UNIVERSITÀ DEGLI STUDI DI PADOVA
DIPARTIMENTO DI MATEMATICA
CORSO DI DOTTORATO DI RICERCA IN:
SCIENZE MATEMATICHE
CURRICULUM: MATEMATICA COMPUTAZIONALE
CICLO *XXX*

BIOLOGICALLY INSPIRED FORMULATION OF OPTIMAL TRANSPORT PROBLEMS

COORDINATORE: Ch.mo Prof. Martino Bardi
SUPERVISOR : Ch.mo Prof. Mario Putti
CO-SUPERVISOR: Ch.mo Prof. Franco Cardin

DOTTORANDO: *Enrico Facca*

Acknowledgments

This thesis summarizes all the mathematical work I carried out in the last years, thus I would like to express my gratitude to all the people who accompanied me throughout academic journey as mathematical scholar and researcher. I will stick to professional acknowledgment because for personal ones there is no point in writing them here: those who have been close to me are aware of my gratefulness.

First of all, I want to thank Univeristà degli Studi di Padova, where I have been studying for the last nine years. My gratitude goes to all the people who trained me as student during my undergraduate studies, and as a researcher in these last three years. In particular, I want to express my sincere gratitude to my advisors, Mario Putti and Franco Cardin, for the constant support and for the suggestions (the mathematical and the non-mathematical ones) given me since we started to work together with my master degree thesis. All the results I achieved were only possible thank to their ability to limit my weaknesses and enhance my strengths as a researcher.

Another person that I want to thank is Peter Knabner, who gave me the opportunity to spend seven months in Erlangen during my second year of the PhD studies. Its support and suggestions have been very important to me.

Another person that I want to thank is Filippo Santambrogio. The ideas developed during my time in Orsay had a fundamental role for in the accomplishment of this thesis.

Abstract

In this thesis we propose a model that we conjecture is a new and original formulation of the Optimal Transport Problem, a recently expanding area of mathematics that studies optimal strategies to move resources from one place to another. The proposed approach is an infinite-dimensional extension of a model describing the dynamics of Physarum Polycephalum (PP), a slime mold with surprising abilities to find the shortest path connecting two food sources. The original model describes the dynamics of the slime mold on a finite planar graph using a pipe-flow analogy whereby mass transfer occurs because of pressure differences with a conductivity coefficient that varies with the flow intensity. This model has been shown to be equivalent to a problem of "optimal transportation" on graphs. Our extension abandons the graph structure and moves to a continuous domain, coupling an elliptic diffusion equation enforcing PP density balance with an ordinary differential equations governing the flow dynamics. We conjecture that the new system of equations presents a time-asymptotic equilibrium connected to solutions of many instances of OTP, including the standard L^1 case and the conjested and branched transport problems.

From a theoretical point of view, we are only able to prove well-posedness of the proposed model for sufficiently small times and under restrictive hypothesis on the the regularity of the diffusion coefficient and the functions describing the initial and final configurations of the transported mass. However, our extensive numerical results show that the approximate solution of our proposed formulation converges at large times to an equilibrium configuration that well compares with the solutions of the different flavors of OTP. In particular, we are able to efficiently recover the numerical solutions that closely resemble the singular and ramified structures typical of branched transport problems. These simualtions provide strong support to our conjectures. Notwithstanding the numerical difficulties related mainly to the ill-conditioning of the algebraic systems, the rather simple approach adopted for the discretization of the proposed formulation resulted highly efficient and robust in terms of convergence and computational speed.

We also propose and tackle several applications to real world problems . In particular, we discuss how our formulation can be applied to model the geomorphology of river networks and the dynamics of plant roots. In addition, based on numerical evidence, we argue that the emergence of robustness-enhancing loops in complex networks can be attributed to nonstationarity of the forcing terms rather than optimality of the network configuration.

Sommario

In questa tesi proponiamo un nuovo modello che congetturiamo rappresenti una nuova formulazione del Problema di Trasporto Ottimo, un'area della matematica notevolmente sviluppatasi negli ultimi ultimi anni e che studia come trasportare in maniera efficiente delle risorse da un luogo ad un altro. La formulazione da noi proposta è l'estensione infinito-dimensionale di un modello nato per descrivere il comportamento di una muffa, dal nome Physarum Polycephalum (PP), capace di trovare il cammino minimo tra due fonti di cibo. Nel modello originale, definito su grafi, il corpo di PP viene schematizzata come un tubo attraverso il quale il trasporto di risorse avviene per mezzo di un flusso dato dal prodotto di un gradiente di pressione per un coefficiente di diffusione. Quest'ultimo varia nel tempo in funzione dell'intensità del flusso stesso, descrivendo in tal modo la dinamica adattativa della muffa. L'equivalenza tra tale modello e la soluzione di problemi di trasporto ottimo su grafi è giá stata dimostrata. Il formulazione da noi proposta abbandona la struttura finito dimensionale del grafo per passare in un ambiente continuo. Il derivante modello è descritto da un sistema composto da un'equazione ellittica con un coefficiente di diffusione e un'equazione differenziale ordinaria per il coefficiente. In questa tesi proponiamo la congettura che quest'ultimo sistema ammetta un equilibrio stazionario legato alla soluzione di problemi di trasporto ottimo, sia per il caso L^1, sia per i problemi di trasporto congestionato e ramificato.

Da un punto di visto teorico, siamo riusciti a provare che il modello é ben posto solo assumendo determinate ipotesi di regolarità del coefficiente di diffusione e delle densità che descrivono la configurazione iniziale e finale delle masse trasportate. Nonostante ciò, numerosi risultati numerici mostrano come la soluzione approssimata del nostro modello converga a soluzioni stazionarie che ben si confrontano con la soluzione dei sopracitati problemi di trasporto ottimo. Riusciamo inoltre ad ottenere soluzioni numeriche che assomigliano fortemente alle strutture singolari del trasporto ramificato, dando ulteriore supporto alle nostre congetture. Nonostante alcune difficoltà numeriche, essenzialmente legate al malcondizionamenteo di sistemi lineari, lo schema numerico utilizzato per la discretizzazione del nostro modello, la cui implementazione risulta relativamente

semplice, si é rivelato estremamente efficiente e robusto, sia da punto di visto delle convergenze numeriche, sia dal punto di vista dell'efficienza computazionale.

Il nostro modello si presta inoltre a numerose applicazioni a problemi reali, come lo studio della morfologia dei fiumi e la modellizzazione dell'evoluzione delle radici delle piante, argomenti discussi nella parte finale della tesi. In ultimo, sulla base di prove numeriche, analizziamo come la presenza di loop in reti complesse, indice della loro robustