

$$\begin{aligned}\int_\Omega (\varrho\mathbf{u})(\tau,\cdot)\cdot\boldsymbol{\psi}(\tau,\cdot)\ \mathrm{d}x &- \int_\Omega \mathbf{m}_0\cdot\boldsymbol{\psi}(0,\cdot)\ \mathrm{d}x\\ &= \int_0^\tau\!\!\int_\Omega (\varrho\mathbf{u}\cdot\partial_t\boldsymbol{\psi} + \varrho\mathbf{u}\otimes\mathbf{u} : \nabla\boldsymbol{\psi})\ \mathrm{d}x\ \mathrm{d}t\\ &\quad + \int_0^\tau\!\!\int_\Omega (\pi_\varepsilon(Z)\mathrm{div}\boldsymbol{\psi} + p(Z)\mathrm{div}\boldsymbol{\psi} - \mathbf{S}(\mathbf{u}) : \nabla\boldsymbol{\psi})\ \mathrm{d}x\ \mathrm{d}t,\end{aligned}$$

holds for all $\boldsymbol{\psi} \in C_c^1([0,T]\times\Omega,\mathbb{R}^3)$,

$$\int_\Omega Z(\tau,\cdot)\varphi(\tau,\cdot)\ \mathrm{d}x - \int_\Omega Z_0\varphi(0,\cdot)\ \mathrm{d}x = \int_0^\tau\!\!\int_\Omega \Big(Z\partial_t\varphi + Z\mathbf{u}\cdot\nabla\varphi\Big)\ \mathrm{d}x\ \mathrm{d}t,$$

holds for all $\varphi \in C^1([0,T]\times\overline{\Omega})$.

(iii) For a.e. $\tau \in (0,T)$, there holds the energy inequality

$$\mathscr{E}(\varrho,\mathbf{u},Z)(\tau) + \int_0^\tau\!\!\int_\Omega \Big(\mu|\nabla\mathbf{u}|^2 + (\mu+\lambda)(\mathrm{div}\mathbf{u})^2\Big)\ \mathrm{d}x\ \mathrm{d}t \le \mathscr{E}\left(\varrho_0, \frac{\mathbf{m}_0}{\varrho_0}, Z_0\right) \tag{5.1.13}$$

where

$$\mathscr{E}(\varrho,\mathbf{u},Z)(\tau) = \int_\Omega \Big(\frac{1}{2}\varrho|\mathbf{u}|^2 + Z\Gamma(Z)\Big)(\tau)\ \mathrm{d}x, \tag{5.1.14}$$

$$\Gamma(Z) = \int_0^Z \frac{\pi_\varepsilon(s)+p(s)}{s^2}\ \mathrm{d}S.$$

Our second goal is to show the convergence of the weak solutions to the system (5.1.10), to the solutions of the limit system:

$$\partial_t\varrho + \mathrm{div}(\varrho\mathbf{u}) = 0, \tag{5.1.15a}$$

$$\partial_t(\varrho\mathbf{u}) + \mathrm{div}(\varrho\mathbf{u}\otimes\mathbf{u}) + \nabla\pi + \nabla p(Z) - \mathrm{div}\mathbf{S}(\mathbf{u}) = \mathbf{0}, \tag{5.1.15b}$$

$$\partial_t Z + \mathrm{div}(Z\mathbf{u}) = 0, \tag{5.1.15c}$$

$$0 \le Z \le 1, \quad c_\star\varrho \le Z \le c^\star\varrho, \tag{5.1.15d}$$

$$\mathrm{div}\mathbf{u} = 0 \text{ in } \{Z=1\}, \tag{5.1.15e}$$

$$\pi \ge 0 \text{ in } \{Z=1\}, \quad \pi = 0 \text{ in } \{Z<1\}. \tag{5.1.15f}$$

The weak solutions to the limit system are defined below.

Definition 5.7 (*Weak solution of the limit system with Z*). A quadruple $(\varrho,\mathbf{u},Z,\pi)$ is called a global finite energy weak solution to (5.1.15), (5.1.4), with the initial data (5.1.11), (5.1.12), if for any $T>0$:

(i) There holds:

$$0 \le c_\star \varrho \le Z \le c^\star \varrho \text{ a.e. in } (0,T)\times\Omega, \quad \text{for} \quad 0 < c_\star \le c^\star < \infty,$$

$$Z \le 1 \text{ a.e. in } (0,T)\times\Omega, \quad \mathbf{u}|_{(0,T)\times\Omega} = \mathbf{0},$$

$$\mathrm{div}\mathbf{u} = 0 \quad \textit{a.e. in } \{\varrho = \varrho^*\}, \qquad (\varrho^* - \varrho)\pi = 0,$$

and

$$\begin{aligned}
&\varrho \in C_w([0,T]; L^\infty(\Omega)),\\
&Z \in C_w([0,T]; L^\infty(\Omega)),\\
&\mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega,\mathbb{R}^3)), \quad \varrho|\mathbf{u}|^2 \in L^\infty(0,T;L^1(\Omega)),\\
&\pi \in \mathscr{M}^+((0,T)\times\Omega).
\end{aligned}$$

(ii) Equations (5.1.15a), (5.1.15b) are satisfied in the weak sense as in Definition 5.6, and (5.1.15b) is satisfied in the weak sense as in Definition 5.1 (with $p\left(\frac{\varrho}{\varrho^*}\right)$ replaced by $p(Z)$).

(iii) The energy inequality from Definition 5.6 holds for $\Gamma(Z) = \int_0^Z \frac{p(s)}{s^2}\,\mathrm{d}S$.

The convergence result reads as follows.

Theorem 5.8. *Let $(\varrho_\varepsilon, \mathbf{u}_\varepsilon, Z_\varepsilon)_{\{\varepsilon>0\}}$ be a a sequence of weak solutions to the approximate system (5.1.10),* (5.1.9)*. Then, for $\varepsilon \to 0$, the sequence $(\varrho_\varepsilon, \mathbf{u}_\varepsilon, Z_\varepsilon)$ converges to the weak solution of* (5.1.15) *in the sense of Definition 5.7*

More precisely,

$$\begin{aligned}
\varrho_\varepsilon \to \varrho \quad &\textit{in} \ \ C_w([0,T]; L^\infty(\Omega)), \quad \textit{and weakly in } L^p((0,T)\times\Omega),\\
Z_\varepsilon \to Z \quad &\textit{in} \ \ C_w([0,T]; L^\infty(\Omega)), \quad \textit{and strongly in } L^p((0,T)\times\Omega),
\end{aligned}$$

for any $p < \infty$, and

$$\mathbf{u}_\varepsilon \to \mathbf{u} \qquad \textit{weakly in } L^2(0,T; W^{1,2}(\Omega,\mathbb{R}^3)).$$

Moreover,

$$\pi_\varepsilon(Z_\varepsilon) \longrightarrow \pi \quad \textit{weakly in} \quad \mathscr{M}^+((0,T)\times\Omega).$$

We see that rewriting the system in terms of the conserved quantities $(\varrho, \varrho\mathbf{u}, Z)$ causes that the limit passage $\varepsilon \to 0$ leads to the switching relation $(1-Z)\pi = 0$. This resembles the homogeneous condition from the works [32, 43, 153]. The novelty of this paper is the proof of the fact that the same relation can be obtained for system with two densities, and that the final relation $(1-Z)\pi = 0$ can be identified with $(\varrho^* - \varrho)\pi = 0$. For this we need to prove that various weak formulations of the limit system are equivalent, and that the formal derivation of (5.1.8a) by ϱ^* leading to equation for Z can be inverted and made rigorous.

5.2 The Existence of Solution for ε Fixed

When ε is fixed, say $\varepsilon = 1$, system (5.1.10) resembles the system considered in [155]. The difference is the presence of the singular pressure. It provides uniform L^∞ estimate for Z, which, however, does not imply the higher integrability of the pressure itself. Indeed, the energy estimate provides uniform bound for a corresponding potential energy whose singularity at $Z = 1$ is one order weaker. Therefore, in order to use the result from [155] to prove the existence of approximate solution to (5.1.10), we need to approximate the singular pressure by a monotone non-singular function of Z. This is obtained in our work by introducing the truncation parameter δ and a parameter κ that improves integrability of Z at the first approximation level. Approximation like this was considered in the previous paper [197], where ϱ^* was assumed to be a function dependent on x only. The difference with respect to [197] is that now, the uniform estimates as well as the effective viscus flux technique (see, for example, [187]) can be applied to the variable Z only, but it does not imply strong convergence of the sequences approximating ϱ nor ϱ^*.

In the following proof of existence of solutions we will recall some elements of the two approaches from [197] and [155] in order to avoid repetitions.

5.2.1 Formulation of the Approximate Problem

The first level of approximation introduces truncation parameter δ in the singular pressure and artificial pressure κZ^K, with K sufficiently large to be determined in the course of the proof. We consider

$$\partial_t \varrho_\delta + \operatorname{div}(\varrho_\delta \mathbf{u}_\delta) = 0, \tag{5.2.1a}$$

$$\partial_t(\varrho_\delta \mathbf{u}_\delta) + \operatorname{div}(\varrho_\delta \mathbf{u}_\delta \otimes \mathbf{u}_\delta) + \nabla \pi_\delta(Z_\delta) + \nabla p_\kappa(Z_\delta) - \operatorname{div}\mathbf{S}(\mathbf{u}_\delta) = \mathbf{0}, \tag{5.2.1b}$$

$$\partial_t Z_\delta + \operatorname{div}(Z_\delta \mathbf{u}_\delta) = 0, \tag{5.2.1c}$$

with $\kappa,\ \delta > 0$, and π_δ, p_κ given by

$$\pi_\delta(Z_\delta) = \begin{cases} \frac{Z_\delta^\alpha}{(1-Z_\delta)^\beta} & \text{if} \quad Z_\delta < 1-\delta, \\ \frac{Z_\delta^\alpha}{\delta^\beta} & \text{if} \quad Z_\delta \geq 1-\delta, \end{cases} \tag{5.2.2}$$

$$p_\kappa(Z_\delta) = \kappa Z_\delta^K + Z_\delta^\gamma.$$

We drop the subindex δ when no confusion can arise, and we introduce the notion of weak solution for the system (5.2.1).

Definition 5.9. A triplet $(\varrho, \mathbf{u}, Z)$ is called a global finite energy weak solution to (5.2.1), (5.1.4), with the initial data (5.1.11), (5.1.12), if for any $T > 0$:

(i) There holds:

$$0 \leq \varrho, \quad 0 \leq Z \text{ a.e. in } (0,T)\times\Omega, \quad \mathbf{u}|_{(0,T)\times\Omega} = \mathbf{0},$$

and

$$\varrho \in C_w([0,T]; L^K(\Omega)),$$
$$\varrho\mathbf{u} \in C_w([0,T]; L^{\frac{2K}{K+1}}(\Omega,\mathbb{R}^3)),\ \mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega,\mathbb{R}^3)),\ \varrho|\mathbf{u}|^2 \in L^\infty(0,T;L^1(\Omega)),$$
$$Z \in C_w([0,T]; L^K(\Omega)).$$

(ii) Equations (5.2.1a), (5.2.1b), (5.2.1b) are satisfied in the weak sense as in Definition 5.6 (with π_ε replaced by π_δ, and p replaced by p_k).

(iii) The energy inequality from Definition 5.6 holds for

$$\Gamma(Z) = \int_0^Z \frac{\pi_\delta(s) + p_\kappa(s)}{s^2}\,\mathrm{d}S. \tag{5.2.3}$$

We have the following existence result for solutions defined by Definition 5.9 (see also [155], Theorem 2).

Theorem 5.10. *Let $\mathbf{S}$ satisfy (5.1.3), $K > 6$, $\beta > 5/2$, $\alpha \geq 0$, $\kappa, \delta, \varepsilon$ be fixed and positive, and the initial data $(\varrho_0, \mathbf{m}_0, Z_0)$ satisfy (5.1.12).*

Then there exists a weak solution $(\varrho, \mathbf{u}, Z)$ to problem (5.2.1), (5.2.2) with boundary conditions (5.1.4), in the sense of Definition 5.9.

Moreover, $(Z, \mathbf{u})$ solves (5.2.1c) in the renormalized sense, i.e. $(Z, \mathbf{u})$, extended by zero outside of Ω, satisfies

$$\partial_t b(Z) + \operatorname{div}(b(Z)\mathbf{u}) + \Big(b'(Z)Z - b(Z)\Big)\operatorname{div}\mathbf{u} = 0, \tag{5.2.4}$$

in the sense of distributions on $(0,T) \times \mathbb{R}^3$, where

$$b \in C^1(\mathbb{R}), \quad b'(z) = 0, \quad \forall z \in \mathbb{R} \text{ large enough.} \tag{5.2.5}$$

In addition,

$$0 \leq c_\star \varrho \leq Z \leq c^\star \varrho \quad \text{a.e. in } (0,T) \times \Omega. \tag{5.2.6}$$

The proof of Theorem 5.10 requires further modification of the system (5.2.1) with two additional approximation levels involving the parabolic regularisation of two continuity equations and the Galerkin approximation of the velocity. This approximation allows, in particular, to deduce inequalities (5.2.6). At this point existence of regular solution can be obtained following the arguments presented in [155]. Also the compactness arguments needed to recover system (5.2.1) are analogous, therefore we skip this part and focus only on the a-priori estimates needed to perform the limit passages $\delta \to 0$, $\kappa \to 0$.

Having the existence of solutions to the system (5.2.1)–(5.2.2), we show that this solution can be used to recover the weak solution to the system (5.1.10), where only the parameter ε is present.

Theorem 5.11. *Let ε be fixed and let $(\varrho_{\kappa,\delta}, \mathbf{u}_{\kappa,\delta}, Z_{\kappa,\delta})$ be a weak solution to the approximate system (5.2.1)–(5.2.2) established in Theorem 5.10. Then, for $\delta, \kappa \to 0$, the sequence $(\varrho_{\kappa,\delta}, \mathbf{u}_{\kappa,\delta}, Z_{\kappa,\delta})_{\kappa,\delta>0}$ converges to the weak solution of (5.1.10) in the sense of Definition 5.9.*

More precisely,

$$\begin{aligned} \varrho_{\kappa,\delta} \to \varrho \quad &\text{in } C_w([0,T]; L^\infty(\Omega)), \quad \text{and weakly in } L^p((0,T)\times\Omega),\\ Z_{\kappa,\delta} \to Z \quad &\text{in } C_w([0,T]; L^\infty(\Omega)), \quad \text{and strongly in } L^p((0,T)\times\Omega), \end{aligned}$$

for any $p < \infty$, and

$$\mathbf{u}_{\kappa,\delta} \to \mathbf{u} \qquad \text{weakly in } L^2(0,T; W^{1,2}(\Omega,\mathbb{R}^3)).$$

Moreover,

$$\pi_\delta(Z_{\kappa,\delta}) \longrightarrow \pi_\varepsilon(Z) \quad \textit{strongly in} \quad L^1((0,T)\times\Omega).$$

In the next two sections we prove Theorem 5.11. Starting from the weak solution to the system (5.2.1), we first derive uniform bounds and then we let $\delta \to 0$. The modification of the reasoning needed to perform the limit passage $\kappa \to 0$ is explained at the end.

5.2.2 Uniform Estimates

In this section we obtain the uniform estimates for the weak solutions to the system (5.2.1)–(5.2.2). This involves three a-priori estimates: the standard energy estimate, and two estimates involving the application of the Bogovskii operator, one of which gives us the uniform bound for the singular pressure and the other the uniform integrability of the pressure. All of the aforementioned estimates are essential for performing the limit passage $\delta,\kappa \to 0$, as well as the last limit passage $\varepsilon \to 0$. However, as we shall see later on, the uniform integrability of the pressure will no longer be valid in this case.

The energy estimate

We present a formal computation that can be made rigorous at the level of the Galerkin approximation of the velocity. Multiplying the momentum equation (5.2.1b) by $\mathbf{u}$ and integrating by parts with respect to space, yield the energy equality

$$\frac{d}{dt}\int_\Omega \frac{1}{2}\varrho|\mathbf{u}|^2 \ \mathrm{d}x + \int_\Omega \nabla\left(\pi_\delta(Z)+p_\kappa(Z)\right)\cdot\mathbf{u} \ \mathrm{d}x + \int_\Omega \mathbf{S}(\mathbf{u}) : \nabla\mathbf{u} \ \mathrm{d}x = 0. \tag{5.2.7}$$

Note that the second term in our case is different than in [197], there is no additional ϱ^* in front of the gradient. This however allows us to proceed straightforwardly, for reader's convenience we repeat here the derivation

$$\begin{aligned}\int_\Omega \nabla\left(\pi_\delta(Z)+p_\kappa(Z)\right)\cdot\mathbf{u} \ \mathrm{d}x &= \int_\Omega \frac{\pi'_\delta(Z)+p'_\kappa(Z)}{Z}\nabla Z\cdot(Z\mathbf{u}) \ \mathrm{d}x\\ = -\int_\Omega Q_{\kappa,\delta}(Z)\mathrm{div}(Z\mathbf{u}) \ \mathrm{d}x &= \int_\Omega Q_{\kappa,\delta}(Z)\partial_t Z \ \mathrm{d}x = \frac{d}{dt}\int_\Omega Z\Gamma_{\kappa,\delta}(Z) \ \mathrm{d}x\end{aligned}$$

where we used the equation (5.2.1c), we denoted $Q'_{\kappa,\delta}(Z) = \frac{\pi'_\delta(Z)+p'_\kappa(Z)}{Z}$, and $\Gamma_{\kappa,\delta}$ is a solution of the following ODE:

$$\Gamma_{\kappa,\delta}(Z) + Z\Gamma'_{\kappa,\delta}(Z) = Q_{\kappa,\delta}(Z).$$

Using the definition of $Q_{\kappa,\delta}$ and $\pi_{\kappa,\delta}$ we can express $\Gamma_{\kappa,\delta}$ as in (5.2.3). Integrating (5.2.7) with respect to time, and using the definition of the stress tensor (5.1.3) we get the following uniform estimates

$$\begin{gathered}\sup_{t\in[0,T]}\left(\|\sqrt{\varrho_\delta}\mathbf{u}_\delta(t)\|_{L^2(\Omega)} + \left\|Z_\delta\Gamma_{\kappa,\delta}(Z_\delta)(t)\right\|_{L^1(\Omega)}\right) \le C,\\ \int_0^T \|\mathbf{u}_\delta\|^2_{W^{1,2}(\Omega,\mathbb{R}^3)} \ \mathrm{d}t \le C.\end{gathered} \tag{5.2.8}$$

The integrability of the pressure

The energy estimate obtained above is insufficient to control the L^1 norm of the pressure, because the singularity appearing in $\Gamma_{\kappa,\delta}(Z)$ at $Z = 1$ is of lower order than the singularity for $\pi_{\kappa,\delta}(Z)$ (take f.i. $\alpha = 2$, $\beta = 1$ in (5.1.9) and use (5.2.3)). Therefore, further estimates are needed. The first of them is obtained by testing the momentum equation by the function

$$\boldsymbol{\psi} = \phi(t)\mathscr{B}\left(Z - \frac{1}{|\Omega|}\int_\Omega Z(t,y)\ \mathrm{d}y\right), \tag{5.2.9}$$

where ϕ is smooth and compactly supported in the interval $(0,T)$, and $\mathscr{B}$ is the Bogovskii operator, i.e. a solution operator $\mathscr{B} : \{f \in L^p(\Omega); \int_\Omega f(y)\,\mathrm{d}y = 0\} \to W_0^{1,p}(\Omega, \mathbb{R}^3)$ to the following problem

$$\mathrm{div}\Phi = f, \quad \Phi|_{\partial\Omega} = \mathbf{0}.$$

The main properties of $\mathscr{B}$ can be found, for example, in [187, Lemma 3.17], and [197, Appendix]. In particular, for $\Phi = \mathscr{B}(f)$ we have the following estimate

$$\|\nabla\Phi\|_{L^p(\Omega,\mathbb{R}^3\times\mathbb{R}^3)} \le c(p,\Omega)\|f\|_{L^p(\Omega)}, \quad 1 < p < \infty.$$

Moreover, if $f = \mathrm{div}g$, with $g \in L^q(\Omega, \mathbb{R}^3)$, $\mathrm{div}g \in L^p(\Omega)$, $1 < q < \infty$, then

$$\|\Phi\|_{L^q(\Omega,\mathbb{R}^3)} \le c(q,\Omega)\|g\|_{L^q(\Omega,\mathbb{R}^3)}.$$

Remark 5.12. Note that, alike for ϱ, the integral of Z over the space is an invariant of the motion, therefore we have

$$\int_\Omega Z(t,x)\ \mathrm{d}x = \int_\Omega Z_0(x)\ \mathrm{d}x < |\Omega|, \tag{5.2.10}$$

according to (5.1.12), at least for sufficiently regular solutions. In what follows we denote $\int_\Omega Z_0(x)\ \mathrm{d}x = M_Z$.

Remark 5.13. Function φ must be of a certain regularity in order to use it as a test function in in the weak formulation of the momentum equation (5.2.1b), see Definition 5.9. In fact, it follows from the construction of the solution in [155], that taking K sufficiently large guarantees admissibility of this function.

Using in the weak formulation of (5.2.1b) the test function (5.2.9) results in the following equality

$$\begin{aligned}
&\int_0^T\!\!\int_\Omega \phi\left(\pi_\delta(Z_\delta) + p_\kappa(Z_\delta)\right)\left(Z_\delta - \frac{1}{|\Omega|}\int_\Omega Z_\delta\ \mathrm{d}y\right)\ \mathrm{d}x\ \mathrm{d}t\\
&\quad = -\int_0^T\!\!\int_\Omega \varrho_\delta\mathbf{u}_\delta\cdot\partial_t\boldsymbol{\psi}\ \mathrm{d}x\ \mathrm{d}t - \int_0^T\!\!\int_\Omega \varrho_\delta\mathbf{u}_\delta\otimes\mathbf{u}_\delta : \nabla\boldsymbol{\psi}\ \mathrm{d}x\ \mathrm{d}t\\
&\qquad + \int_0^T\!\!\int_\Omega \mathbf{S}(\mathbf{u}_\delta) : \nabla\boldsymbol{\psi}\ \mathrm{d}x\ \mathrm{d}t,
\end{aligned}$$

whose r.h.s. can be bounded using the uniform estimates (5.2.8) together with (5.2.6), provided that K is sufficiently large, say $K > 4$. Therefore, one gets the following estimate, which is now uniform with respect to κ and δ

$$\int_0^T\!\!\int_\Omega \phi\left(\pi_\delta(Z_\delta) + p_\kappa(Z_\delta)\right)\left(Z_\delta - \frac{1}{|\Omega|}\int_\Omega Z_\delta \ \mathrm{d}y\right)\ \mathrm{d}x\ \mathrm{d}t \leq C. \tag{5.2.11}$$

We then consider two complementary subsets of $(0,T)\times\Omega$: $\Sigma_1 = \{Z_\delta(t,x) < Z^*\}$ and $\Sigma_2 = \{Z_\delta(t,x) \geq Z^*\}$, for $\frac{M_Z}{|\Omega|} < Z^* < 1$. The l.h.s. of (5.2.11) can be easily controlled on Σ_1, because on this subset Z_δ stays far away from the singularity, for Σ_2 we have

$$\begin{aligned}&\int_0^T\!\!\int_\Omega \phi\pi_\delta(Z_\delta)\left(Z_\delta - \frac{1}{|\Omega|}\int_\Omega Z_\delta \ \mathrm{d}y\right)\mathbf{1}_{\Sigma_2}\ \mathrm{d}x\ \mathrm{d}t\\ &\geq \left(Z^* - \frac{M_Z}{|\Omega|}\right)\int_0^T\!\!\int_\Omega \phi\pi_\delta(Z_\delta)\mathbf{1}_{\Sigma_2}\ \mathrm{d}x\ \mathrm{d}t,\end{aligned}$$

and so, from (5.2.11) it follows that

$$\|\pi_\delta(Z_\delta)\|_{L^1((0,T)\times\Omega)} + \|Z_\delta\pi_\delta(Z_\delta)\|_{L^1((0,T)\times\Omega)} \leq C, \tag{5.2.12}$$

as well as

$$\kappa\|Z_\delta\|^{K+1}_{L^{K+1}((0,T)\times\Omega)} \leq C. \tag{5.2.13}$$

The estimate (5.2.11) implies that the L^{K+1} norm of Z_δ is bounded uniformly in δ, but not uniformly in κ, which suggests that the passage to the limit $\delta \to 0$ should be performed as first.

The equi-integrability of the singular pressure

Because π_δ is a nonlinear function of Z, identification of the limit in this term, after letting $\delta \to 0$ cannot be justified only by the uniform L^1 bound. Following the idea from [104], we can prove that the pressure π_δ enjoys some additional estimate near to the singularity $Z = 1$. Indeed, for $K > 6$, $\beta > 5/2$ and

$$\eta_\delta(s) = \begin{cases} -\log(1-s) & \text{if } s \leq 1-\delta,\\ -\log(\delta) & \text{if } s > 1-\delta,\end{cases} \tag{5.2.14}$$

uniformly with respect to δ one has

$$\int_0^T\!\!\int_\Omega \pi_\delta(Z)\eta_\delta(Z)\ \mathrm{d}x\ \mathrm{d}t \leq C. \tag{5.2.15}$$

The proof follows by testing the momentum equation (5.2.1b) by the function of the form

$$\boldsymbol{\psi} = \phi(t)\mathscr{B}\left(\eta_\delta(Z_\delta) - \frac{1}{|\Omega|}\int_\Omega \eta_\delta(Z_\delta)\ \mathrm{d}y\right), \tag{5.2.16}$$

where ϕ is smooth and compactly supported in the interval $(0,T)$. For the details of this estimate we refer to [197], for $\beta > 3$ and to [103] for $\beta > \frac{5}{2}$. One of the difficulties in the proof of analogue of (5.2.15) presented in [197] concerned the renormalization of the equation for $\frac{\varrho}{\varrho^*}$. Here this problem does not appear anymore, since $(Z_\delta, \mathbf{u}_\delta)$ is by definition a distributional solution of the renormalized transport equation (5.2.4).

5.2.3 Passage to the Limit $\delta \to 0$

Convergences Following from the Uniform Estimates

Using the uniform estimates (5.2.6), (5.2.8), and the Hölder inequality, we can deduce that up to a subsequence

$$\begin{aligned} Z_\delta \to Z \qquad & \text{weakly-* in } L^\infty(0,T;L^K(\Omega)),\\ \varrho_\delta \to \varrho \qquad & \text{weakly-* in } L^\infty(0,T;L^K(\Omega)),\\ \mathbf{u}_\delta \to \mathbf{u} \qquad & \text{weakly in } L^2(0,T;W^{1,2}(\Omega,\mathbb{R}^3)). \end{aligned} \tag{5.2.17}$$

Using the continuity equation (5.2.1a) and (5.2.1c), the first two convergences can be strengthened to

$$\begin{aligned} Z_\delta \to Z \qquad & \text{in } \ C_w([0,T];L^K(\Omega)),\\ \varrho_\delta \to \varrho \qquad & \text{in } \ C_w([0,T];L^K(\Omega)), \end{aligned} \tag{5.2.18}$$

which together with the weak convergence of the velocity gradient, after using the momentum equation (5.2.1b) implies that

$$\varrho_\delta \mathbf{u}_\delta \to \varrho \mathbf{u} \quad \text{in } \ C_w([0,T];L^{\frac{2K}{K+1}}(\Omega,\mathbb{R}^3)). \tag{5.2.19}$$

This in turn, assuming that K is sufficiently large so that the imbedding of $L^{\frac{2K}{K+1}}(\Omega,\mathbb{R}^3)$ to $W^{-1,2}(\Omega,\mathbb{R}^3)$ is compact, implies that

$$\varrho_\delta \mathbf{u}_\delta \otimes \mathbf{u}_\delta \to \varrho \mathbf{u} \otimes \mathbf{u} \quad \text{weakly in } L^q((0,T)\times\Omega,\mathbb{R}^3\times\mathbb{R}^3), \tag{5.2.20}$$

for some $q > 1$. Moreover, the uniform bound (5.2.12) together with the growth condition $\beta > 5/2$ in (5.2.2) imply that

$$Z \le 1 \quad \text{a.a. } (t,x) \in (0,T)\times\Omega, \tag{5.2.21}$$

and thus also

$$\varrho \le \frac{1}{c_\star} \quad \text{a.a.}(t,x) \in (0,T)\times\Omega,$$

on the account of (5.2.6). Finally, from the uniform bounds (5.2.12) we can extract the subsequence such that

$$\begin{aligned} \pi_\delta(Z_\delta) \to \overline{\pi(Z)} \quad & \text{weakly in} \quad \mathscr{M}^+((0,T)\times\Omega),\\ Z_\delta\pi_\delta(Z_\delta) \to \overline{Z\pi(Z)} \quad & \text{weakly in} \quad \mathscr{M}^+((0,T)\times\Omega), \end{aligned} \tag{5.2.22}$$

for some $\overline{\pi(Z)}$, $\overline{Z\pi(Z)}$ that need to be determined. Note, however, that (5.2.15) together with De La Vallée-Poussin criterion allow us to deduce the equi-integrability of the pressure, therefore the first limit can be strengthened to

$$\pi_\delta(Z_\delta) \to \overline{\pi(Z)} \quad \text{weakly in} \quad L^1((0,T)\times\Omega). \tag{5.2.23}$$

As for the second convergence, we cannot say that much immediately. However, using the fact that $\pi_{\delta^1}(\cdot) \geq \pi_{\delta^2}(\cdot)$ provided $\delta^1 \leq \delta^2$, we can estimate for any smooth, nonnegative, compactly supported function $\phi(x,t)$

$$\begin{aligned}
&\liminf_{\delta\to 0} \int_0^T\!\!\int_\Omega \phi\left(\pi_\delta(Z_\delta)Z_\delta - \overline{\pi(Z)Z}\right)\ \mathrm{d}x\ \mathrm{d}t \\
&\geq \liminf_{\delta\to 0} \int_0^T\!\!\int_\Omega \phi\left(\pi_{\delta^*}(Z_\delta)Z_\delta - \overline{\pi(Z)Z}\right)\ \mathrm{d}x\ \mathrm{d}t \\
&\geq \liminf_{\delta\to 0} \int_0^T\!\!\int_\Omega \phi\left(\overline{\pi_{\delta^*}(Z)} - \overline{\pi(Z)}\right) Z\ \mathrm{d}x\ \mathrm{d}t
\end{aligned} \tag{5.2.24}$$

where the last inequality is a consequence of the fact that $\cdot \mapsto \pi_{\delta^*}(\cdot)$ is non-decreasing function for fixed δ^*. Using again equi-integrability of the pressure and letting $\delta^* \to 0$ we verify that the r.h.s. of (5.2.24) vanishes and thus

$$\overline{\pi(Z)} \geq Z\overline{\pi(Z)} \tag{5.2.25}$$

in the sense of distributions. In order to say something more, we need to investigate the strong convergence of the sequence Z_δ, which is the purpose of the next section.

Strong convergence of Z_δ

It was observed in [155], that the strong convergence of Z_δ does not imply the strong convergence of ϱ_δ. The proof of the strong convergence of Z_δ requires an analogue of the effective flux equality for the barotropic compressible Navier-Stokes equations. In our case it can be written as follows.

Lemma 5.14. *Let $\varrho_\delta, \mathbf{u}_\delta, Z_\delta$ be the sequence of approximate solutions enjoying the properties from above. Then, at least for a subsequence*

$$\begin{aligned}
&\lim_{\delta\to 0^+} \int_0^T\!\!\int_\Omega \psi\phi\Big(\pi_\delta(Z_\delta) + p_\kappa(Z_\delta) - (\lambda+2\mu)\mathrm{div}\mathbf{u}_\delta\Big)Z_\delta\ \mathrm{d}x\ \mathrm{d}t \\
&\quad = \int_0^T\!\!\int_\Omega \psi\phi\Big(\overline{\pi(Z)} + \overline{p_\kappa(Z)} - (\lambda+2\mu)\mathrm{div}\mathbf{u}\Big)Z\ \mathrm{d}x\ \mathrm{d}t
\end{aligned} \tag{5.2.26}$$

for any $\psi \in C_c^\infty((0,T))$ and $\phi \in C_c^\infty(\Omega)$.

The proof of this fact can be seen as a special case of an analogous result proven in [155, cf. Lemma 11]. On the account of (5.2.25) and the monotonicity of $p_\kappa(\cdot)$ we obtain from (5.2.26) that

$$\lim_{\delta\to 0^+} \int_0^T\!\!\int_\Omega \psi\phi Z_\delta\,\mathrm{div}\mathbf{u}_\delta\ \mathrm{d}x\ \mathrm{d}t - \int_0^T\!\!\int_\Omega \psi\phi Z\,\mathrm{div}\mathbf{u}\ \mathrm{d}x\ \mathrm{d}t \geq 0, \tag{5.2.27}$$

for any $\psi\phi \geq 0$. Recall that Z_δ satisfies the renormalized continuity equation (5.2.4) with b specified in (5.2.5). By density argument and standard approximation technique, we may extend the validity of (5.2.4) to functions $b \in C([0,\infty) \cap C^1((0,\infty))$ such that

$$\begin{aligned}
&|b'(t)| \leq Ct^{-\lambda_0}, \qquad \lambda_0 < -1, \quad t \in (0,1], \\
&|b'(t)| \leq Ct^{\lambda_1}, \qquad -1 < \lambda_1 \leq \frac{q}{2} - 1, \quad t \geq 1.
\end{aligned}$$

The renormalization technique applied to the barotropic Navier-Stokes system is due to DiPerna and Lions [84], and the above extension can be found, for example, in [101].

We can now write the renormalized continuity equation with $b(Z_\delta) = Z_\delta \log Z_\delta$:

$$\partial_t(Z_\delta \log Z_\delta) + \operatorname{div}(Z_\delta \log Z_\delta \mathbf{u}_\delta) = -Z_\delta \operatorname{div}\mathbf{u}_\delta \quad \text{in} \quad \mathscr{D}'((0,T)\times\mathbb{R}^3).$$

Passing to the limit $\delta \to 0^+$, we hence obtain

$$\partial_t(\overline{Z\log Z}) + \operatorname{div}(\overline{Z\log Z}\mathbf{u}) = -\overline{Z\operatorname{div}\mathbf{u}} \quad \text{in} \quad \mathscr{D}'((0,T)\times\mathbb{R}^3).$$

Writing an analogous equation for the limit function Z and subtracting it from the above, we obtain

$$\partial_t(\overline{Z\log Z} - Z\log Z) + \operatorname{div}[(\overline{Z\log Z} - Z\log Z)\mathbf{u}] = Z\operatorname{div}\mathbf{u} - \overline{Z\operatorname{div}\mathbf{u}}, \tag{5.2.28}$$

satisfied in the sense of distributions on $(0,T)\times\mathbb{R}^3$. Integrating the above equation with respect to time and space, and using the convexity of the function $s \mapsto s\log s$ we get from (5.2.28) that

$$\int_0^T\!\!\int_\Omega \overline{Z\operatorname{div}\mathbf{u}}\ \mathrm{d}x\ \mathrm{d}t \le \int_0^T\!\!\int_\Omega Z\operatorname{div}\mathbf{u}\ \mathrm{d}x\ \mathrm{d}t,$$

which is an opposite to (5.2.27). Therefore, recalling (5.2.28) we see that $\overline{Z\log Z}(t,x) = Z\log Z(t,x)$ almost everywhere in $(t,x) \in (0,T)\times\Omega$, which yields the strong convergence of Z_δ in $L^p((0,T)\times\Omega)$ for any $p < K+1$. With this at hand, we verify that (5.2.23) can be replaced by

$$\pi_\delta(Z_\delta) \to \pi(Z) \quad \text{strongly in} \quad L^1((0,T)\times\Omega),$$

similarly

$$p_\kappa(Z_\delta) \to p_\kappa(Z) \quad \text{strongly in} \quad L^1((0,T)\times\Omega).$$

In order to complete the proof of Theorem 5.11, we have to show that we are allowed to let $\kappa \to 0$. Note that all the uniform estimates obtained above stay in force independently of κ. Indeed, the only issue here is to see that the uniform L^{K+1} bound on Z_κ, that was previously obtained from (5.2.12) follows now directly from (5.2.21). On the account of (5.2.6), the same is true also for the sequence ϱ_κ. The proof of Theorem 5.11 is now complete. □

5.3 Passage to the Limit $\varepsilon \to 0$

The purpose of this section is to prove our main Theorem 5.4. For technical reasons we first perform the limit $\varepsilon \to 0$ in the auxiliary system (5.1.10) proving Theorem 5.8 and then we prove equivalence between systems (5.1.15) and (5.1.1) in the certain class of solutions.

5.3.1 Convergence Following from the Uniform Estimates

The estimates performed in the previous section give rise to several estimates that are uniform with respect to ε. Indeed, passing to the limit in the energy estimate we obtain

$$\begin{gathered}\sup_{t\in[0,T]}\Big(\|\sqrt{\varrho_\varepsilon}\mathbf{u}_\varepsilon(t)\|_{L^2(\Omega)}+\|Z_\varepsilon\Gamma_\varepsilon(Z_\varepsilon)(t)\|_{L^1(\Omega)}\Big)\le C,\\ \int_0^T\|\mathbf{u}_\varepsilon\|^2_{W^{1,2}(\Omega,\mathbb{R}^3)}\ \mathrm{d}t\le C.\end{gathered}\tag{5.3.1}$$

Passing to the limit in (5.2.21) and in (5.2.6) we obtain

$$0\le Z_\varepsilon\le 1,\quad 0\le c_\star\varrho_\varepsilon\le Z_\varepsilon\le c^\star\varrho_\varepsilon,\tag{5.3.2}$$

in particular both sequences Z_ε, ϱ_ε are uniformly bounded in $L^p((0,T)\times\Omega)$ for $p\le\infty$. Therefore, by means of the arguments from the previous section we get, up to the subsequence, that

$$\begin{aligned}&\mathbf{u}_\varepsilon\to\mathbf{u} &&\text{weakly in } L^2(0,T;W^{1,2}(\Omega,\mathbb{R}^3)),\\ &Z_\varepsilon\to Z &&\text{in } C_w([0,T];L^\infty(\Omega)),\\ &\varrho_\varepsilon\to\varrho &&\text{in } C_w([0,T];L^\infty(\Omega)).\end{aligned}\tag{5.3.3}$$

This information allows us to pass to the limit in all terms of the system (5.1.10), apart from the nonlinear pressure terms $p(Z)$ and $\pi_\varepsilon(Z)$. Repeating the Bogovskii type estimate with the test function (5.2.9) we obtain

$$\|Z_\varepsilon p(Z_\varepsilon)\|_{L^1((0,T)\times\Omega)}+\|\pi_\varepsilon(Z_\varepsilon)\|_{L^1((0,T)\times\Omega)}+\|Z_\varepsilon\pi_\varepsilon(Z_\varepsilon)\|_{L^1((0,T)\times\Omega)}\le C,\tag{5.3.4}$$

however, the estimate (5.2.15) does not hold anymore. Therefore the convergence in the sense of measures is the most we can hope for, we have

$$\begin{aligned}&\pi_\varepsilon(Z_\varepsilon)\to\pi\quad\text{weakly in}\quad\mathscr{M}^+((0,T)\times\Omega),\\ &Z_\varepsilon\pi_\varepsilon(Z_\varepsilon)\to\pi_1\quad\text{weakly in}\quad\mathscr{M}^+((0,T)\times\Omega).\end{aligned}\tag{5.3.5}$$

For the background pressure, due to (5.3.2), we have

$$p(Z_\varepsilon)\to\overline{p(Z)}\quad\text{weakly in}\quad L^p((0,T)\times\Omega),$$

for any $p<\infty$. At this point, we can identify the second limit in (5.3.5) using the explicit form of the pressure (5.1.9). We have

$$Z_\varepsilon\pi_\varepsilon(Z_\varepsilon)=\varepsilon\frac{1}{(1-Z_\varepsilon)^\beta}=\pi_\varepsilon(Z_\varepsilon)-\varepsilon\frac{1}{(1-Z_\varepsilon)^{\beta-1}},\tag{5.3.6}$$

thus letting $\varepsilon\to0$ and observing that the last term converges to zero strongly, we obtain the relation

$$\pi_1=\pi\tag{5.3.7}$$

in the sense of the measures. The recovery of the constraint condition $(1-Z)\pi=0$ and the identification of the limit $\overline{p(Z)}=p(Z)$ require stronger information about the convergence of Z_ε.

5.3.2 Strong Convergence of Z_ε

The nowadays well known technique of proving the strong convergence of the density in the compressible barotropic Navier-Stokes equations involves the study of propagation of the so called oscillation defect measure [100]. Such level of precision is not needed in our case, because, due to the singularity of the pressure argument Z, we have sufficiently high integrability of the pressure in order to apply the DiPerna-Lions technique [84]. Nevertheless, a variant of effective viscous flux equality is still needed. It can be derived the same way as in Lemma 5.14, after observing that the inverse divergence operator $\nabla\Delta^{-1}[1_\Omega Z]$ is regular enough to be used as a test function in the limiting momentum equation.

With this information the statement of Lemma 5.14 adapted to the ε-labelled sequences gives rise to the equality

$$\begin{aligned}\lim_{\varepsilon\to 0^+} \int_0^T\!\!\int_\Omega \psi\phi\Big(\pi_\varepsilon(Z_\varepsilon) + p(Z_\varepsilon) - (\lambda+2\mu)\mathrm{div}\mathbf{u}_\varepsilon\Big)Z_\varepsilon \ \mathrm{d}x \ \mathrm{d}t \\ = \int_0^T\!\!\int_\Omega \psi\phi\Big(\pi + \overline{p(Z)} - (\lambda+2\mu)\mathrm{div}\mathbf{u}\Big)Z \ \mathrm{d}x \ \mathrm{d}t,\end{aligned} \tag{5.3.8}$$

for any $\psi \in C_c^\infty((0,T))$ and $\phi \in C_c^\infty(\Omega)$.

Let us now explain the meaning of the product $Z\pi$ on the r.h.s. of (5.3.8). To this end, one needs to come back to the limiting momentum equation

$$\partial_t(\varrho\mathbf{u}) + \mathrm{div}(\varrho\mathbf{u}\otimes\mathbf{u}) + \nabla\pi + \nabla\overline{p(Z)} - \mathrm{div}\mathbf{S}(\mathbf{u}) = \mathbf{0},$$

and use the bounds (5.3.1), (5.3.2) to justify that π is in fact more regular than it follows just from (5.3.5). Indeed, we have

$$\pi \in W^{-1,\infty}(0,T;W^{1,2}(\Omega)) \cup L^p(0,T;L^q(\Omega)) \quad p,q>1. \tag{5.3.9}$$

Moreover from the equation for Z, we easily get

$$Z \in C_w([0,T];L^\infty(\Omega)) \cap C^1([0,T];W^{-1,2}(\Omega)). \tag{5.3.10}$$

Regularizing in space and time the limits Z and π using the standard multipliers ω_n, $Z_n = Z*\omega_n$, $\pi_n = \pi * \omega_n$, we can clarify the meaning of $Z\pi$ by writing

$$Z\pi = Z_n\pi_n + (Z - Z_n)\pi_n + Z(\pi - \pi_n), \tag{5.3.11}$$

and by passing to the limit with the support of mollifying kernel, see [197] for more details.

From (5.3.8) it follows that

$$\begin{aligned}&(\lambda+2\mu)\int_0^T\!\!\int_\Omega \psi\phi\left(\overline{Z\mathrm{div}\mathbf{u}} - Z\mathrm{div}\mathbf{u}\right) \ \mathrm{d}x \ \mathrm{d}t \\ &= \int_0^T\!\!\int_\Omega \psi\phi\,(\pi_1 - Z\pi) \ \mathrm{d}x \ \mathrm{d}t + \int_0^T\!\!\int_\Omega \psi\phi\left(\overline{Zp(Z)} - Z\overline{p(Z)}\right) \ \mathrm{d}x \ \mathrm{d}t \\ &\geq \int_0^T\!\!\int_\Omega \psi\phi\,(1-Z)\,\pi \ \mathrm{d}x \ \mathrm{d}t \geq 0,\end{aligned} \tag{5.3.12}$$

where to get to the r.h.s. of the above we have used subsequently: monotonicity of $p(\cdot)$, (5.3.7), and the limit of (5.3.2). Since both pairs $(Z_\varepsilon, \mathbf{u}_\varepsilon)$ and $(Z, \mathbf{u})$ satisfy the renormalized continuity equation, we can use the renormalization in the form $b(z) = z \log z$ to justify that

$$Z_\varepsilon \to Z \quad \text{strongly in } L^p((0,T)\times\Omega), \quad \forall p<\infty.$$

Note however, that this property is not transferred to the sequence ϱ_ε, for which we only have (5.3.3). Nevertheless, using this information and formula (5.3.11) we can justify that

$$\pi_1 = Z\pi,$$

which together with (5.3.7) implies (5.1.15f).

It remains to show the condition (5.1.15e), or rather its compatibility with the other conditions in system (5.1.15). This follows from the following lemma proven by Lions and Masmoudi in [153], that we recall here without the proof.

Lemma 5.15. *[153, Lemma 2.1]*
Let $\mathbf{u} \in L^2(0,T; W_0^{1,2}(\Omega,\mathbb{R}^3))$ and $f \in L^2((0,T)\times\Omega)$ such that

$$\partial_t f + \operatorname{div}(f\mathbf{u}) = 0 \quad \textit{in } (0,T)\times\Omega, \quad f(0,x) = f_0(x) \quad \textit{in } \Omega,$$

$$0 \le f_0 \le 1, \quad f_0 \not\equiv 0, \; f_0 \not\equiv 1,$$

then the following two assertions are equivalent

(i) $\operatorname{div}\mathbf{u} = 0$ *a.e. on* $\{f=1\}$,

(ii) $0 \le f(t,x) \le 1$.

Applying this lemma for $f = Z$, we conclude the proof of Theorem 5.8. □

5.4 Recovery of the Original System

In Section 5.3 we proved that system (5.1.15) possesses a weak solution in the sense of Definition 5.7. Our next aim is to prove that this solution can be identified with the solution to the original problem (5.1.1). In other words, we need to deduce the existence of $\varrho^* \in L^\infty((0,T)\times\Omega)$ satisfying the transport equation, such that the measure in the momentum equation vanishes for $\varrho = \varrho^*$.

We proceed similarly to [155]. Note that

$$\frac{\varrho_0}{Z_0}\bigg|_{\{\varrho_0=0\}} = \frac{\varrho_0}{Z_0}\bigg|_{\{Z_0=0\}} = \widetilde{\varrho^*} > 0.$$

When extended by 0 outside Ω the couples $(\varrho, \mathbf{u})$ and $(Z, \mathbf{u})$ satisfy the renormalized continuity equations, due to uniform L^∞ bounds for both ϱ_ε and Z_ε. Therefore, we may test the equations (5.1.15a), (5.1.15c) by $\omega_n(x-\cdot)$, where ω_n is a standard mollifier, which leads to

$$\partial_t \varrho_n + \operatorname{div}(\varrho_n \mathbf{u}) = r_n^1, \tag{5.4.1}$$

$$\partial_t Z_n + \operatorname{div}(Z_n \mathbf{u}) = r_n^2, \tag{5.4.2}$$

satisfied a.e. in $(0,T)\times\mathbb{R}^3$, where by a_n we denoted $a*\omega_n$. It follows from the Friedrichs commutator lemma, see e.g. [101, Lemma 10.12], that r_n^1, and r_n^2 converge to 0 strongly in $L^1((0,T)\times\mathbb{R}^3)$ as $n\to\infty$.

We now multiply (5.4.1) by $\frac{1}{Z_n+\lambda}$, and (5.4.2) by $-\frac{\varrho_n+\lambda\widetilde{\varrho^*}}{(Z_n+\lambda)^2}$, with $\lambda>0$, and obtain, after some algebraic transformations, that

$$\partial_t\left(\frac{\varrho_n+\lambda\widetilde{\varrho^*}}{Z_n+\lambda}\right)+\operatorname{div}\left[\left(\frac{\varrho_n+\lambda\widetilde{\varrho^*}}{Z_n+\lambda}\right)\mathbf{u}\right]-\left[\frac{(\varrho_n+\lambda\widetilde{\varrho^*})Z_n}{(Z_n+\lambda)^2}+\frac{\lambda\widetilde{\varrho^*}}{Z_n+\lambda}\right]\operatorname{div}\mathbf{u}$$
$$=r_n^1\frac{1}{Z_n+\lambda}-r_n^2\frac{\varrho_n+\lambda\widetilde{\varrho^*}}{(Z_n+\lambda)^2}.$$

By passing with $n\to\infty$, we get

$$\partial_t\left(\frac{\varrho+\lambda\widetilde{\varrho^*}}{Z+\lambda}\right)+\operatorname{div}\left[\left(\frac{\varrho+\lambda\widetilde{\varrho^*}}{Z+\lambda}\right)\mathbf{u}\right]-\left[\frac{(\varrho+\lambda\widetilde{\varrho^*})Z}{(Z+\lambda)^2}+\frac{\lambda\widetilde{\varrho^*}}{Z+\lambda}\right]\operatorname{div}\mathbf{u}=0. \tag{5.4.3}$$

We distinguish two cases:

Case 1. For $Z=0$, from (5.1.15d) it follows that $\varrho=0$ and therefore $\frac{\varrho+\lambda\widetilde{\varrho^*}}{Z+\lambda}=\widetilde{\varrho^*}$, and $\frac{(\varrho+\lambda\widetilde{\varrho^*})Z}{(Z+\lambda)^2}+\frac{\lambda\widetilde{\varrho^*}}{Z+\lambda}=\widetilde{\varrho^*}$, thus (5.4.3) becomes trivial.

Case 2. For $Z>0$, we notice that $\frac{\varrho+\lambda\widetilde{\varrho^*}}{Z+\lambda}\le\max\{\widetilde{\varrho^*},\frac{1}{c_\star}\}$. By means of the strong convergence of $\varrho_\lambda=\varrho+\lambda$ and $Z_\lambda=Z+\lambda$ to ϱ and Z, respectively, we can now let $\lambda\to0$ in (5.4.3) to obtain

$$\partial_t\left(\frac{\varrho}{Z}\right)+\operatorname{div}\left(\frac{\varrho}{Z}\mathbf{u}\right)-\frac{\varrho}{Z}\operatorname{div}\mathbf{u}=0.$$

Obviously, ϱ^* defined as $\frac{\varrho}{Z}$ satisfies $\varrho^*\in\left\{\min\{(c^\star)^{-1},\widetilde{\varrho^*}\},\max\{(c_\star)^{-1},\widetilde{\varrho^*}\}\right\}$ a.e. in $(0,T)\times\Omega$, and thus $Z=\frac{\varrho}{\varrho^*}$ a.e. in $(0,T)\times\Omega$. This leads to the conclusion that the condition $(1-Z)\pi=0$ can be replaced by $\left(1-\frac{\varrho}{\varrho^*}\right)\pi=0$, or, equivalently by $(\varrho^*-\varrho)\pi=0$, where the product is defined as in (5.3.11). This finishes the proof of Theorem 5.4. □

5.5 Numerical Scheme

Numerical simulation of two-phase flows with free boundary requires to design a method that captures phase transition and the limit behaviour. In our case, the main difficulty is to propose a scheme that is independent of singular pressure parameter ε (5.1.9). This property is referred to as the Asymptotic Preserving (AP) property, see e.g.[73]. The passage with $\varepsilon\to0$ resembles the low Mach number limit problem, where one observes incompressible behaviour in regions where the Mach number approaches 0. However, in the contrary to the low Mach number limit, in our model the singularity is embedded in the definition of the singular pressure π.

We adapt numerical method from [78] where the Euler system with constant maximal density constraint has been studied. One dimensional version of *the Direct method* is modified and extended to capture variable density constraint and the viscosity term in the momentum equation.

In what follows we focus only on the new elements of our approach, for the detailed description of the other parts we refer to [78] and references therein. Following this rule, we present our technique on the time semi-discrete level, the discretization in space is omitted for brevity.

5.5.1 Discretization Scheme

To find an approximate solution to the system (5.1.8) we propose a splitting algorithm. Time is discretized by one step finite difference (with fixed time step Δt) and a finite volume method is used in space. At each time step the set of the equations is decomposed into three parts which are solved subsequently in three sub-steps.

Step 1: Hyperbolic Part

The numerical solution of the Euler part of the system follows the strategy presented in [78]. The flux in the mass balance and the singular pressure are treated implicitly:

$$\frac{\varrho^{n+1} - \varrho^n}{\Delta t} + \mathrm{div}(\varrho^{n+1}\mathbf{u}^*) = 0, \tag{5.5.1a}$$

$$\frac{(\varrho^{n+1}\mathbf{u}^*) - (\varrho^n\mathbf{u}^n)}{\Delta t} + \mathrm{div}(\varrho^n\mathbf{u}^n \otimes \mathbf{u}^n) + \nabla\pi_\varepsilon\left(\frac{\varrho^{n+1}}{\varrho^{*n}}\right) + \nabla p\left(\frac{\varrho^n}{\varrho^{*n}}\right) = \mathbf{0}. \tag{5.5.1b}$$

The system (5.5.1) is reformulated on a discrete level in terms of singular pressure. Into (5.5.1a) we substitute implicit mass flux from (5.5.1b) and obtain an elliptic equation for the singular pressure

$$\varrho^{n+1}(\pi_\varepsilon) - (\Delta t)^2\Delta\pi_\varepsilon\left(\frac{\varrho^{n+1}}{\varrho^{*n}}\right) = \phi(\varrho^n, \varrho^{*n}, \mathbf{u}^n), \tag{5.5.2}$$

where the right hand side of (5.5.2) reads

$$\phi(\varrho^n, \varrho^{*n}, \mathbf{u}^n) = \varrho^n - (\Delta t)\mathrm{div}(\varrho^n\mathbf{u}^n) + (\Delta t)^2\mathrm{div}\left(\mathrm{div}(\varrho^n\mathbf{u}^n \otimes \mathbf{u}^n) + \nabla p\left(\frac{\varrho^n}{\varrho^{*n}}\right)\right).$$

The singular pressure π_ε is computed by solving (5.5.2) by means of the Newton method with numerical Jacobian. In the next step we invert singular pressure to get the density. The purpose of this approach is to ensure that the density constrain is satisfied ($\varrho \le \varrho^*$), which now follows from the definition of π_ε.

After the new density is obtained we directly update the momentum. This approach is called *the Direct method*, see [78, Section 4.1]. The second approach presented in the literature is referred to as *the Gauge method* [73] that is based on the decomposition of the momentum $\varrho\mathbf{u} = \mathbf{a} + \nabla\varphi$, $\mathrm{div}\mathbf{a} = 0$ into a divergence free part $\mathbf{a}$, and the irrotational part φ. As reported in [78] the Direct method indicates oscillations of the velocity in congested part, while the Gauge method is diffusive in uncongested region. Since the first method does not introduce any additional numerical dissipation, we adapt it for this work. For detailed description of the space discretization we refer to [78].

We would like to emphasize that (5.5.1) is a strictly hyperbolic problem, with characteristic wave speeds $\lambda_{1,2} = \mathbf{u} \pm \sqrt{\frac{\partial p}{\partial(\varrho/\varrho^*)}}$. By the definition, the Courant–Friedrichs–Lewy (CFL) condition for the explicit part is equal to $\max(|\lambda_{1,2}|) \leq \sigma\frac{\Delta x}{\Delta t}$, with the Courant number σ. Splitting for implicit singular and explicit background pressure (5.5.1b) provides that CFL condition is satisfied uniformly in ε.

Step 2: Diffusion

For the sake of numerical simulations, we consider (5.1.3) in a simplified form $\mathbf{S}(\mathbf{u}) = 2\mu\Delta\mathbf{u}$. We treat the diffusion term implicitly to avoid additional stability restrictions:

$$\frac{(\varrho^{n+1}\mathbf{u}^{n+1}) - (\varrho^{n+1}\mathbf{u}^*)}{\Delta t} + 2\mu\Delta\mathbf{u}^{n+1} = 0. \tag{5.5.3}$$

The presence of the diffusion term is important from analytical reasons only. In fact, the presented numerical scheme has been designed to solve the Euler system, therefore one can take arbitrary viscosity, such that $\mu \geq 0$. The viscosity coefficient is fixed independently of ε and small enough to recover compressible/incompressible transition. Equation (5.5.3) is discretized in space by cell-centered finite volume scheme.

Step 3: Congestion Transport

The transport of the congested density is undoubtedly a main new feature of the model (5.1.8) and so of the presented numerical scheme. Having the new velocity $\mathbf{u}^{n+1}$ we compute the congested density as follows

$$\frac{\varrho^{*n+1} - \varrho^{*n}}{\Delta t} + \mathbf{u}^{n+1}\nabla\varrho^{*n} = 0,$$

where a cell-centered finite volume scheme together with upwind is used in space.

5.5.2 Numerical Results

In this section we present four numerical examples that demonstrate behaviour of the proposed model in one-dimensional periodic setting. As a consequence of finite volume framework the proposed scheme conserves mass. As for domain we take the unit interval with the mesh size $\Delta x = 10^{-3}$ and the time-step $\Delta t = 10^{-4}$. In the following we choose singular pressure parameter $\varepsilon = 10^{-4}$ with the exponents $\alpha = \beta = 2$, (5.1.9), and background pressure (5.1.2) with the exponent $\gamma = 2$, if not stated differently.

The test cases are:

- Case 1 (constant congestion):

$$\begin{cases} \varrho(x,0) &= 0.7, \\ \varrho^*(x,0) &= 1.0, \\ \mathbf{u}(x,0) &= \begin{cases} 0.8 \text{ if } 0.2 < x < 0.6 \\ -0.8 \text{ otherwise} \end{cases} \end{cases},$$

- Case 2:

$$\begin{cases} \varrho(x,0) & = 0.7, \\ \varrho^*(x,0) & = 0.8 + 0.15\left(\tanh(50(x-0.4)) - \tanh(50(x-0.6))\right), \\ \mathbf{u}(x,0) & = \begin{cases} 0.8 \text{ if } 0.25 < x < 0.5 \\ -0.8 \text{ if } 0.5 < x < 0.75 \\ 0.0 \text{ otherwise} \end{cases}, \end{cases}$$

- Case 3:

$$\begin{cases} \varrho(x,0) & = \begin{cases} 0.8 \text{ if } 0.3 < x < 0.7 \\ 0.1 \text{ otherwise} \end{cases}, \\ \varrho^*(x,0) & = 0.34 + 0.3(\tanh(50(x-0.275)) - \tanh(50(x-0.725))), \\ \mathbf{u}(x,0) & = \begin{cases} 0.8 \text{ if } 0.1 < x < 0.7 \\ 0.0 \text{ otherwise} \end{cases}, \end{cases}$$

- Case 4:

$$\begin{cases} \varrho(x,0) & = 0.6, \\ \varrho^*(x,0) & = 0.9 + 0.05(\cos(10\pi x) - \cos(6\pi x) + \cos(134\pi x) + \cos(24\pi x)), \\ \mathbf{u}(x,0) & = \begin{cases} 0.8 \text{ if } 0.3 < x < 0.7 \\ -0.8 \text{ otherwise} \end{cases}, \end{cases}$$

Case 1 illustrates shock and rarefaction of the density for constant initial congestion density. The initial value of the congestion density stays the same for all times, due to transport. The congestions and rarefactions are created solely due to opposite initial velocities, exactly as in the analogous case from [78]. In Figure 5.1 we moreover present the behaviour of unknowns for different values of parameter in the singular pressure: $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. These numerical results show that the algorithm indeed satisfies the Asymptotic Preserving property. Note that with ε decreasing to zero we approach incompressible limit more thoroughly. However, the exact value of the maximal density constraint can never be reached by the numerically computed density.

Case 2 and Case 3 show the main feature of presented model, namely variable congestion density. For both cases the initial maximal density is set to “smooth hat”. In the first of them the initial velocity describes the velocities of two groups of individuals that want to go in the opposite directions. The individuals close to the contact line are more willing to compress (as the congestion density is higher). We see that initially the individuals at the rear of the groups press so intensively onto the front members, so that the whole “hat” is tightly filled. When this happens, we see a similar effect as for elastic collision: part of individuals start to move in the opposite direction to initially intended.

Case 3 describes a situation, when the well organized crowd moving in one direction with the same velocity approaches a barrier ahead, being the group of individuals that move much slower and prefer to keep bigger distances between each other. We observe how the faster individuals behind push the slower group to speed up, by filling the all available gaps between the individuals (this is where the congestion occurs at position $x \approx 0.8$). This kind of behaviour could be observed, for example, at airports or in the groups of marathon runners.

Case 4 illustrates shock and rarefaction of the density when the maximal density constrain consists of a sum of cosines (periodic setting) with different frequencies. This example mimics randomness in the individual preferences of the members of population. We observe that congested regions “freeze” maximal density due to the zero velocity, which is consistent with the theoretical prediction. As expected from the properties of the limiting system (5.1.1f), for $\varepsilon \ll 1$, the singular pressure π_ε is activated only in the congested region. In the theoretical part of the paper, this pressure is merely a nonnegative measure in the limit and our simulations seem to confirm this lack of regularity.

Another interesting feature observed in this case is the travelling wave-like behaviour of the density of the crowd. Note that taking time derivative of (5.1.1a) and substituting $\partial_t(\varrho\mathbf{u})$ from (5.1.1b) we obtain wave-like equation for density. We observe this effect on the Figure 5.4 between time $t = 0.25$ and $t = 0.5$ at $x = 0.2$, where two "crowds" interfere with each other. This leads to reaching the congestion density and propagation of the congestion in the opposite directions.

Looking at the Figure 5.4 it seems that the scheme for the transport of the congestion density is quite diffusive. The high spatial frequency oscillations present at the beginning are very quickly washed away and there only subsists the small frequency components. Thus, the numerical examples presented above should be treated just as the illustration of the behaviour of solutions to the approximation (5.1.8).

The thorough numerical discussion as well as the study of two-dimensional case, has been addressed in [79]. It has been shown that the presented model, however without diffusion, exhibits typical crowd behaviour like: stop-and-go waves and faster goes slower effect.

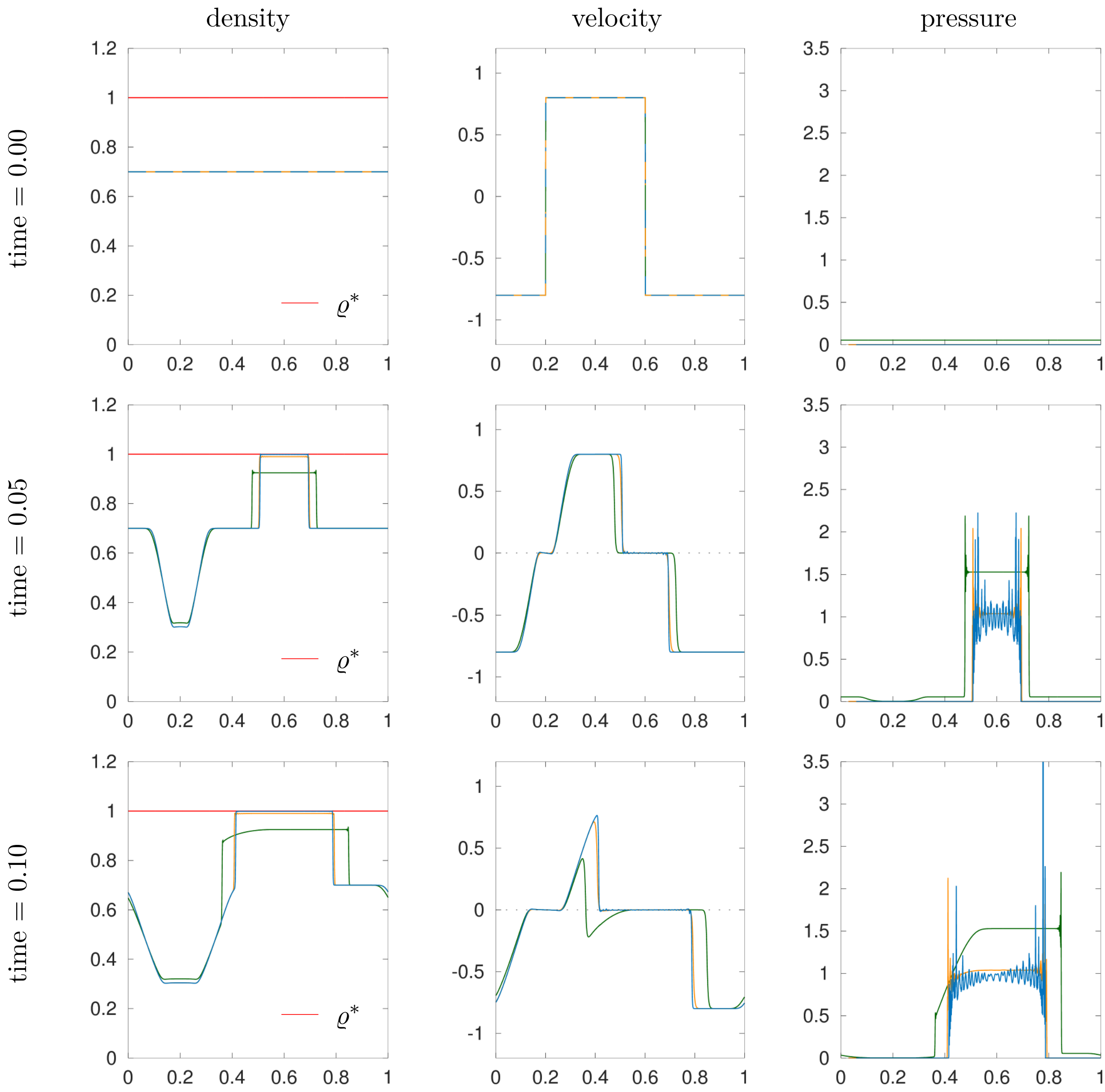


Figure 5.1: Case 1: density, velocity, and singular pressure for $\varepsilon = 10^{-2}$(green), $\varepsilon = 10^{-4}$(yellow), $\varepsilon = 10^{-6}$(blue). A video available at https://youtu.be/zB0Czbj1bsw.

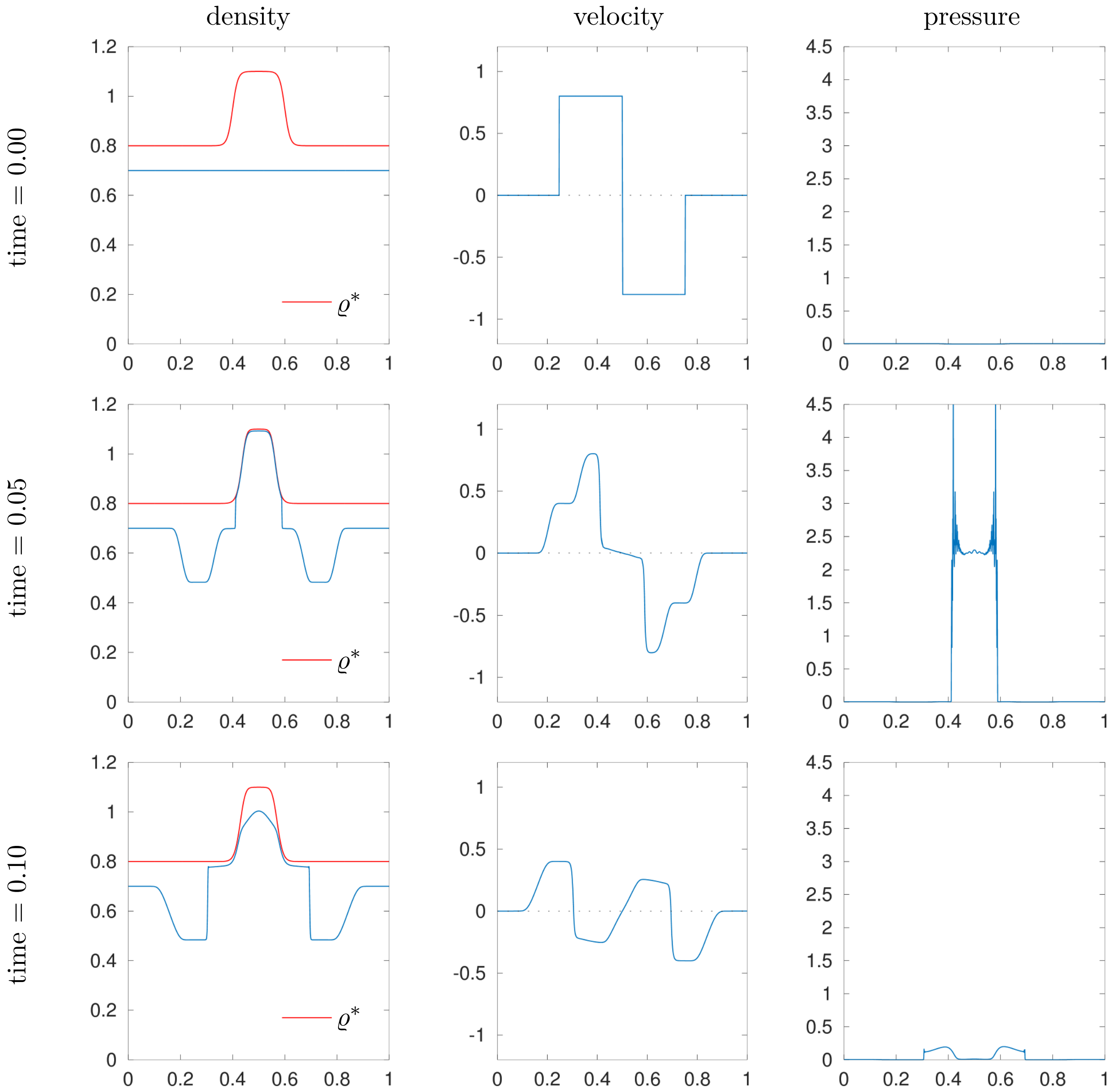


Figure 5.2: Case 2: density, velocity and singular pressure. A video of available at https://youtu.be/5zLualVGdBs.

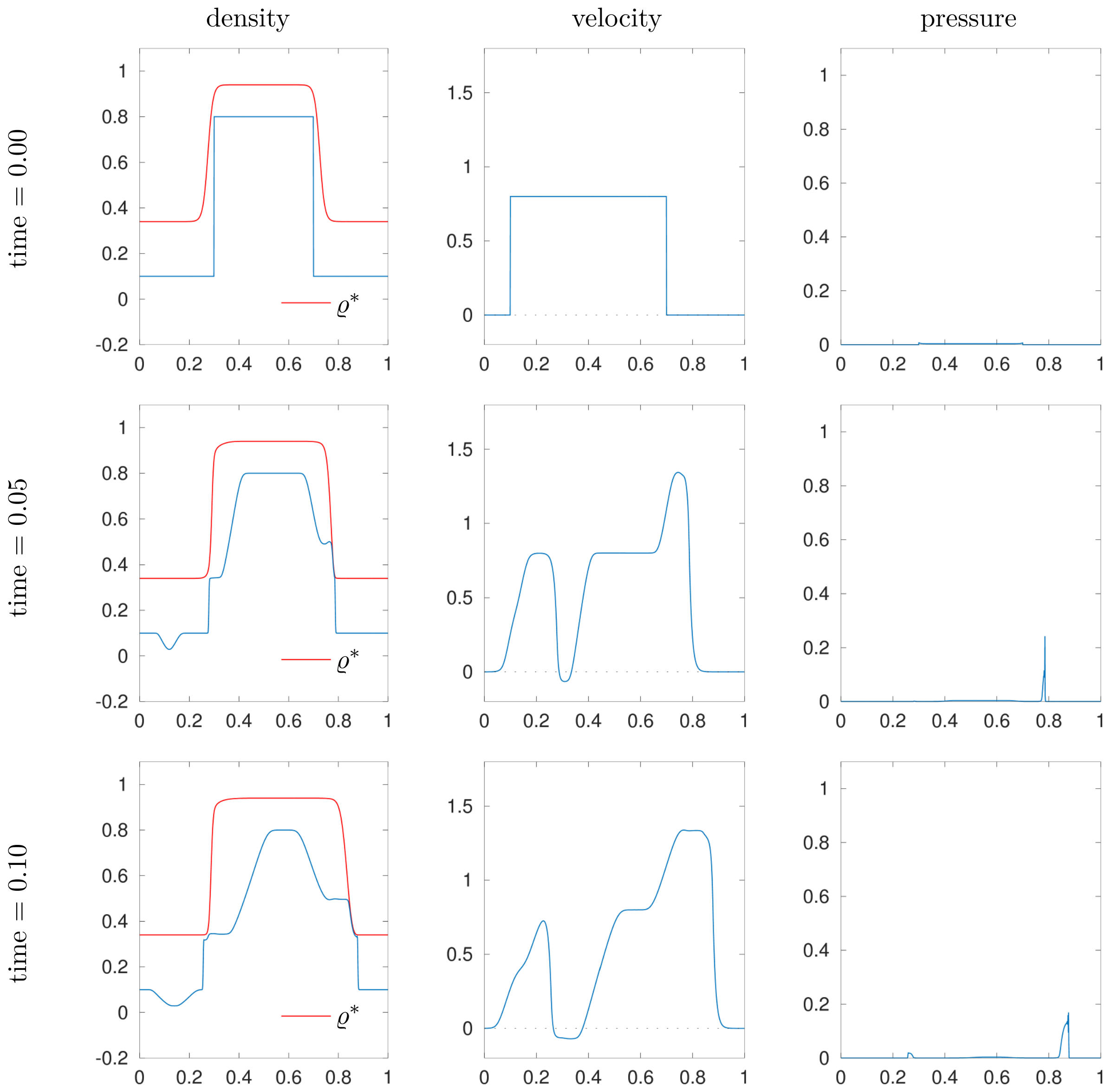


Figure 5.3: Case 3: density, velocity and singular pressure. A video available at `https://youtu.be/GhhZ0c4xk6k`.

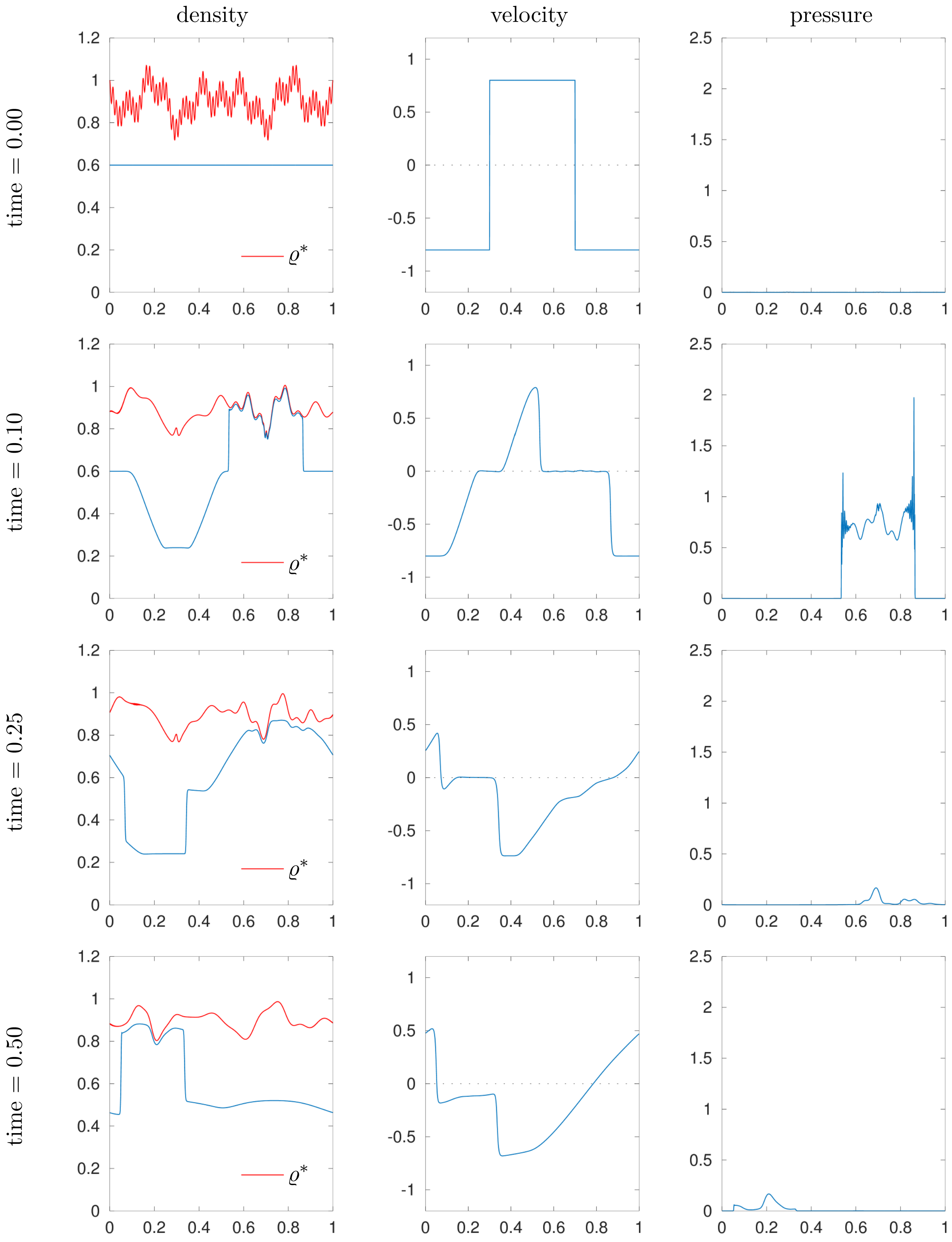


Figure 5.4: Case 4: density, velocity and singular pressure. A video available at https://youtu.be/Ql2R2qEn9lM.

CHAPTER 6

Finite Volume Approximations of the Euler System with Variable Congestion

The content of this chapter is joint work with Pierre Degond, Ewelina Zatorska, and Laurent Navoret and is published in the paper

- P. Degond, P. Minakowski, L. Navoret and E. Zatorska. ‘Finite Volume Approximations of the Euler System with Variable Congestion’. In: *Computers & Fluids* 169 (2018). Recent progress in nonlinear numerical methods for time-dependent flow & transport problems, pp. 23–39. doi: 10.1016/j.compfluid.2017.09.007

Chapter Summary. We are interested in the numerical simulations of the Euler system with variable congestion encoded by a singular pressure [80]. This model describes for instance the macroscopic motion of a crowd with individual congestion preferences. We propose an asymptotic preserving (AP) scheme based on a conservative formulation of the system in terms of density, momentum and density fraction. A second order accuracy version of the scheme is also presented. We validate the scheme on one-dimensionnal test-cases and compare it with a scheme previously proposed in [80] and extended here to higher order accuracy. We finally carry out two dimensional numerical simulations and show that the model exhibit typical crowd dynamics.

Chapter Organisation. In Section 6.2 we present our numerical schemes using the two formulations (6.1.1) and (6.1.8). They are referred to as $(\varrho, \boldsymbol{q})$-method/SL and $(\varrho, \boldsymbol{q}, Z)$-method, respectively. In Section 6.2.1 we describe the first-order semi-discretization in time and the full discretization for the $(\varrho, \boldsymbol{q}, Z)$-method. Then, in Section 6.2.2, we discuss the second order scheme for the $(\varrho, \boldsymbol{q}, Z)$-method. At last, in Section 6.2.3 we present the $(\varrho, \boldsymbol{q})$-method/SL for the system written in terms of the physical variables (6.1.1). Section 6.3 is devoted to validation of the schemes on the Riemann problem whose solutions are described in 6.5. Finally, in Section 6.4 we discuss the two-dimensional numerical results: in Section 6.4.1 we present how these schemes work for three different initial congestion densities, and in Section 6.4.2 we present an application of $(\varrho, \boldsymbol{q})$-method/SL to model crowd behaviour in the evacuation scenario.

Contents of Chapter

6.1 Introduction

In this work we study two phase compressible/incompressible Euler system with variable congestion:

$$\partial_t \varrho + \operatorname{div}(\varrho \mathbf{u}) = 0, \tag{6.1.1a}$$

$$\partial_t(\varrho \mathbf{u}) + \operatorname{div}(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla \pi + \nabla p\left(\frac{\varrho}{\varrho^*}\right) = \mathbf{0}, \tag{6.1.1b}$$

$$\partial_t \varrho^* + \mathbf{u} \cdot \nabla \varrho^* = 0, \tag{6.1.1c}$$

$$0 \leq \varrho \leq \varrho^*, \tag{6.1.1d}$$

$$\pi(\varrho^* - \varrho) = 0, \quad \pi \geq 0, \tag{6.1.1e}$$

with the initial data

$$\varrho(0,x) = \varrho_0(x) \geq 0, \quad \mathbf{u}(0,x) = \mathbf{u}_0(x), \quad \varrho^*(0,x) = \varrho_0^*(x), \quad \varrho_0 < \varrho_0^*, \tag{6.1.2}$$

where the unknowns are: $\varrho = \varrho(t,x)$ – the mass density, $\mathbf{u} = \mathbf{u}(t,x)$ – the velocity, $\varrho^* = \varrho^*(t,x)$ – the congestion density, and π – the congestion pressure. The barotropic pressure p is an explicit function of the density fraction $\frac{\varrho}{\varrho^*}$

$$p\left(\frac{\varrho}{\varrho^*}\right) = \left(\frac{\varrho}{\varrho^*}\right)^\gamma, \quad \gamma > 1, \tag{6.1.3}$$

and plays the role of the background pressure.

The congestion pressure π appears only when the density ϱ satisfying (6.1.1d) achieves its maximal value, the congestion density ϱ^*. Therefore ϱ^* can be referred to as the barrier or the threshold density. It was observed in [153], and then generalized in [80], that the restriction on the density (6.1.1d) is equivalent with the condition

$$\mathrm{div}\mathbf{u} = 0 \text{ in } \{\varrho = \varrho^*\}, \tag{6.1.4}$$

if only $\varrho, \mathbf{u}, \varrho^*$ are sufficiently regular solutions of the continuity equation (6.1.1a) and the transport equation (6.1.1c). For that reason, system (6.1.1) can be seen as a free boundary problem for the interface between the compressible (uncongested) regime $\{\varrho < \varrho^*\}$ and the incompressible (congested) regime $\{\varrho = \varrho^*\}$.

The main purpose of this work is to analyze (6.1.1) numerically, i.e. to propose the numerical scheme capturing the phase transition. To this end we use the fact that (6.1.1) can be obtained as a limit when $\varepsilon \to 0$ of the compressible Euler system:

$$\partial_t \varrho + \mathrm{div}(\varrho \mathbf{u}) = 0, \tag{6.1.5a}$$

$$\partial_t(\varrho \mathbf{u}) + \mathrm{div}(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla \pi_\varepsilon + \nabla p \left(\frac{\varrho}{\varrho^*}\right) = \mathbf{0}, \tag{6.1.5b}$$

$$\partial_t \varrho^* + \mathbf{u} \cdot \nabla \varrho^* = 0, \tag{6.1.5c}$$

with the singular approximation π_ε of the congestion pressure:

$$\pi_\varepsilon \left(\frac{\varrho}{\varrho^*}\right) = \varepsilon \left(\frac{\frac{\varrho}{\varrho^*}}{1 - \frac{\varrho}{\varrho^*}}\right)^\alpha, \quad \alpha > 0. \tag{6.1.6}$$

The singularity of the pressure π_ε implies that for every $\varepsilon > 0$ fixed $\varrho_\varepsilon \leq \varrho^*$. Note that for fixed $\varepsilon > 0$, $\pi_\varepsilon \to \infty$ when $\varrho \to \varrho^*$. Therefore, at least formally, for $\varepsilon \to 0$, π_ε converges to a measure supported on the set of singularity, i.e. $\{(x,t) \in \Omega \times (0,T) : \varrho(x,t) = \varrho^*(x,t)\}$. The rigorous proof of this fact is an open problem, at least for the Euler type of systems. There have been, however, several results for a viscous version of the model, see [43] for the one-dimensional case, [197] for multi-dimensional domains and space-dependent congestion $\varrho^*(x)$ and [80] for the case of congestion density satisfying the transport equation (6.1.1c). The last of mentioned results requires a technical assumption $\alpha > 5/2$ for the 3-dimensional domain. Intuitively, the value of parameter α indicates the strength of singularity of the pressure close to $\varrho = \varrho^*$. However, since taking the limit $\varepsilon \to 0$ magnifies this singularity, the value of $\alpha > 0$ might be arbitrary small for sufficiently small ε. An alternative approximation leading to a similar two-phase system was considered first by P.-L. Lions and N. Masmoudi [153], and more recently for the model of tumour growth [198]. The advantage of approximation (6.1.6) considered here lies in the fact that for each ε fixed, the solutions to the approximate system stay in the physical regime, i.e. $\varrho \leq \varrho^*$. This feature is especially important for the numerical purposes, see for example [161] for further discussion on this subject.

System (6.1.1) is a generalization of the pressureless Euler system with the maximal density constraint

$$\partial_t \varrho + \operatorname{div}(\varrho \mathbf{u}) = 0, \tag{6.1.7a}$$

$$\partial_t(\varrho \mathbf{u}) + \operatorname{div}(\varrho \mathbf{u} \otimes \mathbf{u}) + \nabla \pi = \mathbf{0}, \tag{6.1.7b}$$

$$0 \leq \varrho \leq 1 \tag{6.1.7c}$$

$$\pi(\varrho - 1) = 0, \quad \pi \geq 0. \tag{6.1.7d}$$

introduced originally by Bouchut et al. [38], who also proposed the first numerical scheme based on an approach developed earlier for the pressureless systems, see for example [40], and the projection argument. The model was studied later on by Berthelin [32, 33] by passing to the limit in the so-called sticky-blocks dynamics, see also [240], and a very interesting recent paper [195] using the Lagrangian approach for the monotone rearrangement of the solution to prove the existence of solutions to (6.1.7) with additional memory effects.

The pressureless Euler equations with the density constraint were originally introduced in order to describe the motion of particles of finite size. Our model extends this concept by including the variance of the size of particles. In system (6.1.1) ϱ^* is given initially and is transported along with the flow.

One can also think of ϱ^* as a congestion preference of individuals moving in the crowd (cars, pedestrians), which is one of the factors determining their final trajectory and the speed of motion. The macroscopic modelling of crowd is one of possible approaches and it allows to determine the averaged quantities such as the density and the mean velocity rather than the precise position of an individual. One of the first models of this kind based on classical mechanics was introduced by Henderson [127]. More sophisticated model was introduced by Hughes [131] where the author considers the continuity equation equipped with a phenomenological constitutive relation between the velocity and the density. For a survey of the crowd models we refer the reader to [62, 27, 201, 160, 134] and to the review paper [28].

As far as the numerical methods are concerned, the macroscopic models of pedestrian flow with condition preventing the overcrowding were studied, for example in [234]. The influence of the maximal density constraint was investigated also in the context of vehicular traffic in [35]. The strategy that we want to adapt in this paper, i.e. to use the singularities of the pressure similar to (6.1.6) has been developed in the past for a number of Euler-like systems for the traffic models [35, 31, 34], collective dynamics [78, 77], or granular flow [159, 196]. In our previous work [80], we have drafted the numerical scheme for system (6.1.1) in the one-dimensional case. We used a splitting algorithm at each time step that consists of three sub-steps. At first, the hyperbolic part is solved with the AP-preserving method presented in [78]. Next the diffusion is solved by means of cell-centered finite volume scheme, and the transport of the congested density is resolved with the upwind scheme.

The extension of this method to two-dimensions is one of the main results of the present paper. We also propose an alternative scheme using different formulation in terms of the *conservative variables*: the density ϱ, the momentum $\boldsymbol{q} = \varrho \mathbf{u}$, and the density fraction $Z = \frac{\varrho}{\varrho^*}$:

$$\partial_t \varrho + \operatorname{div} \boldsymbol{q} = 0, \tag{6.1.8a}$$

$$\partial_t \boldsymbol{q} + \operatorname{div}\left(\frac{\boldsymbol{q} \otimes \boldsymbol{q}}{\varrho} + \pi_\varepsilon(Z)\boldsymbol{I} + p(Z)\boldsymbol{I} \right) = \mathbf{0}, \tag{6.1.8b}$$

$$\partial_t Z + \operatorname{div}\left(Z\frac{\boldsymbol{q}}{\varrho}\right) = 0, \tag{6.1.8c}$$

with the initial data

$$\varrho(0,x) = \varrho_0(x), \quad \boldsymbol{q}(0,x) = \mathbf{q}_0(x), \quad Z(0,x) = Z_0(x), \tag{6.1.8d}$$

where $Z_0 = \frac{\varrho_0}{\varrho_0^*}$, and $\mathbf{q}_0 = \varrho_0\mathbf{u}_0$. $\boldsymbol{I}$ denotes the identity tensor. This is a strictly hyperbolic system whose wave speeds in the x_1-direction are given by:

$$\begin{aligned}
\lambda_1^\varepsilon(\varrho, q_1, Z) &= \frac{q_1}{\varrho} - \sqrt{\frac{Z}{\varrho}p_\varepsilon'(Z)},\\
\lambda_2^\varepsilon(\varrho, q_1, Z) &= \frac{q_1}{\varrho},\\
\lambda_3^\varepsilon(\varrho, q_1, Z) &= \frac{q_1}{\varrho} + \sqrt{\frac{Z}{\varrho}p_\varepsilon'(Z)},
\end{aligned} \tag{6.1.9}$$

where $p_\varepsilon = p + \pi_\varepsilon$, and q_1 denotes the component of $\boldsymbol{q}$ in the x_1 direction. Consequently, in region where the density ϱ is closely congested, i.e. Z is close to 1, the characteristic speeds of the system are extremely large. This corresponds to the nearly incompressible dynamics.

The first scheme uses the *physical variables* $\varrho, \mathbf{u}, \varrho^*$ as in (6.1.1), while the second one uses the *conservative variables*: the density ϱ, the momentum $\boldsymbol{q} = \varrho\mathbf{u}$, and the density fraction $Z = \frac{\varrho}{\varrho^*}$.

6.2 Numerical Schemes

In this section, we first introduce a numerical scheme based on system (6.1.8) using the conservative variables. In order to use large time steps not restricted by too drastic CFL condition, implicit-explicit (IMEX) type methods need to be designed. The scheme can be solved through the following steps: first an elliptic equation on the density fraction Z is solved, and then we update $\boldsymbol{q}$ and ϱ, respectively.

Such scheme is compared with an extension of the method introduced in [80], where the congestion density is advected separately from the update of ϱ and $\boldsymbol{q}$. For the sake of completeness, a description of the scheme is given in Section 6.2.3.

Note that the scheme is unable to deal with vacuum. In what follows we require that $\varrho_0 > 0$ (vacuum is not allowed in the initial data). However, the effect of the background pressure (6.1.3) is to smear out the vacuum regions.

6.2.1 The First Order (ϱ, q, Z)-Method

Discretization in time We adopt the previous work [78] to introduce a method treating implicitly the stiff congestion pressure $\pi_\varepsilon(Z)$. We consider a constant time step $\Delta t > 0$ and ϱ^n, $\boldsymbol{q}^n$, Z^n, ϱ^{*n} denote the approximate solution at time $t^n = n\Delta t$, $\forall n \in \mathbb{N}$. We thus consider the following semi-implicit time discretization:

$$\frac{\varrho^{n+1}-\varrho^n}{\Delta t} + \nabla_x \cdot \boldsymbol{q}^{n+1} = 0, \tag{6.2.1a}$$

$$\frac{\boldsymbol{q}^{n+1}-\boldsymbol{q}^n}{\Delta t} + \nabla_x \cdot \left(\frac{\boldsymbol{q}^n \otimes \boldsymbol{q}^n}{\varrho^n} + p(Z^n)\boldsymbol{I}\right) + \nabla_x(\pi_\varepsilon(Z^{n+1})) = 0, \tag{6.2.1b}$$

$$\frac{Z^{n+1}-Z^n}{\Delta t} + \nabla_x \cdot \left(Z^n \frac{\boldsymbol{q}^{n+1}}{\varrho^n}\right) = 0. \tag{6.2.1c}$$

Note that in the flux term in equation (6.2.1c), the momentum is taken implicitly. Inserting (6.2.1b) into (6.2.1c), we obtain:

$$\begin{aligned}&\frac{Z^{n+1}-Z^n}{\Delta t} + \nabla_x \cdot \left(Z^n \frac{\boldsymbol{q}^n}{\varrho^n}\right) \\ &\quad - \Delta t\, \nabla_x \cdot \left(\frac{Z^n}{\varrho^n}\nabla_x \cdot \left(\frac{\boldsymbol{q}^n \otimes \boldsymbol{q}^n}{\varrho^n} + p(Z^n)\boldsymbol{I}\right) + \frac{Z^n}{\varrho^n}\nabla_x(\pi_\varepsilon(Z^{n+1}))\right) = 0,\end{aligned}$$

This is an elliptic equation on the unknown Z^{n+1}, that can be written as:

$$Z^{n+1} - \Delta t^2\, \nabla_x \cdot \left(\frac{Z^n}{\varrho^n}\nabla_x\left(\pi_\varepsilon(Z^{n+1})\right)\right) = \phi(\varrho^n,\ q^n, Z^n), \tag{6.2.2}$$

where

$$\begin{aligned}&\phi(\varrho^n,\ q^n, Z^n) \\ &\quad = Z^n + \Delta t^2\, \nabla_x \cdot \left(\frac{Z^n}{\varrho^n}\nabla_x \cdot \left(\frac{\boldsymbol{q}^n \otimes \boldsymbol{q}^n}{\varrho^n} + p(Z^n)\boldsymbol{I}\right)\right) - \Delta t\, \nabla_x \cdot \left(Z^n \frac{\boldsymbol{q}^n}{\varrho^n}\right).\end{aligned}$$

The n-th time step of the scheme is decomposed into three parts: first get Z^{n+1} when solving (6.2.2), then compute $\boldsymbol{q}^{n+1}$ thanks to (6.2.1b) and then ϱ^{n+1} from (6.2.1a).

Discretization in space We only derive the fully discrete scheme in the one-dimensional case; the two-dimensional formula are given in 6.6. We consider the computational domain $[0,1]$ and a spatial space step $\Delta x = 1/N_x > 0$, with $N_x \in \mathbb{N}$: the mesh points are thus $x_i = i\Delta x$, $\forall i \in \{0,\dots,N_x\}$. Let ϱ_i^n, $\boldsymbol{q}_i^n$, Z_i^n, $\varrho_i^{*\,n}$ denote the approximate solution at time t^n on mesh cell $[x_i, x_{i+1}]$. The spatial discretization have to capture correctly the entropic solutions of the hyperbolic system. To derive the fully discrete scheme, we thus make the same algebra on the following fully discrete system:

$$\frac{\varrho_i^{n+1}-\varrho_i^n}{\Delta t} + \frac{1}{\Delta x}(F^{n+1}_{i+\frac{1}{2}} - F^{n+1}_{i-\frac{1}{2}}) = 0, \tag{6.2.3a}$$

$$\frac{q_i^{n+1}-q_i^n}{\Delta t} + \frac{1}{\Delta x}(G^n_{i+\frac{1}{2}} - G^n_{i-\frac{1}{2}}) + \frac{\pi_\varepsilon(Z^{n+1}_{i+1}) - \pi_\varepsilon(Z^{n+1}_{i-1})}{2\Delta x} = 0, \tag{6.2.3b}$$

$$\frac{Z_i^{n+1}-Z_i^n}{\Delta t} + \frac{1}{\Delta x}(H^{n+1}_{i+\frac{1}{2}} - H^{n+1}_{i-\frac{1}{2}}) = 0. \tag{6.2.3c}$$

where the stiff pressure is discretized by the centered finite difference and the numerical fluxes F^{n+1}, G^n, H^{n+1} (we denote implicit-explicit fluxes by current timestep $n+1$ and fully explicit fluxes by previous timestep n) are splitted into centered part and the upwinded part:

$$F^{n+1}_{i+\frac{1}{2}} = \frac{1}{2}\Big(q^{n+1}_{i+1} + q^{n+1}_i\Big) - (D_\varrho)^n_{i+\frac{1}{2}}, \tag{6.2.4}$$

$$G^{n}_{i+\frac{1}{2}} = \frac{1}{2}\bigg(\frac{(q^n_{i+1})^2}{\varrho^n_{i+1}} + \frac{(q^n_i)^2}{\varrho^n_i} + p(Z^n_{i+1}) + p(Z^n_i)\bigg) - (D_q)^n_{i+\frac{1}{2}}, \tag{6.2.5}$$

$$H^{n+1}_{i+\frac{1}{2}} = \frac{1}{2}\bigg(\frac{Z^n_{i+1}}{\varrho^n_{i+1}}q^{n+1}_{i+1} + \frac{Z^n_i}{\varrho^n_i}q^{n+1}_i\bigg) - (D_Z)^n_{i+\frac{1}{2}}. \tag{6.2.6}$$

The upwinded parts are given explicitly. They can be given by the diagonal Rusanov (or local Lax-Friedrichs) upwindings:

$$(D_w)^n_{i+\frac{1}{2}} = \frac{1}{2}c^n_{i+\frac{1}{2}}\Big(w^n_{i+1} - w^n_i\Big), \tag{6.2.7}$$

for any conserved quantities w, where $c^n_{i+\frac{1}{2}}$ is the maximal characteristic speed (in absolute value):

$$c^n_{i+\frac{1}{2}} = \max\left\{\left|\lambda^0_k\Big(\varrho^n_{i+1}, q^n_{i+1}, Z^n_{i+1}\Big)\right|, \left|\lambda^0_k\Big(\varrho^n_i, q^n_i, Z^n_i\Big)\right|, \quad k=1,2,3\right\}, \tag{6.2.8}$$

where λ^0_k are given by eq. (6.1.9) with $\varepsilon = 0$ (no congestion pressure). These correspond to the eigenvalues of the hyperbolic system taken explicitly in (6.2.1). One could also consider less diffusive numerical fluxes like the Polynomial upwind scheme [75].

Like in the semi-discrete case, we now obtain the fully discrete elliptic equation on Z by replacing the implicit momentum terms appearing in the flux H (6.2.6) by their expressions given by the momentum equation (6.2.3b). We get:

$$\begin{aligned}
&Z^{n+1}_i - Z^n_i + \frac{\Delta t}{\Delta x}(\bar{H}^n_{i+1/2} - \bar{H}^n_{i-1/2}) \\
&- \frac{\Delta t^2}{\Delta x^2}\frac{1}{2}\bigg(\frac{Z^n_{i+1}}{\varrho^n_{i+1}}(G^n_{i+\frac{3}{2}} - G^n_{i+\frac{1}{2}}) - \frac{Z^n_{i-1,j}}{\varrho^n_{i-1}}(G^n_{i-\frac{1}{2}} - G^n_{i-\frac{3}{2}})\bigg) \\
&- \frac{\Delta t^2}{\Delta x^2}\frac{1}{2}\bigg(\frac{Z^n_{i+1}}{\varrho^n_{i+1}}\Big(\pi_\varepsilon(Z^{n+1}_{i+2}) - \pi_\varepsilon(Z^{n+1}_i)\Big) - \frac{Z^n_{i-1}}{\varrho^n_{i-1}}\Big(\pi_\varepsilon(Z^{n+1}_i) - \pi_\varepsilon(Z^{n+1}_{i-2})\Big)\bigg) = 0,
\end{aligned}$$

where $\bar{H}^n$ denotes the same expression as (6.2.6) where all quantities are taken explicitly:

$$\bar{H}^n_{i+\frac{1}{2}} = \frac{1}{2}\bigg(\frac{Z^n_{i+1}}{\varrho^n_{i+1}}q^n_{i+1} + \frac{Z^n_i}{\varrho^n_i}q^n_i\bigg) - (D_Z)^n_{i+\frac{1}{2}}.$$

As explained in the introduction the main advantage of approximating the system (6.1.1) by (6.1.5) with the singular pressure (6.1.6) is that it allows to keep the physical constraint $Z \leq 1$ on each level of approximation. In fact, for $\varepsilon > 0$ fixed, our numerical scheme provides that $Z < 1$ in the whole domain. For this to hold we solve first this elliptic equation with respect to the congestion pressure variable π_ε:

$$\begin{aligned}
Z^{n+1}_i((\pi_\varepsilon)^{n+1}_i) - \frac{\Delta t^2}{\Delta x^2}\frac{1}{2}\bigg(&\frac{Z^n_{i+1}}{\varrho^n_{i+1}}\Big[(\pi_\varepsilon)^{n+1}_{i+2} - (\pi_\varepsilon)^{n+1}_i\Big] \\
&- \frac{Z^n_{i-1}}{\varrho^n_{i-1}}\Big[(\pi_\varepsilon)^{n+1}_i - (\pi_\varepsilon)^{n+1}_{i-2}\Big]\bigg) = \phi(\varrho^n, q^n, Z^n)_i,
\end{aligned} \tag{6.2.9}$$

where the right-hand side is given by:

$$\begin{aligned}\phi(\varrho^n, q^n, Z^n)_i =& Z_i^n - \frac{\Delta t}{\Delta x}(H^n_{i+1/2} - H^n_{i-1/2}) \\ &+ \frac{\Delta t^2}{\Delta x^2}\frac{1}{2}\Big(\frac{Z^n_{i+1}}{\varrho^n_{i+1}}(G^n_{i+\frac{3}{2}} - G^n_{i+\frac{1}{2}}) - \frac{Z^n_{i-1,j}}{\varrho^n_{i-1}}(G^n_{i-\frac{1}{2}} - G^n_{i-\frac{3}{2}})\Big).\end{aligned} \tag{6.2.10}$$

This equation is supplemented by periodic or Dirichlet boundary conditions. After solving the equation for π_ε, we take $Z(\pi_\varepsilon) = \frac{\sqrt{\pi_\varepsilon/\varepsilon}}{1+\sqrt{\pi_\varepsilon/\varepsilon}}$ as the inverse function of $\pi_\varepsilon(Z)$, the non-linear equation is solved using the Newton iterations.

The $(n+1)$-th time step of the algorithm thus consists in getting Z^{n+1} by solving (6.2.9)-(6.2.10) and then obtaining q^{n+1} from (6.2.3b) and ϱ^{n+1} from (6.2.3a).

Stability Since the singular pressure π_ε is treated implicitly, the scheme remains stable even for small ε. The stability condition only depends on the wave speeds of the explicit part of the scheme, that is under the Courant-Friedrichs-Levy (CFL) condition:

$$\Delta t \leqslant \frac{\Delta x}{\max\limits_{j=1,2,3;\, x\in[0,1], t\in[0,T]} \Big\{|\lambda_j^0(x,t)|\Big\}}, \tag{6.2.11}$$

where λ_j^0, given by eq. (6.1.9), denotes the eigenvalues of the hyperbolic system with no congestion pressure ($\varepsilon = 0$). The scheme is asymptotically stable with respect to ε.

Discrete energy Like in the viscous version of system (5-6) (see [80]), an energy is conserved in time. Due to the numerical dissipation, our scheme does not preserve the energy at the discrete level even for smooth solutions. However, we can point out that, on discontinuous solutions, the local Lax-Friedrichs scheme selects a viscosity solution of the system with a decreasing energy.

6.2.2 The Second Order $(\varrho,\, q, Z)$-Method

Discretization in time The second-order discretization in time is based on the combined Runge-Kutta 2 / Crank-Nicolson (RK2CN) method as described in [65]: it consists of replacing Euler explicit by Runge-Kutta 2 solver and Euler Implicit by Crank-Nicolson solver in semi-discretization(6.2.1). Note that the second order convergence in time follows from the theorey of partitioned Runge-Kutta methods. Both methods are of second order and so called coupling conditions are satisfied. We here only detail the semi-discretized scheme. However, to be unambiguous, we will denote by $\mathscr{D}_\varrho$, $\mathscr{D}_q$, and $\mathscr{D}_Z$ the numerical diffusion terms resulting from the upwinding terms and the divergence operators will be replaced by centered fluxes. We thus consider the following scheme:

First step (half time step): get $\varrho^{n+1/2}$, $\boldsymbol{q}^{n+1/2}$ and $Z^{n+1/2}$ from

$$\frac{\varrho^{n+1/2}-\varrho^n}{\Delta t/2}+\nabla_x\cdot\boldsymbol{q}^{n+1/2}-\mathscr{D}_\varrho^n=0, \tag{6.2.12a}$$

$$\frac{\boldsymbol{q}^{n+1/2}-\boldsymbol{q}^n}{\Delta t/2}+\nabla_x\cdot\left(\frac{\boldsymbol{q}^n\otimes\boldsymbol{q}^n}{\varrho^n}+p(Z^n)\boldsymbol{I}\right)-\mathscr{D}_q^n+\nabla_x(\pi_\varepsilon(Z^{n+1/2}))=0, \tag{6.2.12b}$$

$$\frac{Z^{n+1/2}-Z^n}{\Delta t/2}+\nabla_x\cdot\left(\frac{Z^n}{\varrho^n}\boldsymbol{q}^{n+1/2}\right)-\mathscr{D}_Z^n=0. \tag{6.2.12c}$$

Second step (full time step): get ϱ^{n+1}, $\boldsymbol{q}^{n+1}$ and Z^{n+1} from

$$\frac{\varrho^{n+1}-\varrho^n}{\Delta t}+\nabla_x\cdot\left(\frac{\boldsymbol{q}^{n+1}+\boldsymbol{q}^n}{2}\right)-\mathscr{D}_\varrho^n=0, \tag{6.2.13a}$$

$$\begin{aligned}\frac{\boldsymbol{q}^{n+1}-\boldsymbol{q}^n}{\Delta t}+\nabla_x\cdot\left(\frac{\boldsymbol{q}^{n+1/2}\otimes\boldsymbol{q}^{n+1/2}}{\varrho^{n+1/2}}+p(Z^{n+1/2})\boldsymbol{I}\right)-\mathscr{D}_q^{n+1/2}&\\ +\nabla_x\left(\frac{\pi_\varepsilon(Z^n)+\pi_\varepsilon(Z^{n+1})}{2}\right)&=0,\end{aligned} \tag{6.2.13b}$$

$$\frac{Z^{n+1}-Z^n}{\Delta t}+\nabla_x\cdot\left(\frac{Z^{n+1/2}}{\varrho^{n+1/2}}\frac{\boldsymbol{q}^{n+1}+\boldsymbol{q}^n}{2}\right)-\mathscr{D}_Z^n=0. \tag{6.2.13c}$$

Like in the first-oder scheme, equations (6.2.12b)-(6.2.12c) and (6.2.13b)-(6.2.13c) result in elliptic equations for π_ε. Solving this equation and inverting the function $\pi_\varepsilon=\pi_\varepsilon(Z)$ allows to find Z satisfying the restriction $Z<1$. In practice, the scheme may fail capturing discontinuities, in particular when small values of ε are concerned. Indeed, the semi-implicit pressure $\Big(\pi_\varepsilon(Z^n)+\pi_\varepsilon(Z^{n+1})\Big)/2$ in (6.2.13b) is constrained to be larger than $\pi_\varepsilon(Z^n)/2$ preventing from having large discontinuities in pressure. One way to overcome this difficulty is to dynamically replace this semi-implicit pressure by an implicit pressure $\pi_\varepsilon(Z^{n+1})$ as soon as the non-linear solver of the elliptic equation detects a pressure lower than half the explicit one.

Discretization in space To get second order accuracy in space, we consider a MUSCL strategy. For any conserved quantity v, it consists in introducing at each mesh interface left and right values w_L and w_R:

$$\begin{aligned}w_{i,L}&=v_i+\frac{1}{2}\,\mathrm{minmod}(w_i-w_{i-1},w_{i+1}-w_i),\\ w_{i,R}&=v_i-\frac{1}{2}\,\mathrm{minmod}(w_i-w_{i-1},w_{i+1}-w_i),\end{aligned}$$

where the minmod function is defined as:

$$\mathrm{minmod}(a,b)=0.5\,(\mathrm{sgn}\,(a)+\mathrm{sgn}\,(b))\min(|a|,|b|).$$

Then all explicit terms in fluxes (6.2.4)-(6.2.5)-(6.2.6) depend on $(\varrho_{i,R}^n,q_{i,R}^n,Z_{i,R}^n)$ and $(\varrho_{i+1,L}^n,q_{i+1,L}^n,Z_{i+1,L}^n)$ instead of $(\varrho_i^n,q_i^n,Z_i^n)$ and $(\varrho_{i+1}^n,q_{i+1}^n,Z_{i+1}^n)$. Implicit terms are unchanged in order to be able to get the elliptic equation.

6.2.3 Congested Euler/Semi-Lagrangian Scheme ((ϱ, q)-method/SL)

Discretization in time We consider a scheme based on the non-conservative form (6.1.1) of the congestion transport. This idea was proposed in [78] in the context of constant congestion and in [80] in the context of variable congestion. The time-discretization reads:

$$\frac{\varrho^{n+1}-\varrho^{n}}{\Delta t}+\nabla_x\cdot\boldsymbol{q}^{n+1}=0, \tag{6.2.14a}$$

$$\frac{\boldsymbol{q}^{n+1}-\boldsymbol{q}^{n}}{\Delta t}+\nabla_x\cdot\left(\frac{\boldsymbol{q}^{n}\otimes\boldsymbol{q}^{n}}{\varrho^{n}}+p\left(\frac{\varrho^{n}}{\varrho^{*n}}\right)\boldsymbol{I}\right)+\nabla_x\pi_\varepsilon\left(\frac{\varrho^{n+1}}{\varrho^{*n}}\right)=0, \tag{6.2.14b}$$

$$\frac{\varrho^{*n+1}-\varrho^{*n}}{\Delta t}+\frac{\boldsymbol{q}^{n+1}}{\varrho^{n+1}}\cdot\nabla_x\varrho^{*n}=0. \tag{6.2.14c}$$

Inserting(6.2.14b) into (6.2.14a) results in

$$\varrho^{n+1}-\Delta t^2\,\Delta_x\Big(\pi_\varepsilon(\varrho^{n+1}/\varrho^{*\,n})\Big)=\\ \varrho^{n}-\Delta t\nabla_x\cdot\boldsymbol{q}^{n}+\Delta t^2\,\nabla_x\cdot\nabla_x\cdot\left(\frac{\boldsymbol{q}^{n}\otimes\boldsymbol{q}^{n}}{\varrho^{n}}+p(\varrho^{n}/\varrho^{*\,n})\boldsymbol{I}\right). \tag{6.2.15}$$

This is an elliptic equation on the density ϱ^{n+1}. The n-th time step of the scheme is decomposed into three parts: first get ϱ^{n+1} when solving (6.2.15), then compute $\boldsymbol{q}^{n+1}$ thanks to (6.2.14b) and then ϱ^{*n+1} from (6.2.14c).

Discretization in space Like for the previous schemes, we restrict the description to the one-dimensional case. Finite volume discretization is used for the spatial discretization of (6.2.14a)-(6.2.14b) as in section 6.2.1, see also [78]. A semi-Lagrangian method is used to solve (6.2.14c) and thus update the congestion density ϱ^*. The congestion density ϱ_i^{*n+1} at node x_i and time t^{n+1} is computed as follows: first we integrate back the characteristic line over one time step and then we interpolate the maximal density ϱ^{*n} at that point. Using Euler scheme for the first step, we obtain:

$$\varrho_i^{*n+1}=\left[\Pi\varrho^{*n}\right](x_i-q_i/\varrho_i\,\Delta t)$$

where $\Pi\varrho^{*n}$ is an interpolation function built from the points (x_i,ϱ_i^{*n}). We here perform a Lagrange interpolation on the $2r+2$ neighboring points:

$$[\Pi\varrho^*]_{|[x_i,x_{i+1}]}=\Pi_{\text{Lagrange}}\Big((x_j,\varrho_j^*),\quad i-r+1\le j\le i+r\Big).$$

resulting in $2r+1$-th spatial accuracy. First ($r=0$) and third ($r=1$) order in space semi-Lagrangian scheme will be used. For more details, we refer to [99].

The second order scheme Extension of the full scheme to second order accuracy in space is made using the MUSCL strategy for the finite volume fluxes. Extension to second order accuracy in time requires a Crank-Nicolson/Runge Kutta 2 method for (ϱ, q) and a second order in time integration of the characteric line for the semi-Lagrangian scheme (with for instance Taylor expansion) combined to a Strang splitting, see 6.7.

6.3 One Dimensional Validation of the Schemes

6.3.1 Riemann Test-Case

We compare the numerical schemes on one-dimensional Riemann test-cases: the initial data is a discontinuity between two constant states and the solutions are given by the superposition of waves separating constant states. In 6.5, we give the form of these solutions with respect to the relative position of left and right states in the phase space. In the case of colliding states, explicit solutions can be numerically obtained. We thus consider the following Riemann test-case:

$$(\varrho_0(x), q_0(x), \varrho_0^*(x)) = \begin{cases} (\varrho_\ell, q_\ell, \varrho_\ell^*) = (0.7, 0.8, 1.2), & \text{if } x \leqslant 0.5, \\ (\varrho_r, q_r, \varrho_r^*) = (0.7, -0.8, 1), & \text{if } x > 0.5. \end{cases} \tag{6.3.1}$$

on the domain $[0, 1]$. The solution is made of two shock waves and an intermediate contact wave, see (6.5.5). The CFL condition (6.2.11) can be estimated by:

$$\Delta t \leqslant \frac{\Delta x}{\max\limits_{x\in[0,1],t\in[0,T]} |v(x,t)| + \sqrt{\gamma/\min \varrho^*(x,t)}}.$$

For the current Riemann test-case with $\gamma = 2$ and $\alpha = 2$, the time step should satisfy $\Delta t \leqslant 0.4\Delta x$.

Comparison of the schemes ($\varepsilon = 10^{-2}$) In Figure 6.1, we represent the solution at time $t = 0.1$ with the different schemes using $\Delta t = 0.1\Delta x$. The $(\varrho, \boldsymbol{q}, Z)$-method refers to the method introduced in Section 6.2.1 for the first order and in Section 6.2.2 for the second order scheme. The (ϱ, q)-method/SL refers to the method described in Section 6.2.3. For the latter scheme, we use the third order semi-Lagrangian scheme for the transport of the congestion density ϱ^*.

We observe that all the methods correctly capture the exact solution. The $(\varrho, \boldsymbol{q})$-method/SL better captures the contact discontinuity at $x \approx 0.487$ since we use a third order accurate scheme for the transport of ϱ^*. Limiters could be used to avoid overshoot and undershoot at this location.

Oscillations in momentum are brought forth at the discontinuity interface of the shock waves. These oscillations are larger for second order schemes due to dispersion effects. In Figure 6.2, we provide a zoom on these oscillations and compare the approximate solution to the exact one. The amplitudes of the oscillations are larger for the $(\varrho, \boldsymbol{q})$-method/SL method. This may be the counterpart of the decoupling of the variables (ϱ, q) and ϱ^*: in the computation of the implicit pressure (see eq. (6.2.15), left-hand side), ϱ and ϱ^* are not taken at the same time. We finally note that, when running the simulation on large time, these oscillations do not increase in magnitude nor in support: this is related to some L^2 stability of the scheme.

Stiff pressure ($\varepsilon = 10^{-4}$) With this value of ε, the intermediate congested state has maximal wave speed equal to $\lambda_{\max} \approx 22$. Hence, taking time step Δt equal to $0.1\Delta x$ does not ensure the resolution of the fast waves.

Figure 6.3 shows the solution at time $t = 0.1$ using the $(\varrho, \boldsymbol{q}, Z)$-method with second order in space accuracy. In the full second order scheme, the scheme switches automatically to a first

		ϱ	q	Z	ϱ^*
$\varepsilon = 10^{-2}$	order 2 in x	8.66×10^{-4}	1.28×10^{-3}	3.03×10^{-4}	5.70×10^{-4}
	order 2	1.17×10^{-3}	3.52×10^{-3}	5.89×10^{-4}	5.77×10^{-4}
$\varepsilon = 10^{-4}$	order 2 in x	9.75×10^{-4}	2.11×10^{-3}	3.70×10^{-4}	5.71×10^{-4}
	order 2	9.89×10^{-4}	3.04×10^{-3}	3.84×10^{-4}	5.77×10^{-4}

Table 6.1: L_1 error between the numerical solutions to Riemann problem (6.3.1) and exact solution at time $t = 0.1$. Numercial solution computed using the $(\varrho, \boldsymbol{q}, Z)$-method. Numerical parameters: $\Delta x = 1 \times 10^{-3}$, $\Delta t = 0.1\,\Delta x$, $\alpha = 2$, $\gamma = 2$.

order in time version of the scheme due to the large discontinuities in pressure, see Section 6.2.2. We observe that the waves are well captured. As previously, oscillations in momentum develop at schock discontinuities and we observe that the second order in time version of the scheme leads to large uppershoots. In Table 6.1, we report the L_1 error between numerical and exact solution: we point out that the numerical errors are of the same order of magnitude independantly of the value of ε. Quite similar results are obtained using the $(\varrho, \boldsymbol{q})$-method/SL.

6.3.2 Numerical Convergence Test-Case

We here consider the following smooth initial data:

$$\begin{aligned} \varrho_0(x) &= 0.6 + 0.2 \exp\Big(-(x-0.5)^2/0.01\Big), \\ q_0(x) &= \exp\Big(-(x-0.5)^2/0.01\Big), \\ \varrho_0^*(x) &= 1.2 + 0.2\Big(1 - \cos\Big(8\pi(x-0.5)\Big)\Big), \end{aligned}$$

on the domain $[0, 1]$ and perdiodic boundary conditions. We compute a reference solution at time $t = 0.05$ using the second order in space (ϱ, q, Z)-method with small space and time steps $\Delta x = 5 \times 10^{-5}$ and $\Delta t = 0.1\,\Delta x$ (see Fig. 6.4).

Figure 6.5 shows the L_1 errors between approximate solutions and the reference solution at time $t = 0.05$ when the space step Δx goes to 0. For first order scheme, time step is set to $\Delta t = 5 \times 10^{-6}$ while for second order schemes, time and space steps satisfy the relation $\Delta t = 0.1\,\Delta x$ and both are varying.

We observe that all the schemes exhibit their expected convergence rates. We point out that $(\varrho, \boldsymbol{q}, Z)$-method and $(\varrho, \boldsymbol{q})$-method/SL have the same level of numerical errors except for variable ϱ^*: ϱ^* is better resolved with $(\varrho, \boldsymbol{q})$-method/SL. This is all the more the case when using the third order semi-Lagrangian scheme (on the right two plots of Fig. 6.5).

6.4 Two-dimensional Numerical Results

In this section we present the results of the numerical simulations in two-dimensions. As for domain we take the unit square with the mesh size $\Delta x = 10^{-3}$ and the time-step $\Delta t = 10^{-4}$. In the following we choose singular pressure (6.1.6) with the parameters $\varepsilon = 10^{-4}$, $\alpha = 2$, and the background pressure (6.1.3) with the exponent $\gamma = 2$, if not stated differently.

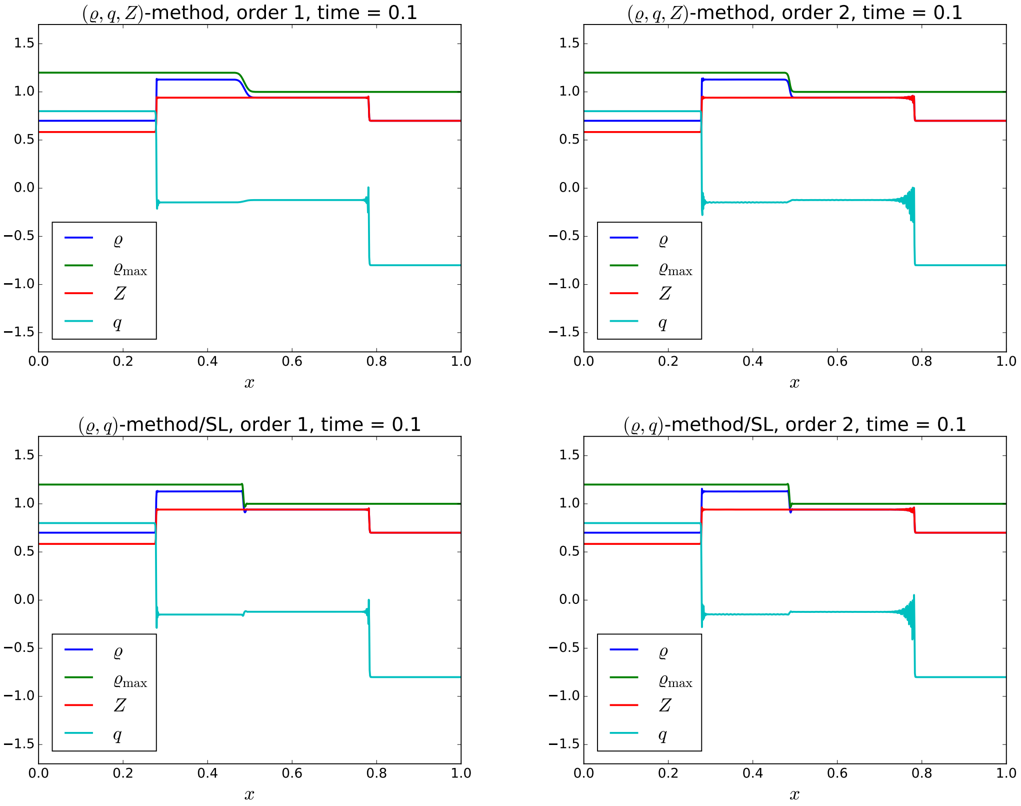


Figure 6.1: Approximate solution to Riemann problem (6.3.1) at time $t = 0.1$. Numerical parameters: $\Delta x = 1 \times 10^{-3}$, $\Delta t = 0.1\,\Delta x$, $\alpha = 2$, $\gamma = 2$, $\varepsilon = 10^{-2}$.

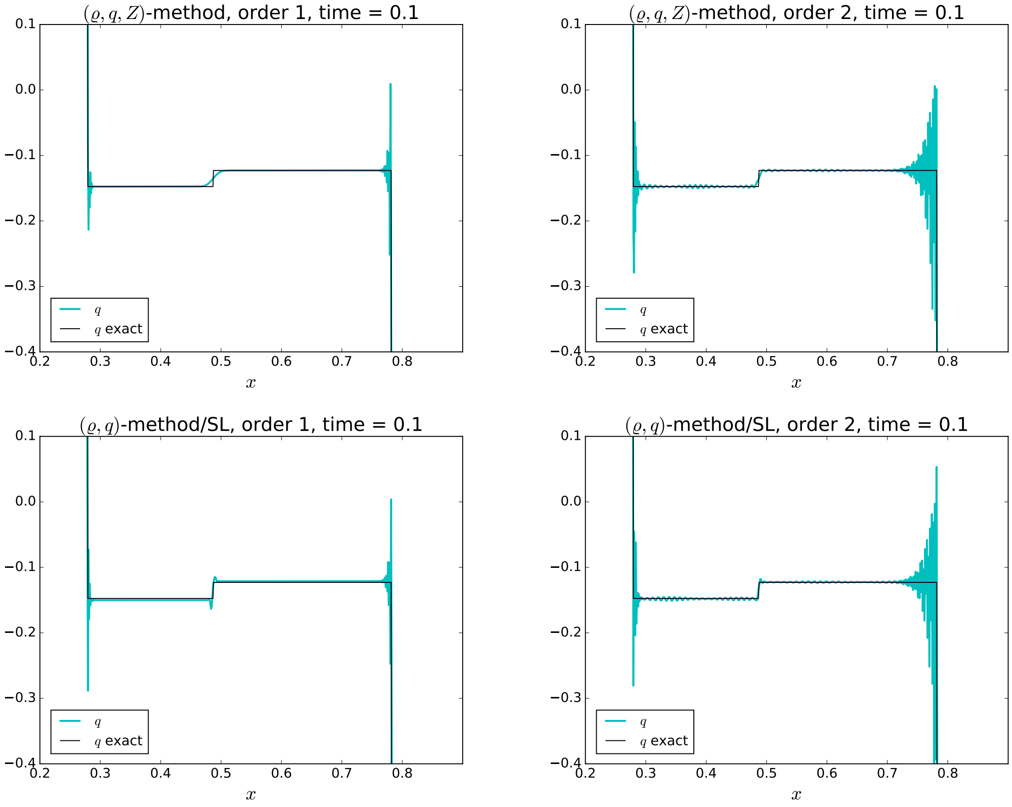


Figure 6.2: Approximate momentum q to Riemann problem (6.3.1) at time $t = 0.1$ and comparison with the exact solution. Numerical parameters: $\Delta x = 1 \times 10^{-3}$, $\Delta t = 0.1\,\Delta x$, $\alpha = 2$, $\gamma = 2$, $\varepsilon = 10^{-2}$.

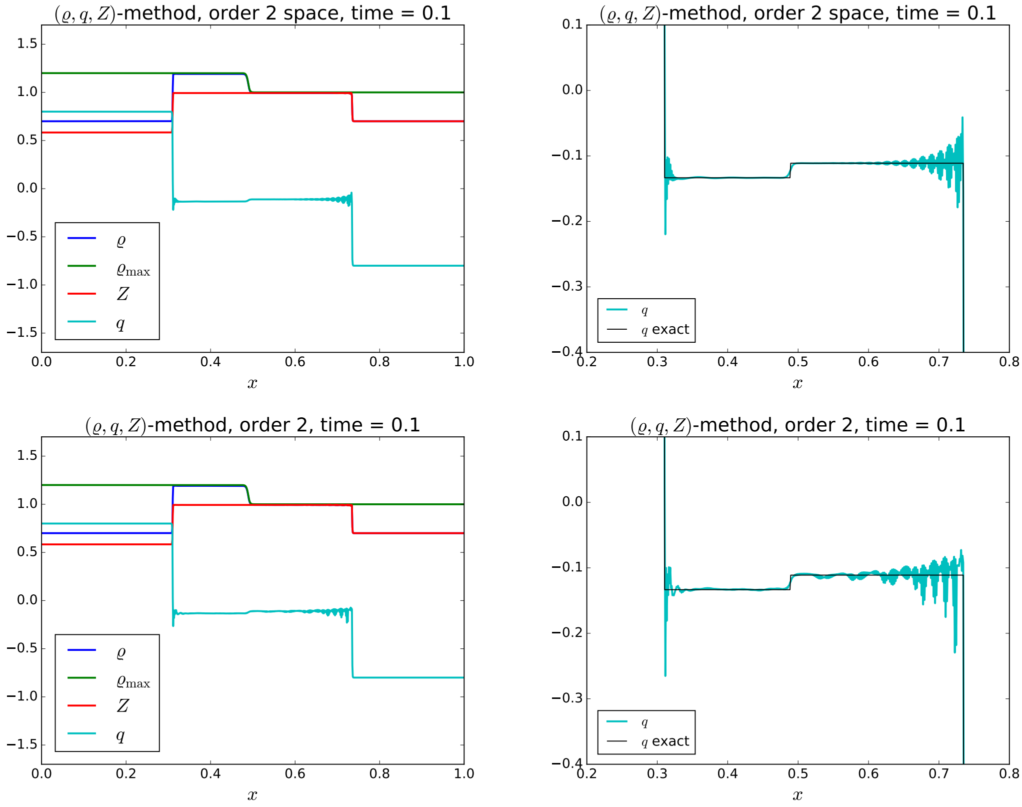


Figure 6.3: Approximate solution to Riemann problem (6.3.1) at time $t = 0.1$. Numerical parameters: $\Delta x = 1 \times 10^{-3}$, $\Delta t = 0.1\,\Delta x$, $\alpha = 2$, $\gamma = 2$, $\varepsilon = 10^{-4}$.

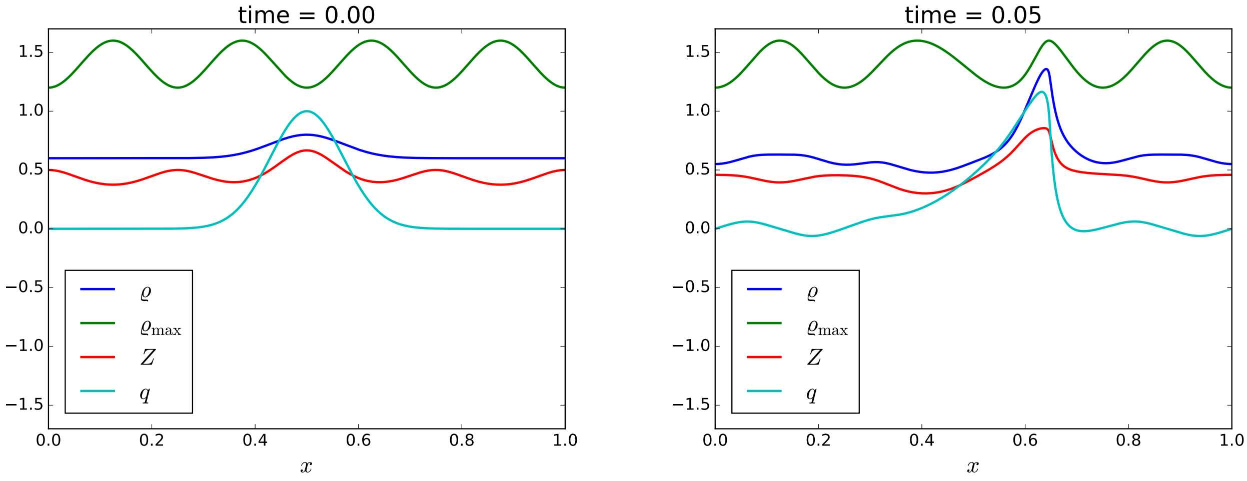


Figure 6.4: Reference solution at initial time (left) and time $t = 0.05$ (right). Numerical parameters: $\Delta x = 5 \times 10^{-5}$, $\Delta t = 0.1\,\Delta x$, $\gamma = 2$, $\varepsilon = 10^{-2}$.

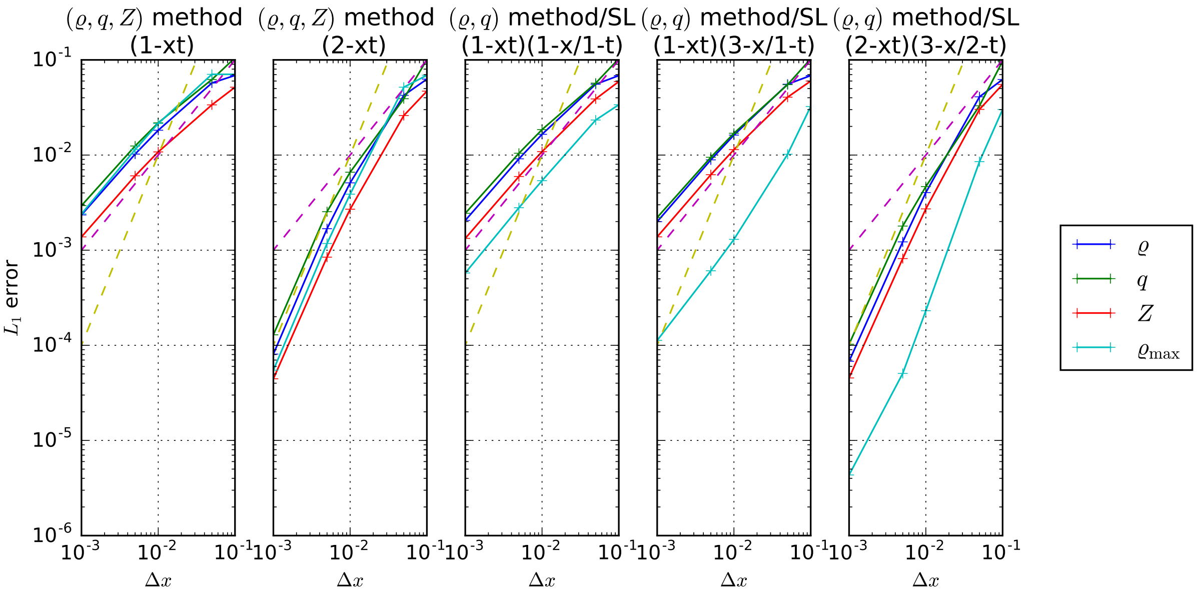


Figure 6.5: L_1 errors for ϱ, q, Z and ϱ^* as function of Δx. Numerical parameters: $\Delta t = 5 \times 10^{-6}$ for first order scheme and $\Delta t = 0.1\,\Delta x$ for second order scheme, $\gamma = 2$, $\varepsilon = 10^{-2}$. $(\varrho, \boldsymbol{q}, Z)$-method: ($k$-xt) k-th order in space and time. $(\varrho, \boldsymbol{q})$-method/SL: ($k$-xt)($m$-x/$n$-t) k-th order in space and time for the $(\varrho, \boldsymbol{q})$-method and m-th order in space and n-th order in time for the advection of ϱ^* by the semi-Lagrangian scheme. In dashed lines: first and second order curves.

First part is devoted to comparison of $(\varrho, \boldsymbol{q}, Z)$-method and $(\varrho, \boldsymbol{q})$-method/SL described in Section 6.2. Second is an application of (ϱ, q)-method/SL to the evacuation scenario. Third order in space semi-Lagrangian scheme is applied.

6.4.1 Collision of 4 Groups with Variable Congestion

In the unit square periodic domain we specify 4 squares, with the centers in points $(x_c, y_c) = \{(0.2, 0.5), (0.5, 0.2), (0.5, 0.8), (0.8, 0.5)\}$. The length of the side l of each square equals 0.2 (for every square we introduce the notation $\text{Square}((x_c, y_c), l)$). We prescribe the initial momentum of 0.5 pointing into the center of the domain provoking a collision. We consider three test cases varying in the initial congestion density, namely:

Case 1: $\varrho^*(x, 0) = 1.0$;

$$\text{Case 2:}\ \ \varrho^*(x,0) = \begin{cases} 0.80 \text{ if } x \in\ \text{Square}((0.2,0.5),0.2) \\ 1.20 \text{ if } x \in\ \text{Square}((0.5,0.2),0.2) \\ 0.80 \text{ if } x \in\ \text{Square}((0.8,0.5),0.2) \\ 1.20 \text{ if } x \in\ \text{Square}((0.5,0.8),0.2) \\ 1.00 \text{ otherwise} \end{cases};$$

Case 3: $\varrho^*(x, 0) = 1 + 0.05(\cos(10\pi x) + \cos(24\pi x))(\cos(6\pi y) + \cos(34\pi y))$.

The results of our simulations for these three cases are presented in Figures 6.6, 6.7, and 6.8. We see that in case of constant congestion density (Case 1, Figure 6.6) the two schemes provide almost identical outcome. The essential difference appears when ϱ_0^* varies. We see in Figure 6.7 that the initial discontinuities of ϱ^* are significantly smoothened by the $(\varrho, \boldsymbol{q}, Z)$-method, while the $(\varrho, \boldsymbol{q})$-method/SL preserves the initial shape, which basically confirms our observations from Section 6.3.2. This is even more visible in Figure 6.8, where the initial oscillations of ϱ^* rapidly decay when simulated by the $(\varrho, \boldsymbol{q}, Z)$-method.

Another interesting observation following from Figures 6.7, and 6.8 when compared to Figure 6.6 is that the preference of the individuals ϱ^* is significant factor to determine the density distribution even far away from the congestion zone.

Moreover, comparing Figure 6.7 with Figure 6.6, we see a clear influence of the density constraint on the velocity of the agents. Indeed, for the Case 2, there is a significant disproportion between the velocities in the x and y directions at time $t = 0.150$ (see Figure 6.7 right). This corresponds to the fact that the agents moving toward the center along y axis have 'more space' to fill since ϱ^* for those groups is higher than the one for the groups moving in the x direction. This results in a certain delay between collisions in two directions.

6.4.2 Application to Crowd Dynamics

In this section we investigate an influence of the variable density ϱ^* on a possible evacuation scenario. For this, we consider an impenetrable room in the shape of unit square, initially filled with uniformly distributed agents. There is an exit located at $x \in [0.4, 0.6]$, $y = 0$ that allows for free outflow. The initial density $\varrho_0 = 0.6$ and the initial momentum is equal to $\mathbf{0}$. The desire of going to the exit is introduced in the system (6.1.1) (6.1.6) by adding the relaxation therm in the momentum equation

$$\partial_t \boldsymbol{q} + \operatorname{div}\left(\frac{\boldsymbol{q} \otimes \boldsymbol{q}}{\varrho} + \pi_\varepsilon\left(\frac{\varrho}{\varrho^*}\right)\boldsymbol{I} + p\left(\frac{\varrho}{\varrho^*}\right)\boldsymbol{I}\right) = \frac{1}{\beta}\left(\boldsymbol{q} - \varrho\boldsymbol{w}\right), \tag{6.4.1}$$

where $\boldsymbol{w}$ is the desired velocity, and β stands for the relaxation parameter. The desired velocity is given by a unit vector field, that points into the centre of the exit,

$$\boldsymbol{w} = \Big(-x/((x-0.5)^2 + y^2), -y/((x-0.5)^2 + y^2)\Big).$$

In the numerical scheme we apply splitting of the momentum equation between the transport and pressure part, and the relaxation (source) part, with the intermediate momentum $\boldsymbol{q}^*$. After the momentum is updated we perform implicit relaxation step, for given density ϱ^{n+1},

$$\frac{\boldsymbol{q}^* - \boldsymbol{q}^n}{\Delta t} + \nabla_x \cdot \left(\frac{\boldsymbol{q}^n \otimes \boldsymbol{q}^n}{\varrho^n} + p\left(\frac{\varrho^n}{\varrho^{*n}}\right)\boldsymbol{I}\right) + \nabla_x \cdot \pi_\varepsilon\left(\frac{\varrho^{n+1}}{\varrho^{*n}}\right) = 0, \tag{6.4.2a}$$

$$\frac{\boldsymbol{q}^{n+1} - \boldsymbol{q}^*}{\Delta t} = \frac{1}{\beta}\left(\boldsymbol{q}^{n+1} - \varrho^{n+1}\boldsymbol{w}\right). \tag{6.4.2b}$$

We use the $(\varrho, \mathbf{q})$-method/SL, which requires to solve the transport equation for ϱ^*. This is especially problematic in the corners of the domain, where the Dirichlet boundary condition is considered. This leads to oscillations of ϱ and ϱ^* close to these points. Nevertheless, we may

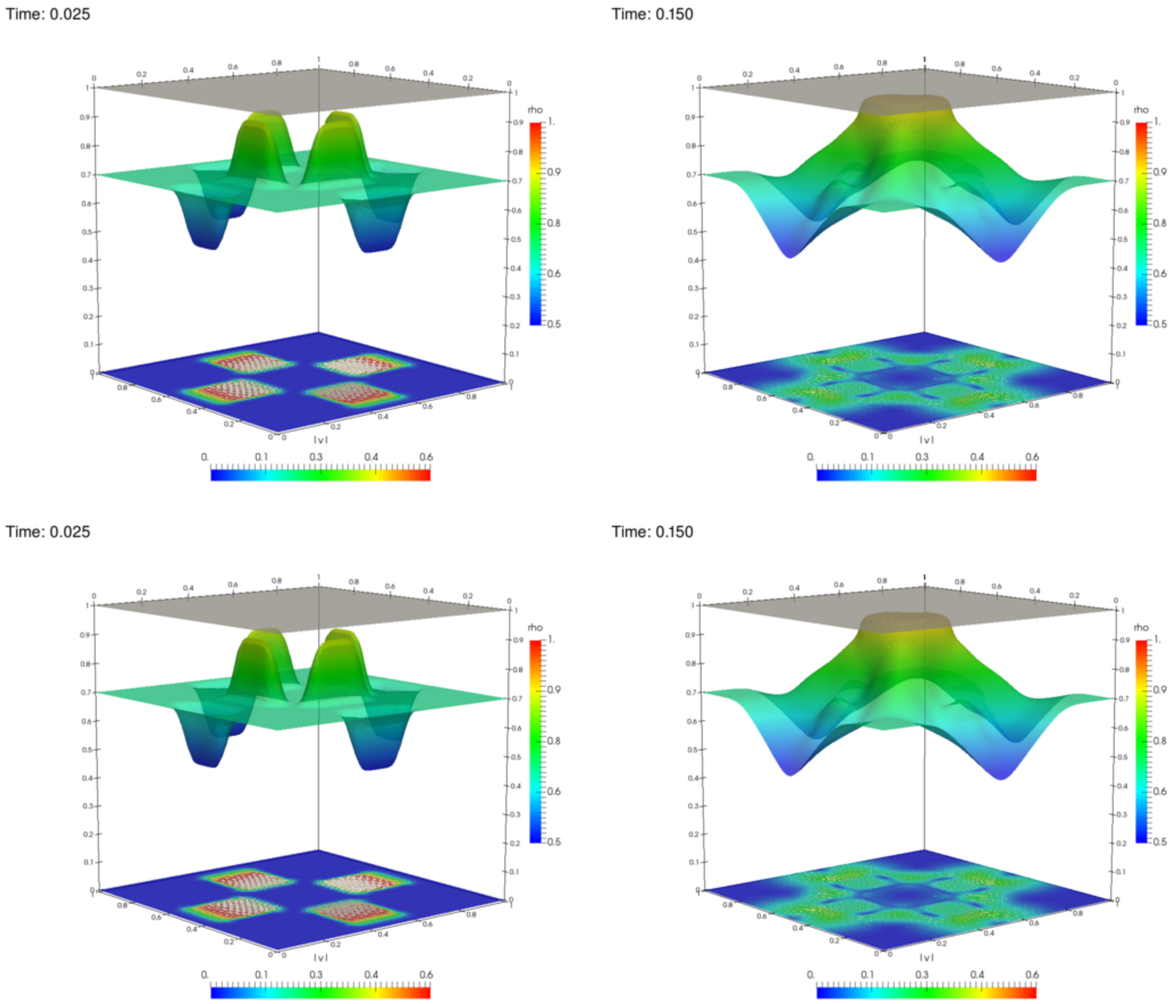


Figure 6.6: Case 1: the comparison of $(\varrho, \boldsymbol{q}, Z)$-method (top) and $(\varrho, \boldsymbol{q})$-method/SL (bottom) at time 0.025 (left), and 0.150 (right). A video available at `https://youtu.be/jbNXJJd4Kbo`.

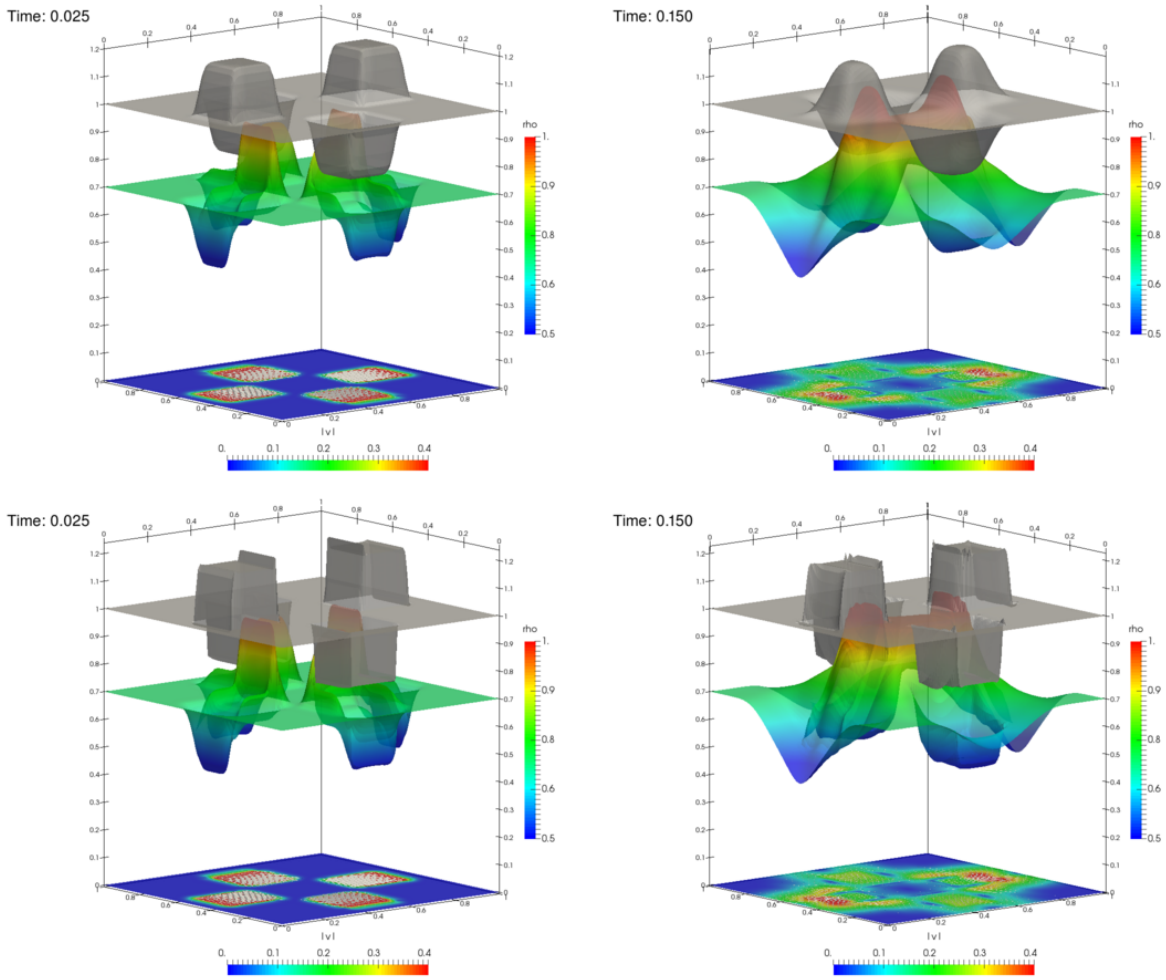


Figure 6.7: Case 2: the comparison of $(\varrho, \boldsymbol{q}, Z)$-method (top) and $(\varrho, \boldsymbol{q})$-method/SL (bottom) at time 0.025 (left), and 0.150 (right). A video available at https://youtu.be/TPEO9wUVPZg.

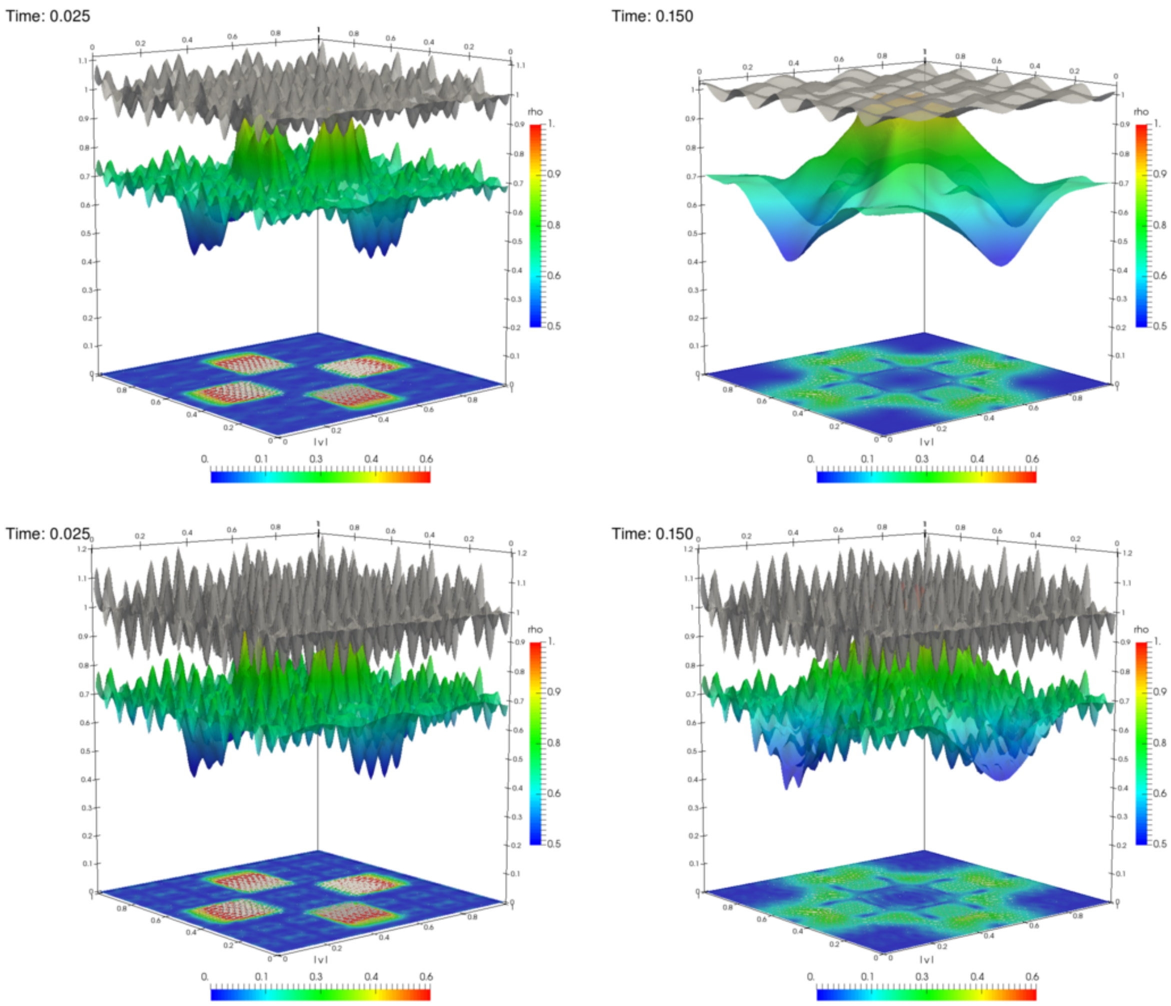


Figure 6.8: Case 3: the comparison of $(\varrho, \boldsymbol{q}, Z)$-method (top) and $(\varrho, \boldsymbol{q})$-method/SL (bottom) at time 0.025 (left), and 0.150 (right). A video available at https://youtu.be/Gvvlgq8bY_E.

observe, see Figure 6.9 and Figure 6.10, the so called *stop-and-go* behaviour, namely distinct high velocity regions in the domain, one in the vicinity of the exit and the second one that propagates in the direction opposite to flow.

This reflects an empirical observation that once a pedestrian arrives to the space of high congestion, he or she slows down or even stops until some space opens up in front. This kind of stop-and-go waves have been described, for example, by Helbing and Johansson in [126]. For the description of the real evacuation experiments we refer to [123], see also [110]. In the last of the mentioned papers the authors provide an experimental demonstration of the so called *faster goes slower* effect. This means that an increase in the density of pedestrians does not necessarily lead to a larger flow rate. Our simulations show that when the parameter ϱ^* is low, the outflow of the individuals is slower. This is especially visible in the third row of Figures 6.9 and 6.10 presenting the evacuation scenario for the initial barrier density in the shape of the step function

$$\varrho_0^*(x,y) = \begin{cases} 1.1 & \text{for} \quad 0.5 < x < 1, \\ 0.9 & \text{for} \quad 0 < x < 0.5. \end{cases} \tag{6.4.3}$$

This observation can be also confirmed in terms of speed of evacuation. Indeed, we performed analogous simulations for 3 cases of constant ϱ_0^* equal to 0.9, 1.0. 1.1 show that the speed of emptying the room is bigger the bigger value of ϱ_0^*. To see this we have measured the mass remaining in the room at time $t = 1$ and it is equal to 0.51030, 0.048037, and 0.457123, respectively. We have moreover observed that evacuation speed of the room with individuals of the average congestion preference equal to 1 initially can be improved by placing the individuals with higher ϱ_0^* closer to the exit. This is illustrated in the Figures 6.9 and 6.10 the second row, for which, the initial congestion preference ϱ_0^* equals

$$\varrho_0^*(x,y) = 1.1 - 0.2y. \tag{6.4.4}$$

The random distribution of preferences of the individuals with expected value equal to 1, on the other hand, corresponds to the increase of the evacuation time (see Figures 6.9 and 6.10 the bottom row).

6.5 Solution to the Riemann Problem

The one-dimensional Riemann problem for the system (6.1.8) is the following initial-value problem:

$$\partial_t \varrho + \partial_x q = 0, \tag{6.5.1a}$$

$$\partial_t q + \partial_x \left(\frac{q^2}{\varrho} + p_\varepsilon(Z) \right) = 0, \tag{6.5.1b}$$

$$\partial_t Z + \partial_x \left(Z \frac{q}{\varrho} \right) = 0, \tag{6.5.1c}$$

where $p_\varepsilon(Z) = \pi_\varepsilon(Z) + p(Z)$, and

$$(\varrho, q, Z)(0,x) = \begin{cases} (\varrho_\ell, q_\ell, Z_\ell) & \text{for} \quad x < 0, \\ (\varrho_r, q_r, Z_r) & \text{for} \quad x > 0. \end{cases} \tag{6.5.2}$$

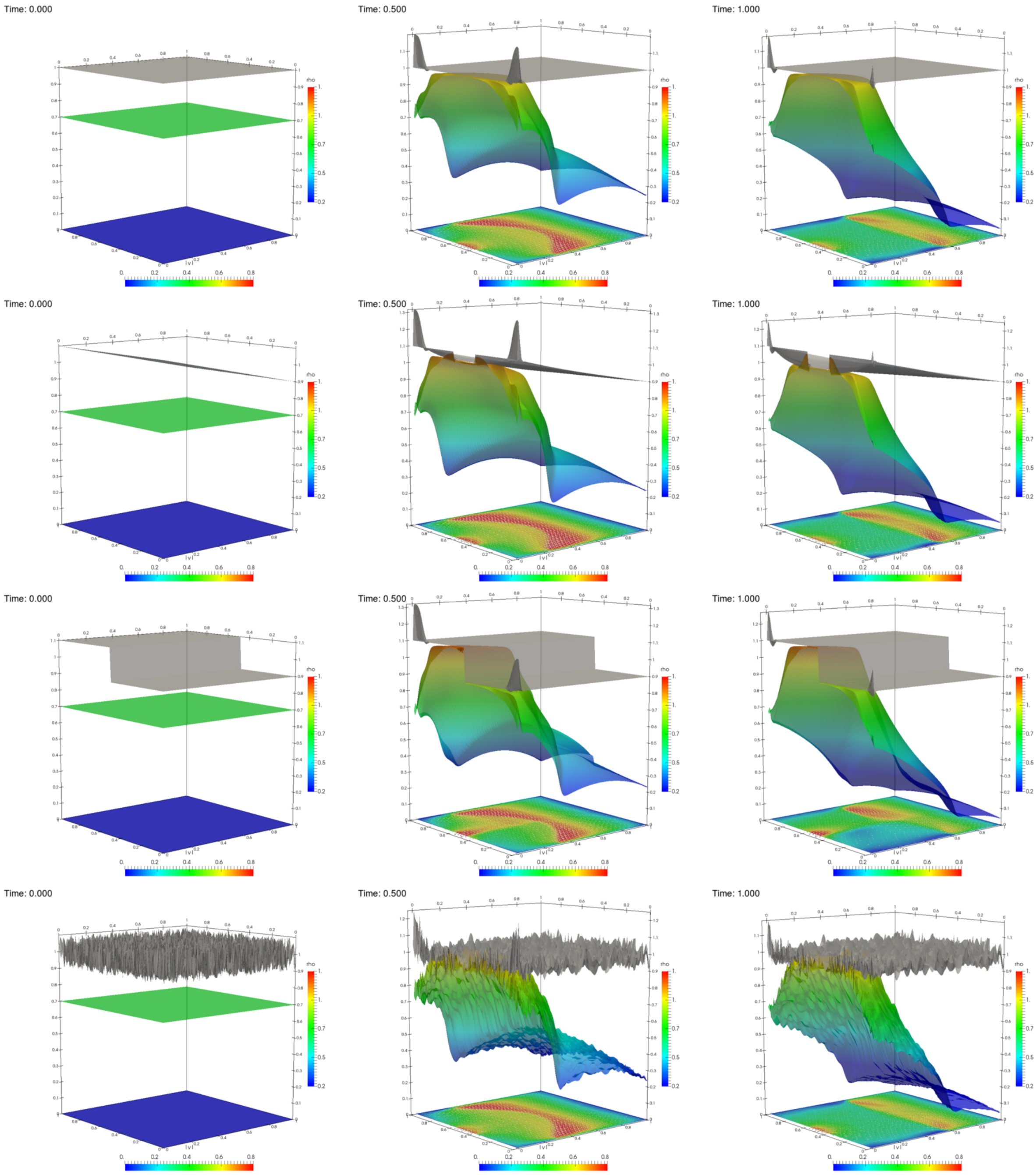


Figure 6.9: Stop-and-go behaviour for the evacuation scenario, with ϱ_0^* being constant, with linear slope in $y-$direction (6.4.4), step-function (6.4.3), and a random function. The congestion density (upper) the density (middle) and the velocity amplitude (bottom) at times $t = 0$ (left column) $t = 0.5$ (middle column), and $t = 1.0$ (right column). A video available at `https://youtu.be/sK9J5BoUmtE`.

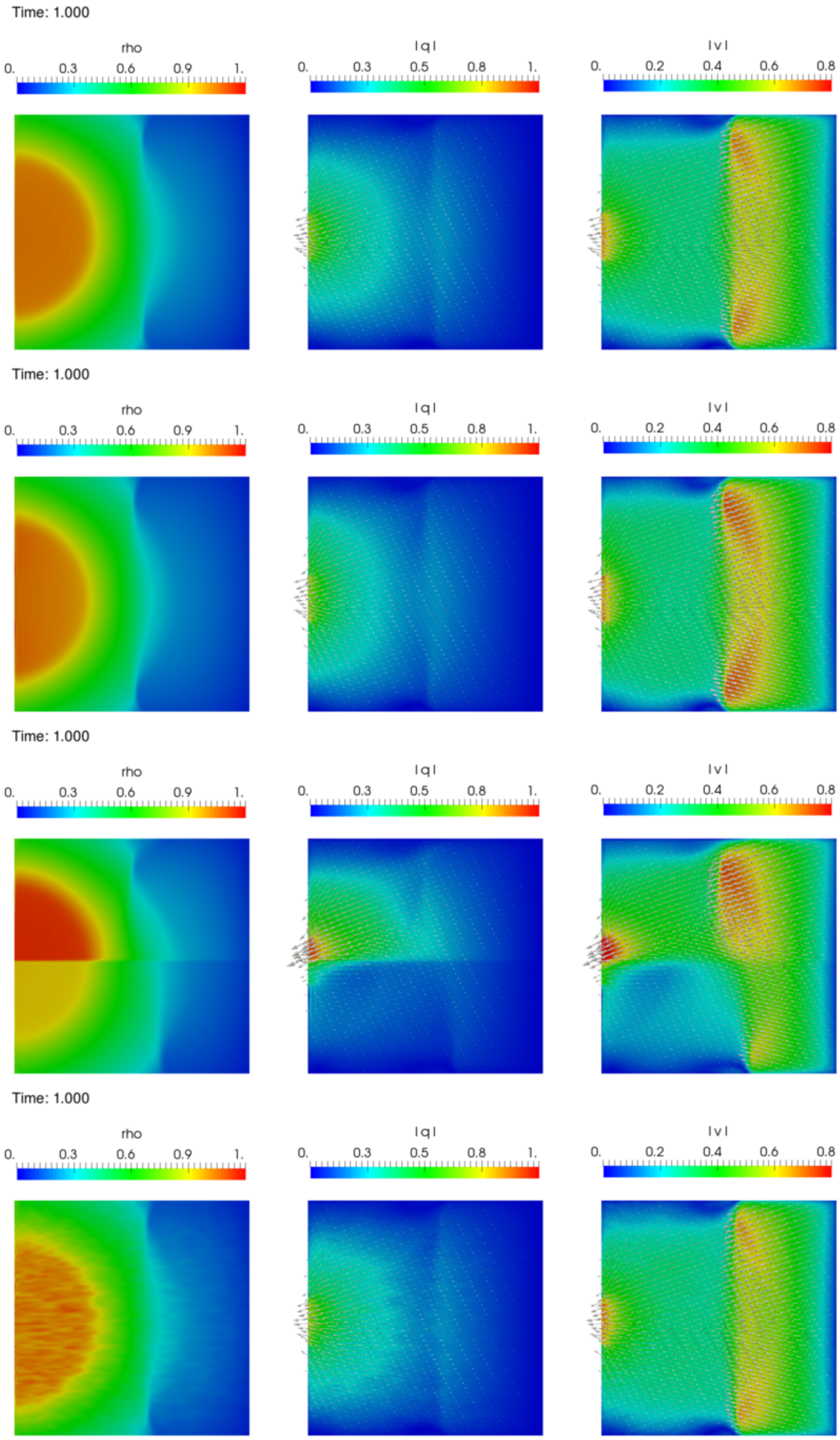


Figure 6.10: The evacuation scenario for ϱ_0^* being constant, with linear slope in $y-$direction (6.4.4), step-function (6.4.3), and a random function. The figures present the values of the density ϱ, the direction momentum $|q|$ and the direction and values of the velocity v at time $t = 1.0$ for different test cases. A video available at https://youtu.be/aoDw2aqvNEQ.

The purpose of this section is to find possible weak solution to (6.5.1) (6.5.2). We will also consider the limit of these solutions as $\varepsilon \to 0$.

As already mentioned in the introduction, the system (6.5.1) is strictly hyperbolic provided $p'_\varepsilon(Z) > 0$, see (6.1.9). The associated characteristic fields are given by:

$$r_1^\varepsilon(\varrho, q, Z) = \begin{bmatrix} 1 \\ v - \sqrt{\frac{Z}{\varrho} p'_\varepsilon(Z)} \\ Z/\varrho \end{bmatrix}, \; r_2^\varepsilon(\varrho, q, Z) = \begin{bmatrix} 1 \\ v \\ 0 \end{bmatrix}, \; r_3^\varepsilon(\varrho, q, Z) = \begin{bmatrix} 1 \\ v + \sqrt{\frac{Z}{\varrho} p'_\varepsilon(Z)} \\ Z/\varrho \end{bmatrix},$$

where $v = q/\varrho$ is the velocity. The second characteristic field is linearly degenerate (since $\nabla\lambda_2 \cdot r_2 = 0$). The two others characteristic field are genuinely non-linear.

We now present the elementary wave solutions of the Riemann problem.

6.5.1 Elementary Waves

Shock Discontinuities A shock wave is a discontinuity between two constant states, (ϱ, q, Z) and $(\widehat{\varrho}, \widehat{q}, \widehat{Z})$, travelling at a constant speed σ. We now fix the left (or right) state $(\widehat{\varrho}, \widehat{q}, \widehat{Z})$ and look for all triples (ϱ, q, Z) that can be connected to $(\widehat{\varrho}, \widehat{q}, \widehat{Z})$ by the shock discontinuity. Across the shock, the Rankine-Hugoniot conditions must be satisfied meaning that:

$$[q] = \sigma[\varrho], \quad \left[\frac{q^2}{\varrho} + p_\varepsilon(Z)\right] = \sigma[q], \quad \left[Z\frac{q}{\varrho}\right] = \sigma[Z],$$

where $[a] := a - \widehat{a}$ denotes the jump of quantity a. Treating ϱ as a parameter, we check that the two admissible states are of the form $(\varrho, q_{h,\pm}(\varrho), Z(\varrho))$ with $q_{h,\pm} = \varrho v_{h,\pm}(\varrho)$ and

$$v_{h,\pm}(\varrho) = \widehat{v} \pm \operatorname{sign}(Z(\varrho) - \widehat{Z}) \frac{1}{\sqrt{\widehat{\varrho}\varrho}} \sqrt{(\varrho - \widehat{\varrho}) \left(p_\varepsilon\left(\frac{\widehat{Z}\varrho}{\widehat{\varrho}}\right) - p_\varepsilon(\widehat{Z})\right)},$$
$$Z(\varrho) = \widehat{Z}\frac{\varrho}{\widehat{\varrho}}.$$

The shock speed therefore equals:

$$\sigma_\pm = \widehat{v} \pm \operatorname{sign}(Z - \widehat{Z}) \sqrt{\frac{\varrho}{\widehat{\varrho}}} \sqrt{\frac{p_\varepsilon(\widehat{Z}\varrho/\widehat{\varrho}) - p_\varepsilon(\widehat{Z})}{(\varrho - \widehat{\varrho})}}.$$

These solutions can also be expressed as functions of Z:

$$\varrho(Z) = Z\frac{\widehat{\varrho}}{\widehat{Z}},$$
$$v_{h,\pm}(Z) = \widehat{v} \pm \operatorname{sign}(Z - \widehat{Z}) \frac{1}{\sqrt{\widehat{\varrho}}} \sqrt{\left(1 - \frac{\widehat{Z}}{Z}\right) \Big(p_\varepsilon(Z) - p_\varepsilon(\widehat{Z})\Big)}.$$

Note that the maximal density ($\varrho^* = \varrho/Z$) does not jump across a shock discontinuity. Expanding $(\varrho(Z), q_{h,\pm}(Z), Z)$ around $Z = \widehat{Z}$, we obtain

$$\begin{aligned}
\varrho(Z) - \widehat{\varrho} &= (Z - \widehat{Z})\frac{\widehat{\varrho}}{\widehat{Z}}, \\
\varrho(Z)v_{h,\pm}(Z) - \widehat{\varrho}\widehat{v} &= (Z - \widehat{Z})\frac{\widehat{\varrho}}{\widehat{Z}}\widehat{v} \pm Z\frac{\widehat{\varrho}}{\widehat{Z}}\text{sign}(Z - \widehat{Z})\sqrt{\frac{1}{\widehat{\varrho}}}\sqrt{(1 - \widehat{Z}/Z)\Big(p_\varepsilon(Z) - p_\varepsilon(\widehat{Z})\Big)} \\
&\approx (Z - \widehat{Z})\frac{\widehat{\varrho}}{\widehat{Z}}\widehat{v} \pm Z\frac{\widehat{\varrho}}{\widehat{Z}}\text{sign}(Z - \widehat{Z})\sqrt{\frac{1}{\widehat{\varrho}Z}}\sqrt{p'_\varepsilon(\widehat{Z})(Z - \widehat{Z})^2} \\
&\approx (Z - \widehat{Z})\frac{\widehat{\varrho}}{\widehat{Z}}\left(\widehat{v} \pm \sqrt{\frac{\widehat{Z}}{\widehat{\varrho}}}\sqrt{p'_\varepsilon(\widehat{Z})}\right), \\
Z - \widehat{Z} &= (Z - \widehat{Z})\frac{\widehat{\varrho}}{\widehat{Z}}\frac{\widehat{Z}}{\widehat{\varrho}}.
\end{aligned}$$

Note that $(\varrho(Z), q_{h,-}(Z), Z)$ is tangent at $(\widehat{\varrho}, \widehat{q}, \widehat{Z})$ to $r_1(\widehat{\varrho}, \widehat{q}, \widehat{Z})$, therefore $v_{h,-}$ corresponds to the 1-characteristic field, analogously $v_{h,+}$ corresponds to the 3-characteristic field. The graph of $Z \mapsto v_{h,-}(Z)$ (resp. $Z \mapsto v_{h,+}(Z)$) is called the 1-Hugoniot curve (resp. 3-Hugoniot curve) issued from $(\widehat{v}, \widehat{Z})$.

To check the admissibility of the discontinuity, we need to check the entropy condition. If $(\widehat{v}, \widehat{Z})$ is the left state, the right states that can be connected to it by an entropic shock wave are those located on the 1-shock curve $\Big\{\Big(v_{h,-}(Z), Z\Big) : Z > \widehat{Z}\Big\}$ or the 3-shock curve $\Big\{(v_{h,+}(Z), Z) : Z < \widehat{Z}\Big\}$. Indeed, on these curves the associated eigenvalue is decreasing. If on the other hand, $(\widehat{v}, \widehat{Z})$ is the right state, the left states that can be connected to it by an entropic shock wave are those located on the 1-shock curve $\Big\{(v_{h,-}(Z), Z) : Z < \widehat{Z}\Big\}$ or the 3-shock curve $\Big\{(v_{h,+}(Z), Z) : Z > \widehat{Z}\Big\}$. Indeed, on these curves the associated eigenvalue is increasing.

Rarefaction Waves The rarefaction waves are continuous self-similar solutions, $(\varrho(t,x), q(t,x), Z(t,x)) = (\varrho(x/t), q(x/t), Z(x/t))$, connecting two constant states (ϱ, q, Z) and $(\widehat{\varrho}, \widehat{q}, \widehat{Z})$. They thus satisfy the following differential equations:

$$\varrho'(s) = 1, \quad q'(s) = \widetilde{v}(s) \pm \sqrt{\frac{Z(s)}{\varrho(s)}p'_\varepsilon(Z(s))}, \quad Z'(s) = Z(s)/\varrho(s), \tag{6.5.3}$$

Denoting $q(s) = \varrho(s)\widetilde{v}_{i,\pm}(s)$ and parametrizing by ϱ, we obtain:

$$\widetilde{v}'_{i,\pm}(\varrho) = \pm\frac{1}{\varrho}\sqrt{\frac{Z(\varrho)}{\varrho}p'_\varepsilon(Z(\varrho))}, \quad Z'(\varrho) = Z(\varrho)/\varrho.$$

From the first and third equation of (6.5.3), we have $(\varrho/Z(\varrho))' = 0$, and so, $\varrho/Z(\varrho) = \widehat{\varrho}/Z(\widehat{\varrho})$. This means that as in the case of shock discontinuities the maximal density ϱ^* does not jump. Denoting $\varrho^* = \varrho/Z(\varrho)$ and making the change of coordinates $v_{i,\pm}(Z) = \widetilde{v}_{i,\pm}(\varrho)$ with $\varrho = \varrho^* Z$, we thus have:

$$v'_{i,\pm}(Z) = \pm\frac{1}{Z}\sqrt{\frac{1}{\varrho^*}p'_\varepsilon(Z)}.$$

Hence, the states satisfy:

$$v_{i,\pm}(Z) = \widehat{v} \pm \Big(F_\varepsilon(Z) - F_\varepsilon(\widehat{Z})\Big), \tag{6.5.4}$$

where F_ε is an antiderivative of $Z \mapsto \frac{1}{Z}\sqrt{\frac{1}{\varrho^*}p'_\varepsilon(Z)}$.

The graph of $Z \mapsto v_{i,+}(Z)$ (resp. $Z \mapsto v_{i,-}(Z)$) is called the 1-integral curve (resp. 3-integral curve) issued from $(\widehat{v}, \widehat{Z})$. If $(\widehat{v}, \widehat{Z})$ is a left state, the right states that can be connected to it by an entropic rarefaction wave are those located on the 1-integral curve $\Big\{(v_{i,-}(Z), Z) : Z < \widehat{Z}\Big\}$ or the 3-integral curve $\Big\{(v_{i,-}(Z), Z) : Z > \widehat{Z}\Big\}$. Indeed, on these curves the associated eigenvalue is increasing. If $(\widehat{v}, \widehat{Z})$ is a right state, the left states that can be connected to it by an entropic rarefaction wave are those located on the 1-integral curve $\Big\{(v_{i,-}(Z), Z) : Z > \widehat{Z}\Big\}$ or the 3-integral curve $\Big\{(v_{i,-}(Z), Z) : Z < \widehat{Z}\Big\}$. Indeed, on these curves the associated eigenvalue is decreasing.

Contact Discontinuities Since the second characteristic field is linearly degenerate, there are linear discontinuities that propagate at velocity $\lambda_2 = \widehat{v}$. Let us write the Rankine-Hugoniot conditions:

$$[q] = \widehat{v}[\varrho], \qquad \left[\frac{q^2}{\varrho} + p_\varepsilon(Z)\right] = \widehat{v}[q], \qquad \left[Z\frac{q}{\varrho}\right] = \widehat{v}[Z].$$

From the first relation, we obtain $v = \widehat{v}$ and then the second relation states that the pressure jump is zero. By strict monotony of the pressure, it implies that $Z = \widehat{Z}$ and the third equation is satisfied. Along this discontinuity, the velocity and the pressure are thus conserved. Note that every density jump is possible.

6.5.2 Solution to Riemann Problem

Let $(\varrho_\ell, q_\ell, Z_\ell)$ and (ϱ_r, q_r, Z_r) be the left and right initial states (6.5.2). The solutions to Riemann problems are determined as follows. First, in the (v, Z) plane, find out the intersection state (v_m, Z_m) of the 1-st integral/Hugoniot curves issued from (v_ℓ, Z_ℓ) and the 3-rd integral/Hugoniot curves issued from (v_r, Z_r). Then, compute the two densities $\varrho_{m,\ell}$ and $\varrho_{m,r}$ so that the congestion density across the two non-linear waves is conserved. Then we connect the two distinct intermediate states by a contact discontinuity. We finally end up with the following solution:

$$\begin{aligned} (\varrho_\ell, q_\ell, Z_\ell) \overset{shock/rarefaction}{\rightarrow} & (\varrho_{m,\ell}, \varrho_{m,\ell}v_m, Z_m) \\ & \overset{contact}{\rightarrow} (\varrho_{m,r}, \varrho_{m,r}v_m, Z_m) \overset{shock/rarefaction}{\rightarrow} (\varrho_r, q_r, Z_r) \end{aligned} \tag{6.5.5}$$

where $\varrho_{m,\ell} = Z_m\varrho_\ell/Z_\ell$ and $\varrho_{m,r} = Z_m\varrho_r/Z_r$. The nature of the non-linear waves (rarefaction or shock) depends on the relative position of the states (v_ℓ, Z_ℓ), (v_r, Z_r) in the (v, Z) plane.

6.5.3 Limit $\varepsilon \to 0$

We are now interested in the asymptotic behaviour, when $\varepsilon \to 0$ of the Hugoniot $v^\varepsilon_{h,\pm}$ and the integral curves $v^\varepsilon_{i,\pm}$ obtained in the previous paragraph for the elementary waves. We have the following result.

Proposition 6.1. *The graph of the Hugoniot curve,* $\Big\{(Z, v^{\varepsilon}_{h,\pm}(Z)) : Z \in [0,1)\Big\}$*, tends to the union of the set* $\Big\{(Z, v^{0}_{h,\pm}(Z)) : Z \in [0,1)\Big\}$ *and the horizontal half straight line* $\Big\{(1, v) : v \in [v^{0}_{h,\pm}(1), +\infty)\Big\}$*.*

The graph of the integral curve, $\Big\{(Z, v^{\varepsilon}_{i,\pm}(Z)) : Z \in [0,1)\Big\}$*, tends to the union of the set* $\Big\{(Z, v^{0}_{i,\pm}(Z)) : Z \in [0,1)\Big\}$ *and the horizontal half straight line* $\Big\{(1, v) : v \in [v^{0}_{i,\pm}(1), +\infty)\Big\}$*.*

The proof of this proposition uses the convexity of the pressure and are similar to the one developed in [74].

Regarding the Riemann problem in the limit $\varepsilon \to 0$, the intersection point of the 1-st integral/Hugoniot curves issued from (v_ℓ, Z_ℓ) and the 3-rd integral/Hugoniot curves issued from (v_r, Z_r), denoted by $(v^{\varepsilon}_m, Z^{\varepsilon}_m)$, has either a limit (v^0_m, Z^0_m) with $0 \leqslant Z^0_m < 1$ or tends to a congested state $(\bar v, 1)$. Then finding a solution can be divided into the following steps:

(1) compute the intersection (v^0_m, Z^0_m) of the 1-st integral/Hugoniot curves and 3-rd integral/Hugoniot curves;

(2a) if $Z^0_m < 1$, the solution is as described in the previous section, it is a usual Riemann solution of the hyperbolic system with no congestion pressure;

(2b) if $Z^0_m \geq 1$, then the congested state is given by the following proposition.

Proposition 6.2 (*Case* $Z^0_m \geq 1$.). *The solution consists in three waves:*

$$(\varrho_\ell, q_\ell, Z_\ell) \overset{shock}{\rightarrow} (\varrho^*_\ell, \varrho^*_\ell \bar v, 1) \overset{contact}{\rightarrow} (\varrho^*_r, \varrho^*_r \bar v, 1) \overset{shock}{\rightarrow} (\varrho_r, q_r, Z_r)$$

where the intermediate velocity $\bar v$ *and pressure* $\bar p$ *satisfy:*

$$\bar v = v_\ell - \sqrt{\frac{1}{\varrho_\ell}}\sqrt{(1-Z_\ell)(\bar p - p_0(Z_\ell))} = v_r + \sqrt{\frac{1}{\varrho_r}}\sqrt{(1-Z_r)(\bar p - p_0(Z_r))},$$

the intermediate densities are given by:

$$\widehat{\varrho}_\ell = \varrho_\ell / Z_\ell = \varrho^*_\ell, \quad \widehat{\varrho}_r = \varrho_r / Z_r = \varrho^*_r,$$

and the shock speeds σ_-*,* σ_+ *are given by:*

$$\sigma_- = v_\ell - \sqrt{\frac{\varrho^*_\ell}{\varrho_\ell(\varrho^*_\ell - \varrho_\ell)}}\sqrt{\bar p - p_0(Z_\ell)}, \quad \sigma_+ = v_r + \sqrt{\frac{\varrho^*_r}{\varrho_r(\varrho^*_r - \varrho_r)}}\sqrt{\bar p - p_0(Z_r)}.$$

This proposition can be proven using similar arguments as in [74].

Below, on Figure 6.11 we present two different solutions to the Riemann problem (6.5.1)-(6.5.2). Depending on the initial location of the left and right states, the intersection state (v_m, Z_m) might be a congested state or not.

6.6 Fully Discrete Scheme in Dimension 2

We consider the computational domain $[0,1] \times [0,1]$ and spatial space steps $\Delta x = 1/N_x, \Delta y = 1/N_y > 0$, with $N_x, N_y \in \mathbb{N}$: the mesh points are thus $\boldsymbol{x}_{i,j} = (i\Delta x, j\Delta y)$, $\forall (i,j) \in \{0, \dots, N_x\} \times$

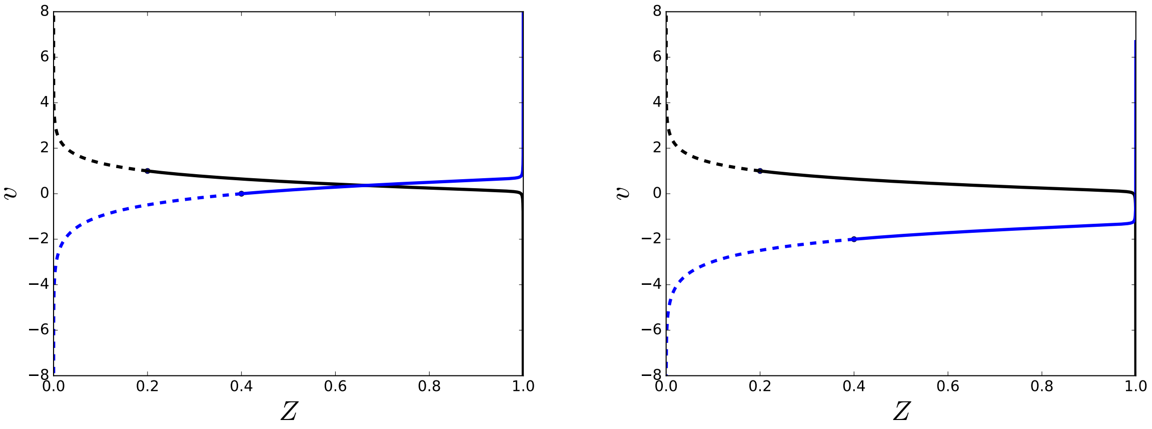


Figure 6.11: Intersection of the 1-integral/Hugoniot curve issued from the left state $(\varrho_\ell, v_\ell, Z_\ell) = (0.8, 1, 0.2)$ and the 3-integral/Hugoniot curve issued from the right state for $\varepsilon = 10^{-3}$. The rarefaction curves are in dashed line and the shock curve in solid line. Left: the right state is given by $(\varrho_r, v_r, Z_r) = (0.8, 0, 0.4)$ and the intermediate state (v^0_m, Z^0_m) is not a congested state. Right: the right state is given by $(\varrho_r, v_r, Z_r) = (0.8, -2, 0.4)$ and the intersection point is very closed to the congested line $Z = 1$.

$\{0, \dots, N_y\}$. Let $\varrho^n_{i,j}$, $\boldsymbol{q}^n_{i,j}$, $Z^n_{i,j}$, $\varrho^{*n}_{i,j}$ denote the approximate solution at time t^n on mesh cell $[i\Delta x, (i+1)\Delta x] \times [j\Delta x, (j+1)\Delta x]$.

The two-dimensional version of (6.2.3) reads:

$$\frac{\varrho^{n+1}_{i,j} - \varrho^n_{i,j}}{\Delta t} + \frac{1}{\Delta x}(F^{n+1}_{(i+\frac{1}{2},j)} - F^{n+1}_{(i-\frac{1}{2},j)}) + \frac{1}{\Delta y}(\widetilde{F}^{n+1}_{(i,j+\frac{1}{2})} - \widetilde{F}^{n+1}_{1,(i,j-\frac{1}{2})}) = 0, \tag{6.6.1}$$

$$\frac{\boldsymbol{q}^{n+1}_{i,j} - \boldsymbol{q}^n_{i,j}}{\Delta t} + \frac{1}{\Delta x}(\boldsymbol{G}^n_{(i+\frac{1}{2},j)} - \boldsymbol{G}^n_{(i-\frac{1}{2},j)}) + \frac{1}{\Delta y}(\widetilde{\boldsymbol{G}}^n_{(i,j+\frac{1}{2})} - \widetilde{\boldsymbol{G}}^n_{(i,j-\frac{1}{2})}) + (\nabla \pi_\varepsilon(Z^{n+1}))_{i,j} = 0, \tag{6.6.2}$$

$$\frac{Z^{n+1}_{i,j} - Z^n_{i,j}}{\Delta t} + \frac{1}{\Delta x}(H^{n+1}_{(i+\frac{1}{2},j)} - H^{n+1}_{(i-\frac{1}{2},j)}) + \frac{1}{\Delta y}(\widetilde{H}^{n+1}_{(i,j+\frac{1}{2})} - \widetilde{H}^{n+1}_{3,(i,j-\frac{1}{2})}) = 0. \tag{6.6.3}$$

where fluxes F^{n+1}, $\boldsymbol{G}^n$, H^{n+1} (in the first spatial direction) are defined:

$$F^{n+1}_{(i+\frac{1}{2},j)} = \frac{1}{2}\Big(q^{n+1}_{1,(i+1,j)} + q^{n+1}_{1,(i,j)}\Big) - (D_\varrho)^n_{i+\frac{1}{2},j}, \tag{6.6.4}$$

$$\boldsymbol{G}^n_{(i+\frac{1}{2},j)} = \frac{1}{2}\Big(\boldsymbol{f}^n_{(i+1,j)} + \boldsymbol{f}^n_{(i,j)}\Big) - (\boldsymbol{D}_{\boldsymbol{q}})^n_{i+\frac{1}{2},j}, \tag{6.6.5}$$

$$H^{n+1}_{(i+\frac{1}{2},j)} = \frac{1}{2}\left(\frac{Z^n_{i+1,j}}{\varrho^n_{i+1,j}} q^{n+1}_{1,(i+1,j)} + \frac{Z^n_{i,j}}{\varrho^n_{i,j}} q^{n+1}_{1,(i,j)}\right) - (D_Z)^n_{i+\frac{1}{2},j}, , \tag{6.6.6}$$

with

$$\boldsymbol{f}^n = \begin{bmatrix} (q^n_1)^2 + p(Z^n) \\ q^n_1 q^n_2 \end{bmatrix}.$$

Fluxes $\widetilde{F}^{n+1}$, $\widetilde{\boldsymbol{G}}^n$, $\widetilde{H}^{n+1}$ in the second spatial direction are defined by:

$$\widetilde{F}^{n+1}_{(i,j+\frac{1}{2})} = \frac{1}{2}\Big(q^{n+1}_{2,(i,j+1)} + q^{n+1}_{2,(i,j)}\Big) - (D_\varrho)^n_{i,j+\frac{1}{2}}, \tag{6.6.7}$$

$$\widetilde{\boldsymbol{G}}^n_{(i,j+\frac{1}{2})} = \frac{1}{2}\Big(\widetilde{\boldsymbol{f}}^n_{(i,j+1)} + \widetilde{\boldsymbol{f}}^n_{(i,j)}\Big) - (\boldsymbol{D_q})^n_{i,j+\frac{1}{2}}, \tag{6.6.8}$$

$$\widetilde{H}^{n+1}_{(i,j+\frac{1}{2})} = \frac{1}{2}\bigg(\frac{Z^n_{i,j+1}}{\varrho^n_{i,j+1}} q^{n+1}_{2,(i,j+1)} + \frac{Z^n_{i,j}}{\varrho^n_{i,j}} q^{n+1}_{2,(i,j)}\bigg) - (D_Z)_{i,j+\frac{1}{2}}, \tag{6.6.9}$$

with

$$\widetilde{\boldsymbol{f}}^n = \begin{bmatrix} q_1^n q_2^n \\ (q_2^n)^2 + p(Z^n) \end{bmatrix}.$$

The upwindings D_ϱ, $\boldsymbol{D_q}$, D_Z are defined similarly as for the one-dimensional case (sse (6.2.7)-(6.2.8)).

The implicit pressure in (6.6.2) is discretized by the centered difference:

$$(\nabla \pi_\varepsilon(Z^{n+1}))_{i,j} = \begin{bmatrix} \dfrac{\pi_\varepsilon(Z^{n+1}_{i+1,j}) - \pi_\varepsilon(Z^{n+1}_{i-1,j})}{2\Delta x} \\ \dfrac{\pi_\varepsilon(Z^{n+1}_{i,j+1}) - \pi_\varepsilon(Z^{n+1}_{i,j-1})}{2\Delta y} \end{bmatrix}.$$

Inserting equation (6.6.2) into (6.6.3), we obtain:

$$\begin{aligned}
&Z^{n+1}_{i,j} - Z^n_{i,j} + \frac{\Delta t}{\Delta x}\Big(\bar{H}^n_{(i+\frac{1}{2},j)} - \bar{H}^n_{(i+\frac{1}{2},j)}\Big) + \frac{\Delta t}{\Delta y}\Big(\bar{\widetilde{H}}^n_{(i+\frac{1}{2},j)} - \bar{\widetilde{H}}^n_{(i+\frac{1}{2},j)}\Big) \\
&- \frac{\Delta t^2}{\Delta x^2}\frac{1}{2}\bigg(\frac{Z^n_{i+1,j}}{\varrho^n_{i+1,j}}\Big(G^n_{(i+\frac{3}{2},j),1} - G^n_{(i+\frac{1}{2},j),1}\Big) - \frac{Z^n_{i-1,j}}{\varrho^n_{i-1,j}}\Big(G^n_{(i-\frac{1}{2},j),1} - G^n_{(i-\frac{3}{2},j),1}\Big)\bigg) \\
&- \frac{\Delta t^2}{\Delta x \Delta y}\frac{1}{2}\bigg(\frac{Z^n_{i+1,j}}{\varrho^n_{i+1,j}}\Big(\widetilde{G}^n_{(i+1,j+\frac{1}{2}),1} - \widetilde{G}^n_{(i+1,j-\frac{1}{2}),1}\Big) \\
&\qquad\qquad\qquad - \frac{Z^n_{i-1,j}}{\varrho^n_{i-1,j}}\Big(\widetilde{G}^n_{(i-1,j+\frac{1}{2}),1} - \widetilde{G}^n_{(i-1,j-\frac{1}{2}),1}\Big)\bigg) \\
&- \frac{\Delta t^2}{\Delta y^2}\frac{1}{2}\bigg(\frac{Z^n_{i,j+1}}{\varrho^n_{i,j+1}}\Big(\widetilde{G}^n_{(i,j+\frac{3}{2}),2} - \widetilde{G}^n_{(i,j+\frac{1}{2}),2}\Big) - \frac{Z^n_{i,j-1}}{\varrho^n_{i,j-1}}\Big(\widetilde{G}^n_{(i,j-\frac{1}{2}),2} - \widetilde{G}^n_{(i,j-\frac{3}{2}),2}\Big)\bigg) \\
&- \frac{\Delta t^2}{\Delta x \Delta y}\frac{1}{2}\bigg(\frac{Z^n_{i,j+1}}{\varrho^n_{i,j+1}}\Big(G^n_{(i+\frac{1}{2},j+1),2} - G^n_{(i-\frac{1}{2},j+1),2}\Big) \\
&\qquad\qquad\qquad - \frac{Z^n_{i,j-1}}{\varrho^n_{i,j-1}}\Big(G^n_{(i+\frac{1}{2},j-1),2} - G^n_{(i-\frac{1}{2},j-1),2}\Big)\bigg) \\
&- \frac{\Delta t^2}{\Delta x^2}\frac{1}{4}\bigg(\frac{Z^n_{i+1,j}}{\varrho^n_{i+1,j}}\Big(\pi_\varepsilon(Z^{n+1}_{i+2,j}) - \pi_\varepsilon(Z^{n+1}_{i,j})\Big) - \frac{Z^n_{i-1,j}}{\varrho^n_{i-1,j}}\Big(\pi_\varepsilon(Z^{n+1}_{i,j}) - \pi_\varepsilon(Z^{n+1}_{i-2,j})\Big)\bigg) \\
&- \frac{\Delta t^2}{\Delta y^2}\frac{1}{4}\bigg(\frac{Z^n_{i,j+1}}{\varrho^n_{i,j+1}}\Big(\pi_\varepsilon(Z^{n+1}_{i,j+2}) - \pi_\varepsilon(Z^{n+1}_{i,j})\Big) - \frac{Z^n_{i,j-1}}{\varrho^n_{i,j-1}}\Big(\pi_\varepsilon(Z^{n+1}_{i,j}) - \pi_\varepsilon(Z^{n+1}_{i,j-2})\Big)\bigg) = 0,
\end{aligned}$$

where terms $\bar{H}^n$ and $\bar{\widetilde{H}}^n$ have the same expressions as (6.6.6)-(6.6.9) but where all quantities are taken explicitly.

6.7 Second Order in Time (ϱ, q)-method/SL

The second order accuracy scheme for the (ϱ, q)-method/SL is based on a Strang splitting between advection of congestion density and advection of $(\varrho, \boldsymbol{q})$. It consists in the following steps:

1. Compute $\varrho^{*\ n+1/2}$ by solving the advection of over $\Delta t/2$

$$\frac{\varrho^{*n+1/2} - \varrho^{*n}}{\Delta t/2} + \frac{\boldsymbol{q}^n}{\varrho^n} \cdot \nabla_x \varrho^{*n} = 0.$$

2. Compute (ϱ^{n+1}, q^{n+1}) with the RK2CN scheme as proposed in [65]:

First step (half time step):

$$\frac{\varrho^{n+1/2} - \varrho^n}{\Delta t/2} + \nabla_x \cdot \boldsymbol{q}^{n+1/2} - \mathscr{D}_\varrho^n = 0,$$
$$\frac{\boldsymbol{q}^{n+1/2} - \boldsymbol{q}^n}{\Delta t/2} + \nabla_x \cdot \left(\frac{\boldsymbol{q}^n \otimes \boldsymbol{q}^n}{\varrho^n} + p(Z^n)\boldsymbol{I}\right) - \mathscr{D}_q^n + \nabla_x(\pi_\varepsilon(\varrho^{n+1/2}/\varrho^{*,n+1/2})) = 0.$$

Second step (full time step):

$$\frac{\varrho^{n+1} - \varrho^n}{\Delta t} + \nabla_x \cdot \left(\frac{\boldsymbol{q}^{n+1} + \boldsymbol{q}^n}{2}\right) - \mathscr{D}_\varrho^n = 0,$$
$$\frac{\boldsymbol{q}^{n+1} - \boldsymbol{q}^n}{\Delta t} + \nabla_x \cdot \left(\frac{\boldsymbol{q}^{n+1/2} \otimes \boldsymbol{q}^{n+1/2}}{\varrho^{n+1/2}} + p(Z^{n+1/2})\boldsymbol{I}\right) - \mathscr{D}_q^{n+1/2}$$
$$+ \nabla_x \left(\frac{\pi_\varepsilon(\varrho^n/\varrho^{*,n+1/2}) + \pi_\varepsilon(\varrho^{n+1}/\varrho^{*,n+1/2})}{2}\right) = 0.$$

where $\mathscr{D}_\varrho$, $\mathscr{D}_q$ denote the numerical diffusion coming from fluxes.

3. Advection of ϱ^* on $\Delta t/2$ time step

$$\frac{\varrho^{*n+1} - \varrho^{*n+1/2}}{\Delta t/2} + \frac{\boldsymbol{q}^{n+1}}{\varrho^{n+1}} \cdot \nabla_x \varrho^{*n+1/2} = 0.$$

A second order in time version of the semi-Lagrangian scheme has to be used. We here consider the second order Taylor approximation of the caracteristic line whose one-dimensional version reads:

$$\varrho_i^{*n+1} = [\Pi\varrho^{*n}]\left(x_i - v_i\Delta t + a_i v_i \frac{\Delta t^2}{2}\right).$$

where $v_i = q_i/\varrho_i$ for all i and a_i is an upwind finite difference approximation of the first derivative of the velocity: $a_i = (v_i - v_{i-1})/\Delta x$ if $v_i > 0$ and $a_i = (v_{i+1} - v_i)/\Delta x$ if $v_i \leqslant 0$.

CHAPTER 7

Density-induced Consensus Protocol

The content of this chapter is a joint work with Piotr Mucha and Jan Peszek, and is published in the paper

P. Minakowski, P. B. Mucha and J. Peszek. 'Density-Induced Consensus Protocol'. In: *Mathematical Models and Methods in Applied Sciences* 30.12 (July 2020), pp. 2389–2415. doi: 10.1142/s0218202520500451

Chapter Summary. We introduce a model of collective behavior where agents receive information only from sufficiently dense crowds in their immediate vicinity. The system is an asymmetric, density-induced version of the Cucker-Smale model with short-range interactions. We prove the basic mathematical properties of the system and concentrate on the presentation of interesting behaviors of the solutions. The results are illustrated by numerical simulations.

Chapter Organisation. After introductory Section 7.1, in Section 7.2 we introduce the main motivations behind the model (Section 7.2.1), we present the main results (Section 7.2.2) and reformulate the problem in terms of graphs (Section 7.2.3), as well as deliver additional remarks and comparison with other relevant models. Section 7.3 and Section 7.4 are dedicated to prove the main results: Theorems 7.5 and 7.6, and to showcase the possible interaction between clusters (No. 1 – No. 3). Finally, in Section 7.5, we present numerical simulations.

Contents of Chapter

7.1 Introduction

We consider an ensemble of N agents with $(x_i(t), v_i(t)) \in \mathbb{R}^{2d}$ denoting the position and the velocity of ith agent at the time $t \geq 0$. The agents follow the density-induced consensus protocol (DI)

$$\begin{cases} \dot{x}_i = v_i, & x_i(0) = x_{i0} \in \mathbb{R}^d, \qquad (7.1.1a) \\ \dot{v}_i = \displaystyle\sum_{k \in \mathscr{N}_i} M(v_k - v_i), & v_i(0) = v_{i0} \in \mathbb{R}^d, \qquad (7.1.1b) \end{cases}$$

where $\mathscr{N}_i$ is a neighbor set of ith agent. The definition of this set requires some care. It is not empty if the number of other agents, in a given ball centered at the ith agent, is big enough. Thus the interaction is engaged only if $\mathscr{N}_i$ is not trivial. In addition, due to technicalities related to well possedness, the communication is delayed in time.

Hence it is defined through the following relation: given positive parameters δ, m and h, for $t \geq h$ we define

$$\begin{aligned} k \in \mathscr{N}_i(t) \ \Leftrightarrow\ & x_k(t-h) \in B(x_i(t-h), \delta) \text{ and} \\ & \#\Big\{k \in \{1, ..., N\} : x_k(t-h) \in B(x_i(t-h), \delta)\Big\} > m \end{aligned} \qquad (7.1.2)$$

where $B(x_i(t-h), \delta)$ is an open ball centered at $x_i(t-h)$ with radius δ and $\#A$ denotes the number of elements of the set A. For $t \in [0, h)$ we take

$$\begin{aligned} k \in \mathscr{N}_i(t) \ \Leftrightarrow\ & x_k(0) \in B(x_i(0), \delta) \text{ and} \\ & \#\Big\{k \in \{1, ..., N\} : x_k(0) \in B(x_i(0), \delta)\Big\} > m. \end{aligned} \qquad (7.1.3)$$

Parameter $0 < h$, negligibly small compared to the rest of parameters, is introduced to ensure that the neighbour sets $\mathscr{N}_i(t)$ are well defined. Indeed, taking $h = 0$ causes instability of the system if a particle is situated at the boundary of $B(x_i(t), \delta)$. A natural interpretation of h is a time step from a discrete in time version of (7.1.1). To grasp the intuition behind the behaviour, we can view h as 0 and focus on the qualitative analysis of the model, which is the main goal of the paper. Further explanation can be found in Remark 7.9. Parameter $M = M(N, \#\mathscr{N}_i)$ is a normalizing factor discussed later; to fix our attention we may assume that $M = \kappa/\#\mathscr{N}_i$, where $\kappa > 0$ is the non-dimensional coupling strength and $\#\mathscr{N}_i$ is the number of elements in $\mathscr{N}_i$.

Condition (7.1.2) introduces a two-step verification of whether k belongs to the set of neighbors of i. First, $x_k(t-h)$ is required to be close to $x_i(t-h)$. Second, the crowd density in the immediate vicinity of i needs to be large enough so that the number of the individuals with $x_k(t-h) \in B(x_i(t-h), \delta)$ is larger than m. If the second condition is not satisfied, the set of neighbors of i is empty. Observe that the rightmost condition in definition (7.1.2) is asymmetric and consequently so is the relation of adjacency $\rightsquigarrow$ defined by $k \rightsquigarrow i \Leftrightarrow k \in \mathscr{N}_i$.

Dividing the right-hand side of (7.1.2) by δ^d (where d is the dimension of the space) identifies $\frac{m}{\delta^d}$ as a threshold imposed on the empirical density of the particles. Hence, the interaction is induced by sufficiently high density of the agents and the DI protocol operates within the following general paradigm:

To influence the behavior of an individual,
communication with a sufficiently dense nearby crowd is required.

The protocol is specific to real-life phenomena related to societal dynamics, where individuals do not interact with separate agents but are highly susceptible to the influence of crowds. It is inspired by such phenomena as emergence of trends in decision-making and viral videos in social media. In such a setting we view x_i and v_i as the opinion and tendency of the individual, respectively (we elaborate it in Section 7.2.1).

Formally (7.1.1) is a second order system which in the kinetic formalism determines the acceleration of particles in terms of their positions and velocities. Indeed, it originates from a highly recognizable model of collective behavior, the classical Cucker-Smale (CS) system

$$\begin{cases} \dot{x}_i = v_i, & x_i(0) = x_{i0} \in \mathbb{R}^d, \qquad (7.1.4a) \\ \dot{v}_i = \dfrac{1}{N}\displaystyle\sum_{k=1}^{N} \psi(|x_i - x_k|)(v_k - v_i), & v_i(0) = v_{i0} \in \mathbb{R}^d, \qquad (7.1.4b) \end{cases}$$

with $\psi(s) = (1+s)^{-\alpha}$.

Comparing the DI and CS models, we observe *the first prominent feature* of the DI protocol: it leads to the emergence of sharply distinguishable, dense clusters. Moreover, we observe local flocking, with nontrivial dependence on the density; particularly the *time* = 150 velocity of singletons varies significantly, see Fig. 7.1.

The second prominent feature of the DI model is a structural asymmetry of the interactions showcased in Fig. 7.2.

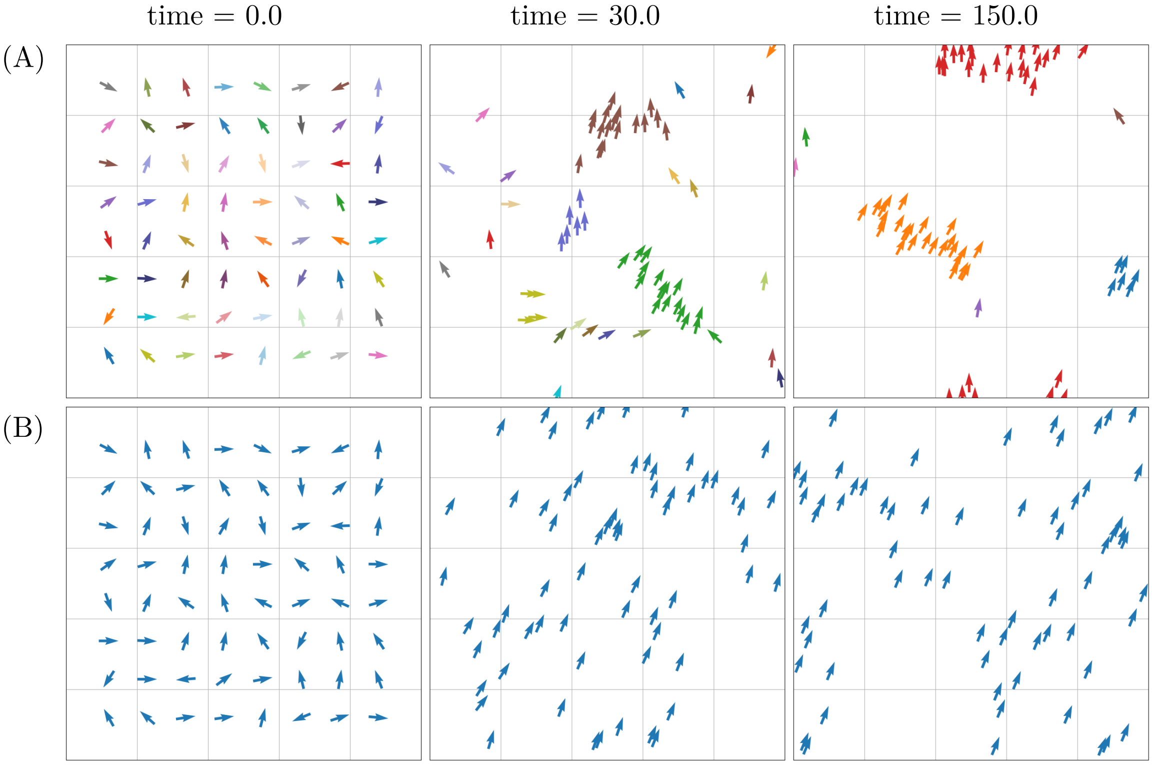


Figure 7.1: Behaviour of particles for DI (A) and CS (B). Identical initial data lead to sharper clusters for the DI model. The color coding represents clusters of indirect communication - multiple clusters for DI and a single cluster for (global) CS. A video available at https://youtu.be/qDFTOQ_8mFo.

In the case (A), at *time* $= 2.0$ the orange singleton already influences the cluster but the cluster does not influence the singleton yet, because the number of the particles in the vicinity of the singleton is insufficient for it to receive communication. Consequently, there is a period of time when the singleton influences the cluster but is yet to be affected by the cluster, which results in a particularly strong asymmetry of the interaction. On the other hand in the case (B), at *time* $= 2.0$ the orange singleton is already influenced by the cluster. Therefore, a singleton can change the direction of an entire crowd and whether it succeeds depends on the spacial distribution of the crowd itself. These phenomena are generally impossible to obtain in most well-known models of consensus.

Notation. In what follows we use bold symbols for points in $\mathbb{R}^{dN}$ or $\mathbb{R}^{d\#\mathscr{A}}$ space, where a cluster $\mathscr{A}$ is any subset of $\{1, ..., N\}$, namely

$$\mathbf{x} = (x_1, ..., x_N) \in \mathbb{R}^{dN}, \quad \boldsymbol{v} = (v_1, ..., v_N) \in \mathbb{R}^{dN}, \quad \mathbf{x}_{\mathscr{A}} = (x_{i_1}, ..., x_{i_{\#\mathscr{A}}}) \in \mathbb{R}^{d\#\mathscr{A}}.$$

This convention applies for instance to the initial data $(\mathbf{x}_0, \boldsymbol{v}_0) \in \mathbb{R}^{2dN}$.

7.2 Preliminaries

7.2.1 Motivations

The main motivation of our DI model and the reason why we decided to make it second-order is so that it can be interpreted as a topological CS model with ”propagative” topology (as opposed to

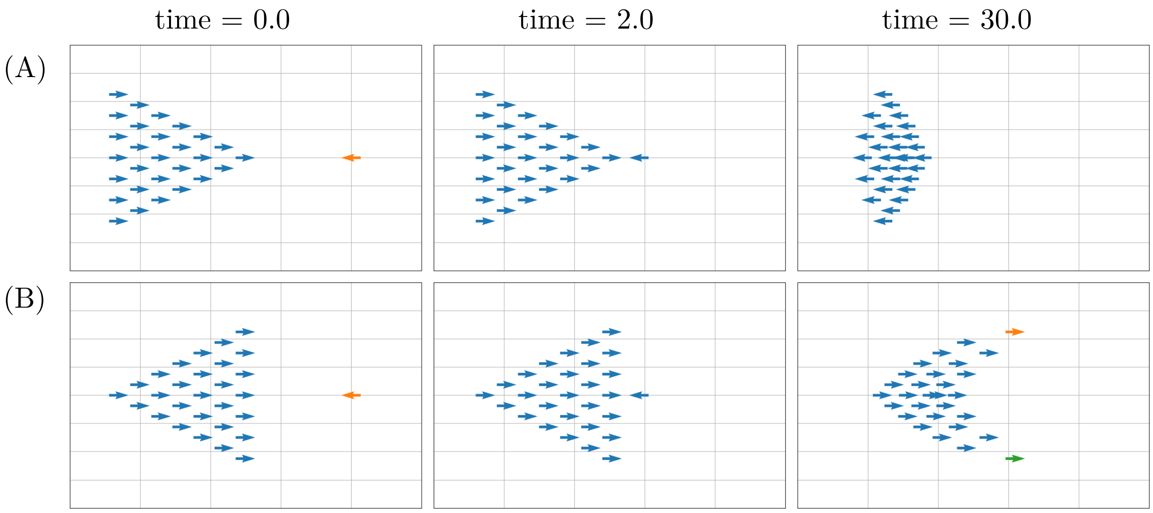


Figure 7.2: Interaction between a group and an individual for the DI model. Initially (A) and (B) differ only by the shape of the cluster. In (A) the individual diverts the cluster while in (B) the cluster diverts the individual. A video available at `https://youtu.be/_y4g-qyu6l0`.

”interference” topology presented by Shvydkoy and Tadmor [223]). Indeed, by taking $M = \kappa/N$ in (7.1.1b) we obtain

$$\dot{v}_i = \frac{\kappa}{N}\sum_{k=1}^{N}\psi_{ik}(v_k - v_i), \tag{7.2.1}$$

with suitable $\mathscr{N}_i$- and x_k-dependent $\psi_{ik} \in \{0, 1\}$, which can be treated as a function of the density of the particles in a neighborhood of x_i. The larger the density around x_i, the more likely it is for ψ_{ik} to be equal to 1. In [223] the ”interference” topological part of the interaction can be roughly expressed as

$$\psi_{ik} = \frac{1}{m_{ik}^{\gamma}}, \quad \gamma > 0, \tag{7.2.2}$$

where m_{ik} is the density of particles between ith and kth particle. In (7.2.1) with the DI interactions the neighboring particles actively enhance the propagation of information while in (7.2.2) they impede it. This motivation is essential to our considerations and we provide more extensive comparison to variants of the CS model in the sequel. Naturally, for the sake of potential applications, unless a very peculiar phenomena are considered, the DI model should be combined with vision-based interactions originating in classical models of animal behavior [68][18], crowd dynamics [72][85] [4] and economics [16][17]. This is however outside of the scope of the paper, which focuses solely on the interactions through clusters.

The second motivation originates in applications to opinion dynamics. Here we interpret x_i as the opinion of ith individual and v_i as the tendency. Interaction propagated exclusively through clusters is characteristic to phenomena that significantly emphasize influence of group-thinking over the individual drive, such as observed in emergence of viral videos and news in social media or in decision-making in consumption. It is also reasonable in phenomena, where clusters have essential interpretation themselves, such as models of organization membership. In this case an individual does not interact with other individuals, the change of the opinion occurs as the singleton interacts with a cluster-organization.

For more information on nonlinear opinion dynamics we refer to the works by Krause [144] and Hegselmann and Krause [122]. We also recommend the survey [6]. Compared, for instance, to the Hegselmann-Krause *bounded confidence* model [122], in which the interactions occur only between similar-minded individuals, the DI protocol puts an emphasis on the individual's aptitude to conform with groups of similar-minded individuals. Of course the DI model is of second order as opposed to more classical models of opinion dynamics.

Interestingly enough, a variant of the DI protocol is widely applied in computer science, particularly in machine learning. The cluster formation is equivalent to a density-based spatial clustering algorithm (DBSCAN) [93]. The protocol has advantages such as: it does not require to a-priori specify the number of clusters, it can find arbitrarily shaped clusters, and it does not affect outliers.

Linear consensus

In what follows $G = (V, \mathscr{E}, \mathbf{\Phi})$ is an N-node digraph generating the ODE system

$$\dot{\boldsymbol{v}} = -\mathbf{A}\boldsymbol{v}, \qquad \mathbf{A} = \operatorname{diag}(d_1, ..., d_N) - \mathbf{\Phi}, \tag{7.2.3}$$

where

$$d_i = \sum_{k=1}^{N} \Phi_{ik}$$

is the out-degree of the ith node. Matrix $\mathbf{A}$ is known as the Laplacian matrix.

Theorem 7.1. *Graph G is strongly connected if and only if* $\operatorname{rank}(\mathbf{A}) = N - 1$.

Theorem 7.2. *Denote the maximum out-degree of G by $d_{max}(G) = \max_i d_i$. Then all the eigenvalues of $\mathbf{A}$ are located in the disk*

$$D(G) = \{z \in \mathbb{C} : |z - d_{max}(G)| \le d_{max}(G)\}.$$

Proposition 7.3. *If G is strongly connected then system (7.2.3) admits an asymptotically stable steady state. This steady state belongs to the convex hull of $\{v_i(0) : i = 1, ..., N\}$.*

The above theorems and proposition are presented in Section 4 of [191] (Theorems 1 and 2 and Corollary 1, respectively). Second statement in Proposition 7.3 is presented in Corollary 2[191] and can be deduced from Perron-Frobenius theorem. Theorem 7.2 is a direct application of Geršgorin circle theorem. We note that stability in Proposition 7.3 is actually exponential, which follows immediately by solving (7.2.3) using Jordan decomposition of $\mathbf{A}$.

7.2.2 Main Results

Hereinafter, we fix our attention by referring to the individuals following the DI protocol as particles with position x_i and velocity v_i (of ith particle). Inter-particle interactions within the DI model highly depend on the local density of the ensemble, and thus the following notion of densely packed clusters is of significant importance.

Definition 7.4. We say that the cluster $\mathscr{A} \subset \{1, ..., N\}$ is r-densely packed at t if

1. the Minkowski sum
$$\mathbf{x}_{\mathscr{A}}(t-h)+B(0,r/2):=\{x+y : x\in\mathbf{x}_{\mathscr{A}}(t-h),\ y\in B(0,r/2)\}$$
is connected,
2. each open ball $B(x_k(t-h),r)$, for $k\in\mathscr{A}$ contains more than m particles,

with convention that $t-h=0$ if $t\le h$.

As we show later in Lemma 7.11, interactions between the particles in any r-densely packed cluster for $r\le\delta$ are propagated to the entire cluster, which serves as a stepping stone to obtain a flocking estimate.

We perform basic qualitative and quantitative analysis of (7.1.1). That includes: the existence and uniqueness of classical solutions, conditional flocking estimates, and conditional cluster stability in terms of cluster density. Moreover, we provide analytical examples signifying the rich dynamics of the system. Finally, we illustrate the variability of possible behaviors by a number of numerical simulations.

The main results of the paper read as follows.

Theorem 7.5. *Given the time interval $[0,T)\subset[0,+\infty)$, parameters $M,m,\delta,h>0$ and initial data $(\mathbf{x}_0,\boldsymbol{v}_0)\in\mathbb{R}^{2dN}$, the system (7.1.1) admits a unique classical $W^{1,\infty}([0,T))$ solution with nonincreasing velocity fluctuations*
$$V(t):=\max_{i,j\in\{1,\ldots,N\}}|v_i(t)-v_j(t)|\le V(s)\le V(0),\qquad 0<s\le t\le T. \tag{7.2.4}$$

In the following theorem M is a positive function of i, N and $\#\mathscr{N}_i$, thus it has a finite set of possible values, and thus it is bounded below by a constant $M_*>0$. It is further discussed in Section 7.2.3.

Theorem 7.6 (*Dense clusters flock*). *Given the time interval $[0,T)\subset[0,+\infty)$, parameters $M,m,\delta,h>0$ and initial data $(\mathbf{x}_0,\boldsymbol{v}_0)\in\mathbb{R}^{2dN}$ suppose that the entire ensemble $\{1,...,N\}$ is r-densely packed with $r<\delta$. For sufficiently large M the ensemble remains at least δ-densely packed for all $t>0$ and the particles flock exponentially fast.*

In other words there exists a constant $\lambda>0$ depending d, N and the initial data, such that
$$V(t)\le 2e^{-M_*\lambda t}V(0),$$
provided that
$$M_*:=\inf M>\frac{2V(0)}{\lambda(\delta-r)}. \tag{7.2.5}$$

Remark 7.7. Theorem 7.6 concerns the whole ensemble $\{1,...,N\}$ but also holds for any cluster $\mathscr{A}\subset\{1,...,N\}$ with the a priori assumption that $\mathscr{A}$ is sufficiently far away from the rest of the particles so that it is never influenced by them.

Remark 7.8. Assumption (7.2.5) is first and foremost an assumption on M; since the DI model has local interactions, similarly to the short-range CS model, flocking occurs with high coupling strength [177]. However it can be viewed also as an assumption on the initial density governed by the constant r. With small r the cluster is initially more densely packed and the right-hand side of (7.2.5) decreases.

To illustrate possible behaviors of particles governed by the DI model, we provide the following analysis of clusters' interactions.

Clusters' interactions. Let us consider two connected to each other clusters $\mathscr{A}$ and $\mathscr{C}$. Depending on the positions and velocities of individual particles within each cluster the following scenarios may occur:

1. Stability: Cluster $\mathscr{A}$ remains connected. Cluster $\mathscr{C}$ eventually detaches from $\mathscr{A}$ and they move separately.
2. Breaking: Cluster $\mathscr{A}$ breaks under the influence of cluster $\mathscr{C}$. Cluster $\mathscr{C}$ detaches from $\mathscr{A}$ together with a number of particles from $\mathscr{A}$.
3. Sticking: Cluster $\mathscr{A}$ is diverted by cluster $\mathscr{C}$. They remain connected indefinitely and their total momentum changes in the direction of the total momentum of $\mathscr{C}$.

The example showcasing all of the above scenarios, with a special choice of clusters $\mathscr{A}$ and $\mathscr{C}$, can be found in Section 7.4.

7.2.3 Comparison with other Second-Order Models

In recent years, it was recognized that the global all-to-all character of interactions in the Cucker-Smale (CS) flocking model not always corresponds to the actual behavior of agents in real-life phenomena, be it flocks of birds, networks of unmanned aerial vehicles or in opinion dynamics. The main issue is that, usually, the range of communication between autonomous agents is finite, the interactions are not symmetric and the structure of the network of interactions is not necessarily immersed in the standard Euclidean geometry. For instance, the range of perception of a bird within a flock tends to be a finite cone-shaped area in front of it. Responding to these issues a number of non-standard alignment CS-type models emerged recently. Among them is the model with short-range interactions (CS$_\delta$ model), the q-closest neighbors model (CS$_q$ model – motivated by the analysis of empirical data in [18]) and the model with interference topology (CS$_t$ model). Below we summarize similarities and differences between these models and the DI model. Further, numerical comparison can be found in Section 7.5.

◇ *Short-range model CS$_\delta$.* The CS$_\delta$ model is a simple modification of the standard CS model in which the smooth and decreasing communication weight ψ_δ is assumed to be compactly supported on the set $[0, \delta]$. For instance, the classical CS weight $\psi(s) = (1+s)^{-\frac{1}{2}}$ could be cut-off by taking $\psi_\delta(s) = \psi(s)\chi_{[0,\delta]}$. With such a modiffication qualitative behavior and particularly asymptotics becomes significantly more difficult to study. Compared to the DI model, CS$_\delta$ model is symmetric and purely geometrical with $j \in \mathscr{N}_i(t)$ if and only if $|x_i(t) - x_j(t)| \leq \delta$. For further information on the CS$_\delta$ model we refer to [177, 135, 116, 223, 83].

◇ *The model with q-closest neighbors CS$_q$.* The CS$_q$ model is a modification of the CS model with the sum in (7.1.4) taken over only those j that are at most q-closest to i in terms of position. This model is both density dependent and non-symmetric, with an opposite influence of the density to the DI model. While in the DI model high density is used to propagate the interactions, in the CS$_q$ model the interactions spread over constant mass of the particles and thus with low density the interactions reach further. For further information on the CS$_q$ model we refer to [67].

◇ *Topological model with interference CS$_t$.* Recently, Shvydkoy and Tadmor introduced a CS-type alignment model, in which the interactions between the particles depend on the mass of the particles belonging to a symmetric area between them. The higher the mass of the particles between x_i and x_j, the lower the interaction, which justifies the interpretation that intermediate particles interfere with the communication. The CS$_t$ model is density-dependent, symmetric and, similarly to CS$_q$, the influence of the density seems to be the opposite to the DI model. For further information on the CS$_t$ model we refer to [223].

The main difference between the DI model and the aforementioned is that the agents interact only with a densely-packed crowd (defined by the threshold m). Singletons, outsiders and agents forming low-density clusters do not interact. As a somewhat surprising consequence we obtain the possibility of a significant influence of the outsider particle on a cluster as showcased in Fig. 7.2. While, interactions of crowds with highly influential individuals were studied in the past [150, 218, 66], we obtain such phenomenon naturally within the framework of the model without artificially boosting the influence of any individual. Whether an individual manages to influence a cluster significantly depends not only on the models parameters but also on the spatial structure of the clusters, c.f. Section 7.5.2.

Further state of the art for related models, including asymptotics [51, 116, 180, 68] and pattern formation [231, 60], collision avoidance [53, 157] including models with singular interactions [199, 200, 247], time delay [59] and the recently emerging thermodynamical consistence [57, 117] can be found in the surveys [58] (CS model with regular interactions) and [166] (CS model with singular interactions). The surveys include also references to kinetic [5, 47, 181] and hydrodynamic [143, 222, 121, 221, 71] limits of the CS model.

On the Normalizing Factor M

There are many ways to normalize a consensus particle system, with the most popular prominently represented in the works by Cucker and Smale [68] and Motsch and Tadmor [179]. The first one, corresponding to $M = \frac{\kappa}{N}$, is the standard "flat" Cucker-Smale normalization, which connects the magnitude of the interactions to the total mass of the particles. The second one, $M = \frac{\kappa}{\#\mathcal{N}_i}$, corresponds to the non-symmetric Motsch-Tadmor normalization. Throughout the paper we shall use a more general assumption that M is any positive function of i, N and $\#\mathcal{N}_i$ i.e.

$$M^* \geq M = M(i, N, \#\mathcal{N}_i) \geq M_* > 0. \tag{7.2.6}$$

Existence of positive constants M^* and M_* above follows from the fact that M is a positive function on a discrete domain (with fixed N). Of course, with M as in (7.2.6) we incorporate both Cucker-Smale and Motsch-Tadmor normalizations.

Communication on Graphs

The analysis of asymptotic or otherwise qualitative behavior of system (7.1.1) is based on two foundations. The first is the analysis of how any particular configuration of particles $\mathbf{x}$ influences their neighbor sets. This issue is closely related to the notion of r-densely packed clusters. The second foundation is focused on the translation of the propagation of interaction to the asymptotic behavior of the particles. This process was widely studied in the framework of graph theory (see for instance the works by Olfati-Saber [191, 190]). This motivates the following reformulation of (7.2.8). Recalling the possible dependence of M on i in (7.2.6) we take

$$\Phi_{ik} = \frac{M_i}{M_*} \quad \text{iff} \quad k \in \mathscr{N}_i, \quad \text{otherwise} \quad \Phi_{ik} = 0, \tag{7.2.7}$$

and we rewrite (7.1.1) as

$$\dot{x}_i = v_i, \qquad \dot{v}_i = M_* \sum_{k=1}^{N} \Phi_{ik}(v_k - v_i). \tag{7.2.8}$$

Note that the lack of symmetry of the adjacency relation between the particles implies that usually $\Phi_{ik} \neq \Phi_{ki}$.

Equation (7.2.8) above, can be further restated as a model on graphs. To this end for any fixed time $t > 0$ let $G = G(t)$ with

$$G = (V, \mathscr{E}, \mathbf{\Phi}) \ \text{ be a digraph (or directed graph),}$$

where V is the set of N nodes corresponding to the N particles and $\mathscr{E} = \mathscr{E}(t) \subset V \times V$ is the set of edges between the nodes, that represents the connectivity of G. Matrix $\mathbf{\Phi} = \mathbf{\Phi}(t) = [\Phi_{ik}]$ describes interaction weights: kth node is interacting with ith node iff $\Phi_{ik} > 0$, or equivalently iff $k \in \mathscr{N}_i$. Then (7.1.1b) can be restated as

$$\dot{\mathbf{x}} = \boldsymbol{v}, \qquad \dot{\boldsymbol{v}} = M_*(\mathbf{\Phi} - \mathbf{D})\boldsymbol{v}, \qquad \mathbf{D} = \operatorname{diag}(d_1, ..., d_N), \tag{7.2.9}$$

where

$$d_i = \frac{\#\mathscr{N}_i M_i}{M_*} = \sum_{k=1}^{N} \Phi_{ik}$$

is the (M-scaled) out-degree of the ith node of G.

The importance of the reformulation (7.2.9) lies in its usefulness in the proof of Theorem 7.6 following a 4-step strategy as presented in the diagram:

$$\begin{tikzcd}
\text{densely packed clusters} \arrow[r] & \text{connectivity of } G \arrow[d] \\
\text{propagation of the density} \arrow[u] & \text{control over velocity} \arrow[l]
\end{tikzcd}$$

It can be summarized as follows. Starting with a densely packed cluster at $t = t_0$ we ensure (see Lemma 7.11 below) that interaction between particles in such a cluster is propagated to every particle. In other words, the graph $G(t_0)$ is strongly connected. Using connectivity of $G(t_0)$, we apply theory developed in [191, 190] to establish local-in-time exponential decay of the relative velocity of the particles (see Lemma 7.12 below). This estimate enables us to control the changes in the positions of the particles and in turn – it ensures the propagation of the density. This leads back to the densely packed clusters at some $t > t_0$ and an indefinite repetition of the scheme.

7.3 Proof of the Main Results

In this section we prove Theorem 7.5 and Theorem 7.6.

7.3.1 Proof of Theorem 7.5

Proof. (Theorem 7.5: Existence.)

By the definition of $\mathscr{N}_i$ the system is elementarily solved on $[0, h]$ – it is a linear system with constant coefficients. Its unique solution is a linear combination of functions of the type $t^k e^{\alpha t}$, with $k \leq N$ and α determined by eigenvalues of the matrix appearing in (7.2.9).

The main difficulty is that neighborhoods $\mathscr{N}_i$ change with time, which causes discontinuities of the right-hand side in (7.2.9). This is not particularly problematic unless the number of discontinuities is infinite. Thus, we want to exclude an oscillatory behavior of solutions.

We proceed by induction. Suppose that the solution exists on $[(k-1)h, kh]$ for some $k = 1, 2, ...$ and the sets of neighbors $\mathscr{N}_i$ change finitely many times for each particle in the time interval $[(k-1)h, kh]$. Then in the time interval $[kh, (k+1)h]$ the system is well defined with a finite number of discontinuities. In particular $[kh, (k+1)h]$ can be decomposed into a finite collection of intervals on which the system is linear. On each of these intervals a unique solution exists and is a linear combination of functions of the type $t^k e^{\alpha t}$. These intervals can be glued together and while the smoothness of the solution may be lost at the endpoints, the right-hand side of (7.1.1) is a bounded function of $\mathbf{x}, \boldsymbol{v}$ and thus the solution belongs at least to $W^{1,\infty}([kh, (k+1)h])$ and is continuous and unique. Then such a solution, as a piecewise linear combination of analytic functions, cannot exhibit an oscillatory behavior and the number of times of discontinuity in $[kh, (k+1)h]$ is finite. By induction, this procedure can be repeated indefinitely and the existence of solutions with $\mathbf{x}$ and $\boldsymbol{v}$ continuous and $\boldsymbol{v} \in W^{1,\infty}([0, T))$ is proved for any $T > 0$.

□

Remark 7.9. The reason to introduce the parameter $h > 0$ is to circumvent the problems arising in the following scenario. Suppose that $h = 0$ and at $t = 0$ we have $|x_1(0) - x_2(0)| = \delta$ and $v_1(0) = v_2(0)$. Then it is unclear whether at $t > 0$ the particles will be within each others range. Introducing $h > 0$ is an easy way to ignore this problem as shown in the above proof.

□

In order to prove the decay of velocity fluctuations we apply the L^∞ method used, for instance in [179].

Proof. (Theorem 7.5: Decay of velocity fluctuations.) Fix $t > 0$ and let the maximal relative velocity between the particles be realized by ith and jth particles i.e.

$$V(t) = |v_i(t) - v_j(t)|.$$

Using the (7.2.8) formulation of the velocity equation, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(v_i - v_j)^2 = (v_i - v_j) \cdot \left(M_* \sum_{k=1}^N \Phi_{ik}(v_k - v_i) - M_* \sum_{k=1}^N \Phi_{jk}(v_k - v_j) \right). \tag{7.3.1}$$

Observe that Φ_{ii} does not play any role in (7.3.1), since it is multiplied by $v_i - v_i = 0$. Thus, denoting $M^* := \sup M$, we redefine it for all $i \in \{1, ..., N\}$ as

$$\Phi_{ii} := \frac{NM^*}{M_*} - \sum_{k \neq i} \Phi_{ik} > 0,$$

so that

$$M_* \sum_{k=1}^N \Phi_{ik} = M^* N \quad \text{for each} \quad i \in \{1, ..., N\}.$$

Then, we rewrite (7.3.1) as

$$\begin{aligned} \frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}(v_i - v_j)^2 &= (v_i - v_j) \cdot M^* N(v_j - v_i) + (v_i - v_j) \cdot M_* \sum_{k=1}^N (\Phi_{ik} v_k - \Phi_{jk} v_k) \\ &= -M^* N V^2 + N(M^* - M_* \eta)(v_i - v_j) \cdot \sum_{k=1}^N (\alpha_{ik} v_k - \alpha_{jk} v_k), \end{aligned} \tag{7.3.2}$$

where

$$\alpha_{ik} := \frac{M_*(\Phi_{ik} - \eta)}{N(M^* - M_*\eta)}, \qquad \eta := \frac{1}{2} \min_{k,l \in \{1,...,N\}} \Phi_{kl} \in [0, M^*/2M_*],$$

$$\alpha_{ik} \geq 0, \qquad \sum_{k=1}^N \alpha_{ik} = 1.$$

Therefore the right-most term in (7.3.2) is a convex combination of elements in $\mathrm{conv}\{v_k : k \in \{1, ..., N\}\} + \mathrm{conv}\{-v_k : k \in \{1, ..., N\}\}$ and thus – it can be bounded from the above by the maximal diameter of such a set, which is V. Consequently, we have

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} V^2 \leq -M_* N \eta V^2.$$

In the worst case scenario $\eta = 0$, so we deduce that V does not increase and the proof is finished. □

Remark 7.10. The L^∞ method applied in the above proof is rather complicated for what it can accomplish for the DI model. However, we chose it to showcase why there is no unconditional exponential flocking, which is precisely because the minimal decay rate η can be 0. On the other hand if we suppose that all particles are at most at distance δ from one another and the number of particles is greater than m, $\Phi_{ik} = \frac{M_i}{M_*}$ for all $i, k \in \{1, ..., N\}$, then in the above proof $\eta = 1/2$

and V decays exponentially with exponent $\lambda = M_* N/2$. This is however of no surprise since then equation (7.1.1b) is reduced to

$$\dot{v}_i = \sum_{k=1}^{N} M_i(v_k - v_i),$$

which is a simple linear model of consensus. The difficulty lies in the propagation of the interaction between the particles if only some of the Φ_{ik} are positive. It is addressed in Theorem 7.6 and particularly in Lemma 7.12, presented below.

7.3.2 Proof of Theorem 7.6

To prove Theorem 7.6 we utilize the 4-step strategy described in Section 7.2. First we show that any r-densely packed cluster with $r \leq \delta$ induces a strongly connected subgraph of G. That is to say, the interaction propagates throughout any r-densely packed cluster of the particles.

Lemma 7.11. *If $\mathscr{A}$ is an r-densely packed cluster at $t \geq 0$ for $r \leq \delta$ then for all $i, j \in \mathscr{A}$ we have*

(i) $i \in \mathscr{N}_j$ iff $j \in \mathscr{N}_i$,

(ii) there exists a sequence $\{l_k\}_{k=1}^{p}$ such that

$$l_1 = i, \quad l_p = j \quad \textit{and} \quad l_k \in \mathscr{N}_{k+1} \quad \textit{for all} \quad k \in \{1, ..., p-1\}.$$

In particular $\mathscr{A}$ treated as a subgraph of $G(t)$ is strongly connected.

Proof. Fix $t \geq 0$. Since, for $r < \delta$ an r-densely packed cluster is also δ-densely packed, we will prove the lemma assuming that $r = \delta$.

First observe that condition 2 in Definition 7.4 implies that for all pairs $i, j \in \mathscr{A}$ we have $i \in \mathscr{N}_j$ if and only if $|x_i(t-h) - x_j(t-h)| < \delta$. Thus, assertion (i) follows and it remains to show that $G(t)$ is connected (then it is automatically also strongly connected by (i)). Now suppose that the lemma is not true and there exists a δ-densely packed cluster $\mathscr{A}$ that is not connected. Then there exist two particles with indexes i and j without a connecting sequence $\{l_k\}_{k=1}^{p}$. Let $\mathscr{A}_i$ and $\mathscr{A}_j$ be the maximal connected clusters including i and j, respectively. Then $\mathrm{dist}(\mathscr{A}_i, \mathscr{A}_j) > \delta$, since otherwise we could find $q_i \in \mathscr{A}_i$ and $q_j \in \mathscr{A}_j$ such that $|x_{q_i}(t-h) - x_{q_j}(t-h)| \leq \delta$ and $\mathscr{A}_i \cup \mathscr{A}_j$ would be connected through the pair (q_i, q_j). However with $\mathrm{dist}(\mathscr{A}_i, \mathscr{A}_j) > \delta$ condition 1 in Definition 7.4 is not satisfied. Contradiction with the assumption that $\mathscr{A}$ is not strongly connected finishes the proof. □

The next lemma uses the propagation of interaction within densely packed clusters, ensured by Lemma 7.11, to provide a local flocking estimate.

Lemma 7.12. *Let $\mathscr{A}$ be a cluster that is strongly connected in the interval $[t_1, t_2)$ with constant neighbor sets $\mathscr{N}_i$ for all $i \in \mathscr{A}$. Assume further that no particle outside of $\mathscr{A}$ influences $\mathscr{A}$. Then there exist $\lambda > 0$ and a steady state $v_{\mathscr{A}} \in \mathrm{conv}\{v_i(t_1) : i \in \mathscr{A}\}$, such that*

$$|v_i(t) - v_{\mathscr{A}}| \leq e^{-tM_*\lambda}|v_i(t_1) - v_{\mathscr{A}}| \qquad \textit{for all} \quad i \in \mathscr{A} \quad \textit{and all} \quad t \in [t_1, t_2),$$

where M_* is the minimal value of the function M, c.f. (7.2.6).

Proof. If $\mathscr{A}$ is a strongly connected graph with no outside influence then it can be treated as a separate ensemble of particles following equation (7.2.9). To finish the proof we apply Proposition 7.3 from the appendix, with $\mathbf{A} = M_*(\mathbf{D} - \mathbf{\Phi})$. Observe that within the assumptions of Lemma 7.12 the interaction matrix is constant.

Finally we prove the following slightly reformulated version of Theorem 7.6.

Proposition 7.13. *Fix* $(\mathbf{x}_0, \boldsymbol{v}_0)$. *Let the initial ensemble of the particles* $\{1, ..., N\}$ *associated with positions* $\mathbf{x}_0$ *and velocities* $\boldsymbol{v}_0$ *be* r*-densely packed with* $r < \delta$. *Then the ensemble remains densely packed for all* $t > 0$ *and the particles flock exponentially fast, provided that*

$$M_* > \frac{2V(0)}{\lambda(\delta - r)}.$$

Proof. (Proposition 7.13 and Theorem 7.6.) Since, initially, the particles are r-densely packed for $r < \delta$ and the maximal velocity of the particles is uniformly bounded, see Theorem 7.5. Thus, there exists a time interval $[0, T]$ such that the particles are δ-densely packed for $t \in [0, T]$. Moreover, the right-hand side of (7.4.3) is piecewise constant in time. Then by Lemma 7.11 and Lemma 7.12, for $t \in [0, T]$, the ensemble is a strongly connected cluster and each velocity v_i converges to a (piecewise constant) steady state $v_{st}(t)$ exponentially fast with exponent $M_*\lambda$. Let

$$T^* := \sup\{T > 0 : \{1, ..., N\} \text{ is } \delta\text{-densely packed for all } t \leq T\} > 0.$$

We shall prove that if M_* is large enough then actually $T^* = \infty$. Assuming the contrary, for all $i, j \in \{1, ..., N\}$ and all $t \in [0, T^*]$, we have

$$\begin{aligned} x_i(t) - x_j(t) &= \int_0^t v_i(s) - v_j(s)\mathrm{d}s + x_i(0) - x_j(0) \\ &= \int_0^t (v_i(s) - v_{st}(s))\mathrm{d}s + \int_0^t (v_{st}(s) - v_j(s))\mathrm{d}s + (x_i(0) - x_j(0)). \end{aligned}$$

which leads to

$$|x_i(t) - x_j(t)| \leq \int_0^t |v_i(s) - v_{st}(s)|\mathrm{d}s + \int_0^t |v_j(s) - v_{st}(s)|\mathrm{d}s + |x_i(0) - x_j(0)|. \tag{7.3.3}$$

Let us fix in (7.3.3) any pair (i, j) such that $|x_i(0) - x_j(0)| \leq r$. Such pairs exist since the ensemble is r-densely packed. On $[0, T^*)$ the ensemble is densely packed and thus, by Lemmas 7.11 and 7.12 we have

$$|x_i(t) - x_j(t)| \leq 2V(0) \int_0^\infty e^{-M_*\lambda t}\mathrm{d}t = \frac{2V(0)}{M_*\lambda} + r. \tag{7.3.4}$$

Inequality (7.3.4) follows from Lemma 7.12, since $v_{st}(t) \in \operatorname{conv}\{v_i(t) : i = 1, ..., N\}$ and thus $|v_{st}(t) - v_i(t)| \leq V(t) \leq V(0)$ by Theorem 7.5. Therefore if

$$M_* > \frac{2V(0)}{\lambda(\delta - r)},$$

then $|x_i(t)-x_j(t)| < r_* \leq \delta$ for all $t \in [0,T^*]$. This implies that any two particles of distance not greater than r initially are of distance at most r_* from one another. It is therefore easy to see that for all $t \in [0,T^*]$ the r_*-neighborhood of the ensemble is connected and the number of particles in each $B(x_i,r_*)$ is larger than m. Thus the ensemble is r_*-densely packed in $[0,T^*]$. In particular, the ensemble is r_*-densely packed at T^* and, again by boundedness of the velocity, we can find $\varepsilon > 0$ such that the ensemble is δ-densely packed in $[0,T^*+\varepsilon)$, which contradicts the assumption that $T^* < +\infty$. Consequently $T^* = \infty$ and the ensemble is δ-densely packed, strongly and symmetrically connected and it flocks exponentially fast to the the steady state v_{st}. The proof is finished. □

7.4 Clusters' Interaction

In what follows we perform the analysis of clusters' interactions as introduced at the end of Section 7.2.2. For this sake we consider a simple yet significant setting of cluster interactions, where one cluster ($\mathscr{C}$) is a single particle (c). In such a case the three scenarios read:

1. Stability: Cluster $\mathscr{A}$ remains connected. Particle c detaches from $\mathscr{A}$.
2. Breaking: A single particle b from $\mathscr{A}$ detaches from $\mathscr{A}$ under the influence of c.
3. Sticking: Cluster $\mathscr{A}$ changes its momentum under the influence of c.

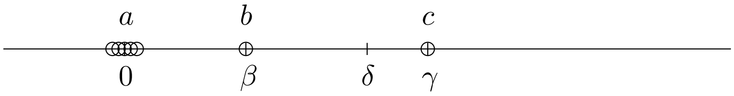


Figure 7.3: Scheme of the considered Cluster setting.

The setting.

Suppose that we have $N+1 > m+1$ particles in $\mathbb{R}^2$ distributed as in Fig. 7.3. All of the particles are initially situated on the horizontal x_1-axis. The left-most N particles form the cluster $\mathscr{A}$. Within this cluster the leftmost $N-1$ particles are initially in the position near zero with zero velocities. The distances between them are significantly smaller then δ. We shall refer to all of these particles as a denoting its approximate position as $x_a(0) = 0$. The rightmost particle (b) in cluster $\mathscr{A}$ is in the position $x_b(0) = (\beta, 0)$ with zero initial velocity and $\beta < \delta$. The last particle under considerations, the singleton (c) is positioned at $x_c(0) = (\gamma, 0)$ with $\gamma > \delta$ and $\gamma - \beta < \delta$; the singleton's initial velocity is $(0, v_c)$. Since we assume that $0 < h \ll \beta, \delta, \gamma, r, 1/N$, for simplicity of the presentation we set $h = 0$.

Note that not only is cluster $\mathscr{A}$ r-densely packed with $r \leq \delta$ but actually each particle in $\mathscr{A}$ interacts with each other particle. Moreover, the only non-zero initial velocity of the entire ensemble is $(0, v_c)$, which implies that any change in the velocity of any one of the particles can only occur in direction parallel to $(0, v_c)$.

To summarize, we have the following picture:

$$\begin{aligned}
&x_a(0) \approx (0,0), \quad v_a(0) = (0,0), \quad \text{it describes the set of } N-1 \text{ particles,}\\
&x_b(0) = (\beta,0), \quad v_b(0) = (0,0), \quad 0 < \beta < \delta, \\
&x_c(0) = (\gamma,0), \quad v_c(0) = (0,c), \quad \gamma > \delta \;\text{ and }\; \gamma - \beta < \delta,
\end{aligned} \tag{7.4.1}$$

with x_a serving as a stand-in for all x_i with i in the left-most cluster a. Then, for $t > 0$ we let the particles follow the protocol (7.1.1). In what follows we successively reformulate and simplify (7.1.1) in the configuration (7.4.1) leading to a formulation that can be expressed in terms of v_c. First note that configuration (7.4.1) leads to a characterization of the neighbor sets:

$$\mathscr{N}_a(0) = A = a \cup \{b\}, \quad \mathscr{N}_b(0) = A \cup \{c\}, \quad \mathscr{N}_c = \emptyset,$$

which implies that (7.1.1) takes the form:

$$\begin{aligned} \dot{v}_a &= M(v_b - v_a), \\ \dot{v}_b &= M(N-1)(v_a - v_b) + M(v_c - v_b), \\ \dot{v}_c &= 0. \end{aligned} \tag{7.4.2}$$

Since the motion can occur only in directions parallel to $(0, v_c)$, we assume, without a loss of generality, that the velocities appearing in (7.4.2) are scalar. Moreover, through a simple scaling argument (e.g. $\frac{\mathrm{d}}{\mathrm{d}t}\bar{v}_a(t) = \frac{\mathrm{d}}{\mathrm{d}t}v_a(Mt)$ etc.) we may further simplify assuming that $M = 1$. Then (7.4.2) reduces to a scalar ODE

$$\begin{aligned} \dot{v}_a &= -(v_a - v_b), \\ \dot{v}_b &= -(N-1)(v_b - v_a) - (v_b - v_c), \end{aligned} \tag{7.4.3}$$

with initial data $v_a(0) = 0, v_b(0) = 0$ and a constant v_c. Then

$$\frac{\mathrm{d}}{\mathrm{d}t}(v_b - v_a) = -N(v_b - v_a) - (v_b - v_c),$$

which, with constant v_c, implies

$$(v_b - v_a)(t) = -e^{-Nt} \int_0^t e^{Ns}(v_b(s) - v_c)\mathrm{d}s. \tag{7.4.4}$$

Inserting (7.4.4) to (7.4.3) yields

$$\dot{v}_b = -(N-1)\left(-e^{Nt} \int_0^t e^{Ns}(v_b(s) - v_c)\mathrm{d}s\right) - (v_b - v_c).$$

Thus v_b satisfies the second order ODE

$$\ddot{v}_b + (N+1)\dot{v}_b + v_b - v_c = 0$$

with initial data $v_b(0) = 0$ and $\dot{v}_b(0) = v_c$. Since v_c is a constant, introducing

$$V_b = v_b - v_c \text{ with } V_b(0) = -v_c \text{ and } \dot{V}_b(0) = v_c,$$

we obtain

$$\ddot{V}_b + (N+1)\dot{V}_b + V_b = 0.$$

By solving the above ODE, we find that

$$V_b(t) = V_1 e^{-\lambda_1 t} + V_2 e^{-\lambda_2 t},$$

where

$$\lambda_1 = \frac{-(N+1) - \sqrt{(N+1)^2 - 4}}{2} \approx -(N+1),$$
$$\lambda_2 = \frac{-(N+1) + \sqrt{(N+1)^2 - 4}}{2} \approx -(N+1)^{-1}.$$

Here $\approx$ is the *asymptotic equality* as $N \to \infty$.

We recover V_1 and V_2 from the initial data by taking $V_1 + V_2 = -v_c$ and $-\lambda_1 V_1 - \lambda_2 V_2 = v_c$. Therefore

$$V_1 = v_c \frac{1-\lambda_2}{\lambda_2 - \lambda_1} \approx v_c \frac{1}{N} \quad \text{and} \quad V_2 = v_c \frac{\lambda_1 - 1}{\lambda_2 - \lambda_1} \approx -v_c(1 + \frac{1}{N}).$$

Hence we obtain the explicit formula on approximate $V_b(t)$:

$$\boxed{V_b(t) \approx v_c \frac{1}{N} e^{-(N+1)t} - v_c \left(1 + \frac{1}{N}\right) e^{-(N+1)^{-1}t} \approx -v_c \left(1 + \frac{1}{N}\right) e^{-(N+1)^{-1}t}.} \tag{7.4.5}$$

The three scenarios.

Our next goal is to obtain three distinct aforementioned scenarios within the configuration expressed by (7.4.5). First we aim to distinguish scenarios No. 1 and No. 2 (the *stability* and *breaking* scenarios), with particle c either failing to drastically change the structure of the cluster $\mathscr{A}$ or scooping particle b out of the cluster $\mathscr{A}$. Assume that N is large. By (7.4.5), For large N, (7.4.5) gives

$$v_b - v_c = V_b \approx -v_c e^{-(N+1)^{-1}t}, \tag{7.4.6}$$

thus for any $T > 0$ we have

$$\begin{aligned}|x_b(T) - x_c(T)| =& |\gamma - \beta| + \left| \int_0^T (v_b(t) - v_c) \mathrm{d}t \right| \\ &\approx |\gamma - \beta| + \left| \int_0^T -v_c e^{-(N+1)^{-1}s} \mathrm{d}s \right| \approx |\gamma - \beta| + |v_c T|\end{aligned}$$

and

$$|x_a(T) - x_b(T)| \approx |\beta| + \left| v_c \int_0^T e^{-Nt} \int_0^t e^{Ns} e^{-(N+1)^{-1}s} \mathrm{d}s \right| \approx |\beta| + \left| \frac{v_c T}{N} \right|.$$

The above equations imply that if $t = T$ is the breaking point for $c \in \mathscr{N}_b(t)$ i.e.

$$T = \sup \Big\{ t : c \in \mathscr{N}_b(s) \quad \text{for all} \quad s \in [0, t) \Big\}$$

or the breaking point for $b \in \mathscr{N}_a(t)$ it needs to satisfy

$$|\gamma - \beta| + |v_c T| = \delta \;\text{ or }\; |\beta| + \left| \frac{v_c T}{N} \right| = \delta, \tag{7.4.7}$$

respectively. Then scenarios No. 1 and No. 2 are distinguished as follows:

1. With $\beta = \delta/2$ and $\gamma = \delta$ so that $\gamma - \beta = \delta/2$ in (7.4.7) we have

$$\delta = |\gamma - \beta| + |v_c T| = \frac{\delta}{2} + |v_c T| > \frac{\delta}{2} + \left|\frac{v_c T}{N}\right| = |\beta| + \left|\frac{v_c T}{N}\right|.$$

The above implies that at T the c particle breaks out from $\mathscr{N}_b$ while cluster $\mathscr{A}$ itself remains unaffected (in particular $b \in \mathscr{N}_a$ and $a \in \mathscr{N}_b$), provided that N is large. Thus we obtain the *stability* scenario No. 1.

2. With $\beta = \delta - \varepsilon$ and $\gamma = \delta$ so that $\gamma - \beta = \varepsilon$ with $0 < \varepsilon \ll \delta$ we have

$$\delta - \varepsilon + \left|\frac{v_c T}{N}\right| = |\beta| + \left|\frac{v_c T}{N}\right| = \delta,$$

if and only if

$$\varepsilon = \left|\frac{v_c T}{N}\right|. \tag{7.4.8}$$

On the other hand, taking the above into the account yields

$$|\beta| + \left|\frac{v_c T}{N}\right| = \delta - \varepsilon + \varepsilon > \varepsilon + |v_c T| = |\gamma - \beta| + |v_c T|$$

which holds if $|v_c|$ satisfies

$$T|v_c|\left(\frac{1}{N} + 1\right) < \delta. \tag{7.4.9}$$

If the above conditions are satisfied, the breaking point for $b \in \mathscr{N}_a$ comes sooner than the breaking point for $c \in \mathscr{N}_b$ and thus the b particle is removed from the cluster $\mathscr{A}$, leading to the *breaking* No. 2 scenario.

3. The last, *sticking* No. 3 scenario requires more preparation. For a fixed N and small v_c, we recall (7.4.6) to see that

$$V_b(t) = v_b(t) - v_c \approx -v_c e^{-\frac{1}{N}t}$$

And recalling from (7.4.4) that

$$v_b(t) - v_a(t) = -e^{-Nt}\int_0^t e^{Ns} V_b(s)ds \approx v_c \frac{1}{N} e^{-\frac{1}{N}t}.$$

Integration of the above two equations leads to the uniform-in-time upper-bounds on the relative distance between the sub-clusters

$$\sup_{t\geq 0} |x_c(t) - x_b(t)| \leq |\gamma - \beta| + \left|\int_0^\infty V_b(t)\mathrm{d}t\right| \approx |\gamma - \beta| + |v_c N|$$

and

$$\sup_{t\geq 0} |x_a(t) - x_b(t)| \leq |\beta| + \left|\int_0^t (v_b - v_a)(s)\mathrm{d}s\right| \approx |\beta| + |v_c|.$$

Hence, preservation of the cluster's structure is ensured by taking

$$|\gamma-\beta|+|v_cN|\le\delta \quad\text{and}\quad |\beta|+|v_c|\le\delta, \tag{7.4.10}$$

which holds for small v_c. Such conditions ensure that the distance between the leftmost sub-cluster a and the middle particle b remains smaller than δ indefinitely; in other words cluster $\mathscr{A}$ is preserved for all $t>0$. Furthermore the distance between the middle particle b and the rightmost singleton c is also smaller than δ for all times. Hence $\mathscr{A}\cup\{c\}$ is a connected cluster. The connectivity of $\mathscr{A}\cup\{c\}$ is weak in the sense that while $\mathscr{A}$ is strongly and symmetrically connected (it is in fact δ-densely packed) and c influences $\mathscr{A}$ through the middle particle b, there is no influence directed from $\mathscr{A}$ to c ($\mathscr{N}_c=\emptyset$).

Remark 7.14. Let us describe how the total momentum (equal to the sum of velocities, since $M=1$) of the $N+1$ particles evolves in each case No. 1 - No. 3. By the above considerations one has:

$$v_b\approx v_c(1-e^{-T/N}) \text{ and } v_a\approx v_c(1-e^{T/N}-\frac{1}{N}e^{-T/N}) \tag{7.4.11}$$

In the first scenario the time of separation is of order $T\approx\frac{\delta}{v_c}$. Thus the momentum of the cluster increases (here $N+1\approx N$):

$$Mom:=v_b+Nv_a=v_c(1-e^{-T/N})+Nv_c(1-e^{-T/N}-\frac{1}{N}e^{-T/N})\approx \\ v_c\frac{T}{N}+Nv_c\frac{T}{N}+v_ce^{-T/N}\approx\delta+v_ce^{-\delta/(Nv_c)} \tag{7.4.12}$$

In the second case by (7.4.8), (7.4.9) $T\approx\frac{\varepsilon N}{v_c}$, but $N\varepsilon<\delta$ and we get almost the same ($T\approx\frac{\delta}{v_c}$)

$$Mom:=\delta+v_ce^{-\delta/(Nv_c)}. \tag{7.4.13}$$

In the last case, the time of separation T is infinity, all velocities are reaching v_c, so by condition (7.4.10) $Mom:=Nv_c\approx\delta$.

Interestingly enough, in each of the scenarios No. 1 - No. 3 the growth of the momentum is almost the same.

□

7.5 Simulations

In this section we present the results of numerical simulation of the DI model in several different cases. The aim is twofold. First we compare the model with related models. Second we illustrate its unique features, and provide insight to the theoretical results of this work.

Non-dimensional scaling Starting with the non-dimensional scaling of (7.1.1)

$$\begin{cases} \dfrac{\bar{L}}{\bar{T}\bar{V}}\dot{x}_i=v_i, & (7.5.1a)\\ \dfrac{1}{\bar{T}\bar{M}}\dot{v}_i=\dfrac{M}{N}\displaystyle\sum_{k\in\mathscr{N}_i}(v_k-v_i), & (7.5.1b)\end{cases}$$

where $\bar{L}$, $\bar{T}$, $\bar{V}$ are characteristic values for the length, the time, the velocity, respectively. The characteristic M is inverse proportional to the characteristic time

$$\bar{M} = \frac{1}{\bar{T}} = \frac{\bar{V}}{\bar{L}}.$$

$$\frac{1}{\bar{T}}\frac{\partial}{\partial t^*} = \frac{\partial}{\partial t}$$

$$\frac{\bar{L}}{\bar{T}}\frac{\partial}{\partial t^*} x_k = \bar{V} v_k$$

Let us recall that the DI model is characterized by the set of parameters (N, M, m, δ). Together with the volume of the domain, which in the case of two-dimensional square equals L^2, we can transform N, m, δ and L to the average particle density ϱ_a and the local minimal interaction density ϱ_m, namely

$$\varrho_a = \frac{N}{L^2}, \quad \varrho_m = \frac{m}{\pi\delta^2}.$$

For the fixed $L = 25$ in the case of cluster formations (Sect. 7.5.1) we choose $(N, m, \delta) = (64, 3, 2)$. In this example $\varrho_m = 2.3\varrho_a$, thus the necessary condition for a particle interaction is to locally reach particle density that is more than two times of the average, however in practice since the values are discrete its in fact three times more.

Computational domain is a two-dimensional square with the periodic boundary condition. Periodicity is implemented by introducing domain extension of the size δ and ghost particle technique.

For the sake of visualisation we construct a directed graph that represents interactions $G = (V, \mathscr{E}, \boldsymbol{\Phi})$, see Section 7.2. In each time-step we identify clusters as a strongly connected components of graph G, and prescribe them a color from predefined list of colors. To this end we apply an algorithm *connected_components* implemented in *scipy.sparse.csgraph* [139].

Numerical Scheme

As a numerical method we choose the classic 4th-order Runge-Kutta method (RK4), that is one of the most widely used method to solve ODEs. The application of RK4 algorithm to system (7.1.1), where we renamed the rhs in (7.1.1b) as a, yield the following numerical scheme

$$\begin{cases} v_i^{n+1} = v_i^n + \dfrac{1}{6}\left(k_{v1} + 2k_{v2} + 2k_{v3} + k_{v4}\right), & (7.5.2a) \\ x_i^{n+1} = x_i^n + \dfrac{1}{6}\left(k_{x1} + 2k_{x2} + 2k_{x3} + k_{x4}\right), & (7.5.2b) \end{cases}$$

where

$$\begin{aligned} k_{v1} &= a(x_i^n, v_i^n)\,\mathrm{d}t, & k_{x1} &= v_i^n\,\mathrm{d}t, \\ k_{v2} &= a(x_i^n + k_{x1}/2, v_i^n + k_{v1}/2)\,\mathrm{d}t, & k_{x2} &= \left(v_i^n + k_{v1}/2\right)\,\mathrm{d}t, \\ k_{v3} &= a(x_i^n + k_{x2}/2, v_i^n + k_{v2}/2)\,\mathrm{d}t, & k_{x3} &= \left(v_i^n + k_{v2}/2\right)\,\mathrm{d}t, \\ k_{v4} &= a(x_i^n + k_{x3}, v_i^n + k_{v3})\,\mathrm{d}t, & k_{x4} &= \left(v_i^n + k_{v3}\right)\,\mathrm{d}t. \end{aligned}$$

7.5.1 Cluster Formations

In the first scenario we show spontaneous behaviour of particles in time. Starting from N particles that are randomly distributed and initially placed with a certain distance from the domain boundary. We observe particle interactions and clusters formations based on local particle density.

Moreover, we present comparison with the Cucker-Smale (CS) type models. Beyond the classical CS model, two local versions of CS model are considered: CS$_\delta$ where any particle interacts only within the ball of radius δ and CS$_q$ where any particle interacts with $q-$closest neighbours. The weight function for CS models is given by $\psi(s) = 1/\sqrt{1+|s|}$.

The initial velocity is equal $v_i(0) = r_i(\cos(\alpha_i), \sin(\alpha_i)) + f_i, \forall i = \{1..N\}$ with r_i and α_i being random numbers taken from the uniform distributions $U([0,1])$ and $U([0,2\pi])$ respectively. Moreover, in order to avoid the average velocity being close to zero, for half of the particles $i \leq N/2$ we introduce the fixed contribution f_i that is equal to $f_i = r_i(0.5, 1)$.

For the sake of the comparison of different models we put M in such a way that it is always equal to one over the number of particles within a cluster, c.f. Section 7.2.3. Tab. 7.1 summarizes considered models with corresponding set of parameters.

	model	N	M	m	δ	q
(A)	DI	64	$1/\#\mathcal{N}_i$	3	2	
(B)	CS$_\delta$	64	$1/\#\mathcal{N}_i$		2	
(C)	CS$_q$	64	$1/q$		∞	3
(D)	CS	64	$1/N$		∞	

Table 7.1: Summary of models considered in Section 7.5.1.

Fig. 7.4 (movie at https://youtu.be/axBgCEW6lNI)

presents the evolution of particles governed by DI, CS$_\delta$, CS$_q$ and CS models at $t = 0$, $t = 30$ and $t = 150$. We prescribe the same colour for particles that interact with each other at least indirectly. In all of the cases we observe clustering and velocity alignment.

For DI and CS$_\delta$ we have the same range of interaction $\delta = 2$, but DI model on the top of the geometrical condition requires more that 3 particles for the activation of interaction. This two models form densely-packed clusters.

Global interaction models CS$_q$ and CS do not form clusters (or more precisely the entire ensemble is a single cluster). Interestingly, for this particular random initial condition, the relatively small number of closest neighbours $q = 3$ is enough to propagate interactions throughout the entire ensemble.

Moreover, in Fig. 7.5 the maximal relative velocity between the particles $V(t)$, c.f. (7.2.4), is presented, showcasing regions of sharp decay for the DI model.

7.5.2 Sharp Change of the Total Momentum

Following Fig. 7.2 we investigate the phenomenon of a single particle diverting an entire cluster and its dependence on the spatial distribution of particles within the cluster. This scenario has

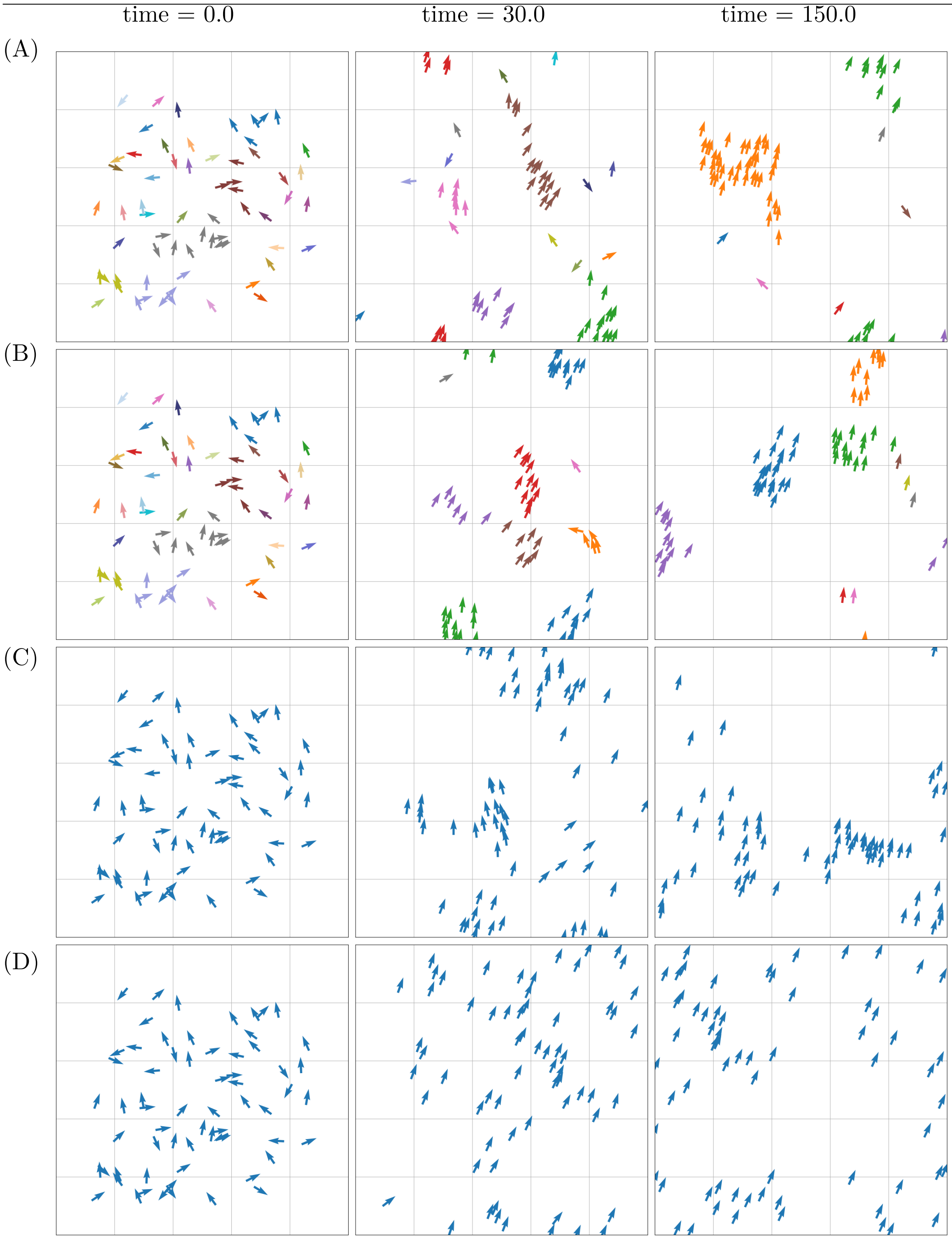


Figure 7.4: Spontaneous behaviour of particles for DI(A), CS$_\delta$(B), CS$_q$(C), and CS(D), see Tab. 7.1. A video available at https://youtu.be/FRIGCH6ziCU.

been already addressed in the introduction. A cluster of 28 particles moves to the right with velocity $(0.1, 0)$, while a single particle moves in the opposite direction with velocity $(2.7, 0)$, facing the cluster. Thus the total momentum (equal to the sum of velocities) of the particles is

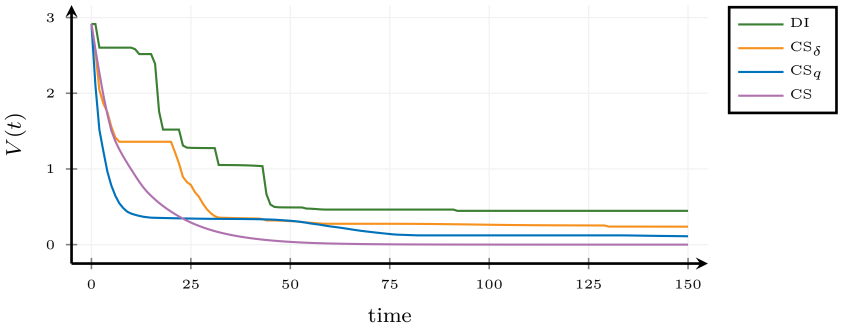


Figure 7.5: The maximal relative velocity between the particles for DI, CS$_\delta$, CS$_q$ and CS models.

0.1.

We compare cases (A) and (B), c.f. Fig. 7.2 and determine whether the total momentum changes its sign for the DI, CS$_\delta$, CS$_q$ and CS models with parameters M, m, δ and q as in Tab. 7.1. The results are showcased in Fig. 7.6. As expected, in the case of the CS model the total momentum is constant in both (A) and (B) cases. The total momentum for the CS$_\delta$ model changes (note that this is due to the non-symmetric normalization of the interaction $M = 1/\#\mathscr{N}_i$) but it remains positive. In the case of the CS$_q$ model the total momentum changes sign in both (A) and (B) cases. Finally, the DI model is the only one exhibiting a sharp distinction between (A) and (B). In (A) the total momentum changes its sign and in (B) – it grows greater than for any other considered model.

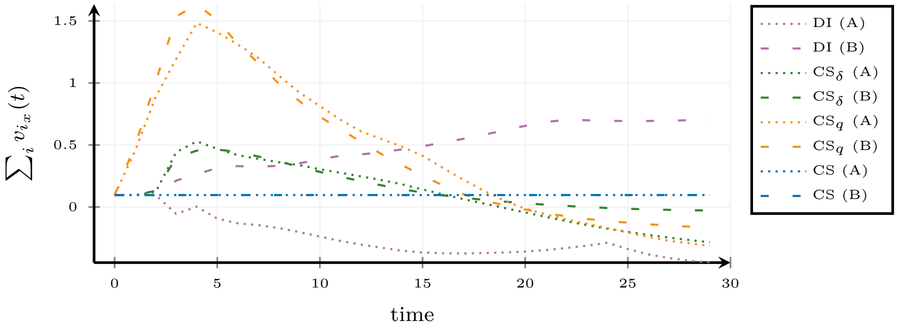


Figure 7.6: Sum of horizontal velocities of particles for DI, CS$_\delta$, CS$_q$ and CS models.

7.5.3 Chain

In this scenario we present an interaction between a chain of particles and an individual. The chain consists of 21 particles that are equally distributed in the vertical direction. The initial velocity of the chain is small in comparison with the initial velocity of single particle and they are directed facing each other.

The initial velocity equals $v_0 = (0.1, 0)$ and $v_0 = (-8, 0)$ for the chain and the individual, respectively. The DI model parameters $m = 3$ and $M = 1$, and we consider three scenarios that differs with parameter δ that equals 2, 3, 4, cf. (A), (B), and (C) in Fig. 7.7 respectively. Since δ describes the interaction range we observe three different behaviours. In the case $\delta = 2$, the cluster splits, since the single particle is causing activation of interaction, and due the velocity difference it takes cluster with itself. For the second scenario, $\delta = 3$, the chain deforms, but is able to turn back the singe agent. In the third example, with $\delta = 4$, the chain is very stiff, and the single particle pushes the entire cluster.

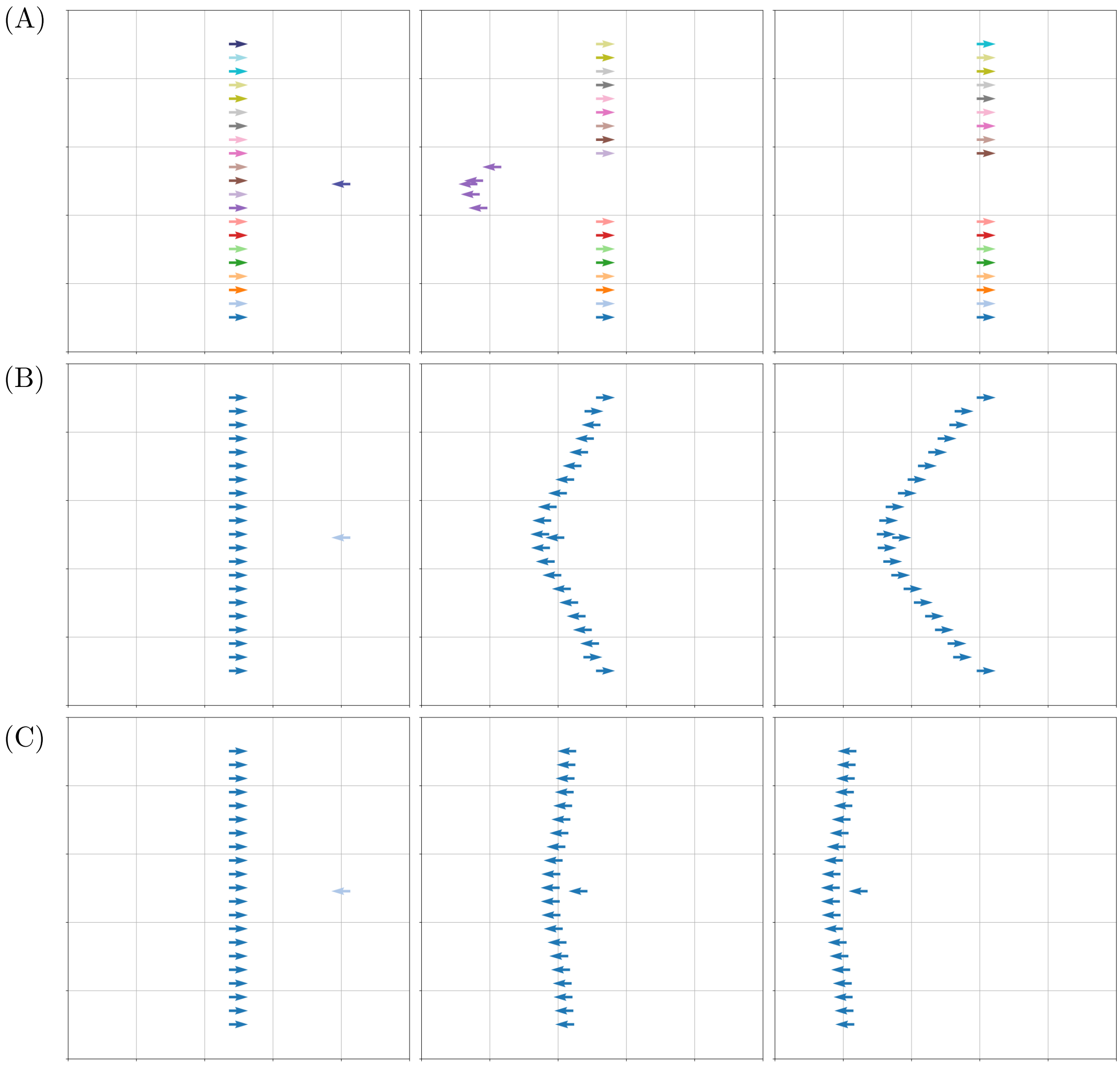


Figure 7.7: Interaction between chain and individual particle for DI$(8, \delta, 1, 3)$ model for $\delta = 2$ (A), $\delta = 3$ (B) and $\delta = 4$ (C). A video available at `https://youtu.be/OdT2-g1Omvs`.

Supplement Materials. Source code to reproduce simulations is available at [164].

CHAPTER 8

Data Clustering as an Emergent Consensus of Autonomous Agents

The content of this chapter is joint work with Jan Peszek and is based on the preprint

- P. Minakowski and J. Peszek. ‘Data Clustering as an Emergent Consensus of Autonomous Agents’. In: (2022). doi: 10.48550/ARXIV.2204.10585

Chapter Summary. We present a novel data segmentation method based on a first-order density-induced consensus protocol. We provide a mathematically rigorous analysis of the consensus model leading to the stopping criteria of the data segmentation algorithm. To illustrate our method, the algorithm is applied on selected images from Berkeley Segmentation Dataset. The method can be seen as a clustering technique for multimodal feature space with lower numerical complexity than other popular density-based algorithms.

Chapter Organisation. In Section 8.2 we compare the DI protocol to previous research on consensus dynamics and density based clustering, focusing on unsupervised color image segmentation. In Section 8.3 we provide an analysis of emergent behavior of the DI protocol based on the interplay between density of the agents and connectivity. Sections 8.4 and 8.5 are dedicated to the presentation of the main segmentation algorithm and its applications to color image segmentation, respectively. Here, we also apply the results of Section 8.3 to derive the stopping criterion. Finally, in Sections 8.6 we provide mathematical proofs related to the results of Section 8.3.

Contents of Chapter

8.1 Introduction

The paper explores the relation between first-order collective dynamics and data clustering, focusing on color image segmentation in image processing. Our main contributions include:

- A novel technique of unsupervised density-based clustering for multimodal feature space.
- The technique inherits the advantages of other density-based clustering methods, with lower average numerical complexity and manageable parameters.
- Applications to color image segmentation.
- Mathematically rigorous analysis of the model leading to the stopping criteria for the algorithm.
- Exploration of connection between data clustering and first-order collective dynamics.

Our approach is based on the recently introduced density-induced consensus protocol (DI protocol) for agent-based collective dynamics [165]. Consider N agents with $x_i(t) \in \mathbb{R}^d$ denoting the position of ith agent in a d-dimensional space at the time $t \geq 0$. The agents follow the DI protocol

$$\dot{x}_i = \kappa \sum_{k \in \mathscr{N}_i} (x_k - x_i), \quad x_i(0) = x_{i0} \in \mathbb{R}^d. \tag{8.1.1}$$

Here $\kappa > 0$ is a fixed coupling strength whose influence amounts, in practice, to time-scaling. The neighbor set $\mathscr{N}_i$ of ith agent is defined through the following relation: given positive parameters

δ and m, for $t \geq 0$ we define,

$$\begin{aligned} k \in \mathcal{N}_i(t) \;\; &\Leftrightarrow \;\; x_k(t) \in B(x_i(t), \delta) \text{ and} \\ \#\Big\{ k \in \{1, ..., N\} &: x_k(t) \in B(x_i(t), \delta) \Big\} > m, \end{aligned} \tag{8.1.2}$$

where $B(x_i(t), \delta)$ is an open ball centered at $x_i(t)$ with radius δ and $\#A$ denotes the cardinal number of A. Thus the communication rule of the DI protocol reads as follows: the ith agent is influenced by jth agent if

- the density of agents in close proximity to the ith agent is substantial enough (otherwise the ith agent is an outlier),
- the jth agent is in close proximity to the ith agent.

Interpreting the positions x_i as data points in a multimodal feature space and evolving the data in time using (8.1.1) results in a data clustering method based on a density-induced protocol (hereinafter referred to as DIPCLUST). To explain this idea, here we fix our attention to 2D data represented as vectors/positions in a Euclidean space. The positions $x_i(t) \in \mathbb{R}^2$ for $i \in \{1, ..., N\}$ evolve according to (8.1.1) eventually leading to a steady state at the time $t = \infty$, typically forming multiple clusters. Finally, the $t = \infty$ clustering is retroactively applied to the initial $t = 0$ state, to establish the segmentation, see Fig. 8.1.

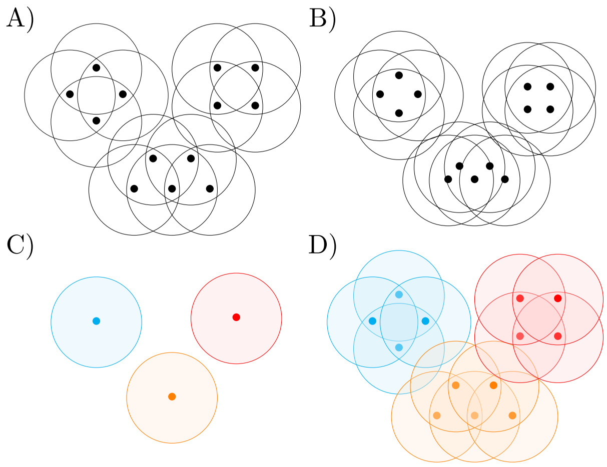


Figure 8.1: An ensemble of 13 agents with visualisation of their range of interactions. A) initial positions of the agents; B) positions evolve towards the steady-state; C) steady state at $t = \infty$ defines the clusters (color coded); D) clustering applied to the initial state.

It is noteworthy that the communication rule of (8.1.1) is equivalent to the density-based spatial clustering algorithm (DBSCAN) [93] widely used in data segmentation [216]. Our method inherits the advantages of the classical density-based clustering algorithms i.e.: no requirement of initial specification of the number of clusters, arbitrary shaped clusters, robustness to outliers and adaptability to various types of data due to the possibility of fine-tuning of the parameters. The main difference is that, we do not need database-oriented range-queries, that lead to computational complexity of DBSCAN. Instead, our procedure is based on local direct connectivity defined by the neighbor sets $\mathcal{N}_i$, and thus it is less computationally demanding.

On the other hand, the DI protocol applies to consensus dynamics as discussed in [165]. In particular it is reminiscent of the bounded-confidence consensus model by Hegselmann and Krause [122]. Similarly, the agents are influenced only by like-minded (i.e. nearby) individuals but in the case of the DI model the agents additionally exhibit a tendency to conform with group-mentality and ignore the outliers. Since our method originates in consensus dynamics it can be particularly useful in segmentation of data related to collective dynamics such as the classification of pedestrians or the users of social networks.

Throughout the chapter we will sometimes interchangeably refer to agents following the DI protocol as nodes, data or pixels depending on the context of applications.

8.2 Previous Research

8.2.1 Consensus Dynamics

Protocol (8.1.1-8.1.2) is a first-order variant of the DI model introduced in [165] which simulates density based alignment in the spirit of the Cucker-Smale model ([68], see also surveys [58] and [167]). First order models similar to (8.1.1) with a variety of interaction laws have been extensively studied from the perspective of both mathematical theory and practical applications, with prominence of applications in opinion dynamics. Opinion dynamics dates back to French's research on social influence [106], and further, to works by De Groot [81] and Lehrer [147]. More modern approach, including nonlinear models, was established by Krause [144] and Hegselmann jointly with Flache [105], among others. We emphasise the aforementioned work by Hegselmann and Krause [122] introducing a bounded confidence opinion dynamics model. We also recommend surveys [6] and [13] for more up-to-date information, and [4] where similar models are presented in applications to vehicular traffic and crowd dynamics.

8.2.2 Density Based Clustering

Since most of the clustering techniques [2] are application dependent, they are not tailored for the classification of an arbitrary feature space. Methods which rely explicitly on the number of clusters or implicitly assume the same shape for all the clusters, are not well suited to analyse data of unknown origin, see [133] for a survey.

Density based clustering is a class of unsupervised learning methods. Points that are in high-density regions in a data space are clustered together and separated from other such clusters by regions of low point density [128, 145].

As already mentioned in the introduction, we group points in a similar way as DBSCAN [93] does. The OPTICS [11] algorithm is an extension of DBSCAN and removes its dependence on parameters. Another related iterative technique is Mean-shift algorithm [56], where each object is assigned to the densest area in its vicinity, based on kernel density estimation.

Unsupervised image segmentation is an important component in many image processing systems. A nonparametric technique for the analysis of a complex multimodal feature space based on mean-shift algorithm with application to image segmentation have been presented in [63].

On the other hand, one can apply AI based approaches. The usage of convolutional neural networks for unsupervised image segmentation have been studied in [140] or [142].

Note that, image segmentation is an ill-defined problem, since there is no unique ground-truth segmentation of an image against which the output of an algorithm may be compared. For the evaluation methods we refer to [235] and survey [246].

The connection between clustering and collective dynamics, has been already explored in the context of particle swarm optimization (PSO) [89]. For density based PSO clustering algorithms we refer to [9] and [112]. Moreover, PSO based image clustering have been developed in [192]. The only similarity of PSO and our DIPCLUST method is the agent based approach itself. However the rules of the agents' evolution are different, and lead to different characteristics of the methods, e.g. PSO requires to a priori specify the number of clusters, and is based on solving global optimisation problem – a computationally expensive process. On the other hand DIPCLUST utilises a simple rule for the evolution of agents (8.1.1) allowing to employ techniques of linear consensus theory [190, 191].

The PSO clustering algorithm aims to simultaneously minimize the distances within one cluster, and to maximize the distances between clusters.

8.3 Emergent Dynamics of the DI Protocol

In this section we analyse the long time dynamics of the DI protocol, focusing on cluster formation. Before we begin, let us introduce the necessary notation.

The DI consensus protocol (8.1.1) can be naturally represented in the language of graph theory. The ensemble $\{1,...,N\}$ is identified with a directed graph (digraph) $(\{1,...,N\},\mathscr{E})$ with edges $\mathscr{E}$ defined by neighborhoods $\mathscr{N}_i$:

$$\mathscr{E}(i,j)=1 \;\Leftrightarrow\; i\in\mathscr{N}_j,$$

cf. (8.1.2). Note that, since $i\in\mathscr{N}_j$ does not imply $j\in\mathscr{N}_i$, the matrix $\mathscr{E}$ is not necessarily symmetric, and thus the graph $(\{1,...,N\},\mathscr{E})$ is directed. We shall refer to all subsets of $\{1,...,N\}$ as clusters. What follows, is a natural inheritance of standard notions from graph theory; for instance a cluster $\mathscr{A}\subset\{1,...,N\}$ is weakly/strongly connected iff its respective graph is weakly/strongly connected. We will also say that a cluster $\mathscr{A}\subset\{1,...,N\}$ is isolated iff it is not connected to any node outside of $\mathscr{A}$.

Next we provide the essential notion of a densely-packed cluster introduced in [165].

Definition 8.1. We say that the cluster $\mathscr{A}\subset\{1,...,N\}$ is r-densely packed at the time t if

1. the set
$$\bigcup_{i\in\mathscr{A}} B(x_i(t),r/2)$$
is connected,
2. each open ball $B(x_i(t),r)$, for $i\in\mathscr{A}$ contains more than m agents.

There are multiple connectivity- and graph-related implications of Definition 8.1. Most of them, are presented in Section 8.6, Lemma 8.7. The main application is the following theorem (also proved in Section 8.6) on sufficient conditions ensuring that the clusters collapse into a single steady state.

Theorem 8.2. *Suppose that at any time t_0 the ensemble $\{1, ..., N\}$ consists of K isolated connected clusters $\mathscr{A}_1, ..., \mathscr{A}_K$ with the property that the convex hulls of $\delta/2$-neighborhoods of the clusters do not intersect, i.e.*

$$\operatorname{conv}\left(\bigcup_{i\in\mathscr{A}_k} B(x_i, \delta/2)\right) \cap \operatorname{conv}\left(\bigcup_{i\in\mathscr{A}_l} B(x_i, \delta/2)\right) = \emptyset \tag{8.3.1}$$

for all $k, l \in \{1, ..., K\}$. Then (8.3.1) persists in time and the clusters remain isolated. Moreover, each r-densely packed cluster $\mathscr{A}$ with r satisfying

$$\frac{r}{\delta}\mathscr{Z}^{\frac{r}{\delta}} \le \frac{m}{6\#\mathscr{A}}\mathscr{Z}, \qquad \mathscr{Z} := e^{\frac{2m}{3(\#\mathscr{A})^3}} \tag{8.3.2}$$

is δ-densely packed indefinitely, and converges to the steady state

$$x_{\mathscr{A}} := \frac{1}{\#\mathscr{A}} \sum_{i\in\mathscr{A}} x_i.$$

Remark 8.3. Condition (8.3.2) may not be clear at the first glance, but is easily viable from the perspective of numerical computations. Intuition behind it is as follows. If $\mathscr{A}$ is r-densely packed with $r \le \delta$, then it is strongly connected (see Lemma 8.7 in Section 8.6 below). Consequently, if r is significantly smaller than δ, the agents have room to move around, before the clusters ceases to be at least δ-densely packed. Thus, the quantity $\frac{r}{\delta}$ represents the cluster's rigidity, i.e. small $\frac{r}{\delta}$ means that the cluster is flexible and the agents may move a lot before connectivity breaks. Condition (8.3.2) requires rigidity to be small compared to quantities related to the cluster's volume and the algebraic connectivity for (8.1.1). In practice we usually have $2m \ll 3(\#\mathscr{A})^3$, for which

$$\mathscr{Z} \approx 1$$

and then (8.3.2) can be reduced to

$$r \approx \frac{\delta m}{6\#\mathscr{A}}. \tag{8.3.3}$$

We use simplification (8.3.3) in numerical computation.

Remark 8.4. Theorem 8.2 states that as soon as convex hulls of the clusters' influence regions are disjoint and any cluster is sufficiently densely packed, then, from the perspective of large-time behaviour, that cluster can be immediately replaced by its center of mass $x_{\mathscr{A}}$. It naturally leads to a stopping criterion for the clustering algorithm presented in Section 8.4.1 below. □

8.4 DIPCLUST Algorithm

The continuous model presented in Section 8.1 can be simulated with a time-discretization scheme. This results in an iterative procedure that, for a given set of agents, performs clustering by advancing their positions in time up to the point when a sufficiently densely packed configuration is reached. The procedure is presented in Algorithm 8.1 with a detailed description of the key steps in sections 8.4.1 - 8.4.8.

```
Input: data(0) and parameters ($\delta$, $m$, $n_{max}$)
Pre-processing (optional);
Build $\delta-$lattice;
for $n = 0$; $n < n_{max}$; $n + +$ do
    Interactions(data(n), $\delta$, m);
    Advance the data in time;
Build $\delta/4-$lattice;
Identify clusters;
Assign colors and outliers to clusters (optional);
Post-processing (optional);
```

Algorithm 8.1: Segmentation Procedure.

8.4.1 Input

The input consists of the given data(0) scaled to the unit cube $[0,1]^d$, parameters δ, m, and the stopping time n_{max}. Parameters δ and m are chosen to fit particular types of data, similar to other unsupervised data segmentation methods. The influence of parameters is investigated in Section 8.5.1.

In order to fix the stopping time n_{max} we test Algorithm 8.1 on sample sets of data. We perform the loop from Algorithm 8.1 until all clusters satisfy (8.3.1) and are r-densely packed with r satisfying (8.3.3). Then we set n_{max} = the number of the last iteration. This process is computationally expensive, but it needs to be performed only once for each type of similar data. For example, in all applications presented in the sequel, we take $n_{max} = 10$.

8.4.2 Build δ-Lattice

By (8.1.1) and (8.1.2) the communication protocol is local and depends on interaction range δ. Therefore, we introduce the division of $[0,1]^d$ into a regular lattice of N_l^d cells, with the edge length $N_l = 1/\delta$. This serves the purpose of reducing the complexity of neighbor queries, in order to compute the right hand side of (8.1.1). The agents within $\delta-$distance are to be found in adjacent cells. Thus, for uniformly distributed data, the complexity scales linearly with the number of agents. This method is applied, for instance, in particle simulations of liquids, c.f. [8].

8.4.3 Interactions

For a fixed $n \in \{0, ..., n_{max}\}$, we establish the connectivity between the agents at the nth iteration, *i.e.* the neighbor sets $\mathscr{N}_i(n)$. It suffices to check conditions (8.1.2) for each $i \in \{1, ..., N\}$ and each k belonging to the same δ-lattice cell as i or to the adjacent cells.

8.4.4 Advance the Data in Time

For temporal discretization we use a scheme from the Runge–Kutta family of explicit methods. We advance in time with a well-known explicit Euler method

$$x_i^{n+1} = x_i^n + (\Delta t)\kappa \sum_{k \in \mathscr{N}_i(n)} (x_k^n - x_i^n), \quad x_i^0 = x_{i0}.$$

Here $\mathscr{N}_i(n)$ is the neighbor set obtained in the previous step. We take $(\Delta t)\kappa = 1/K_n$, where

$$K_n = \max_i \Big\{ \max\{\#\mathscr{N}_i(n), m\} \Big\}.$$

Naturally, other choices of $(\Delta t)\kappa$ are possible, but one should make sure that the time-scale is small, whenever the maximal local density of the agents is large.

8.4.5 Identify Clusters

After completing the n_{max} iteration we construct a $\delta/4$-lattice. We refer to it as the lattice, but the reader should contextually distinguish it from the previously introduced δ-lattice. Note that, the maximal iteration n_{max} has been chosen so that we expect all isolated connected clusters to be r-densely packed with r satisfying (8.3.3). In such a case the lattice cells emulate r-densely packed clusters. The cluster assignment is performed as follows:

1. Cells of the lattice are divided into 2 groups: core cells with more than m agents inside and outlier cells with m or less agents inside.
2. Any core cell forms a cluster along with all of adjacent core cells and, transitively, all subsequent adjacent core cells etc..
3. Agents belonging to core cells are then assigned to the respective clusters.
4. Agents belonging to the outlier cells become outliers and are considered noise.

Finally the cluster assignment is retroactively applied to the initial values i.e. to data(0), as showcased in Fig. 8.1.

Remark 8.5. Clearly, the scale of the lattice $\delta/4$ is much greater than the density threshold (8.3.3). However, at n_{max} we expect all of the connected isolated clusters to be r-densely packed and asymptotically stable, by Theorem 8.2. Therefore the only inadequacy introduced by the procedure described in Section 8.4.5, compared to what we would obtain with a finer r-lattice, amounts to the merging of clusters. Naturally, other scales of the clustering lattice can be chosen, but $\delta/4$ provides a good balance between numerical complexity, accuracy of the algorithm and resistance to the curse of dimensionality.

8.4.6 Assign Colors and Outliers to Clusters

Optionally, for instance in the applications to color image segmentation, at this point we assign values (colors) to each of the clusters obtained in Step 8.4.5. To do so, we simply take the average value of the data within each cluster. Thus, if $\mathscr{A}$ is one of the clusters, it is assigned with the average value $x_{\mathscr{A}} = \sum_{i\in\mathscr{A}} x_i/\#\mathscr{A}$. Then, each outlier cell (and the agents within) is assigned to the cluster with the closest average value $x_{\mathscr{A}}$.

In the case of color image segmentation, this step is responsible for defining the color scheme and dealing with the outliers.

8.4.7 Complexity

For each time-step the most computational expensive part of the algorithm is to establish the connectivity between the agents described in Section 8.4.3. Taking advantage of the δ-lattice structure, for each agent we need to examine neighbors in adjacent lattice cells. In that case only limited number of agents must be visited multiple times, therefore the average complexity of $O(N)$ is obtained. Moreover, we perform constant number of iterations, since $n_{max} \ll N =$ *number of agents.*

As described in Section 8.4.5, identification of clusters is based on the $\delta/4$-lattice. We denote by k the number of core cells in the $\delta/4$-lattice. Consequently, identifying clusters procedure requires $O(k\log k)$ operations. Therefore an overall average runtime complexity of $O(N) + O(k\log k)$ is obtained. If

$$k\log k \leq N, \tag{8.4.1}$$

then we obtain complexity of order $O(N)$. By the definition of the core cells, $km \leq N$ and taking $\frac{m}{\delta^d} \approx \frac{N}{1}$ leads to $k \lesssim \frac{1}{\delta^d}$. Thus (8.4.1) is ensured if

$$\frac{1}{\delta^d}\log\frac{1}{\delta^d} \leq N. \tag{8.4.2}$$

This is the key point of our contribution, namely the development of a method of low complexity, c.f. [242]. It outperforms mid complexity methods, e.g. DBSCAN, whose complexity is driven by distance queries. Even if an indexing structure is used, a neighbourhood query executes in $O(\log N)$ leading to an overall average complexity of DBSCAN of order $O(N\log N)$, with the worst case of $O(N^2)$, c.f. [109].

Remark 8.6. The improvement of the algorithm's numerical complexity hinges on assumption (8.4.1), which we achieve by ensuring that δ satisfies (8.4.2) for any fixed N. It creates a lower bound $\delta_0(N,d)$ for the set of admissible parameters δ.

Observe, that taking $n_{max} = 0$, i.e. immediately skipping to step 8.4.5 of the algorithm, we actually perform a simplified variant of the DBSCAN segmentation with $\delta/4$ as the neighborhood size. However, in a high-dimensional case $\delta_0(N,d)$ can be relatively large resulting in far too few clusters. In other words, proper clustering with the DBSCAN algorithm requires δ to be small enough, so that we cannot ensure (8.4.1) and we do not gain much in terms of numerical complexity. We visualize this phenomenon in the case of color image segmentation in Fig. 8.2. Indeed, the image in column C and row 1 presents such a scenario: Algorithm 8.1 performed with $n_{max} = 0$. It leads to the emergence of only 1 cluster and a poor representation of the original image in column A and row 1. It indicates that the key point of our method, i.e. the evolution of the agents' positions in time, which breaks large clusters, is essential.

□

8.4.8 Pre- and Post-processing

In order to improve clustering, we may perform pre- or post-processing. For instance, in color image segmentation, we pre-process images with Gaussian blur. Post-processing might be needed if the resulting segmentation produces a fairly large number of clusters. One can merge clusters or run Identify clusters with a lattice of larger scale.

8.5 Application to Color Image Segmentation

In order to illustrate the DIPCLUST method, we apply Algorithm 8.1 to selected pictures from Berkeley Segmentation Dataset (BSD) [158]. Size of the images is 481 times 321 what results in $N = 154401$ pixels.

The position $x_i(t) \in \mathbb{R}^5$ represents the ith pixel in a 5D feature space, where the modality is realised by combining 2D spatial positions of pixels in an image and their 3D representations in a color space (e.g. in RGB). In the feature space we define the combined distance $d(i,j)$ between pixels i and j, the spatial distance $d_s(i,j)$ and color distance $d_c(i,j)$ as follows:

$$\begin{aligned} d_s(i,j)^2 &= (y_i^1 - y_j^1)^2 + (y_i^2 - y_j^2)^2, \\ d_c(i,j)^2 &= (r_i - r_j)^2 + (g_i - g_j)^2 + (b_i - b_j)^2, \\ d(i,j) &= \sqrt{d_s(i,j)^2 + d_c(i,j)^2}, \end{aligned}$$

where y^1 and y^2 are spatial coordinates and r, g, b are color values in the RGB color space. All variables are normalised to $[0,1]$, therefore we work in a 5D unit cube.

Segmentation follows Algorithm 8.1. We apply Gaussian blur as pre-processing and for post-processing we assign colors and outliers as described in Section 8.4.6. This process is illustrated in Figure 8.2.

Column A of Figure 8.2 consists of 4 images with original positions of the pixels and colors inherited from 0, 3, 7 and 10 iteration in each row, respectively. For instance, the image in column A and row 2 presents the 5D pixels with combined coordinates $x_i^A(3) = (y_i^1(0), y_i^2(0), r_i(3), g_i(3), b_i(3))$. Column B represents the projection of each pixel x_i^A onto the 3D RGB space, e.g. for $x_i^A(3)$ we have $(r_i(3), g_i(3), b_i(3))$ in the image in column B and row 2. Column C consists of images with original positions and colors obtained through the application of the cluster identification steps of Algorithm 8.1 (c.f. Sections 8.4.5 and 8.4.6) with $n_{max} = 0, 3, 7, 10$, respectively. Column D represents positions of clusters in the RGB space.

Observe the interesting case of image in Column C and row 1. Since no iterations of the loop in Algorithm 8.1 was performed, the pixels are not sufficiently densely packed and only one cluster, say $\mathscr{A}_1$, is identified by the procedure in Section 8.4.5. Then, according to Section 8.4.6, the color of each pixel in $\mathscr{A}_1$ is changed to the average color of all pixels in $\mathscr{A}_1$, which happens to be blue. All remaining pixels outside of $\mathscr{A}_1$ are outliers and are then assigned the same blue color.

In order to show the parameters' influence, we apply Algorithm 8.1 with various m and δ to a picture of peppers, Figure 8.3.

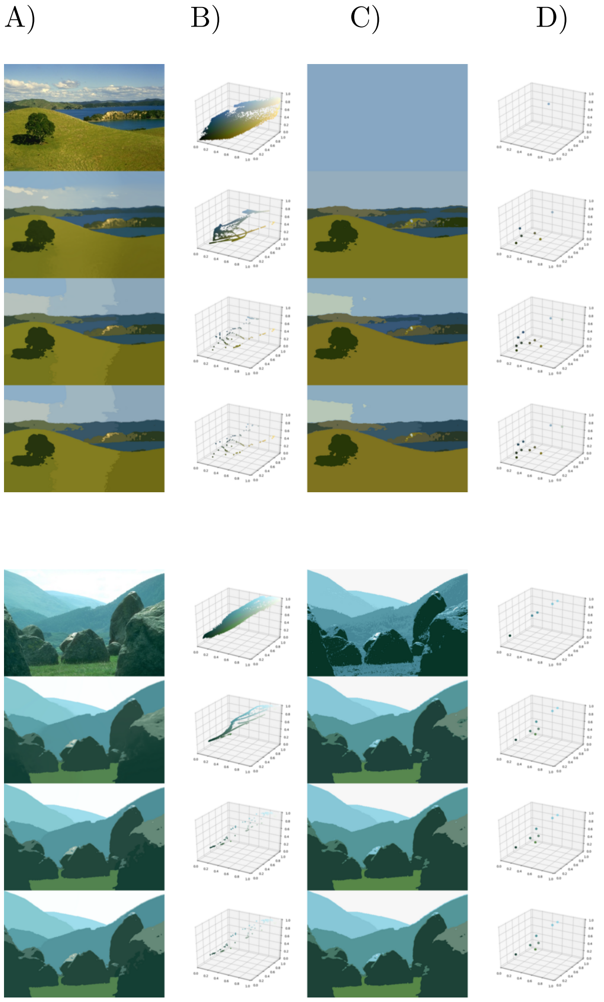


Figure 8.2: Illustration of segmentation process, for parameters $\delta = 0.15$, $m = 40$ and $n_{max} = 0, 3, 7, 10$.

We pick the interaction range $\delta \in \{0.1, 0.2\}$. This choice is connected to a number of pixels N and the dimension d, $\delta \approx \sqrt[5]{\frac{1}{N}} \approx 0.1$. This corresponds to the length of an edge of a $5-$dimensional cube with volume $1/N$ (average volume for one pixel).

The parameter m is set to be equivalent to the minimal cluster density being η times greater than an average pixel density, with $\eta \in \{1, 5\}$.

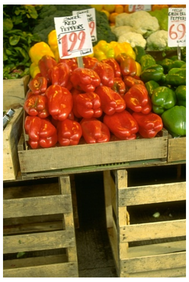


Figure 8.3: Peppers, input picture for parameter discussion.

8.5.1 Parameters

Figure 8.4, presents: fully segmented image (Column C) with $n_{max} = 10$, and two projections onto the 3D *RGB* space (Columns B and D), as described above for Figure 8.1. We observe that the number of clusters is directly connected to the interaction range δ, but seems mostly independent of parameter m. Instead, parameter m, drives the dynamics of agents, c.f. Column B, and thus – the final cluster assignment. Indeed, comparing rows 1 and 2 in Fig. 8.4, we observe that the clusters in Column D are almost the same, but the segmented images in Column C differ.

Note, that in extreme cases, for large δ and small m, one will obtain only one cluster. On the other hand, with small δ and large m each point will be placed into its own cluster, like in the input image.

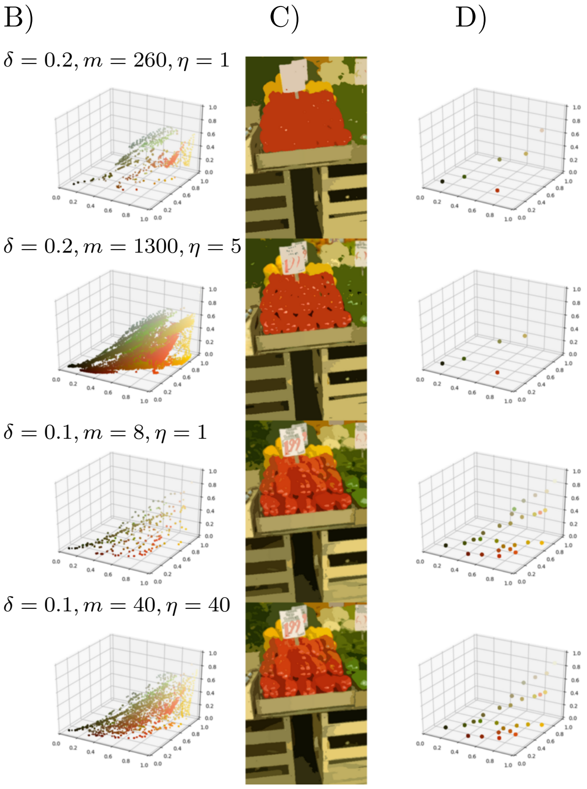


Figure 8.4: Illustration of the influence of parameters.

Unsupervised color image segmentation is an image processing task without a unique result. The BSD provides several human-performed segmentations for each picture. As pointed out in [246] one cannot guarantee that any manually-generated segmentation image is better than another. This is especially the case of nature images. Thus, the dependence on parameters can be seen as an advantage of the presented method, allowing to encapsulate different results of color image segmentation, similar to the manually-generated cases.

Note that the provided parameters are closely related to average density of pixels. In practise, taking δ and m from small ranges of values leads to relatively small differences in the resulting segmented pictures, which corresponds to the slight variation between human-generated segmentations. Thus, restriction of the set of admissible parameters to values related to average density of the pixels, while ensuring that bounds established in Section 8.4.7 are satisfied, makes finding the proper values of parameters manageable. This is further showcased by Fig. 8.5 below, where multiple different pictures are segmented using DIPCLUST with the same values of parameters leading to mostly reasonable results.

8.5.2 Segmentation Results on BSDS500

Next, we present segmentation results on selected images from BSD, see Fig. 8.5. Images are from both the test and the train set. Each pair consists of the original image on the left and the segmented image on the right. Parameters are $\delta = 0.15$, $m = 308$, $\eta = 5$, and $n_{max} = 10$. The selected interaction range, results in a relatively small number of clusters. Note that, it depends on the type of picture, typically images of nature exhibit similar colour, resulting in fewer clusters.

Edge detection

Finally, we briefly showcase the impact of DIPCLUST segmentation on contour detection. In Fig. 8.6 we present, image and edge detection for input (left) and segmented (right) image, respectively. Contours are detected by the method of Canny [48]. We observe, the improvement of edge detection for the segmented image compared to the direct application of Canny method to the original image. The viability of DIPCLUST in contour detection requires further study.

8.6 Mathematical Proofs

In this section we prove Theorem 8.2. The starting point is Lemma 8.7, which connects the notion of r-densely packed clusters with their connectivity. We recall the notation established in Section 8.3.

Throughout this section we denote

$$\mathbf{x} := (x_1, ..., x_N) \in \mathbb{R}^{dN}.$$

Lemma 8.7. *Suppose that* $\mathbf{x}$ *is a smooth solution to* (8.1.1) *with constant coefficients in* $[t_0, t_1)$. *Assume further that* $\mathscr{A}$ *is an isolated* r*-densely packed cluster at each* $t \in [t_0, t_1)$ *for* $r \le \delta$. *Then*

1. *a cluster* $\mathscr{A}$ *treated as a graph is strongly connected and undirected,*
2. *The graph diameter* $d_{\mathscr{A}}$ *of cluster* $\mathscr{A}$, *i.e. the maximal shortest path between any two nodes, is upper-bounded by*

$$d_{\mathscr{A}} \le \begin{cases} \#\mathscr{A} & \text{if } m = 1, \\ \left\lceil \frac{3\#\mathscr{A}}{m+1} \right\rceil - 1 & \text{if } m \ge 2. \end{cases} \tag{8.6.1}$$

3. *The spatial diameter* $\mathscr{D}_{\mathscr{A}} := \max\{|x_i - x_j| : i, j \in \mathscr{A}\}$ *of a cluster* $\mathscr{A}$ *is upper-bounded by*

$$\mathscr{D}_{\mathscr{A}} \le r d_{\mathscr{A}}.$$

4. There exists $\lambda > 0$ such that

$$|x_i(t) - x_{\mathscr{A}}| \leq e^{-\lambda(t-t_0)}|x_i(t_0) - x_{\mathscr{A}}|, \ \forall i \in \mathscr{A}, \ t \in [t_0, t_1),$$

where $x_{\mathscr{A}}$ is the center of mass of $\mathscr{A}$. Moreover the exponent λ satisfies

$$\lambda \geq \frac{4\kappa}{d_{\mathscr{A}} \# \mathscr{A}}. \tag{8.6.2}$$

Proof. We refer to [165, Lemma 1 and Lemma 2], where variants of assertions 1 and 4 are proved by modifying the theory of linear consensus developed in [190] and [191]. The only difference here is that cluster $\mathscr{A}$ is undirected due to the symmetry of weight in (8.1.1), which further implies that the steady state $x_{\mathscr{A}}$ is indeed the center of mass of $\mathscr{A}$. Assertion 1 implies that $\mathscr{A}$ is an isolated, connected undirected graph with a minimal degree of m and assertion 2 holds true due to Erdös, Pach, Pollack and Tuza's classical work [90, Theorem 1]. To prove assertion 3, consider the graph $\mathscr{G} := (\mathscr{A}, \mathscr{E}_r)$, with $\mathscr{E}_r(i,j) = 1$ iff $\mathscr{E}(i,j) = 1$ and $|x_i - x_j| \leq r$. Then $\mathscr{G}$ is an r-densely packed subgraph of $(\mathscr{A}, \mathscr{E})$ with the same nodes and fewer edges. Any two nodes i and j in $\mathscr{A}$ are connected by a path $\pi(i,j) \subset \mathscr{E}_r$ represented by a sequence of nodes $i = k_1, ..., k_l = j$ with

$$l \leq d_{\mathscr{A}} \text{ and } |x_{k_\alpha} - x_{k_{\alpha+1}}| \leq r \quad \text{for all } \alpha \in \{1, ..., l-1\}.$$

Consequently, $|x_i - x_j| \leq r d_{\mathscr{A}}$.

Finally, since $\lambda(t)$ is the algebraic connectivity of the graph $(\mathscr{A}, \mathscr{E})$ multiplied by κ, inequality (8.6.2) follows from McKay's result published by Mohar in [173, Theorem 4.2], which states that the algebraic connectivity of the graph is lower-bounded by 4 divided by the graphs diameter and its number of nodes.

We proceed with the following useful proposition.

Proposition 8.8. *For any smooth solution* $\mathbf{x}$ *to system* (8.1.1) *and all* $t \geq t_0 \geq 0$ *we have*

$$\mathscr{C}(t) := \operatorname{conv}\{x_i(t) : i \in \{1, ..., N\}\} \subset \mathscr{C}(t_0). \tag{8.6.3}$$

Moreover the maximal velocity of each agent is uniformly bounded, i.e.

$$V_i(t) := \max_{t \geq t_0} |\dot{x}_i(t)| \leq \kappa \# \mathscr{N}_i \delta. \tag{8.6.4}$$

Proof. Fix any $s \geq t_0$ and let $x_i(s)$ belong to the boundary of $\mathscr{C}(s)$ and assume without a loss of generality that $x_i(s) = 0$. It suffices to show that $\dot{x}_i(s)$ belongs to the cone $\bigcup_{\tau > 0} \tau \mathscr{C}(s)$. With $\chi_{\{k \in \mathscr{N}_i\}}$ denoting the characteristic function of the event that $k \in \mathscr{N}_i(t)$, we have

$$\begin{aligned} \dot{x}_i(s) &= \kappa \sum_{k \in \mathscr{N}_i} (x_k(s) - x_i(s)) \\ &= \kappa \# \mathscr{N}_i \sum_{k=1}^{N} \frac{\chi_{\{k \in \mathscr{N}_i\}}}{\# \mathscr{N}_i} x_k(s) \in \kappa \# \mathscr{N}_i \mathscr{C}(s), \end{aligned} \tag{8.6.5}$$

since the sum on the right-hand side of (8.6.5) is a convex combination of elements belonging to $\mathscr{C}(s)$.

The proof of (8.6.4) follows immediately from (8.1.1) and (8.1.2) after noting that $|x_i - x_k| \leq \delta$ for $k \in \mathscr{N}_i$. □

The following lemma is the essential part of the proof of Theorem 8.2.

Lemma 8.9. *Let* $\mathbf{x}$ *be a smooth solution to* (8.1.1) *and let* $\mathscr{A}$ *be a cluster in* $\{1, ..., N\}$ *that remains isolated at all times. Then there exists* $r^* < \delta$ *such that if* $\mathscr{A}$ *is* r*-densely packed with* $r \leq r^*$*, then* $\mathscr{A}$ *is at least* δ*-densely packed for all* $t \geq t_0$*.*

Proof. Any r-densely packed cluster is r^*-densely packed whenever $r \leq r^*$, hence for the remainder of the proof we will assume that $r = r^*$. Since, at $t = t_0$, the agents are r-densely packed for $r < \delta$ and the maximal velocity of the agents is uniformly bounded (see Proposition 8.8), there exists a time interval $[t_0, T)$, such that the agents are δ-densely packed for $t \geq t_0$.

Cluster $\mathscr{A}$ is isolated and system (8.1.1) is piecewise linear with finitely many possible right-hand sides, and thus, Lemma 8.7 holds locally at each $t \geq t_0$. Therefore there exists a piecewise constant exponent $\lambda(t) > 0$ as in assertion 4 of Lemma 8.7. By assertion 2 as well as inequality (8.6.2) from Lemma 8.7 the exponent has the uniform lower bound

$$\lambda(t) \geq \frac{4\kappa}{d_{\mathscr{A}} \# \mathscr{A}} \geq \frac{4\kappa m}{3(\#\mathscr{A})^2} =: \lambda_*.$$

Therefore for all $t \geq t_0$ and all $i \in \mathscr{A}$ we have

$$\begin{aligned} |x_i(t) - x_{\mathscr{A}}(t)| &\leq e^{-\lambda_*(t-t_0)} \mathscr{D}_{\mathscr{A}}(t_0), \\ \mathscr{D}_{\mathscr{A}}(s) &:= \sup_{i,j \in \mathscr{A}} |x_i(s) - x_j(s)|. \end{aligned} \tag{8.6.6}$$

Let

$$T := \sup\{t > t_0 : \mathscr{A} \text{ is } \delta\text{-densely packed in } [t_0, t)\} > 0.$$

We shall chose r small enough so that $T = \infty$. For all $i, j \in \mathscr{A}$ and all $t \in [t_0, T)$, we have

$$|x_i(t) - x_j(t)| \leq |x_i(t) - x_{\mathscr{A}}(t)| + |x_j(t) - x_{\mathscr{A}}(t)|. \tag{8.6.7}$$

Let us fix in (8.6.7) any pair (i, j) such that $|x_i(t_0) - x_j(t_0)| \leq r$. Such pairs exist since the ensemble is r-densely packed initially. Applying (8.6.6) to the right-hand side of (8.6.7) leads to

$$|x_i(t) - x_j(t)| \leq 2e^{-\lambda_*(t-t_0)} \mathscr{D}_{\mathscr{A}}(t_0). \tag{8.6.8}$$

Alternatively $|x_i(t) - x_j(t)|$ can be upper-bounded using equation (8.1.1) and Proposition 8.8 yielding

$$\begin{aligned} |x_i(t) - x_j(t)| \leq{}& |x_i(t) - x_i(t_0)| + |x_j(t) - x_j(t_0)| \\ &+ |x_i(t_0) - x_j(t_0)| \leq 2\kappa \# \mathscr{A} \delta (t - t_0) + r, \end{aligned}$$

which ensures that

$$|x_i(t) - x_j(t)| < \delta$$

at least as long as

$$t - t_0 < \frac{1}{2\kappa \# \mathscr{A}} \left(1 - \frac{r}{\delta}\right).$$

However for $t - t_0 \geq \frac{1}{2\kappa \# \mathscr{A}} \left(1 - \frac{r}{\delta}\right)$, by (8.6.8) we have

$$|x_i(t) - x_j(t)| \leq 2e^{-\frac{\lambda_*}{2\kappa \# \mathscr{A}}\left(1 - \frac{r}{\delta}\right)} \mathscr{D}_{\mathscr{A}}(t_0)$$

and we require

$$\widetilde{r} := 2e^{-\frac{\lambda_*}{2\kappa \# \mathscr{A}}\left(1 - \frac{r}{\delta}\right)} \mathscr{D}_{\mathscr{A}}(t_0) < \delta. \tag{8.6.9}$$

Then assertions 2 and 3 of Lemma 8.7 imply that

$$\mathscr{D}_{\mathscr{A}} < 3r \frac{\# \mathscr{A}}{m}.$$

Combining the above bounds yields

$$\widetilde{r} < 6re^{-\frac{2m}{3(\# \mathscr{A})^3}\left(1 - \frac{r}{\delta}\right)} \frac{\# \mathscr{A}}{m} \leq \delta,$$

which can be further rearranged into (8.3.2). Condition (8.3.2) holds for any sufficiently small $r > 0$.

In conclusion, any pair i and j, initially of distance at most r from each other, remains of distance $\widetilde{r} < \delta$ from each other throughout $[t_0, T)$. Thus $\mathscr{A}$ is $\widetilde{r}$-densely packed in $[t_0, T)$ with $\widetilde{r} < \delta$ and, by continuity, it is at least $\frac{\widetilde{r}+\delta}{2}$-densely packed at $t = T$. Then, exactly as at the beginning of the proof, $\mathscr{A}$ remains at least δ-densely packed past the time T which stands in contradiction with the definition of T, unless $T = \infty$. The proof is finished.

Lemma 8.9 ensures that $\mathscr{A}$ remains δ-densely packed as long as (8.3.2) holds and $\mathscr{A}$ is an isolated cluster. We are now ready to finalise the proof of Theorem 8.2.

Proof of Theorem 8.2: First observe that, by Proposition 8.8, condition (8.3.1) persists in time. Hence, each of the clusters $\mathscr{A}_1, ..., \mathscr{A}_K$ is isolated, and thus Lemma 8.9 applies with r^* satisfying condition (8.3.2). Therefore, each r^*-densely packed cluster is at least δ-densely packed indefinitely. Consequently, for such clusters, assumptions of Lemma 8.7 are satisfied, and inequality (8.6.6) holds for all $t \geq t_0$ thereby ensuring the convergence in Theorem 8.2. The proof is finished. ■

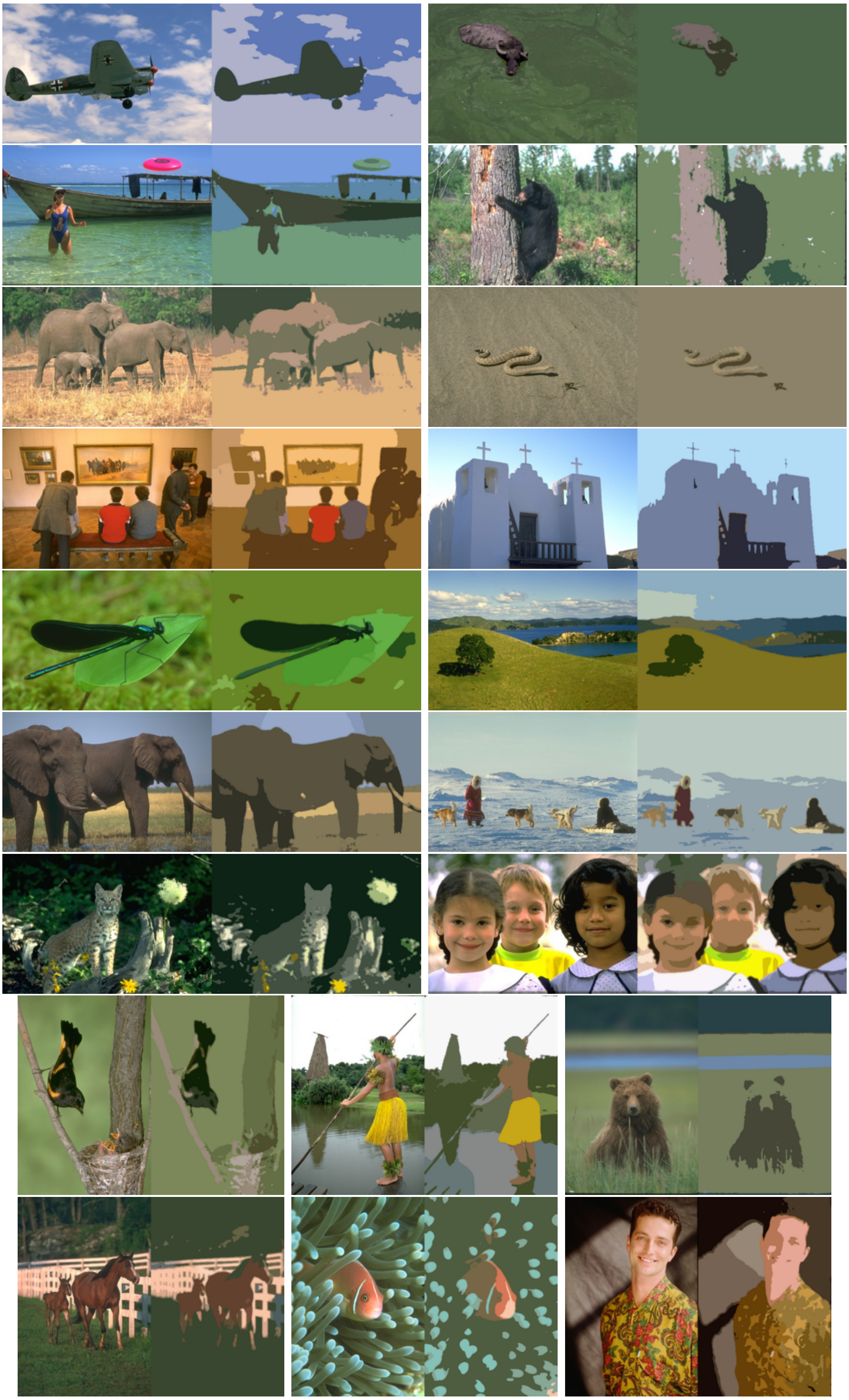


Figure 8.5: Additional segmentation results on the BSD.

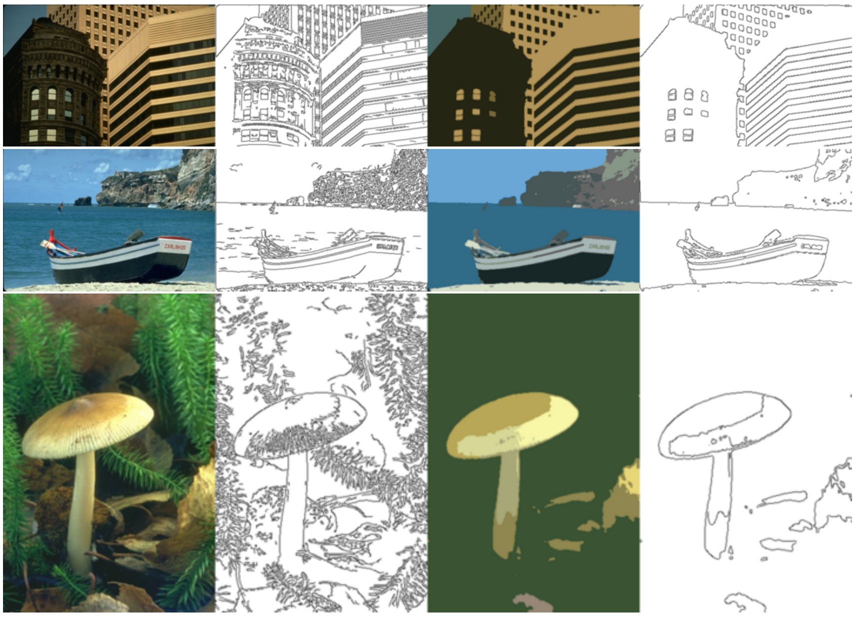


Figure 8.6: Edge detection on original and segmented images.

CHAPTER 9

Dual Weight Residual Error Estimates for Neural Network Solutions of Partial Differential Equations

The content of this chapter is joint work with Thomas Richter, and is published in the preprint

P. Minakowski and T. Richter. *Error Estimates for Neural Network Solutions of Partial Differential Equations.* 2021. arXiv: 2107.11035 [math.NA].

Chapter Summary. We develop an error estimator for neural network approximations of PDEs. The proposed approach is based on *dual weighted residual estimator* (DWR). It is destined to serve as stopping criterion, that guarantees the accuracy of the solution independently of the design of the neural network training. The result is equipped with computational examples for Laplace and Stokes problems.

Chapter Organisation. In Section 9.3 we quickly recapitulate the DWR method in its easiest form and extend it to estimate the network error. Later we briefly present network architecture in Section 9.4. Section 9.5 demonstrates the accuracy of the estimator for different applications and shows how the estimator can be integrated as a stopping criterion in training.

Contents of Chapter

9.1 Introduction

In recent years, the emerging field of (deep) neural networks has reached the numerical approximation of partial differential equations (PDE). Several approaches have been proposed that aim at directly representing the solution to the PDE by a deep neural network. For this purpose the network is considered as (differentiable) function $\mathcal{N} : \Omega \to \mathbb{R}^c$, where $\Omega \subset \mathbb{R}^d$ is the computational domain of dimension $d \in \mathbb{N}$ and $c \in \mathbb{N}$ is the size of the differential system.

The network is trained by integrating the differential equation (and the boundary conditions) into the loss function. Several different approaches based on different realizations have been presented:

The *Deep Ritz* method by E and Yu [88] aims at minimizing the energy functional and it can be applied to symmetric problems. For the Laplace equation, $-\Delta u = f$ in Ω with $u = g$ on $\partial\Omega$, this means to minimize

$$E(\mathcal{N}) = \frac{1}{2}\int_\Omega |\nabla \mathcal{N}(x)|^2 \,\mathrm{d}x - \int_\Omega \mathcal{N}(x) \cdot f(x) \,\mathrm{d}x + \lambda \int_{\partial\Omega} |\mathcal{N}(x) - g(x)|^2 \,\mathrm{d}x,$$

with a parameter $\lambda > 0$, where the integrals are approximated by Monte-Carlo integration. *Training data* is generated by picking random integration points. See [215] for an overview and further examples.

Another approach, denoted as *DeepXDE* (Lu, Meng, Mao and Karniadakis [154]), *Unified Deep Artificial Network* [29] or *DGM* (Deep Galerkin Method) [224], see also [10], minimizes the strong residual of the equation, either as collocation method in randomly picked points within the domain and on the boundary (formulated once more for the Laplace problem)

$$E(\mathcal{N}) = \frac{1}{N_1}\sum_{i=1}^{N_1} |f(x_i) + \Delta \mathcal{N}(x_i)|^2 + \frac{\lambda}{N_2}\sum_{j=1}^{N_2} |\mathcal{N}(x_j) - g(x_j)|^2,$$

or by Monte-Carlo integration of the (strong) residual

$$E(\mathcal{N}) = \|f + \Delta\mathcal{N}\|^2_{L^2(\Omega)} + \lambda \|g - \mathcal{N}\|_{L^2(\partial\Omega)}.$$

This second approach naturally extends to non-symmetric and nonlinear problems and we refer to the above mentioned literature for examples.

Finally, a third variant, variational physics-informed neural network *VPINN* (Kharazmi, Zhang Karniadakis) [141] approach is based on the variational formulation

$$E(\mathcal{N}) = \Big| \int_\Omega \nabla\mathcal{N}(x) \cdot \nabla\phi_k(x) \,\mathrm{d}x - \int_\Omega f(x) \cdot \nabla\phi_k(x) \,\mathrm{d}x \Big|^2$$

and *training data* is generated by choosing specific test functions ϕ_k.

Common to all these approaches is that they aim at solving one specific PDE problem. In opposition there exists the large area of deep neural network approaches that consider parameterized PDEs, which we will however not discuss here.

Instead, we will focus on the different approaches discussed above. The common rationale for the three different approaches is the excellent approximation property of neural networks, in particular Pinkus [205] proved the capability of deep neural networks to uniformly and simultaneously approximate differential functions and their derivatives. In the context of PDEs Petersen and coworkers [113] showed approximation results in Sobolev spaces and also gave convergence rates in the number of layers, neurons and weights. In particular for high dimensional differential equations, deep neural network based approaches promise to be superior [224, 154]. On the other hand, it must be noted that the previous approaches, applied to common, low-dimensional ($d = 1, 2, 3$) problems, cannot compete with established methods in terms of efficiency. While algorithms of $O(N)$ complexity exist for finite element or finite difference approximations of elliptic problems, the training of the deep neural network is a by far more challenging task.

In this contribution we will tackle the question of reliability of deep neural network approaches, i.e. to derive a posteriori error bounds for a trained network and also we aim at introducing adaptivity into the training of the neural networks. Both aspects will be approached within the concept of the *dual weighted residual estimator* (DWR) that has been introduced by Becker and Rannacher [26]. We will handle the error between real solution $u(x)$ and neural network solution $\mathscr{N}(x)$ as a mixture of Galerkin error and consistency error. What we derive is not a rigorous bound, but an efficient computational tool that can be used to validate neural network solutions, that serves as an estimate in stopping criteria during the network training and that can be localized to steer the training process in an adaptive procedure.

Training points can be chosen in several ways: fix points before training, e.g. grid points on a lattice or random points, in each training epoch, one could select different training points, or points can be chosen adaptively during training.

In [154] the authors propose Residual-Based Adaptive Refinement (RAR), in order to improve the distribution of residual points during the training process. After training for a fixed number of epochs on the initial set of residual points, the mean PDE residual is estimated by Monte Carlo integration, i.e., by the average of values at a set of randomly sampled dense locations. Then the training set is extended by new points with the largest residuals until the specified threshold for the mean is reached. Similar strategy called adaptive collocation has been proposed in [10]. Starting from the coarse set of points, the additional points are added to the training set based on the evaluation of the residual.

Note, that evaluating the model at a larger number of points is quite inexpensive computationally, while the number of training points impacts the performance significantly. These two approaches are based on strong formulation.

The stopping criteria of the training process have not been studied. A Commonly fixed number of epochs is performed, see e.g. [88] or train until mean residual reaches specific threshold [154]. According to the authors knowledge this result is the first that introduces error estimation based stopping criterion.

Deep learning techniques can be applied in the context of numerical simulation, as an extension of existing CFD codes to increase their efficiency. One can generalize existing numerical methods as artificial neural networks, with a set of trainable parameters.

In the context of solving time-dependent ODEs and PDEs. The authors recast finite volume schemes as NN and train the underlying parameters to improve accuracy on coarse grids [171]. This approach was extended to finite element methods in [44].

9.2 Finite Element and Neural Network Approximations

To keep the notation simple we will focus on Laplace problem. Let $\Omega \subset \mathbb{R}^d$ be a d-dimensional domain. We find the weak solution $u \in \mathscr{V} := H_0^1(\Omega)$ to

$$-\Delta u = f \text{ in } \Omega, \quad u = 0 \text{ on } \partial\Omega, \tag{9.2.1}$$

where $f \in L^2(\Omega)$ is the right hand side. By $\mathscr{V} = H_0^1(\Omega)$ we denote the space of L^2-functions with first weak derivative in L^2 with vanishing trace on $\partial\Omega$. The solution $u \in \mathscr{V}$ is characterized by the variational problem

$$(\nabla u, \nabla v) = (f, \phi) \quad \forall \phi \in \mathscr{V}, \tag{9.2.2}$$

where we denote by $(\cdot, \cdot)$ the L^2-inner product on Ω. Further, the solution is also equivalently characterized as minimizer of the functional

$$E(u) \le E(v) := \frac{1}{2}\|\nabla v\|^2 - (f, v) \quad \forall v \in \mathscr{V}, \tag{9.2.3}$$

where $\|\cdot\|$ is the L^2-norm.

9.2.1 Finite Element Approximation

Now, let Ω_h be a triangulation of Ω into open triangular or quadrilateral (in 2d) elements satisfying usual regularity requirements on the structure and the form of the elements. For an element $T \in \Omega_h$ we denote by $h_T = \operatorname{diam}(T)$ the element size and by

$$h = \max_{T \in \Omega_h} h_T$$

the maximum mesh size of the discretization which serves as a parameter for measuring the fineness.

Error analysis By $V_h \subset \mathscr{V}$ we will denote the finite dimensional (finite element) subspace of $H_0^1(\Omega)$. Then, let $u_h \in V_h$ be the approximation to $u \in \mathscr{V}$ given by

$$(\nabla u_h, \nabla v_h) = (f, v_h) \quad \forall v_h \in V_h. \tag{9.2.4}$$

It holds

$$\|\nabla(u - u_h)\| \le \|\nabla(u - \phi_h)\| \quad \forall \phi_h \in V_h,$$

such that the finite element error $u - u_h$ is bound by the interpolation error yielding the standard estimate

$$\|\nabla(u - u_h)\| \le ch^r \|f\|_{H^{r-1}(\Omega)},$$

where r is the polynomial degree of the finite element space and using the notation $H^0(\Omega) := L^2(\Omega)$ in the case of linear finite elements, $r = 1$. Naturally, this estimate requires sufficient regularity of the right hand side $f \in H^{r-1}(\Omega)$ and also of the domains boundary, i.e. $\partial\Omega$ must be convex polygonal for $r = 1$ or locally parameterizable by a C^{r+1}-function for $r \ge 1$.

9.2.2 Deep Ritz Approximation of the Laplace Problem

In principle, the *Deep Ritz* method as proposed by E and Yu [88] is based on minimizing the energy functional (9.2.3) by representing the unknown solution $u_{\mathscr{N},\mathbf{u}} : \mathbb{R}^d \to \mathbb{R}$ by a neural network instead of a finite element function. Here, we denote by $\mathscr{N}$ the topology of the neural network and by $\mathbf{u} \in \mathbb{R}^N$ the parameters of the network, where $N = \#\mathscr{N}$ is the total number of free parameters. Finally, $u_{\mathscr{N},\mathbf{u}}$ is the function that is realized by this specific combination of network topology and parameter choice. Mostly we will simply use the notation $u_{\mathscr{N}}$ and skip the indication of the parameter vector $\mathbf{u}$ unless it is of relevance in the given context.

E and Yu [88] considered a network layout $\mathscr{N}$ with residual connection and differentiable activation functions, but we could also use different layouts. The framework of the *Deep Ritz* method requires differentiability of the network, i.e. differentiable activation functions. Fig. 9.1 shows the layout of the deep neural network as chosen by E and Yu, but also a simpler feedforward network that can be used.

Having a certain network topology $\mathscr{N}$ in mind, the *neural network approximation space* $W_{\mathscr{N}}$ used in the *Deep Ritz* method is then given by

$$W_{\mathscr{N}} := \{u_{\mathscr{N},\mathbf{w}} : \mathbb{R}^d \to \mathbb{R}, \mid \mathbf{w} \in \mathbb{R}^N\}. \tag{9.2.5}$$

If N is finite and if the activation functions are differentiable it holds $V_{\mathscr{N}} \subset H^1(\Omega)$. We can however not expect $u_{\mathscr{N}} = 0$ on $\partial\Omega$ for $u_{\mathscr{N}} \in V_{\mathscr{N}}$, hence $V_{\mathscr{N}} \not\subset \mathscr{V} = H^1_0(\Omega)$. In [29], the authors discussed a modified setup of the neural network that strongly satisfies the homogeneous Dirichlet condition.

Here we consider the penalized energy functional (compare [88])

$$E_\lambda(v) := \frac{1}{2}\|\nabla v\|^2 - (f, v) + \frac{\lambda}{2}|v|^2_{\partial\Omega}, \tag{9.2.6}$$

where $\lambda \in \mathbb{R}_+$ is a parameter and $|\cdot|_{\partial\Omega}$ is the L^2-norm on the boundary of the domain. The additional penalty term forces v to be close to zero along the boundary. The minimizer of (9.2.6) in $u_\lambda \in H^1(\Omega)$ is characterized by the variational problem

$$(\nabla u_\lambda, \nabla v) + \lambda\langle u_\lambda, v\rangle = (f, v) \quad \forall v \in H^1(\Omega), \tag{9.2.7}$$

where $\langle\cdot,\cdot\rangle$ denote the L^2-inner product on the boundary $\partial\Omega$. The weak solution $u_\lambda \in H^1(\Omega)$ solves the Laplace problem with a disturbed Robin boundary condition, i.e.

$$-\Delta u_\lambda = f \text{ in } \Omega, \quad u_\lambda + \lambda^{-1}\partial_n u_\lambda = 0 \text{ on } \partial\Omega. \tag{9.2.8}$$

The penalized energy functional does hence introduce an additional modelling error term $\|u - u_\lambda\|$ that will depend on the parameter λ and that will converge to zero for $\lambda \to \infty$.

Training of the neural network is then by minimizing the modified energy functional (9.2.6) using Monte Carlo integration. To be precise: $N^i \in \mathbb{N}$ inner quadrature points $\mathbf{x}^i \in \Omega^{N^i}$ and $N^b \in \mathbb{N}$ boundary quadrature points $\mathbf{x}^b \in \partial\Omega^{N^b}$ are chosen, either randomly or based on a mesh of the domain. The loss function is given by

$$l(u_{\mathscr{N}}; \mathbf{x}^i, \mathbf{x}^b) := \frac{|\Omega|}{N^i}\sum_{h=1}^{N^i}\left(\frac{1}{2}|\nabla u_{\mathscr{N}}(x^i_j)|^2 - f(x^i_j)\cdot u_{\mathscr{N}}(x^i_j)\right) + \frac{|\partial\Omega|}{N^b}\sum_{j=1}^{N^b}\frac{\lambda}{2}|u_{\mathscr{N}}(x^b_j)|^2. \tag{9.2.9}$$

Details on this training process and also a study on the convergence of the neural network *Deep Ritz* approximation is presented in Section 9.4.1.

Minimizing 9.2.9 will identify the weights $\mathbf{u} \in \mathbb{R}^N$ and result in an approximation $u_{\mathscr{N},\mathbf{u}}$ to the optimal network realization $u_{\mathscr{N},\widetilde{\mathbf{u}}}$.

Error analysis An error analysis of the *Deep Ritz* approach is more delicate as compared to the case of finite elements. First, a modelling error is introduced as the energy functional (9.2.6) is distorted by the boundary penalty term and the minimizer $u_\lambda \in H^1(\Omega)$ does not exactly satisfy Dirichlet conditions, i.e. there is the error $u - u_\lambda$. Next, the approximation error $u_\lambda - u_{\mathscr{N}}$ of the neural network space $W_{\mathscr{N}}$ enters considering exact integration of the energy functional. This is followed by the error introduced by Monte Carlo integration $u_{\mathscr{N}} - u_{\mathscr{N},\widetilde{\mathbf{u}}}$ considering that $\widetilde{\mathbf{u}} \in \mathbb{R}^N$ represents an exact minimum and finally, $u_{\mathscr{N},\widetilde{\mathbf{u}}} - u_{\mathscr{N},\mathbf{u}}$ is the training error. All together, four distinct contributions can be identified.

$$\|u - u_{\mathscr{N},\mathbf{u}}\| \le \|u - u_\lambda\| + \|u_\lambda - u_{\mathscr{N}}\| + \|u_{\mathscr{N}} - u_{\mathscr{N},\widetilde{\mathbf{u}}}\| + \|u_{\mathscr{N},\widetilde{\mathbf{u}}} - u_{\mathscr{N},\mathbf{u}}\|.$$

The first we call *model error*, the second is the *network approximation error*, the third the *quadrature error* or *generalization error* and finally, the fourth is the *optimization* or *training error.*

Lemma 9.1 (*Model error*). *Let $f \in L^2(\Omega)$, $\lambda \in \mathbb{R}$ with $\lambda > 0$ and Ω be such that the solutions $u \in H_0^1(\Omega)$ and $u_\lambda \in H^1(\Omega)$ to*

$$(\nabla u, \nabla \phi) = (f, \phi) \quad \forall \phi \in H_0^1(\Omega), \quad (\nabla u_\lambda, \nabla \phi_\lambda) + \lambda \langle u_\lambda, \phi_\lambda \rangle_{\partial\Omega} = (f, \phi_\lambda) \quad \forall \phi_\lambda \in H^1(\Omega)$$

satisfy $\|u\|_{H^2(\Omega)} \le c_s \|f\|$ and $\|u_\lambda\|_{H^2(\Omega)} \le c_s \|f\|$. It holds

$$\|\nabla(u - u_\lambda)\| \le \frac{c}{\lambda}\|f\|,$$

where $c > 0$ depends on the domain Ω only.

Proof. Let $z \in H_0^1(\Omega)$ be the solution to the adjoint problem

$$-\Delta z = \frac{\nabla(u - u_\lambda)}{\|\nabla(u - u_\lambda)\|} \text{ in } \Omega, \quad z = 0 \text{ on } \partial\Omega. \tag{9.2.10}$$

Since the right hand side of this problem is in $L^2(\Omega)$, $z \in H^2(\Omega)$ and $\|z\|_{H^2(\Omega)} \le c_s$ follows for sufficiently smooth (or convex polygonal) domain Ω.

Multiplication of (9.2.10) with the error $u - u_\lambda$ and integration over the domain gives the error identity

$$\|\nabla(u - u_\lambda)\| = (\nabla z, \nabla(u - u_\lambda)) - \langle \partial_n z, u - u_\lambda \rangle_{\partial\Omega}.$$

As $u = 0$ and $z = 0$ on $\partial\Omega$ this gives

$$\|\nabla(u - u_\lambda)\| = \underbrace{(\nabla(u - u_\lambda), \nabla z) + \overbrace{\lambda \langle u - u_\lambda, z \rangle_{\partial\Omega}}^{=0}}_{=(f-f,z)=0} + \langle \partial_n z, u_\lambda \rangle_{\partial\Omega}.$$

Finally, with (9.2.8) and using both the trace inequality and the regularity of adjoint and primal solution gives

$$\|\nabla(u-u_\lambda)\| \leq |\partial_n z|_{L^2(\partial\Omega)}|u_\lambda|_{L^2(\partial\Omega)} \leq \frac{c}{\lambda}\|z\|_{H^2}\|u_\lambda\|_{H^2(\Omega)} \leq \frac{c}{\lambda}\|f\|.$$

□

The approximation properties of neural networks are already extensively studied in literature [21, 113] and they show convergence $\|u_\lambda - u_{\mathscr{N}}\| \to 0$ for an increasing size of the neural networks. For example, [113, Theorem 4.1] states that there for all $\varepsilon > 0$ the bound

$$\|\nabla(u_\lambda - u_{\mathscr{N}})\| = \mathscr{O}(\varepsilon)$$

is obtainable (given $u_\lambda \in H^2(\Omega)$) with a neural network consisting of $L = \log_2(\varepsilon^{-2})$ layers and $N = L \cdot 2^L$ weights and neurons. This approximately corresponds to

$$\|\nabla(u_\lambda - u_{\mathscr{N}})\| \approx \mathscr{O}\Big(\frac{\log(N)}{\sqrt{N}}\Big)$$

which, in terms of unknown N, is comparable to the linear finite element approximation.

The choice of the numerical quadrature points gives rise to the generalization error of the neural network representation: does the training, based on a certain set of quadrature points also gives an approximation on the complete domain? There are no a priori bounds for such kind of errors but it has been shown [182, 217] that such generalization bounds are equivalent to the stability of the training under perturbation of the training data, e.g. in the context of *Deep Ritz*, by perturbing or leaving out single quadrature points. In the context of PINNs [219] show convergence of neural network solution, to solutions of linear elliptic and parabolic PDEs as the number of training samples is increased. Moreover [172] provides rigorous upper bounds on the generalization error. Under stability assumptions of the PDEs, the authors show that the generalization error is bounded by the training error and quadrature error.

Finally, the optimization error remains for which there is also no a priori error bound. Further, there usually is no guarantee that we obtain a global minimum.

In the following sections we will present an a posteriori error estimator that will be able to give predictions on the complete error $u - u_{\mathscr{N},\mathbf{u}}$ in goal functionals. While all error contributions are included we will however not be able to distinguish between the different sources.

9.2.3 Deep Ritz Approximation of the Stokes Equations

As a second example we consider the Stokes equation on a two dimensional domain $\Omega \subset \mathbb{R}^2$, i.e. we find velocity $\mathbf{v} \in \mathscr{V}^2 := H_0^1(\Omega)^2$ and pressure $p \in \mathscr{L} := L^2(\Omega) \setminus \mathbb{R}$ such that

$$\text{div }\mathbf{v} = 0, \quad -\Delta\mathbf{v} + \nabla p = \mathbf{f} \text{ in } \Omega, \quad \mathbf{v} = 0 \text{ on } \partial\Omega, \tag{9.2.11}$$

where we denote by $\mathbf{f} \in L^2(\Omega)^2$ a given right hand side. Considering a discrete pair of subspaces $V_h \times L_h \subset \mathscr{V} \times \mathscr{L}$, the finite element solution is defined by

$$(\text{div }\mathbf{v}_h, \xi_h) + (\nabla\mathbf{v}_h, \nabla\phi_h) - (p_h, \nabla\cdot\phi_h) = (\mathbf{f}_h, \phi_h) \quad \forall(\phi_h, \xi_h) \in V_h \times L_h. \tag{9.2.12}$$

Assuming inf-sup stability of the discrete finite element pair the solution exists uniquely and standard best approximation results hold, e.g. for the $P^2 - P^1$ Taylor-Hood element it holds

$$\|\nabla(\mathbf{v} - \mathbf{v}_h)\| + \|p - p_h\| \leq ch^2 \|\mathbf{f}\|_{H^1(\Omega)}, \tag{9.2.13}$$

or, for equal-order linear finite elements for pressure and velocity, the solution to the stabilized formulation

$$(\text{div } \mathbf{v}_h, \xi_h) + (\nabla \mathbf{v}_h, \nabla \phi_h) - (p_h, \nabla \cdot \phi_h) + h^2(\nabla p_h, \nabla \xi_h) = (\mathbf{f}_h, \phi_h) \quad \forall (\phi_h, \xi_h) \in V_h \times L_h \tag{9.2.14}$$

satisfies the estimate

$$\|\nabla(\mathbf{v} - \mathbf{v}_h)\| + \|p - p_h\| \leq ch\|\mathbf{f}\|. \tag{9.2.15}$$

We refer to the literature, e.g. the monography of John [138] for these and further aspects on the finite element approximations to the Stokes equations.

Having a saddle-point structure the Stokes system is not directly associated to an energy form. Instead we realize the *Deep Ritz* method by introducing a penalty term to enforce the divergence condition, e.g.

$$E_{\lambda,\alpha}(\mathbf{v}) := \frac{1}{2}\|\nabla \mathbf{v}\|^2 - (\mathbf{f}, \mathbf{v}) + \frac{\alpha}{2}\| \text{ div } \mathbf{v}\|^2 + \frac{\lambda}{2}|\mathbf{v}|^2_{\partial\Omega}, \tag{9.2.16}$$

where $\alpha, \lambda > 0$ are two parameters controlling the balance between minimizing the energy and satisfying the divergence constraint and the boundary values. The solution is characterized by the variational problem $\mathbf{v}_{\lambda,\alpha} \in H^1(\Omega)^2$

$$(\nabla \mathbf{v}_{\lambda,\alpha}, \nabla \phi) + \alpha(\text{div } \mathbf{v}_{\lambda,\alpha}, \text{div } \phi) + \lambda \langle \mathbf{v}_{\lambda,\alpha}, \phi \rangle_{\partial\Omega} = (\mathbf{f}, \phi) \quad \forall \phi \in H^1(\Omega). \tag{9.2.17}$$

This variational problem in term is corresponding to the following classical formulation which also reveals the disturbed boundary condition

$$-\Delta \mathbf{v}_{\lambda,\alpha} - \alpha \nabla \text{ div } \mathbf{v}_{\lambda,\alpha} = \mathbf{f} \text{ in } \Omega, \quad \lambda \mathbf{v}_{\lambda,\alpha} + \alpha \vec{n} \text{ div } \mathbf{v}_{\lambda,\alpha} - \partial_n \mathbf{v}_{\lambda,\alpha} = 0 \text{ on } \partial\Omega. \tag{9.2.18}$$

Hereby, similar to Lemma 9.1 we get

Lemma 9.2 (*Stokes model error*). *Let* $\mathbf{f} \in L^2(\Omega)^2$, $\lambda, \alpha \in \mathbb{R}$ *with* $\lambda, \alpha > 0$ *and* Ω *be such that the solutions* $(\mathbf{v}, p) \in \mathscr{V} \times \mathscr{L}$ *and* $\mathbf{v}_{\lambda,\alpha} \in H^1(\Omega)^2$ *to (9.2.11) and (9.2.17), respectively, satisfy* $\|\mathbf{v}\|_{H^2(\Omega)} + \|p\|_{H^1(\Omega)} \leq c_s\|\mathbf{f}\|$ *and* $\|u_{\lambda,\alpha}\|_{H^2(\Omega)} \leq c_s\|\mathbf{f}\|$. *It holds*

$$\|\nabla(\mathbf{v} - \mathbf{v}_{\lambda,\alpha})\| \leq c\Big(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{\alpha}}\Big)\|\mathbf{f}\|,$$

where $c > 0$ *depends on the domain* Ω *only.*

Proof. Due to similarity to Lemma 9.1 we just give a sketch, considering the adjoint $(\mathbf{z}, p) \in \mathscr{V} \times \mathscr{L}$

$$\frac{\nabla(\mathbf{v} - \mathbf{v}_{\lambda,\alpha})}{\|\nabla(\mathbf{v} - \mathbf{v}_{\lambda,\alpha})\|} = -\Delta \mathbf{z} - \nabla q, \quad \text{div } \mathbf{z} = 0,$$

which gives the error identity

$$\begin{aligned}\|\nabla(\mathbf{v} - \mathbf{v}_{\lambda,\alpha})\| &= (\nabla z, \nabla(\mathbf{v} - \mathbf{v}_{\lambda,\alpha})) + (q, \text{div}\,(\mathbf{v} - \mathbf{v}_{\lambda,\alpha})) + \langle \partial_n \mathbf{z} + q\vec{n}, \mathbf{v}_{\lambda,\alpha}\rangle_{\partial\Omega} \\ &\quad \underbrace{-\alpha(\text{div } \mathbf{v}_{\lambda,\alpha}, \text{div } \mathbf{z}) - \lambda\langle \mathbf{v}_{\lambda,\alpha}, \mathbf{z}\rangle_{\partial\Omega} - (p, \text{div } \mathbf{z})}_{=0} \\ &= \frac{1}{\lambda}\langle \partial_n z + q\vec{n}, \mathbf{v}_{\lambda,\alpha}\rangle_{\partial\Omega} - (\text{div } \mathbf{v}_{\lambda,\alpha}, q),\end{aligned} \tag{9.2.19}$$

to be estimated as

$$\|\nabla(\mathbf{v}-\mathbf{v}_{\lambda,\alpha})\| \le c\Big(\underbrace{\|\mathbf{z}\|_{H^2(\Omega)} + \|q\|_{H^1(\Omega)}}_{\le c_s}\Big)\|\mathbf{f}\|\Big(\|\operatorname{div}\mathbf{v}_{\lambda,\alpha}\| + |\mathbf{v}_{\lambda,\alpha}|_{\partial\Omega}\Big).$$

Diagonal testing of (9.2.17) then gives the claimed estimate

$$\|\nabla(\mathbf{v}-\mathbf{v}_{\lambda,\alpha})\| \le c\Big(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{\alpha}}\Big)\|\mathbf{f}\|.$$

□

For optimal scaling, the two parameters α and λ should be chosen similarly. This penalized energy minimization formulation does not produce an approximation to the pressure.

9.3 A Posteriori Error Estimation for Neural Network Solutions

9.3.1 The Dual Weighted Residual Method

We start by giving a very short description of the dual weighted residual method such as presented in [25, 26] for the most simple case of the Laplace problem $-\Delta u = f$ with homogeneous Dirichlet data $u = 0$. As introduced above, $\mathscr{V} = H_0^1(\Omega)$ and $V_h \subset \mathscr{V}$ is a discrete subspace.

Now, let $J : \mathscr{V} \to \mathbb{R}$ be a linear functional and let $z \in \mathscr{V}$ be the solution to the adjoint problem

$$-\Delta z = J \text{ in } \Omega, \quad z = 0 \text{ on } \partial\Omega, \tag{9.3.1}$$

in variational formulation given as

$$(\nabla v, \nabla z) = J(v) \quad \forall v \in \mathscr{V}. \tag{9.3.2}$$

This already gives the *primal error identity*

$$J(u-u_h) = (\nabla(u-u_h), \nabla z) = (f, z) - (\nabla u_h, \nabla z), \tag{9.3.3}$$

and, by using Galerkin orthogonality $(\nabla(u-u_h), \nabla v_h) = 0$ for all $v_h \in V_h$ the corresponding *dual error identity*

$$J(u-u_h) = (\nabla(u-u_h), \nabla(z-z_h)) = (\nabla u, \nabla(z-z_h)) = J(u) - (\nabla u, \nabla z_h), \tag{9.3.4}$$

where $z_h \in V_h \subset \mathscr{V}$ is the discrete solution to the adjoint problem.

Both simple error identities, (9.3.3) and (9.3.4) cannot be used in practice since the adjoint solution $z \in \mathscr{V}$ and the primal solution $u \in \mathscr{V}$ are not known. Applying the DWR method calls for an approximation of the primal or adjoint solutions in a subspace $V_{hh} \subset \mathscr{V}$ which is no subspace of the discrete space, e.g. $V_{hh} \not\subset V_h$ and to replace the error identities by

$$J(u-u_h) \approx (f, z_{hh}) - (\nabla u_h, \nabla z_{hh}) \approx J(u_{hh}) - (\nabla u_{hh}, \nabla z_h) \tag{9.3.5}$$

Various approaches are discussed in [26, Sec. 5] or [212, Sec. 3] and in general, the higher-order postprocessing of same-space approximations $u_h, v_h \in V_h$, e.g. by polynomial of double degree on a coarser mesh has been proven to be both reliable and efficient. Anyway, this reconstruction of higher order information is an approximation only such that the DWR method usually does not give a rigorous error bound but only an approximation to the error. This however is usually highly accurate.

The accuracy of the error estimator can be numerically validated by considering the *effectivity index*, e.g. the quotient of estimator value by real error, which in the case of the primal formulation reads

$$\mathrm{eff}_h := \frac{(F, z_{hh}) - (\nabla u_h, \nabla z_{hh})}{J(u - u_h)}, \tag{9.3.6}$$

which of course requires a guess of the *true error* $J(u - u_h)$. For linear elliptic problems one usually observes effectiveness going to 1 (for $h \to 0$).

Besides error estimation the DWR method can be used for *mesh adaptivity*, e.g. for steering an adaptive computation. This requires the *localization* of the error estimator, i.e. the identification of local values η_T which attribute parts of the error to the element T of the mesh (or to each node, edge, etc.). We refer to [212] for details. The basic step is to use Galerkin orthogonality and to introduce interpolations $I_h z \in V_h$ and $I_h u \in V_h$ to the error identities (9.3.3), (9.3.4) to derive at

$$J(u - u_h) = (f, z - I_h z) - (\nabla u_h, \nabla(z - I_h z)) = J(u - I_h u) - (\nabla z_h, \nabla(u - I_h u)), \tag{9.3.7}$$

followed by the approximation of the *weights* $z - I_h z \approx z_{hh} - z_h$ and $u - I_h u \approx u_{hh} - u_h$ as mentioned above. Finally, local error indicators

$$J(u - u_h) \approx \sum_{i=1}^{N} \eta_i(u_h, z_h)$$

are derived and their absolute values $|\eta_i(u_h, z_h)|$ are considered for refinement. A localization is effective, if the localization index

$$\mathrm{loc}_h := \frac{\sum_{i=1}^{N} |\eta_i(u_h, z_h)|}{J(u - u_h)}, \tag{9.3.8}$$

is close to 1. (A good localization must allow for introducing the absolute values, i.e. it must cancel out local oscillations in the error estimator). We refer to [212, Sec. 4] for details.

The DWR method is easily extended to nonlinear problems, to systems of differential equations and to time dependent problems. All this, further extensions and various applications have already been demonstrated by Becker and Rannacher [26]. The fundamental problem that is still open is a reliable and efficient procedure for approximating the weights and, in the case of nonlinear problems, bounds on a higher order remainder that must usually be dropped.

9.3.2 Estimating the Network Error for the Laplace Equation

Within this framework we now aim at estimating the functional error of the neural network solution $u_{\mathcal{N}} \in V_{\mathcal{N}}$ obtained with the *Deep Ritz* approach. Since the network minimizer $u_{\mathcal{N}} \notin \mathcal{V}$ does not satisfy the Dirichlet condition $u = 0$, a consistency term remains when multiplying the error $u - u_{\mathcal{N}}$ with the adjoint problem

$$J(u - u_{\mathcal{N}}) = (\nabla(u - u_{\mathcal{N}}), \nabla z) + \langle \partial_n z, u_{\mathcal{N}} \rangle. \tag{9.3.9}$$

The remaining steps are as in (9.3.3) and we derive the error identity

$$J(u-u_{\mathscr{N}}) = (f,z) - (\nabla u_{\mathscr{N}}, \nabla z) + \langle \partial_n z, u_{\mathscr{N}}\rangle. \tag{9.3.10}$$

Again, we must approximate $z \in \mathscr{V}$ by a discrete solution which is accurate, efficiently achievable and which does not fall into the vicinity of Galerkin orthogonality, which, in the case of the neural network error $u - u_{\mathscr{N}}$ imposes the condition $z_h \notin V_{\mathscr{N}}$. The neural network spaces introduced in Section 9.4 have no similarity to local finite element spaces. Hence, we will approximate the adjoint solution in as coarse as possible finite element space $z_h \in V_h \subset \mathscr{V}$ and approximate

$$\eta(u_{\mathscr{N}}, z_h) = (f, z_h)_\Omega - (\nabla u_{\mathscr{N}}, \nabla z_h) + \langle \partial_n z_h, u_{\mathscr{N}}\rangle. \tag{9.3.11}$$

This error estimator is efficiently evaluated on the finite element mesh using a numerical quadrature rule within the domain and along the boundaries. The effectivity of the estimate is measured by means of the effectivity index 9.3.6. We finally note that the error estimator 9.3.11 is not specific to the *Deep Ritz* method, instead it could also be used in the context of *DeepXDE*, see [154], or for any other approximation technique that yields a H^1-conforming solution.

9.3.3 Estimating the Network Error for the Stokes Equations

The estimate can directly be transferred to the Stokes equations, where we approximate the solution based on the penalized energy form (9.2.16) such as described in Section 9.2.3. For a goal functional $J : H_0^1(\Omega)^2 \to \mathbb{R}$ we introduce the adjoint solution

$$\operatorname{div}\,\mathbf{z} = 0,\; -\Delta\mathbf{z} - \nabla q = J \text{ in } \Omega, \quad \mathbf{z} = 0 \text{ on } \partial\Omega. \tag{9.3.12}$$

The error identity for the network solution $\mathbf{v}_{\mathscr{N}}$ minimizing (9.2.16) is then derived as

$$\begin{aligned} J(\mathbf{v}-\mathbf{v}_{\mathscr{N}}) &= \Big(\nabla\mathbf{z}, \nabla(\mathbf{v}-\mathbf{v}_{\mathscr{N}})\Big) + \Big(q, \operatorname{div}(\mathbf{v}-\mathbf{v}_{\mathscr{N}})\Big) - \langle \partial_n\mathbf{z} + q\vec{n}, \mathbf{v}-\mathbf{v}_{\mathscr{N}}\rangle_{\partial\Omega} \\ &= (\mathbf{f},\mathbf{z}) - \Big(\nabla\mathbf{v}_{\mathscr{N}}, \nabla\mathbf{z}\Big) - \Big(\operatorname{div}\,\mathbf{v}_{\mathscr{N}}, q\Big) + \langle \mathbf{v}_{\mathscr{N}}, \partial_n\mathbf{z} + q\vec{n}\rangle_{\partial\Omega}. \end{aligned}$$

To evaluate and approximate this error identity we compute a coarse finite element approximation $(\mathbf{z}_h, q_h) \in V_h \times L_h$

$$-(\operatorname{div}\,\mathbf{z}_h, \xi_h) + (\nabla\mathbf{z}_h, \nabla\phi_h) + (q_h, \operatorname{div}\,\phi_h) = J(\phi_h) \quad \forall(\phi_h,\xi_h) \in V_h \times L_h,$$

and define the Stokes error estimate as

$$\eta(\mathbf{v}_{\mathscr{N}}, \mathbf{z}_h, q_h) := (\mathbf{f},\mathbf{z}_h) - \Big(\nabla\mathbf{v}_{\mathscr{N}}, \nabla\mathbf{z}_h\Big) - \Big(\operatorname{div}\,\mathbf{v}_{\mathscr{N}}, q_h\Big) + \langle \partial_n\mathbf{z}_h + q_h\vec{n}, \mathbf{v}_{\mathscr{N}}\rangle_{\partial\Omega}. \tag{9.3.13}$$

9.4 Network Architecture and Training

Let us recall from the introduction, that the network is considered as function $\mathscr{N} : \Omega \to \mathbb{R}^c$. More precisely, $\mathscr{N}$ is a L-layer neural network with N_l neurons in the l-th layer. We denote the weight matrix and bias vector in l-th layer by $\mathbf{W}^l \in \mathbb{R}^{N_l \times N_{l-1}}$ and $\mathbf{b}^l \in \mathbb{R}^{N_l}$, respectively. Given an activation function σ, which is applied elementwise.

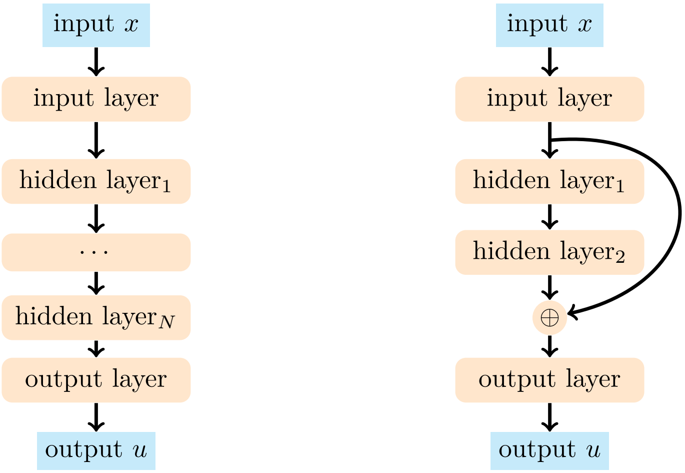


Figure 9.1: Feed Forward Neural Network and Residual Neural Network architectures.

We consider two different architectures, feed forward neural network (FFNet):

$$\begin{aligned}
&\mathscr{N}^0(x) = x \in \Omega,\\
&\mathscr{N}^l(x) = \sigma(\mathbf{W}^l \mathscr{N}^{l-1}(x) + \mathbf{b}^l) \in \mathbb{R}^{N_l}, \text{ for } 1 \le l \le L\\
&\mathscr{N}^{L+1}(x) = \mathbf{W}^l \mathscr{N}^{L-1}(x) + \mathbf{b}^l \in \mathbb{R}^c,
\end{aligned}$$

and residual neural network (ResNet):

$$\begin{aligned}
&\mathscr{N}^0(x) = x \in \Omega,\\
&\mathscr{N}^k(x) = \sigma\left(\mathbf{W}^l\left(\sigma(\mathbf{W}^{l-1}\mathscr{N}^{l-2}(x) + \mathbf{b}^{l-1})\right) + \mathbf{b}^l\right) + \mathscr{N}^{k-2}(x) \in \mathbb{R}^{N_l}, \text{ for } 1 \le l \le L/2\\
&\mathscr{N}^{L+1}(x) = \mathbf{W}^l \mathscr{N}^{L-1}(x) + \mathbf{b}^l \in \mathbb{R}^c.
\end{aligned}$$

We the employed notation $\mathscr{N}^0$ is an input layer with $l_0 = d$ and $\mathscr{N}^{L+1}$ is an output layer with $N_{l+1} = c$. All hidden layers are of the same size H, i.e. $l = H$ for $1 \le l \le L$

Remark 9.3. In [88] the authors used a ResNet architecture, however we observed in that in practise even simpler neural networks gives good results.

9.4.1 Training

In this section we present some insights into the training process for the *Deep Ritz* method applied to Laplace problem on a L-shaped domain. We refer to Section 9.5.1 for a precise definition of the test case. Here, we study the effect of the network architecture, i.e. a Feed Forward Neural Network (FFNet) and a Residual Neural Network (ResNet) on the training.

In Figure 9.2 we present training progress for residual networks of various size. In the left sketch we show the loss function, i.e. the value of the penalized and approximated energy functional (9.2.9) and on the right, we show the L^2 error of the resulting approximations $\|u - u_{\mathscr{N},\mathbf{u}}\|$ during training. In general the bigger the network the fever epochs are needed to reach a certain error. However, this is not always the case. We observe a certain threshold for

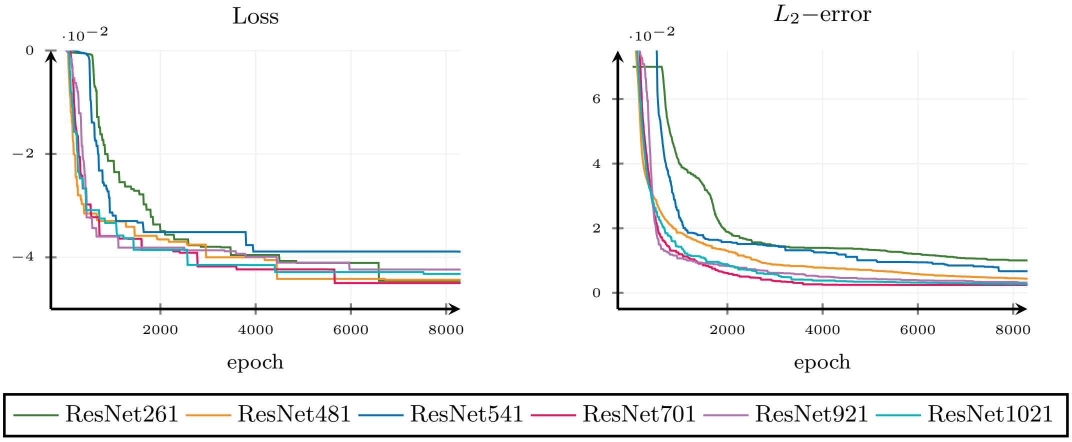


Figure 9.2: Training progress for networks of various size.

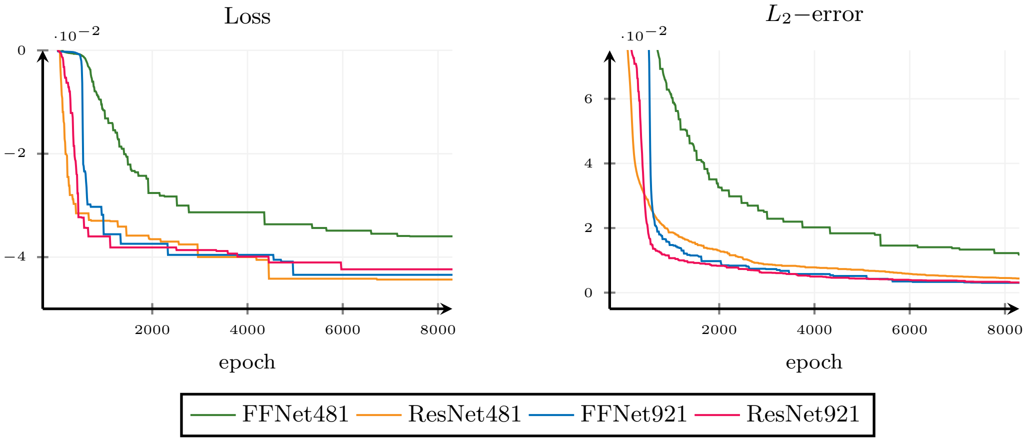


Figure 9.3: Training progress for networks of similar size.

the number of network parameters above which increasing the size of the network does not improve the solution. In general, the slope of the loss function is similar to the progress of the L^2-error. Naturally, once low loss levels are reached, larger networks are able to yield better approximations.

In Figure 9.3 we present the training and approximation progress of networks with Feed Forward and Residual architectures and same sizes. To be precise, for each architecture we consider a small network with 481 parameters and a larger one with 921 parameters. The residual network is faster to train, but this discrepancy gets smaller for larger networks. The advantage of residual networks was already mentioned by E and Yu [88].

The above consideration show the need of for a quality measure of the solution that works across architectures and training methods. In the following section we will present numerical examples that demonstrate the usability of the error estimator for controlling the approximation error during training. This estimate can then be used as stopping criterion once a sufficiently low error level is reached.

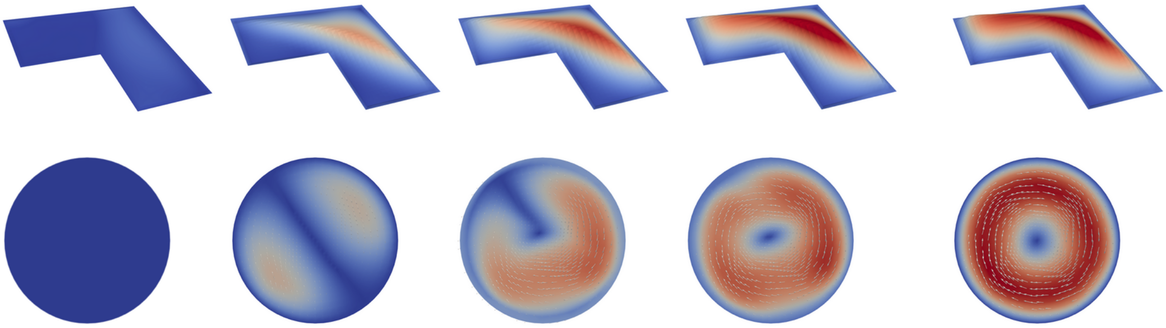


Figure 9.4: Solution of Laplace (top) and Stokes (bottom) during training progress, epochs=$\{0, 2500, 5000, 7500, 10000\}$ (Laplace), epochs=$\{0, 2000, 5000, 10000, 25000\}$ Stokes.

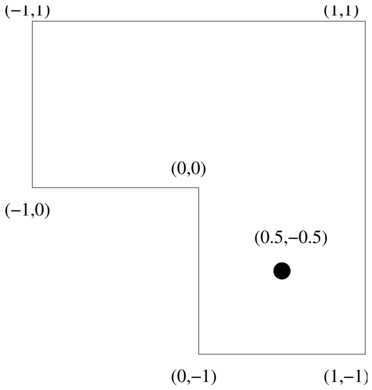


Figure 9.5: L-shaped domain and evaluation point $x_a = (0.5, -0.5)$ to define test case 1. The Laplace problem is solved with homogeneous Dirichlet data and constant right hand side $f = 1$ such that a corner singularity evolves.

9.5 Numerical Examples

In this section we will consider two test cases, the Laplace equation on a L-shaped domain that has already been considered previously and the Stokes equations on a disc. We start by showing the neural network *Deep Ritz* solutions to both test cases in Figure 9.4. We give the approximation at different stages of the training procedure, at the beginning and after a certain number of epochs (indicated in the figure caption).

9.5.1 Test Case 1. Laplace Equation

As first test case we consider the Poisson equation on the L-shaped domain $\Omega_L = (-1,1)^2 \setminus [0,1]^2$ shown in Figure 9.5. The functional of interest is to evaluate the solution in the point $x_a = (0.5, -0.5)$

$$-\Delta u = 1 \text{ in } \Omega_L, \quad u = 0 \text{ on } \partial\Omega_L, \quad J(u) = u(x_a). \tag{9.5.1}$$

We note that $J \notin H^{-1}(\Omega_L)$ is not an admissible functional. The point evaluation should be replaced by averaging over a small neighbourhood of the point x_a. On the other hand it is well documented that the non-regularized functional gives optimal performance in the context of the

	L=1			L=2			L=4		
epoch	error	estimate	eff$_h$	error	estimate	eff$_h$	error	estimate	eff$_h$
500	-0.008480	-0.014786	0.57	0.010748	0.009397	1.14	0.018578	0.019248	0.96
1000	-0.003403	-0.006700	0.51	0.005855	0.007324	0.80	-0.004695	-0.004719	0.99
1500	-0.008228	-0.008022	1.03	-0.007460	-0.003477	2.15	-0.009242	-0.009126	1.01
2000	-0.006424	-0.004946	1.23	-0.009138	-0.011058	0.83	-0.006117	-0.006393	0.96

Table 9.1: The values of error and estimator with effectivity index eff$_h$, see (9.3.6), for different refinement levels of dual solution $h = 2^{-L}$.

dual weighted residual method, see [26, 212]. For comparison, we first determine a reference value by resolved finite element simulations on very fine meshes. We identify it as

$$J_{ref} = 0.1023612 \pm 0.001953125.$$

First we demonstrate the performance of the DWR estimator

$$\eta(u_{\mathcal{N}}, z_h) = (f, z_h)_\Omega - (\nabla u_{\mathcal{N}}, \nabla z_h)_\Omega + \langle \partial_n z_h, u_{\mathcal{N}} \rangle_{\partial\Omega}$$

as presented in Section 9.3.2. The adjoint solution z_h will be computed as finite element approximation on very coarse meshes. Training results and estimator values are shown for neural network solutions obtained with the *Deep Ritz* method and using the strong formulation. Both network architectures of FFNet type and of ResNet type are considered. The complete set of parameters is summarized as follows:

- FFNet : $H = 20$, $L = 4$, $\sigma(x) = \text{ELU}(x)$,
- ResNet: $H = 20$, $L = 2$, $\sigma(x) = \max(x^3, 0)$.

Note, since ResNet block consists of two layers, both architectures have same number of parameters and the Exponential Linear Unit (ELU) is defined as

$$\text{ELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ e^x - 1 & \text{if } x < 0. \end{cases}$$

For the *Deep Ritz* approach and the strong formulation, all gradients are computed both with automatic differentiation and finite difference approximation, respectively. We perform 8000 epochs and the estimator is evaluated every 100 epochs, see Figure 9.6.

One can observe that independently of the applied method, the estimator follows the error. Consequently we study the dependence on the coarse mesh size $h \in \{0.5, 0.25, 0.0625\}$ used to approximate the adjoint solution and show the results in Figure 9.7. Increasing the level of refinement improves the exactness of the estimator. Note however, that results are relatively good even for extremely coarse meshes and the evaluation of estimator is computationally cheap. The choice $h = 0.5$ corresponds to only 12 quadrilateral elements, the finest mesh with $h = 0.0625$ corresponds to just 768 elements. Further, many degrees of freedom reside on the boundary of the domain such that the number of unknowns to approximate the adjoint solution ranges from 5 on the coarsest mesh to 640 on the finest mesh and from the results we observe that the intermediate mesh with $h = 0.25$ comprising 16 unknowns is sufficiently accurate.

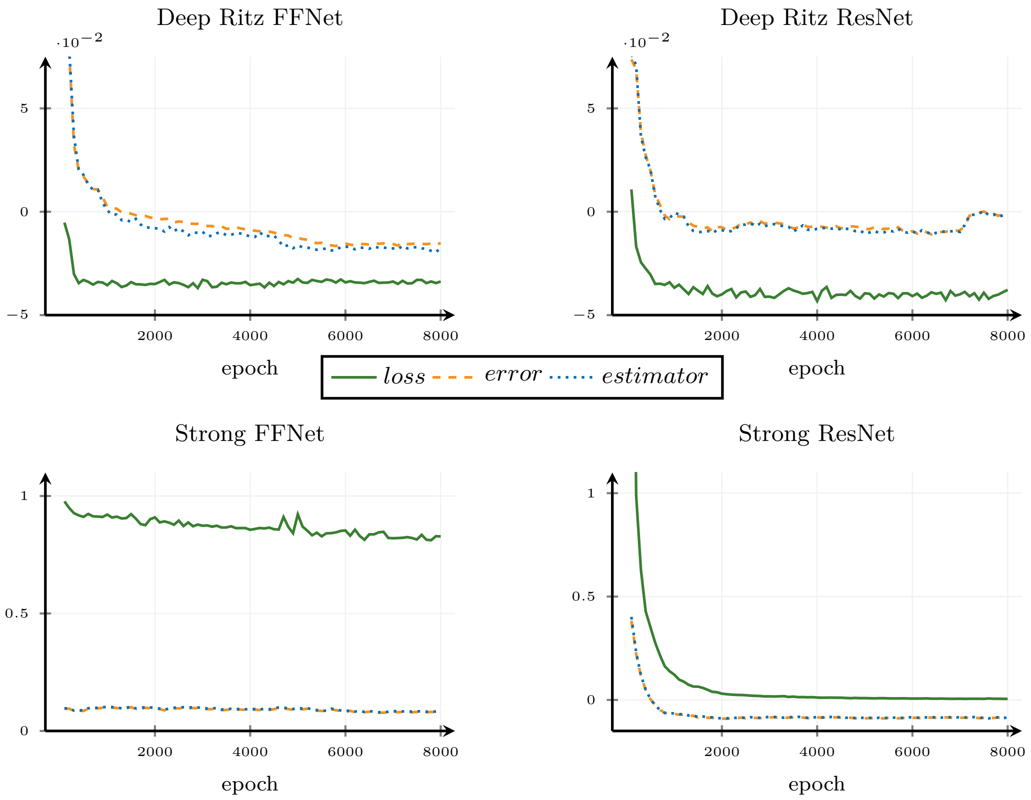


Figure 9.6: Loss, error and estimator for different network architectures and Loss functions.

9.5.2 Test Case 2. Stokes Equations

For the second test case we consider the Stokes equations on a unit circle $\Omega = B((0,0),1)$. We prescribe an analytical solution for comparison with the neural network approximation given by

$$\mathbf{v}(x,y) = \cos\left(\frac{\pi}{2}(x^2+y^2)\right)\begin{pmatrix} y \\ -x \end{pmatrix},$$

and compute the corresponding forcing term as

$$\mathbf{f}(x,y) = \pi\cos\left(\frac{\pi}{2}(x^2+y^2)\right)\begin{pmatrix} y(x^2+y^2)\pi + 4(y-x)\tan\left(\frac{\pi}{2}(x^2+y^2)\right) \\ -x(x^2+y^2)\pi - \ 4(x+y)\tan\left(\frac{\pi}{2}(x^2+y^2)\right) \end{pmatrix}.$$

The functional of interest $J(\mathbf{v})$ is an integral of a y-component of the velocity on a line segment $[0,1]$

$$J(\mathbf{v}) := \int_0^1 \mathbf{v}_y(x,0)\mathrm{d}x, \quad J_{ref} = -\frac{1}{\pi}.$$

In Figure 9.8 we present the loss function and the functional error as well as the error estimator. The training of the Deep Ritz method is performed for 25000 epochs, with the Feedforward Neural Network (FFNet $d = 2$, $c = 2$, $H = 10$, $L = 20$, $\sigma(x) = \mathrm{ELU}(x)$). The adjoint Stokes problem is approximated with equal order finite elements using pressure stabilization on a coarse mesh level $L = 3$ that corresponds to $h \approx 0.04375$. The results for some selected epochs together with effectivity index are summarized in Table 9.2.

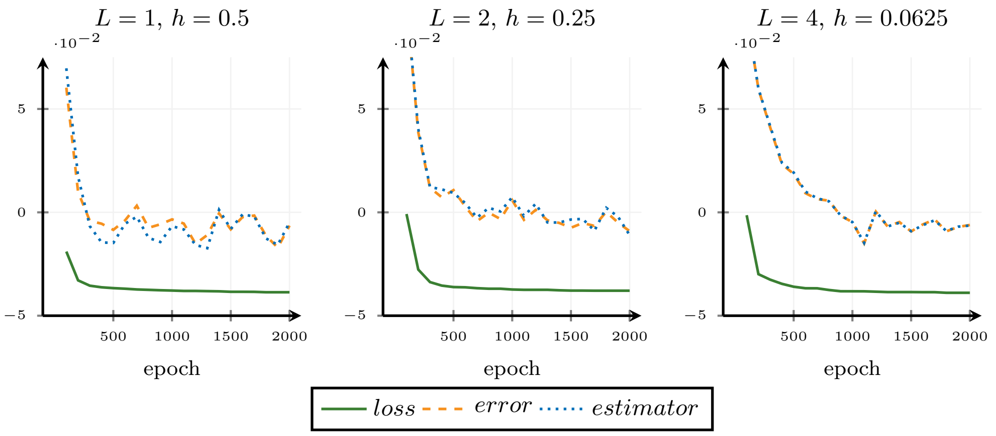


Figure 9.7: Loss, error and estimator for different refinement levels of dual solution $h = 2^{-L}$.

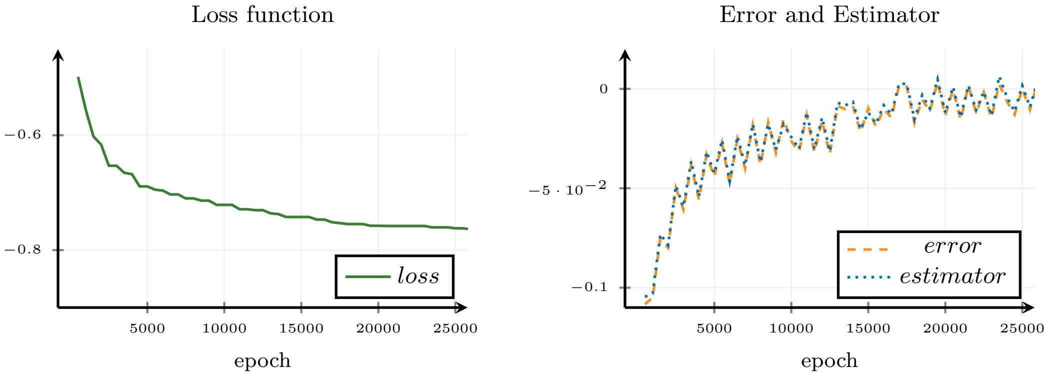


Figure 9.8: Loss, error and estimator for Stokes problem.

epoch	error	estimate	eff_h
5000	-0.0442546	-0.0425343	1.04
10000	-0.0262763	-0.0241991	1.08
15000	-0.0096646	-0.0118802	0.81
20000	-0.0134781	-0.0121668	1.10

Table 9.2: The values of error and estimator with effectivity index eff_h, compare (9.3.6).

CHAPTER 10

Conclusions and Outlook

We have presented new mathematical models for several continuum mechanics problems with their analysis and numerical treatment. The applications include the Navier-Stokes system, Fluid-structure interactions, the Euler system, and collective dynamics.

We now summarise the most significant achievements in this thesis and discuss several open problems relating to the presented results, which we consider interesting for future research.

10.1 Theoretical and Practical Aspects of Blood Flow Simulations

In Chapter 2, we have demonstrated that small boundary perturbations have a crucial impact on the result of the finite element simulations. The developed error estimates are linear with respect to the maximal distance between the real and the approximated domains, cf. Theorem 2.6 for the Laplace problem. We have illustrated the sharp nature of this bound in the computations performed in Section 2.4.

Particularly, in the case of first and second-order approximation we observe how the relation between the mesh size and aforementioned maximal distance impact the resulting L^2- and H^1-errors. The same behaviour has been demonstrated numerically for the Stokes system.

In practice, we do not have control over the accuracy of the domain reconstruction. We have shown that it is worth taking into account the geometric uncertainty when deciding on the mesh-size in order to avoid unnecessary computational effort.

In Chapter 3, we have studied a prototypical geometry that represents a large, curved artery. Three different configurations are considered: a healthy vessel and two cases where stenosis is given in the curved area of the vessel. We consider the stenosis centred around the vessel centerline and a shifted variant where it is concentrated on the inside of the curve. The blood flow is driven by a time-dependent flow rate with clinically relevant values. For all configurations, we study the effect of the elasticity in the vessel walls, i.e. we compare a pure Navier-Stokes simulation with a fully coupled fluid-structure interaction system.

Three different indicators that are relevant in clinical decision making are investigated: the distribution of the wall shear stress that is responsible for stenosis growth and risk of rupture, the computational fractional flow reserve, which is used to estimate the severity of stenosis and the amplitude of the pressure oscillation which also measures the severity of stenosis. In all cases, we observe that the simple Navier-Stokes model cannot depict the effect of the plaque. In particular, the pressure lines are nearly identical for all three geometrical configurations. The FSI model is, however able to replicate clinical observations. For instance, the energy, in terms of pressure oscillations, is fully preserved throughout the curved region, if elasticity of the vessel walls is taken into account.

Although the study covers only an idealised geometry and boundary conditions, the proposed regime corresponds to a physiological blood flow. Therefore, we stress that the compliance of the vessel has a significant impact on clinical hemodynamical factors. In medical practice, one has to be aware of the difference in the results from CFD with FSI and NS models and the limitations of the latter.

Chapter 4 presented a project dedicated to simulating blood flow in the cephalic arch for a patient with arteriovenous fistula. The algorithms developed within the project combine medical diagnostics, imaging methods, and computer simulations. This enables the development of new treatment methods that can contribute to individual therapy.

This research topic will be further investigated in theoretical and practical directions. On the one hand, we will develop finite element estimates on the perturbed boundary, e.g. for hemodynamical factors as the wall shear stress. On the other, knowledge gained in the project enables further development of computer models towards a commercially usable digital twin.

10.2 Variable Congestion in Two-Phase Compressible/Incompressible Flows

In Chapter 5 we analysed the two-phase macroscopic model that can be used to model the congestions in a large group of individuals in a bounded area. The motion of the individuals is described by the averaged quantities: the density of the individuals, the mean velocity, and the threshold density by means of the fluid equations. Similar models have been intensively studied recently in the context of flows through an irregular channel, modelling the tumour growth or description of the granular media.

The model encompasses the flow in the uncongested regime (compressible) and the congested one (incompressible), with the free boundary separating the two phases. The congested regime appears when the density in the uncongested regime achieves a threshold value that describes the comfort zone of individuals. This quantity is prescribed initially and transported along with the flow.

We showed that this system can be approximated by the fully compressible Navier-Stokes system with a singular pressure, supplemented with the transport equation for the congestion density. For this model, we proved that there exists a sequence of weak solutions that converges to the weak solution of the two-phase flow with the free boundary. We also presented the application of this approximation for the purposes of numerical simulations in the one-dimensional domain.

To the best of our knowledge, it is the first analytical and numerical result for the two-phase compressible/incompressible flow with the barrier that can depend both on the space and time variable.

In Chapter 6, we presented the numerical simulation of the Euler system with a singular pressure modelling variable congestion. As the stiffness of the pressure increases (ε tends to 0), the model tends to a free boundary transition between compressible (non-congested) and incompressible (congested) dynamics.

To numerically simulate the asymptotic dynamics, we propose an asymptotic preserving scheme based on a conservative formulation of the system. We also propose a second-order accuracy extension of the scheme. The one-dimensional solutions to Riemann test cases are studied together with their asymptotic limits that validate the code. We compare the results with those obtained with the scheme proposed in Chapter 5. This latter scheme enables to better approximate the congestion density at the contact wave as soon as we use high accuracy in its advection. On the other hand, the former scheme seems to better preserve the maximum principle on that variable. In two-dimensional simulations, we finally show the influence of this variable congestion density on the dynamics and show that the model exhibits stop-and-go behavior.

The two schemes generate oscillations in momentum variable at discontinuities between congested and non-congested domains, in particular in the case of the second-order accuracy schemes. Specific methods should be designed to cure this artefact. Another direction of improvement, that will be addressed in future work, concerns the treatment of the vacuum regions by the numerical scheme.

Future research is to extend the considered model by non-local interaction kernels that model repulsive-attractive and alignment forces. Moreover, the maximal density constraint in the form of the singular pressure plays a similar role as the yield stress threshold in plasticity. Thus, we aim to develop a new method based on singular pressure to capture perfect plasticity.

10.3 Density Induced Consensus Protocol

In Chapter 7 we proposed a model for collective behavior, where the agents interact only if they have a sufficiently large number of neighbours. This model is inspired by the Cucke-Smale model. We investigated the impact of the proposed interaction rule and compared it with previously proposed models. Particularly, the density-induced interaction produces non-symmetric, short-range interactions, and it allows for the existence of highly influential individuals, where just a single individual can change the behaviour of an entire group. Interestingly, this last effect is not imposed in the model but arises naturally from the density-induced interaction.

We studied the existence of solutions of the proposed model, the stability of dense clusters and the convergence to consensus under some assumption of densely packed clusters. We also investigated possible ways in which an individual can influence a whole cluster and presented numerical comparisons with the short-range Cucker-Smale model, the model of q-closest neighbours and the general Cucker-Smale model.

Chapter 8 explored the relation between density-based data segmentation and collective behavior, introducing a variant of the density based clustering. We provided a rigorous mathematical foundation for the method as well as an illustration in the case of color image segmentation. Various aspects of the algorithm were discussed: influence of its parameters, stopping time and numerical complexity. In particular, we achieved a linear average numerical complexity with parameters that led to the emergence of an appropriate number of clusters. The key point is the observation that an evolution of the date according to a density-based first-order ODE system has a low computational cost and breaks unreasonably large clusters that would be obtained using other methods.

Future research will be dedicated to explore connections between collective dynamics and clustering algorithms or more general machine learning. In particular, we will further develop a class of unsupervised data segmentation methods based on collective dynamics.

Moreover, we will explore mathematical similarities between residual neural networks and kinetic theory.

10.4 Dual Weight Residual Error Estimates for Neural Network Solutions of Partial Differential Equations

Based on the dual weighted residual method, the error estimator of neural network approximations of PDEs is developed. We derive the estimator for a functional of interest and demonstrate its performance for the Laplace and the Stokes equations. The method is independent of the design of the neural network and the training procedure. The evaluation on a very coarse mesh already shows very good accuracy, such that little computational overhead is introduced. The estimator can be used as a simple and an accurate stopping criterion during the training process. As a result of this, we gain a first validation of the neural network approximation, and the error controlled training helps to reduce the computational effort by avoiding excessive training epochs.

Besides error estimation, the DWR method can be used for mesh adaptivity, e.g. for steering an adaptive computation. This requires the localization of the error estimator, i.e., identifying local values of the interest functional that attribute parts of the error to the element of the mesh (or to each node, edge, etc.).

In future work, we aim to develop mathematical foundations and implement numerical techniques that combine classical discretization schemes with neural networks for simulating differential equations while conserving the physical structure of the underlying problem.

Bibliography

[1] *ADAN WEB*. URL: http://hemolab.lncc.br/adan-web/ (visited on 30/01/2020).

[2] C. C. Aggarwal and C. K. Reddy. *Data Clustering: Algorithms and Applications*. 1st. Chapman & Hall/CRC, 2013. ISBN: 1466558210.

[3] S. J. M. Akherat, K. Cassel, M. Boghosian, M. Hammes and F. Coe. 'A Predictive Framework to Elucidate Venous Stenosis: Cfd & Shape Optimization'. In: *Computer Methods in Applied Mechanics and Engineering* 321.Supplement C (2017), pp. 46–69. DOI: 10.1016/j.cma.2017.03.036.

[4] G. Albi, N. Bellomo, L. Fermo, S.-Y. Ha, J. Kim, L. Pareschi, D. Poyato and J. Soler. 'Vehicular Traffic, Crowds, and Swarms: From Kinetic Theory and Multiscale Methods to Applications and Research Perspectives'. In: *Mathematical Models and Methods in Applied Sciences* 29.10 (2019), pp. 1901–2005. DOI: 10.1142/S0218202519500374.

[5] G. Albi, M. Herty and L. Pareschi. 'Kinetic Description of Optimal Control Problems and Applications to Opinion Consensus'. In: *Commun. Math. Sci.* 13.6 (2015), pp. 1407–1429. DOI: 10.4310/CMS.2015.v13.n6.a3.

[6] G. Albi, L. Pareschi, G. Toscani and M. Zanella. 'Recent Advances in Opinion Modeling: Control and Social Influence'. In: *Active particles. Vol. 1. Advances in theory, models, and applications*. Model. Simul. Sci. Eng. Technol. Birkhäuser/Springer, Cham, 2017, pp. 49–98.

[7] G. Allaire and C. Dapogny. 'A Deterministic Approximation Method in Shape Optimization under Random Uncertainties'. en. In: *Journal of computational mathematics* 1 (2015), pp. 83–143. DOI: 10.5802/smai-jcm.5.

[8] M. P. Allen and D. J. Tildesley. *Computer Simulation of Liquids*. Oxford University Press, Nov. 2017. DOI: 10.1093/oso/9780198803195.001.0001.

[9] M. Alswaitti, M. Albughdadi and N. A. M. Isa. 'Density-based Particle Swarm Optimization Algorithm for Data Clustering'. In: *Expert Systems with Applications* 91 (Jan. 2018), pp. 170–186. DOI: 10.1016/j.eswa.2017.08.050.

[10] C. Anitescu, E. Atroshchenko, N. Alajlan and T. Rabczuk. 'Artificial Neural Network Methods for the Solution of Second Order Boundary Value Problems'. In: *Computers, Materials & Continua* 59.1 (2019), pp. 345–359. DOI: 10.32604/cmc.2019.06641.

[11] M. Ankerst, M. M. Breunig, H.-P. Kriegel and J. Sander. 'OPTICS'. In: *ACM SIGMOD Record* 28.2 (June 1999), pp. 49–60. DOI: 10.1145/304181.304187.

[12] E. Aulisa, S. Bna and G. Bornia. 'A Monolithic ALE Newton-Krylov Solver with Multigrid-richardson-schwarz Preconditioning for Incompressible Fluid-structure Interaction'. In: *Computers & Fluids* 174 (2018), pp. 213–228.

[13] A. Aydoğdu, M. Caponigro, S. McQuade, B. Piccoli, N. Pouradier Duteil, F. Rossi and E. Trélat. 'Interaction Network, State Space, and Control in Social Dynamics'. In: *Active particles. Vol. 1. Advances in theory, models, and applications*. Model. Simul. Sci. Eng. Technol. Birkhäuser/Springer, Cham, 2017, pp. 99–140.

[14] I. Babuška and J. Chleboun. 'Effects of Uncertainties in the Domain on the Solution of Neumann Boundary Value Problems in Two Spatial Dimensions'. In: *Mathematics of Computation* 71.240 (2002), pp. 1339–1370. URL: http://www.jstor.org/stable/4099954.

[15] I. Babuška and J. Chleboun. 'Effects of Uncertainties in the Domain on the Solution of Dirichlet Boundary Value Problems'. In: *Numerische Mathematik* 93.4 (2003), pp. 583–610. DOI: 10.1007/s002110200400.

[16] H.-O. Bae, S.-y. Cho, S.-h. Lee, J. Yoo and S.-B. Yun. 'A Particle Model for the Herding Phenomena Induced by Dynamic Market Signals'. In: *J. Stat. Phys.* 177.2 (2019), pp. 365–398. DOI: 10.1007/s10955-019-02371-8.

[17] H.-O. Bae, S.-Y. Ha, Y. Kim, S.-H. Lee, H. Lim and J. Yoo. 'A Mathematical Model for Volatility Flocking with a Regime Switching Mechanism in a Stock Market'. In: *Math. Models Methods Appl. Sci.* 25.7 (2015), pp. 1299–1335. DOI: 10.1142/S0218202515500335.

[18] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale and V. Zdravkovic. 'Interaction Ruling Animal Collective Behavior Depends on Topological Rather Than Metric Distance: Evidence from a Field Study'. In: *Proceedings of the National Academy of Sciences* 105.4 (2008), pp. 1232–1237. doi: 10.1073/pnas.0711437105.

[19] D. Balzani, S. Deparis, S. Fausten, D. Forti, A. Heinlein, A. Klawonn, A. Quarteroni, O. Rheinbach and J. Schröder. 'Numerical Modeling of Fluid–structure Interaction in Arteries with Anisotropic Polyconvex Hyperelastic and Anisotropic Viscoelastic Material Models at Finite Strains'. In: *International Journal for Numerical Methods in Biomedical Engineering* 32.10 (2016). e02756 cnm.2756, e02756. doi: 10.1002/cnm.2756.

[20] G. R. Barrenechea and C. González. 'A Stabilized Finite Element Method for a Fictitious Domain Problem Allowing Small Inclusions'. In: *Numerical Methods for Partial Differential Equations* 34.1 (Aug. 2017), pp. 167–183. doi: 10.1002/num.22190.

[21] A. Barron. 'Universal Approximation Bounds for Superpositions of a Sigmoidal Function'. English. In: *IEEE Transactions on Information Theory* 39.3 (1993), pp. 930–945. doi: 10.1109/18.256500.

[22] Y. Bazilevs, M.-C. Hsu, Y. Zhang, W. Wang, T. Kvamsdal, S. Hentschel and J. Isaksen. 'Computational Vascular Fluid–structure Interaction: Methodology and Application to Cerebral Aneurysms'. In: *Biomechanics and modeling in mechanobiology* 9.4 (2010), pp. 481–498.

[23] Y. Bazilevs, V. M. Calo, T. J. Hughes and Y. Zhang. 'Isogeometric Fluid-structure Interaction: Theory, Algorithms, and Computations'. In: *Computational mechanics* 43.1 (2008), pp. 3–37.

[24] R. Becker and M. Braack. 'A Finite Element Pressure Gradient Stabilization for the Stokes Equations Based on Local Projections'. In: *Calcolo* 38.4 (2001), pp. 173–199.

[25] R. Becker and R. Rannacher. 'Weighted a Posteriori Error Control in FE Methods'. In: *ENUMATH'97*. Ed. by e. a. H. G. Bock. World Sci. Publ., Singapore, 1995.

[26] R. Becker and R. Rannacher. 'An Optimal Control Approach to a Posteriori Error Estimation in Finite Element Methods'. In: *Acta Numerica* (2001), pp. 1–225.

[27] N. Bellomo and C. Dogbe. 'On the Modelling Crowd Dynamics from Scaling to Hyperbolic Macroscopic Models'. In: *Mathematical Models and Methods in Applied Sciences* 18.supp01 (2008), pp. 1317–1345. doi: 10.1142/S0218202508003054.

[28] N. Bellomo and C. Dogbe. 'On the Modeling of Traffic and Crowds: A Survey of Models, Speculations, and Perspectives'. In: *SIAM review* 53.3 (2011), pp. 409–463.

[29] J. Berg and K. Nyström. 'A Unified Deep Artificial Neural Network Approach to Partial Differential Equations in Complex Geometries'. In: *Neurocomputing* 317 (Nov. 2018), pp. 28–41. doi: 10.1016/j.neucom.2018.06.056.

[30] C. Bernardi. 'Optimal Finite-element Interpolation on Curved Domains'. In: *SIAM Journal on Numerical Analysis* 26.5 (1989), pp. 1212–1240. doi: 10.1137/0726068.

[31] F. Berthelin, P. Degond, V. Le Blanc, S. Moutari, M. Rascle and J. Royer. 'A Traffic-flow Model with Constraints for the Modeling of Traffic Jams'. In: *Math. Models Methods Appl. Sci.* 18.suppl. (2008), pp. 1269–1298. doi: 10.1142/S0218202508003030.

[32] F. Berthelin. 'Existence and Weak Stability for a Pressureless Model with Unilateral Constraint'. In: *Math. Models Methods Appl. Sci.* 12.2 (2002), pp. 249–272. doi: 10.1142/S0218202502001635.

[33] F. Berthelin. 'Theoretical Study of a Multidimensional Pressureless Model with Unilateral Constraint'. In: *SIAM Journal on Mathematical Analysis* 49.3 (Jan. 2017), pp. 2287–2320. doi: 10.1137/16m1074515.

[34] F. Berthelin, P. Degond, M. Delitala and M. Rascle. 'A Model for the Formation and Evolution of Traffic Jams'. In: *Arch. Ration. Mech. Anal.* 187.2 (2008), pp. 185–220. doi: 10.1007/s00205-007-0061-9.

[35] F. Berthelin and D. Broizat. 'A Model for the Evolution of Traffic Jams in Multi-lane'. In: *Kinet. Relat. Models* 5.4 (2012), pp. 697–728. doi: 10.3934/krm.2012.5.697.

[36] P. J. Blanco, C. A. Bulant, L. O. Müller, G. D. M. Talou, C. G. Bezerra, P. A. Lemos and R. A. Feijóo. 'Comparison of 1d and 3d Models for the Estimation of Fractional Flow Reserve'. In: *Scientific Reports* 8.1 (Nov. 2018). doi: 10.1038/s41598-018-35344-0.

[37] D. Bluestein. 'Utilizing Computational Fluid Dynamics in Cardiovascular Engineering and Medicine-what You Need to Know. Its Translation to the Clinic/bedside'. In: *Artificial Organs* 41.2 (Feb. 2017), pp. 117–121. doi: 10.1111/aor.12914.

[38] F. Bouchut, Y. Brenier, J. Cortes and J.-F. Ripoll. 'A Hierarchy of Models for Two-phase Flows'. In: *J. Nonlinear Sci.* 10.6 (2000), pp. 639–660. doi: 10.1007/s003320010006.

[39] M. Braack, R. Becker, D. Meidner, T. Richter and B. Vexler. *The Finite Element Toolkit Gascoigne*. Version v1.01. Oct. 2021. doi: 10.5281/zenodo.5574969.

[40] Y. Brenier. 'Averaged Multivalued Solutions for Scalar Conservation Laws'. In: *SIAM J. Numer. Anal.* 21 (1984), pp. 1013–1037.

[41] D. Bresch and P.-E. Jabin. 'Global existence of weak solutions for compressible Navier–Stokes equations: Thermodynamically unstable pressure and anisotropic viscous stress tensor'. In: *Annals of Mathematics* 188.2 (Sept. 2018), pp. 577–684. doi: 10.4007/annals.2018.188.2.4.

[42] D. Bresch, P. B. Mucha and E. Zatorska. 'Finite-Energy Solutions for Compressible Two-Fluid Stokes System'. In: *Archive for Rational Mechanics and Analysis* 232.2 (Nov. 2018), pp. 987–1029. doi: 10.1007/s00205-018-01337-6.

[43] D. Bresch, C. Perrin and E. Zatorska. 'Singular Limit of a Navier-Stokes System Leading to a Free/congested Zones Two-phase Model'. In: *C. R. Math. Acad. Sci. Paris* 352.9 (2014), pp. 685–690. doi: 10.1016/j.crma.2014.06.009.

[44] I. Brevis, I. Muga and K. G. van der Zee. 'A Machine-learning Minimal-residual (ML-MRes) Framework for Goal-oriented Finite Element Discretizations'. In: *Computers & Mathematics with Applications* (Sept. 2020). doi: 10.1016/j.camwa.2020.08.012.

[45] A.-P. Calderón. 'Lebesgue Spaces of Differentiable Functions and Distributions'. In: *Proc. Sympos. Pure Math.* Vol. IV. AMS, 1961, pp. 33–49.

[46] A. Cangiani, E. H. Georgoulis and Y. A. Sabawi. 'Adaptive Discontinuous Galerkin Methods for Elliptic Interface Problems'. In: *Mathematics of Computation* 87.314 (Feb. 2018), pp. 2675–2707. doi: 10.1090/mcom/3322.

[47] J. A. Cañizo, J. A. Carrillo and J. Rosado. 'A Well-posedness Theory in Measures for Some Kinetic Models of Collective Motion'. In: *Math. Models Methods Appl. Sci.* 21.3 (2011), pp. 515–539. doi: 10.1142/S0218202511005131.

[48] J. Canny. 'A Computational Approach to Edge Detection'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-8.6 (1986), pp. 679–698. doi: 10.1109/TPAMI.1986.4767851.

[49] C. Canuto and T. Kozubek. 'A Fictitious Domain Approach to the Numerical Solution of Pdes in Stochastic Domains'. In: *Numerische Mathematik* 107.2 (2007), p. 257. doi: 10.1007/s00211-007-0086-x.

[50] T. E. Carew, R. N. Vaishnav and D. J. Patel. 'Compressibility of the Arterial Wall'. In: *Circulation research* 23.1 (1968), pp. 61–68.

[51] J. A. Carrillo, M. Fornasier, J. Rosado and G. Toscani. 'Asymptotic Flocking Dynamics for the Kinetic Cucker-Smale Model'. In: *SIAM J. Math. Anal.* 42.1 (2010), pp. 218–236. doi: 10.1137/090757290.

[52] J. A. Carrillo, S. Martin and M.-T. Wolfram. 'An Improved Version of the Hughes Model for Pedestrian Flow'. In: *Math. Models Methods Appl. Sci.* 26.4 (2016), pp. 671–697. doi: 10.1142/S0218202516500147.

[53] J. A. Carrillo, Y.-P. Choi, P. B. Mucha and J. Peszek. 'Sharp Conditions to Avoid Collisions in Singular Cucker-Smale Interactions'. In: *Nonlinear Anal. Real World Appl.* 37 (2017), pp. 317–328. doi: 10.1016/j.nonrwa.2017.02.017.

[54] J. M. Carson, S. Pant, C. Roobottom, R. Alcock, P. J. Blanco, C. A. Bulant, Y. Vassilevski, S. Simakov, T. Gamilov, R. Pryamonosov, F. Liang, X. Ge, Y. Liu and P. Nithiarasu. 'Non-invasive Coronary CT Angiography-derived Fractional Flow Reserve: A Benchmark Study Comparing the Diagnostic Performance of Four Different Computational Methodologies'. In: *International Journal for Numerical Methods in Biomedical Engineering* 35.10 (Aug. 2019). doi: 10.1002/cnm.3235.

[55] J. Cebral, F. Mut, J. Weir and C. Putman. 'Quantitative Characterization of the Hemodynamic Environment in Ruptured and Unruptured Brain Aneurysms'. In: *American Journal of Neuroradiology* 32.1 (Dec. 2010), pp. 145–151. doi: 10.3174/ajnr.a2419.

[56] Y. Cheng. 'Mean Shift, Mode Seeking, and Clustering'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17.8 (1995), pp. 790–799.

[57] Y.-P. Choi, S.-Y. Ha and J. Kim. 'Propagation of Regularity and Finite-time Collisions for the Thermomechanical Cucker-Smale Model with a Singular Communication'. In: *Netw. Heterog. Media* 13.3 (2018), pp. 379–407. doi: 10.3934/nhm.2018017.

[58] Y.-P. Choi, S.-Y. Ha and Z. Li. 'Emergent Dynamics of the Cucker-Smale Flocking Model and Its Variants'. In: *Active particles. Vol. 1. Advances in theory, models, and applications.* Model. Simul. Sci. Eng. Technol. Birkhäuser/Springer, Cham, 2017, pp. 299–331.

[59] Y.-P. Choi and J. Haskovec. 'Hydrodynamic Cucker-Smale Model with Normalized Communication Weights and Time Delay'. In: *SIAM J. Math. Anal.* 51.3 (2019), pp. 2660–2685. doi: 10.1137/17M1139151.

[60] Y.-P. Choi, D. Kalise, J. Peszek and A. A. Peters. 'A Collisionless Singular Cucker–Smale Model with Decentralized Formation Control'. In: *SIAM Journal on Applied Dynamical Systems* 18.4 (Jan. 2019), pp. 1954–1981. doi: 10.1137/19m1241799.

[61] P. Ciarlet. *The Finite Element Method for Elliptic Problems.* Vol. 40. Classics. SIAM, 2002.

[62] R. M. Colombo and M. D. Rosini. 'Pedestrian Flows and Non-classical Shocks'. In: *Mathematical Methods in the Applied Sciences* 28.13 (2005), pp. 1553–1567. doi: 10.1002/mma.624.

[63] D. Comaniciu and P. Meer. 'Mean Shift: A Robust Approach toward Feature Space Analysis'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24.5 (2002), pp. 603–619. doi: 10.1109/34.1000236.

[64] *Computational Life.* url: https://www.computational-life.com/ (visited on 01/10/2021).

[65] F. Cordier, P. Degond and A. Kumbaro. 'An Asymptotic-preserving All-speed Scheme for the Euler and Navier–stokes Equations'. In: *J. Comput. Phys.* 231 (2012), pp. 5685–5704.

[66] F. Cucker and J.-G. Dong. 'On the Critical Exponent for Flocks under Hierarchical Leadership'. In: *Math. Models Methods Appl. Sci.* 19.suppl. (2009), pp. 1391–1404. doi: 10.1142/S0218202509003851.

[67] F. Cucker and J.-G. Dong. 'On Flocks Influenced by Closest Neighbors'. In: *Math. Models Methods Appl. Sci.* 26.14 (2016), pp. 2685–2708. doi: 10.1142/S0218202516500639.

[68] F. Cucker and S. Smale. 'Emergent Behavior in Flocks'. In: *IEEE Trans. Automat. Control* 52.5 (2007), pp. 852–862. doi: 10.1109/TAC.2007.895842.

[69] M. D'Elia, L. Mirabella, T. Passerini, M. Perego, M. Piccinelli, C. Vergara and A. Veneziani. 'Applications of Variational Data Assimilation in Computational Hemodynamics'. In: *Modeling of Physiological Flows.* Ed. by D. Ambrosi, A. Quarteroni and G. Rozza. Milano: Springer Milan, 2012, pp. 363–394. isbn: 978-88-470-1935-5. doi: 10.1007/978-88-470-1935-5_12.

[70] M. Dambrine, H. Harbrecht and B. Puig. 'Computing Quantities of Interest for Random Domains with Second Order Shape Sensitivity Analysis'. In: *ESAIM: Mathematical Modelling and Numerical Analysis* 49.5 (Aug. 2015), pp. 1285–1302. doi: 10.1051/m2an/2015012.

[71] R. Danchin, P. B. Mucha, J. Peszek and B. Wróblewski. 'Regular Solutions to the Fractional Euler Alignment System in the Besov Spaces Framework'. In: *Mathematical Models and Methods in Applied Sciences* (2019). doi: 10.1142/S0218202519500040.

[72] P. Degond, C. Appert-Rolland, M. Moussaïd, J. Pettré and G. Theraulaz. 'A Hierarchy of Heuristic-based Models of Crowd Dynamics'. In: *J. Stat. Phys.* 152.6 (2013), pp. 1033–1068. doi: 10.1007/s10955-013-0805-x.

[73] P. Degond, S. Jin and J.-G. Liu. 'Mach-number Uniform Asymptotic-preserving Gauge Schemes for Compressible Flows'. In: *Bulletin of the Institute of Mathematics Academia Sinica (New Series)* 7.4 (2007), pp. 851–892.

[74] P. Degond, L. Navoret, R. Bon and D. Sanchez. 'Congestion in a Macroscopic Model of Self-driven Particles Modeling Gregariousness'. In: *J. Stat. Phys.* 138.1 (2010), pp. 85–125.

[75] P. Degond, P. F. Peyrard, G. Russo and P. Villedieu. 'Polynomial Upwind Schemes for Hyperbolic Systems'. In: *C. R. Acad. Sci. Paris Sér. I Math.* 328.6 (1999), pp. 479–483.

[76] P. Degond, A. Frouvelle, J.-G. Liu, S. Motsch and L. Navoret. 'Macroscopic Models of Collective Motion and Self-organization'. In: *Séminaire Laurent Schwartz—Équations aux dérivées partielles et applications. Année 2012–2013.* Sémin. Équ. Dériv. Partielles. École Polytech., Palaiseau, 2014, Exp. No. I, 27.

[77] P. Degond and J. Hua. 'Self-organized Hydrodynamics with Congestion and Path Formation in Crowds'. In: *J. Comput. Phys.* 237 (2013), pp. 299–319. doi: 10.1016/j.jcp.2012.11.033.

[78] P. Degond, J. Hua and L. Navoret. 'Numerical Simulations of the Euler System with Congestion Constraint'. In: *J. Comput. Phys.* 230.22 (2011), pp. 8057–8088. doi: 10.1016/j.jcp.2011.07.010.

[79] P. Degond, P. Minakowski, L. Navoret and E. Zatorska. 'Finite Volume Approximations of the Euler System with Variable Congestion'. In: *Computers & Fluids* 169 (2018). Recent progress in nonlinear numerical methods for time-dependent flow & transport problems, pp. 23–39. doi: 10.1016/j.compfluid.2017.09.007.

[80] P. Degond, P. Minakowski and E. Zatorska. 'Transport of Congestion in Two-phase Compressible/Incompressible Flows'. In: *Nonlinear Analysis: Real World Applications* 42 (2018), pp. 485–510. doi: 10.1016/j.nonrwa.2018.02.001.

[81] M. H. Degroot. 'Reaching a Consensus'. In: *Journal of the American Statistical Association* 69.345 (1974), pp. 118–121. doi: 10.1080/01621459.1974.10480137.

[82] A. Delfino, N. Stergiopulos, J. Moore Jr and J.-J. Meister. 'Residual Strain Effects on the Stress Field in a Thick Wall Finite Element Model of the Human Carotid Bifurcation'. In: *Journal of biomechanics* 30.8 (1997), pp. 777–786.

[83] H. Dietert and R. Shvydkoy. 'On Cucker–Smale dynamical systems with degenerate communication'. In: *Analysis and Applications* (May 2020), pp. 1–23. doi: 10.1142/s0219530520500050.

[84] R. J. DiPerna and P.-L. Lions. 'Ordinary Differential Equations, Transport Theory and Sobolev Spaces'. In: *Invent. Math.* 98.3 (1989), pp. 511–547. doi: 10.1007/BF01393835.

[85] C. Dogbe. 'On the Modelling of Crowd Dynamics by Generalized Kinetic Models'. In: *J. Math. Anal. Appl.* 387.2 (2012), pp. 512–532. doi: 10.1016/j.jmaa.2011.09.007.

[86] C. Dogbé. 'Modeling Crowd Dynamics by the Mean-field Limit Approach'. In: *Math. Comput. Modelling* 52.9-10 (2010), pp. 1506–1520. doi: 10.1016/j.mcm.2010.06.012.

[87] J. Donea. 'An Arbitrary Lagrangian-eulerian Finite Element Method for Transient Dynamic Fluid-structure Interactions'. In: *Computer Methods in Applied Mechanics and Engineering* 33 (1982), pp. 689–723.

[88] W. E and B. Yu. 'The Deep Ritz Method: A Deep Learning-based Numerical Algorithm for Solving Variational Problems'. In: *Communications in Mathematics and Statistics* 6.1 (Feb. 2018), pp. 1–12. doi: 10.1007/s40304-018-0127-z.

[89] R. Eberhart and J. Kennedy. 'A New Optimizer Using Particle Swarm Theory'. In: *MHS'95. Proceedings of the Sixth International Symposium on Micro Machine and Human Science*. IEEE, 1995. doi: 10.1109/mhs.1995.494215.

[90] P. Erdős, J. Pach, R. Pollack and Z. Tuza. 'Radius, Diameter, and Minimum Degree'. In: *J. Combin. Theory Ser. B* 47.1 (1989), pp. 73–79. doi: 10.1016/0095-8956(89)90066-X.

[91] I. Ergatoudis, B. Irons and O. Zienkiewicz. 'Curved, Isoparametric "quadrilateral" Elements for Finite Element Analysis'. In: *I. J. of Solids and Structures* 4.1 (1968), pp. 31–42.

[92] A. Ern and J.-L. Guermond. *Theory and Practice of Finite Elements*. Applied Mathematical Sciences, 159, Springer, 2004.

[93] M. Ester, H.-P. Kriegel, J. Sander and X. Xu. 'A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise'. In: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. KDD'96. Portland, Oregon: AAAI Press, 1996, pp. 226–231.

[94] L. Evans. *Partial Differential Equations*. Providence, R.I.: American Mathematical Society, 2010. isbn: 0821849743.

[95] L. Failer and T. Richter. 'A Newton Multigrid Framework for Optimal Control of Fluid-structure Interactions'. In: *Optimization and Engineering* (2020). doi: 10.1007/s11081-020-09498-8.

[96] L. Failer and T. Richter. 'A Parallel Newton Multigrid Framework for Monolithic Fluid-structure Interactions'. In: *Journal of Scientific Computing* 82.2 (Jan. 2020), p. 28. doi: 10.1007/s10915-019-01113-y.

[97] L. Failer, P. Minakowski and T. Richter. 'On the Impact of Fluid Structure Interaction in Blood Flow Simulations'. In: *Vietnam Journal of Mathematics* (Jan. 2021). doi: 10.1007/s10013-020-00456-6.

[98] L. Failer, P. Minakowski and T. Richter. *Gascoigne Documented Example: Blood Flow and Virtual Stenting*. url: https://gascoigne.math.uni-magdeburg.de/index.php?-show=documentedexamples_05bloodflowsandvirtualstenting (visited on 01/10/2021).

[99] M. Falcone and R. Ferretti. *Semi-lagrangian Approximation Schemes for Linear and Hamilton—jacobi Equations*. SIAM, 2013.

[100] E. Feireisl. 'On Compactness of Solutions to the Compressible Isentropic Navier-Stokes Equations When the Density Is Not Square Integrable'. In: *Comment. Math. Univ. Carolin.* 42.1 (2001), pp. 83–98.

[101] E. Feireisl and A. Novotný. *Singular Limits in Thermodynamics of Viscous Fluids*. Advances in Mathematical Fluid Mechanics. Basel: Birkhäuser Verlag, 2009, pp. xxxvi+382. isbn: 978-3-7643-8842-3. doi: 10.1007/978-3-7643-8843-0.

[102] E. Feireisl, R. Klein, A. Novotný and E. Zatorska. 'On Singular Limits Arising in the Scale Analysis of Stratified Fluid Flows'. In: *Math. Models Methods Appl. Sci.* 26.3 (2016), pp. 419–443. doi: 10.1142/S021820251650007X.

[103] E. Feireisl, Y. Lu and J. Málek. 'On PDE Analysis of Flows of Quasi-incompressible Fluids'. In: *ZAMM Z. Angew. Math. Mech.* 96.4 (2016), pp. 491–508. doi: 10.1002/zamm.201400229.

[104] E. Feireisl, H. Petzeltová, E. Rocca and G. Schimperna. 'Analysis of a Phase-field Model for Two-phase Compressible Fluids'. In: *Math. Models Methods Appl. Sci.* 20.7 (2010), pp. 1129–1160. doi: 10.1142/S0218202510004544.

[105] A. Flache and R. Hegselmann. 'Understanding Complex Social Dynamics: a Plea for Cellular Automata Based Modelling'. In: *Journal of Artificial Societies and Social Simulation* 1.3 (1998), pp. 1–1. url: https://ideas.repec.org/a/jas/jasssj/1998-5-1.html.

[106] J. R. P. French. 'A Formal Theory of Social Power'. In: *Psychological Review* 63.3 (1956), pp. 181–194. doi: 10.1037/h0046123.

[107] Y.-C. Fung. *Biomechanics: Mechanical Properties of Living Tissues*. Springer Science & Business Media, 1993.

[108] S. W. Funke, M. Nordaas, Ø. Evju, M. S. Alnaes and K. A. Mardal. 'Variational Data Assimilation for Transient Blood Flow Simulations: Cerebral Aneurysms As an Illustrative Example'. In: *International Journal for Numerical Methods in Biomedical Engineering* 35.1 (2019). e3152 cnm.3152, e3152. doi: 10.1002/cnm.3152.

[109] J. Gan and Y. Tao. 'Dbscan Revisited: Mis-claim, Un-fixability, and Approximation'. In: *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*. SIGMOD '15. Melbourne, Victoria, Australia: Association for Computing Machinery, 2015, pp. 519–530. isbn: 9781450327589. doi: 10.1145/2723372.2737792.

[110] A. Garcimartín, I. Zuriguel, J. Pastor, C. Martín-Gómez and D. Parisi. 'Experimental Evidence of the "faster Is Slower" Effect'. In: *Transportation Research Procedia* 2 (2014), pp. 760–767.

[111] M. W. Gee, C. Förster and W. Wall. 'A Computational Strategy for Prestressing Patient-specific Biomechanical Problems under Finite Deformation'. In: *International Journal for Numerical Methods in Biomedical Engineering* 26.1 (2010), pp. 52–72.

[112] C. Guan, K. K. F. Yuen and F. Coenen. 'Particle Swarm Optimized Density-based Clustering and Classification: Supervised and Unsupervised Learning Approaches'. In: *Swarm and Evolutionary Computation* 44 (Feb. 2019), pp. 876–896. doi: 10.1016/j.swevo.2018.09.008.

[113] I. Gühring, G. Kutyniok and P. Petersen. 'Error Bounds for Approximations with Deep ReLU Neural Networks in $W^{s,p}$ Norms'. In: *Analysis and Applications* (Aug. 2019), pp. 1–57. doi: 10.1142/s0219530519410021.

[114] R. Y. Guo and H. J. Huang. 'A Mobile Lattice Gas Model for Simulating Pedestrian Evacuation'. In: *Physica A* 387 (2008), pp. 580–586.

[115] M. E. Gurtin, E. Fried and L. Anand. *The Mechanics and Thermodynamics of Continua*. Cambridge University Press, 2010.

[116] S.-Y. Ha, J. Kim, J. Park and X. Zhang. 'Complete Cluster Predictability of the Cucker-Smale Flocking Model on the Real Line'. In: *Arch. Ration. Mech. Anal.* 231.1 (2019), pp. 319–365. doi: 10.1007/s00205-018-1281-x.

[117] S.-Y. Ha and T. Ruggeri. 'Emergent Dynamics of a Thermodynamically Consistent Particle Model'. In: *Arch. Ration. Mech. Anal.* 223.3 (2017), pp. 1397–1425. doi: 10.1007/s00205-016-1062-3.

[118] M. Hammes, M. Boghosian, K. Cassel, S. Watson, B. Funaki, T. Doshi, S. M. J. Mahmoudzadeh Akherat, J. Hines and F. Coe. 'Increased Inlet Blood Flow Velocity Predicts Low Wall Shear Stress in the Cephalic Arch of Patients with Brachiocephalic Fistula Access'. In: *PLoS ONE* 11.4 (Mar. 2016), e0152873. doi: 10.1371/journal.pone.0152873.

[119] H. Harbrecht, M. Peters and M. Siebenmorgen. 'Analysis of the Domain Mapping Method for Elliptic Diffusion Problems on Random Domains'. In: *Numerische Mathematik* 134.4 (2016), pp. 823–856. doi: 10.1007/s00211-016-0791-4.

[120] H. Harbrecht, R. Schneider and C. Schwab. 'Sparse Second Moment Analysis for Elliptic Problems in Stochastic Domains'. In: *Numerische Mathematik* 109.3 (2008), pp. 385–414. doi: 10.1007/s00211-008-0147-9.

[121] S. He and E. Tadmor. 'Global Regularity of Two-dimensional Flocking Hydrodynamics'. In: *C. R. Math. Acad. Sci. Paris* 355.7 (2017), pp. 795–805. doi: 10.1016/j.crma.2017.05.008.

[122] R. Hegselmann and U. Krause. 'Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation'. In: *Journal of Artificial Societies and Social Simulation* 5.3 (2002), pp. 1–24.

[123] D. Helbing, I. Farkas and T. Vicsek. 'Simulating Dynamical Features of Escape Panic'. In: *Nature* 407 (2000), pp. 487–490.

[124] D. Helbing and P. Molnár. 'Self-organization of Complex Structures. from Individual to Collective Dynamics'. In: *Proceedings of the International Conference held in Berlin* 24–28 (Sept. 1997). Ed. by F. Schweitzer.

[125] D. Helbing. 'A Fluid-dynamic Model for the Movement of Pedestrians'. In: *Complex Systems* 6.5 (1992), pp. 391–415.

[126] D. Helbing and A. Johansson. 'Pedestrian, Crowd and Evacuation Dynamics'. In: *Extreme Environmental Events: Complexity in Forecasting andÂ Early Warning*. Ed. by R. A. Meyers. New York, NY: Springer New York, 2011, pp. 697–716. isbn: 978-1-4419-7695-6. doi: 10.1007/978-1-4419-7695-6_37.

[127] L. Henderson. 'The Statistics of Crowd Fluids'. In: *nature* 229 (1971), pp. 381–383.

[128] M. Herbin, N. Bonnet and P. Vautrot. 'A Clustering Method Based on the Estimation of the Probability Density Function and on the Skeleton by Influence Zones. Application to Image Processing'. In: *Pattern Recognition Letters* 17.11 (1996), pp. 1141–1150. doi: 10.1016/0167-8655(96)00085-2.

[129] J. Heywood, R. Rannacher and S. Turek. 'Artificial Boundaries and Flux and Pressure Conditions for the Incompressible Navier-Stokes Equations'. In: *International Journal for Numerical Methods in Fluids* 22 (1992), pp. 325–352.

[130] G. A. Holzapfel, T. C. Gasser and R. W. Ogden. 'A New Constitutive Framework for Arterial Wall Mechanics and a Comparative Study of Material Models'. In: *Journal of elasticity and the physical science of solids* 61.1-3 (2000), pp. 1–48.

[131] R. L. Hughes. 'A Continuum Theory for the Flow of Pedestrians'. In: *Transportation Research Part B: Methodological* 36.6 (2002), pp. 507–535. doi: 10.1016/S0191-2615(01)00015-7.

[132] R. L. Hughes. 'The Flow of Human Crowds'. In: *Annual review of fluid mechanics, Vol. 35.* Vol. 35. Annu. Rev. Fluid Mech. Annual Reviews, Palo Alto, CA, 2003, pp. 169–182. doi: 10.1146/annurev.fluid.35.101101.161136.

[133] A. K. Jain, R. P. W. Duin and J. Mao. 'Statistical Pattern Recognition: A Review'. In: *IEEE Trans. Pattern Anal. Mach. Intell.* 22.1 (Jan. 2000), pp. 4–37. doi: 10.1109/34.824819.

[134] Y. Jiang, P. Zhang, S. Wong and R. Liu. 'A Higher-order Macroscopic Model for Pedestrian Flows'. In: *Physica A: Statistical Mechanics and its Applications* 389.21 (2010), pp. 4623–4635. doi: 10.1016/j.physa.2010.05.003.

[135] C. Jin. 'Flocking of the Motsch-Tadmor Model with a Cut-off Interaction Function'. In: *J. Stat. Phys.* 171.2 (2018), pp. 345–360. doi: 10.1007/s10955-018-2006-0.

[136] S. Jin, J. Oshinski and D. P. Giddens. 'Effects of Wall Motion and Compliance on Flow Patterns in the Ascending Aorta'. In: *Journal of Biomechanical Engineering* 125.3 (June 2003), pp. 347–354. doi: 10.1115/1.1574332.

[137] D. Jodlbauer, U. Langer and T. Wick. 'Parallel Block-preconditioned Monolithic Solvers for Fluid-structure Interaction Problems'. In: *International Journal for Numerical Methods in Engineering* 117.6 (2019), pp. 623–643.

[138] V. John. *Finite Element Methods for Incompressible Flow Problems.* Vol. 51. Computational Mathematics. Springer, 2016.

[139] E. Jones, T. Oliphant, P. Peterson et al. *SciPy: Open Source Scientific Tools for Python.* 2001. url: http://www.scipy.org/.

[140] A. Kanezaki. 'Unsupervised Image Segmentation by Backpropagation'. In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, Apr. 2018. doi: 10.1109/icassp.2018.8462533.

[141] E. Kharazmi, Z. Zhang and G. E. Karniadakis. *Variational Physics-Informed Neural Networks For Solving Partial Differential Equations.* 2019. arXiv: 1912.00873 [cs.NE].

[142] W. Kim, A. Kanezaki and M. Tanaka. 'Unsupervised Learning of Image Segmentation Based on Differentiable Feature Clustering'. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 8055–8068. doi: 10.1109/tip.2020.3011269.

[143] A. Kiselev and C. Tan. 'Global Regularity for 1D Eulerian Dynamics with Singular Interaction Forces'. In: *SIAM J. Math. Anal.* 50.6 (2018), pp. 6208–6229. doi: 10.1137/17M1141515.

[144] U. Krause. 'A Discrete Nonlinear and Non-autonomous Model of Consensus Formation'. In: *Communications in difference equations (Poznan, 1998).* Gordon and Breach, Amsterdam, 2000, pp. 227–236.

[145] H.-P. Kriegel, P. Kröger, J. Sander and A. Zimek. 'Density-based Clustering'. In: *WIREs Data Mining and Knowledge Discovery* 1.3 (Apr. 2011), pp. 231–240. doi: 10.1002/widm.30.

[146] A. Lachapelle and M.-T. Wolfram. 'On a Mean Field Game Approach Modeling Congestion and Aversion in Pedestrian Crowds'. In: *Transportation research part B: methodological* 45.10 (2011), pp. 1572–1589.

[147] K. Lehrer. 'Social Consensus and Rational Agnoiology'. In: *Synthese* 31.1 (1975), pp. 141–160. url: http://www.jstor.org/stable/20115060.

[148] M. Lenoir. 'Optimal Isoparametric Finite Elements and Error Estimates for Domains Involving Curved Boundaries'. In: *SIAM Journal on Numerical Analysis* 23.3 (1986), pp. 562–580.

[149] Y. Li and Y. Sun. 'Global weak solutions and long time behavior for 1D compressible MHD equations without resistivity'. In: *Journal of Mathematical Physics* 60.7 (July 2019), p. 071511. doi: 10.1063/1.5082758.

[150] Z. Li. 'Effectual Leadership in Flocks with Hierarchy and Individual Preference'. In: *Discrete Contin. Dyn. Syst.* 34.9 (2014), pp. 3683–3702. doi: 10.3934/dcds.2014.34.3683.

[151] F. Liang, S. Takagi, R. Himeno and H. Liu. 'Multi-scale Modeling of the Human Cardiovascular System with Applications to Aortic Valvular and Arterial Stenoses'. In: *Medical & Biological Engineering & Computing* 47.7 (July 2009), pp. 743–755. doi: 10.1007/s11517-009-0449-9.

[152] P. Libby, J. E. Buring, L. Badimon, G. K. Hansson, J. Deanfield, M. S. Bittencourt, L. Tokgözoğlu and E. F. Lewis. 'Atherosclerosis'. In: *Nature Reviews Disease Primers* 5.1 (Aug. 2019). doi: 10.1038/s41572-019-0106-z.

[153] P.-L. Lions and N. Masmoudi. 'On a Free Boundary Barotropic Model'. In: *Ann. Inst. H. Poincaré Anal. Non Linéaire* 16.3 (1999), pp. 373–410. doi: 10.1016/S0294-1449(99)80018-3.

[154] L. Lu, X. Meng, Z. Mao and G. E. Karniadakis. 'DeepXDE: A Deep Learning Library for Solving Differential Equations'. In: *SIAM Review* 63.1 (Jan. 2021), pp. 208–228. doi: 10.1137/19m1274067.

[155] D. Maltese, M. Michalek, P. B. Mucha, A. Novotny, M. Pokorny and E. Zatorska. 'Existence of Weak Solutions for Compressible Navier-Stokes Equations with Entropy Transport'. In: *J. Differential Equations* (). url: http://www.sciencedirect.com/science/article/pii/S0022039616301656.

[156] N. Margenberg, P. Minakowski and T. Richter. *GascoignePytorch*. url: https://kosinus.math.uni-magdeburg.de/gascoigne/gascoignepytorch (visited on 01/10/2021).

[157] I. Markou. 'Collision-avoiding in the Singular Cucker-Smale Model with Nonlinear Velocity Couplings'. In: *Discrete Contin. Dyn. Syst.* 38.10 (2018), pp. 5245–5260. doi: 10.3934/dcds.2018232.

[158] D. Martin, C. Fowlkes, D. Tal and J. Malik. 'A Database of Human Segmented Natural Images and Its Application to Evaluating Segmentation Algorithms and Measuring Ecological Statistics'. In: *Proc. 8th Int'l Conf. Computer Vision*. Vol. 2. July 2001, pp. 416–423.

[159] B. Maury. 'A Gluey Particle Model'. In: *ESAIM: Proc.* 18 (2007), pp. 133–142. doi: 10.1051/proc:071811.

[160] B. Maury, A. Roudneff-Chuoin and F. Santambrogio. 'A Macroscopic Crowd Motion Model of Gradient Flow Type'. In: *Mathematical Models and Methods in Applied Sciences* 20.10 (2010), pp. 1787–1821. doi: 10.1142/S0218202510004799.

[161] B. Maury. 'Prise En Compte De La Congestion Dans Les Modeles De Mouvements De Foules'. In: *Actes des colloques Caen* (2012).

[162] M. Michálek. 'Stability Result for Navier-Stokes Equations with Entropy Transport'. In: *J. Math. Fluid Mech.* 17.2 (2015), pp. 279–285. doi: 10.1007/s00021-015-0205-x.

[163] P. Minakowski and T. Richter. 'A priori and a posteriori error estimates for the Deep Ritz method applied to the Laplace and Stokes problem'. In: *Journal of Computational and Applied Mathematics* 421 (2023), p. 114845. issn: 0377-0427. doi: 10.1016/j.cam.2022.114845.

[164] P. Minakowski. *Density-induced Consensus Protocol (source code)*. Nov. 2019. doi: 10.5281/zenodo.3551842.

[165] P. Minakowski, P. B. Mucha and J. Peszek. 'Density-Induced Consensus Protocol'. In: *Mathematical Models and Methods in Applied Sciences* 30.12 (July 2020), pp. 2389–2415. doi: 10.1142/s0218202520500451.

[166] P. Minakowski, P. B. Mucha, J. Peszek and E. Zatorska. 'Singular Cucker–Smale Dynamics'. In: *Active Particles, Volume 2: Advances in Theory, Models, and Applications*. Ed. by N. Bellomo, P. Degond and E. Tadmor. Cham: Springer International Publishing, 2019, pp. 201–243. doi: 10.1007/978-3-030-20297-2_7.

[167] P. Minakowski, P. B. Mucha, J. Peszek and E. Zatorska. 'Singular Cucker-Smale Dynamics'. In: *Active particles, Vol. 2*. Model. Simul. Sci. Eng. Technol. Birkhäuser/Springer, Cham, 2019, pp. 201–243.

[168] P. Minakowski and J. Peszek. 'Data Clustering as an Emergent Consensus of Autonomous Agents'. In: (2022). doi: 10.48550/ARXIV.2204.10585.

[169] P. Minakowski and T. Richter. 'Finite Element Error Estimates on Geometrically Perturbed Domains'. In: *Journal of Scientific Computing* 84.2 (July 2020). doi: 10.1007/s10915-020-01285-y.

[170] P. Minakowski and T. Richter. *Error Estimates for Neural Network Solutions of Partial Differential Equations*. 2021. arXiv: 2107.11035 [math.NA].

[171] S. Mishra. 'A Machine Learning Framework for Data Driven Acceleration of Computations of Di Erential Equations'. In: *Mathematics in Engineering* 1.1 (2018), pp. 118–146. doi: 10.3934/mine.2018.1.118.

[172] S. Mishra and R. Molinaro. *Estimates on the Generalization Error of Physics Informed Neural Networks (PINNS) for Approximating PDEs*. 2020. arXiv: 2006.16144 [math.NA].

[173] B. Mohar. 'Eigenvalues, Diameter, and Mean Distance in Graphs'. In: *Graphs Combin.* 7.1 (1991), pp. 53–64. doi: 10.1007/BF01789463.

[174] P. Moireau, N. Xiao, M. Astorino, C. A. Figueroa, D. Chapelle, C. A. Taylor and J.-F. Gerbeau. 'External Tissue Support and Fluid–structure Simulation in Blood Flows'. In: *Biomechanics and Modeling in Mechanobiology* 11.1-2 (Feb. 2011), pp. 1–18. doi: 10.1007/s10237-011-0289-z.

[175] J. A. Moore, D. A. Steinman, D. W. Holdsworth and C. R. Ethier. 'Accuracy of Computational Hemodynamics in Complex Arterial Geometries Reconstructed from Magnetic Resonance Imaging'. In: *Annals of Biomedical Engineering* 27.1 (1999), pp. 32–41. doi: 10.1114/1.163.

[176] J. Moore, D. Steinman and C. Ethier. 'Computational Blood Flow Modelling: Errors Associated with Reconstructing Finite Element Models from Magnetic Resonance Images'. In: *Journal of Biomechanics* 31.2 (1997), pp. 179–184. doi: 10.1016/S0021-9290(97)00125-5.

[177] J. Morales, J. Peszek and E. Tadmor. 'Flocking with Short-range Interactions'. In: *J. Stat. Phys.* 176.2 (2019), pp. 382–397. doi: 10.1007/s10955-019-02304-5.

[178] P. D. Morris, F. N. van de Vosse, P. V. Lawford, D. R. Hose and J. P. Gunn. '"virtual" (computed) Fractional flow reserve'. In: *JACC: Cardiovascular Interventions* 8.8 (July 2015), pp. 1009–1017. doi: 10.1016/j.jcin.2015.04.006.

[179] S. Motsch and E. Tadmor. 'A New Model for Self-organized Dynamics and Its Flocking Behavior'. In: *J. Stat. Phys.* 144.5 (2011), pp. 923–947. doi: 10.1007/s10955-011-0285-9.

[180] S. Motsch and E. Tadmor. 'Heterophilious Dynamics Enhances Consensus'. In: *SIAM Rev.* 56.4 (2014), pp. 577–621. doi: 10.1137/120901866.

[181] P. B. Mucha and J. Peszek. 'The Cucker–smale Equation: Singular Communication Weight, Measure-valued Solutions and Weak-atomic Uniqueness'. In: *Archive for Rational Mechanics and Analysis* 227.1 (Aug. 2017), pp. 273–308. doi: 10.1007/s00205-017-1160-x.

[182] S. Mukherjee, P. Niyogi, T. Poggio and R. Rifkin. 'Learning Theory: Stability Is Sufficient for Generalization and Necessary and Sufficient for Consistency of Empirical Risk Minimization'. In: *Advances in Computational Mathematics* 25.1 (2006), pp. 161–193. doi: 10.1007/s10444-004-7634-z.

[183] L. O. Müller and E. F. Toro. 'A Global Multiscale Mathematical Model for the Human Circulation with Emphasis on the Venous System'. In: *International journal for numerical methods in biomedical engineering* 30.7 (2014), pp. 681–725.

[184] F. Nobile and C. Vergara. 'An Effective Fluid-structure Interaction Formulation for Vascular Dynamics by Generalized Robin Conditions'. In: *SIAM Journal on Scientific Computing* 30.2 (2008), pp. 731–763. doi: 10.1137/060678439.

[185] F. Nobile, M. Pozzoli and C. Vergara. 'Inexact Accurate Partitioned Algorithms for Fluid–structure Interaction Problems with Finite Elasticity in Haemodynamics'. In: *Journal of Computational Physics* 273 (Sept. 2014), pp. 598–617. doi: 10.1016/j.jcp.2014.05.020.

[186] D. Nolte and C. Bertoglio. 'Reducing the Impact of Geometric Errors in Flow Computations Using Velocity Measurements'. In: *International Journal for Numerical Methods in Biomedical Engineering* (Apr. 2019), e3203. doi: 10.1002/cnm.3203.

[187] A. Novotný and I. Straškraba. *Introduction to the Mathematical Theory of Compressible Flow.* Vol. 27. Oxford Lecture Series in Mathematics and its Applications. Oxford: Oxford University Press, 2004, pp. xx+506. isbn: 0-19-853084-6.

[188] W. Oberkampf and M. Barone. 'Measures of Agreement between Computation and Experiment: Validation Metrics'. In: *Journal of Computational Physics* 217.1 (Sept. 2006), pp. 5–36. doi: 10.1016/j.jcp.2006.03.037.

[189] W. Oberkampf and C. Roy. *Verification and Validation in Scientific Computing.* Cambridge University Press, 2010. doi: 10.1017/CBO9780511760396.

[190] R. Olfati-Saber. 'Flocking for Multi-agent Dynamic Systems: Algorithms and Theory'. In: *IEEE Trans. Automat. Control* 51.3 (2006), pp. 401–420. doi: 10.1109/TAC.2005.864190.

[191] R. Olfati-Saber and R. M. Murray. 'Consensus Problems in Networks of Agents with Switching Topology and Time-delays'. In: *IEEE Trans. Automat. Control* 49.9 (2004), pp. 1520–1533. doi: 10.1109/TAC.2004.834113.

[192] M. Omran, A. P. Engelbrecht and A. Salman. 'Particle Swarm Optimization Method for Image Clustering'. In: *International Journal of Pattern Recognition and Artificial Intelligence* 19.03 (May 2005), pp. 297–321. doi: 10.1142/s0218001405004083.

[193] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai and S. Chintala. 'PyTorch: An Imperative Style, High-Performance Deep Learning Library'. In: *Advances in Neural Information Processing Systems 32.* Curran Associates, Inc., 2019, pp. 8024–8035.

[194] K. Perktold and G. Rappitsch. 'Computer Simulation of Local Blood Flow and Vessel Mechanics in a Compliant Carotid Artery Bifurcation Model'. In: *Journal of Biomechanics* 28.7 (1995), pp. 845–856. doi: 10.1016/0021-9290(95)95273-8.

[195] C. Perrin and M. Westdickenberg. 'One-Dimensional Granular System with Memory Effects'. In: *SIAM Journal on Mathematical Analysis* 50.6 (Jan. 2018), pp. 5921–5946. doi: 10.1137/17m1121421.

[196] C. Perrin. 'Pressure-dependent Viscosity Model for Granular Media Obtained from Compressible Navier–Stokes Equations'. In: *Applied Mathematics Research eXpress* (2016). doi: 10.1093/amrx/abw004.

[197] C. Perrin and E. Zatorska. 'Free/congested Two-phase Model from Weak Solutions to Multi-dimensional Compressible Navier-Stokes Equations'. In: *Comm. Partial Differential Equations* 40.8 (2015), pp. 1558–1589. doi: 10.1080/03605302.2015.1014560.

[198] B. Perthame and N. Vauchelet. 'Incompressible Limit of a Mechanical Model of Tumour Growth with Viscosity'. In: *Philos. Trans. A* 373.2050 (2015), pp. 20140283, 16. doi: 10.1098/rsta.2014.0283.

[199] J. Peszek. 'Existence of Piecewise Weak Solutions of a Discrete Cucker–Smale's Flocking Model with a Singular Communication Weight'. In: *J. Differential Equations* 257.8 (2014), pp. 2900–2925. doi: 10.1016/j.jde.2014.06.003.

[200] J. Peszek. 'Discrete Cucker-Smale Flocking Model with a Weakly Singular Weight'. In: *SIAM J. Math. Anal.* 47.5 (2015), pp. 3671–3686. doi: 10.1137/15M1009299.

[201] B. Piccoli and A. Tosin. 'Pedestrian Flows in Bounded Domains with Obstacles'. In: *Continuum Mechanics and Thermodynamics* 21.2 (2009), pp. 85–107. doi: 10.1007/s00161-009-0100-x.

[202] B. Piccoli and A. Tosin. 'Time-evolving Measures and Macroscopic Modeling of Pedestrian Flow'. In: *Arch. Ration. Mech. Anal.* 199.3 (2011), pp. 707–738. doi: 10.1007/s00205-010-0366-y.

[203] G. Pichler, F. Martinez, A. Vicente, E. Solaz, O. Calaforra and J. Redon. 'Pulse Pressure Amplification and Its Determinants'. In: *Blood Pressure* 25.1 (Sept. 2015), pp. 21–27. doi: 10.3109/08037051.2015.1090713.

[204] D. D. Pietro and A. Ern. *Mathematical Aspects of Discontinuous Galerkin Methods*. Springer Berlin Heidelberg, 2012. doi: 10.1007/978-3-642-22980-0.

[205] A. Pinkus. 'Approximation Theory of the MLP Model in Neural Networks'. In: *Acta Numerica* 8 (1999), pp. 143–195. doi: 10.1017/S0962492900002919.

[206] A. Quarteroni, A. Manzoni and C. Vergara. 'The Cardiovascular System: Mathematical Modelling, Numerical Algorithms and Clinical Applications'. In: *Acta Numerica* 26 (May 2017), pp. 365–590. doi: 10.1017/s0962492917000046.

[207] A. Quarteroni, M. Tuveri and A. Veneziani. 'Computational Vascular Fluid Dynamics: Problems, Models and Methods'. In: *Computing and Visualization in Science* 2.4 (2000), pp. 163–197. doi: 10.1007/s007910050039.

[208] D. K. Rajan, T. I. Clark and D. S. Baumann. 'Cephalic Arch Stenosis'. In: *Essentials of Percutaneous Dialysis Interventions*. Ed. by D. Rajan. New York, NY: Springer New York, 2011. Chap. 19, pp. 341–356. isbn: 978-1-4419-5657-6. doi: 10.1007/978-1-4419-5657-6_19.

[209] T. Richter. *Fluid-structure Interactions. Models, Analysis and Finite Elements*. Vol. 118. Lecture notes in computational science and engineering. Springer, 2017.

[210] T. Richter and J. Mizerski. 'The Candy Wrapper Problem - a Temporal Multiscale Approach for PDE/PDE Systems'. In: *ENUMATH 2019*. Springer, 2020.

[211] T. Richter and T. Wick. 'On Time Discretizations of Fluid-structure Interactions'. In: *Multiple Shooting and Time Domain Decomposition Methods*. Ed. by T. Carraro, M. Geiger, S. Körkel and R. Rannacher. Vol. 9. Contributions in Mathematical and Computational Science. Springer, 2015, pp. 377–400.

[212] T. Richter and T. Wick. 'Variational Localizations of the Dual Weighted Residual Estimator'. In: *Journal of Comp. Appl. Math.* 279 (2015), pp. 192–208.

[213] T. Richter. 'Goal-oriented Error Estimation for Fluid–structure Interaction Problems'. In: *Computer Methods in Applied Mechanics and Engineering* 223-224 (June 2012), pp. 28–42. doi: 10.1016/j.cma.2012.02.014.

[214] T. Richter. 'A Monolithic Geometric Multigrid Solver for Fluid-structure Interactions in ALE Formulation'. In: *International Journal for Numerical Methods in Engineering* 104.5 (May 2015), pp. 372–390. doi: 10.1002/nme.4943.

[215] E. Samaniego, C. Anitescu, S. Goswami, V. Nguyen-Thanh, H. Guo, K. Hamdia, X. Zhuang and T. Rabczuk. 'An Energy Approach to the Solution of Partial Differential Equations in Computational Mechanics Via Machine Learning: Concepts, Implementation and Applications'. In: *Computer Methods in Applied Mechanics and Engineering* 362 (Apr. 2020), p. 112790. doi: 10.1016/j.cma.2019.112790.

[216] E. Schubert, J. Sander, M. Ester, H. P. Kriegel and X. Xu. 'Dbscan Revisited, Revisited: Why and How You Should (still) Use Dbscan'. In: *ACM Trans. Database Syst.* 42.3 (July 2017). doi: 10.1145/3068335.

[217] S. Shalev-Shwartz, O. Shamir, N. Srebro and K. Sridharan. 'Learnability, Stability and Uniform Convergence'. In: *J. Mach. Learn. Res.* 11 (2010), pp. 2635–2670.

[218] J. Shen. 'Cucker-Smale Flocking under Hierarchical Leadership'. In: *SIAM J. Appl. Math.* 68.3 (2007), pp. 694–719. doi: 10.1137/060673254.

[219] Y. Shin. 'On the Convergence of Physics Informed Neural Networks for Linear Second-order Elliptic and Parabolic Type PDEs'. In: *Communications in Computational Physics* 28.5 (June 2020), pp. 2042–2074. doi: 10.4208/cicp.oa-2020-0193.

[220] M. Shojima, M. Oshima, K. Takagi, R. Torii, M. Hayakawa, K. Katada, A. Morita and T. Kirino. 'Magnitude and Role of Wall Shear Stress on Cerebral Aneurysm'. In: *Stroke* 35.11 (Nov. 2004), pp. 2500–2505. doi: 10.1161/01.str.0000144648.89172.0f.

[221] R. Shvydkoy. 'Global Existence and Stability of Nearly Aligned Flocks'. In: *Journal of Dynamics and Differential Equations* 31.4 (Aug. 2018), pp. 2165–2175. doi: 10.1007/s10884-018-9693-8.

[222] R. Shvydkoy and E. Tadmor. 'Eulerian Dynamics with a Commutator Forcing'. In: *Transactions of Mathematics and Its Applications* 1.1 (2017). doi: 10.1093/imatrm/tnx001.

[223] R. Shvydkoy and E. Tadmor. *Topologically-based fractional diffusion and emergent dynamics with short-range interactions.* 2020. arXiv: 1806.01371 [math.AP].

[224] J. Sirignano and K. Spiliopoulos. 'DGM: A Deep Learning Algorithm for Solving Partial Differential Equations'. In: *Journal of Computational Physics* 375 (Dec. 2018), pp. 1339–1364. doi: 10.1016/j.jcp.2018.08.029.

[225] D. Steinman. 'Assumptions in Modelling of Large Artery Hemodynamics'. In: *Modeling of Physiological Flows.* Ed. by D. Ambrosi, A. Quarteroni and G. Rozza. Milano: Springer Milan, 2012, pp. 1–18.

[226] M. Sykora, M. Pavelka, I. Peshkov, P. Minakowski, V. Klika and E. Romenski. 'Comparison of the Symmetric Hyperbolic Thermodynamically Compatible framework with Hamiltonian mechanics of binary mixtures'. In: (2022). doi: 10.48550/ARXIV.2201.04460.

[227] D. Tartakovsky and D. Xiu. 'Stochastic Analysis of Transport in Tubes with Rough Walls'. In: *Journal of Computational Physics* 217.1 (2006), pp. 248–259. doi: 10.1016/j.jcp.2006.02.029.

[228] C. A. Taylor, M. T. Draney, J. P. Ku, D. Parker, B. N. Steele, K. Wang and C. K. Zarins. 'Predictive Medicine: Computational Techniques in Therapeutic Decision-making'. In: *Computer Aided Surgery* 4.5 (1999), pp. 231–247. doi: 10.3109/10929089909148176.

[229] C. A. Taylor, T. A. Fonte and J. K. Min. 'Computational Fluid Dynamics Applied to Cardiac Computed Tomography for Noninvasive Quantification of Fractional Flow Reserve'. In: *Journal of the American College of Cardiology* 61.22 (June 2013), pp. 2233–2241. doi: 10.1016/j.jacc.2012.11.083.

[230] P. A. Tonino, B. D. Bruyne, N. H. Pijls, U. Siebert, F. Ikeno, M. van 't Veer, V. Klauss, G. Manoharan, T. Engstrom, K. G. Oldroyd, P. N. V. Lee, P. A. MacCarthy and W. F. Fearon. 'Fractional Flow Reserve Versus Angiography for Guiding Percutaneous Coronary Intervention'. In: *New England Journal of Medicine* 360.3 (Jan. 2009), pp. 213–224. doi: 10.1056/nejmoa0807611.

[231] C. M. Topaz and A. L. Bertozzi. 'Swarming Patterns in a Two-dimensional Kinematic Model for Biological Groups'. In: *SIAM J. Appl. Math.* 65.1 (2004), pp. 152–174. doi: 10.1137/S0036139903437424.

[232] R. Torii, J. Keegan, N. B. Wood, A. W. Dowsey, A. D. Hughes, G.-Z. Yang, D. N. Firmin, S. A. M. Thom and X. Y. Xu. 'MR Image-based Geometric and Hemodynamic Investigation of the Right Coronary Artery with Dynamic Vessel Motion'. In: *Annals of Biomedical Engineering* 38.8 (Aug. 2010), pp. 2606–2620. doi: 10.1007/s10439-010-0008-4.

[233] A. Treuille, S. Cooper and Z. Popović. 'Continuum Crowds'. In: *ACM Trans. Graph.* 25 (2006), pp. 1160–1168.

[234] M. Twarogowska, P. Goatin and R. Duvigneau. 'Macroscopic Modeling and Simulations of Room Evacuation'. In: *Applied Mathematical Modelling* 38.24 (2014), pp. 5781–5795. doi: 10.1016/j.apm.2014.03.027.

[235] R. Unnikrishnan, C. Pantofaru and M. Hebert. 'Toward Objective Evaluation of Image Segmentation Algorithms'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (2007), pp. 929–944. doi: 10.1109/TPAMI.2007.1046.

[236] A. Vasseur, H. Wen and C. Yu. 'Global weak solution to the viscous two-fluid model with finite energy'. In: *Journal de Mathématiques Pures et Appliquées* 125 (May 2019), pp. 247–282. doi: 10.1016/j.matpur.2018.06.019.

[237] N. Vauchelet and E. Zatorska. 'Incompressible Limit of the Navier-Stokes Model with a Growth Term'. In: *Nonlinear Anal.* 163 (2017), pp. 34–59. doi: 10.1016/j.na.2017.07.003.

[238] H. von Wahl, T. Richter, C. Lehrenfeld, J. Heiland and P. Minakowski. *Numerical Benchmarking of Fluid-rigid Body Interactions.* Version v1. June 2019. doi: 10.5281/zenodo.3253455.

[239] H. von Wahl, T. Richter, C. Lehrenfeld, J. Heiland and P. Minakowski. 'Numerical Benchmarking of Fluid-rigid Body Interactions'. In: *Computers & Fluids* 193 (Oct. 2019), p. 104290. doi: 10.1016/j.compfluid.2019.104290.

[240] G. Wolansky. 'Dynamics of a System of Sticking Particles of Finite Size on the Line'. In: *Nonlinearity* 20.9, 2175 (2007).

[241] D. Xiu and D. Tartakovsky. 'Numerical Methods for Differential Equations in Random Domains'. In: *SIAM Journal on Scientific Computing* 28.3 (2006), pp. 1167–1185. doi: 10.1137/040613160.

[242] D. Xu and Y. Tian. 'A Comprehensive Survey of Clustering Algorithms'. In: *Annals of Data Science* 2.2 (June 2015), pp. 165–193. doi: 10.1007/s40745-015-0040-1.

[243] G. Xu, B. Mourrain, R. Duvigneau and A. Galligo. 'Parameterization of Computational Domain in Isogeometric Analysis: Methods and Comparison'. In: *Computer Methods in Applied Mechanics and Engineering* 200.23 (2011), pp. 2021–2031. doi: 10.1016/j.cma.2011.03.005.

[244] G. Xu, B. Mourrain, R. Duvigneau and A. Galligo. 'Constructing Analysis-suitable Parameterization of Computational Domain from Cad Boundary by Variational Harmonic Method'. In: *Journal of Computational Physics* 252 (2013), pp. 275–289. doi: 10.1016/j.jcp.2013.06.029.

[245] Y. Yang, W. Jäger, M. Neuss-Radu and T. Richter. 'Mathematical Modeling and Simulation of the Evolution of Plaques in Blood Vessels'. In: *J. of Math. Biology* 72.4 (2016), pp. 973–996.

[246] H. Zhang, J. E. Fritts and S. A. Goldman. 'Image Segmentation Evaluation: A Survey of Unsupervised Methods'. In: *Computer Vision and Image Understanding* 110.2 (2008), pp. 260–280. doi: 10.1016/j.cviu.2007.08.003.

[247] X. Zhang and T. Zhu. 'Complete classification of the asymptotical behavior for singular C-S model on the real line'. In: *Journal of Differential Equations* 269.1 (June 2020), pp. 201–256. doi: 10.1016/j.jde.2019.12.004.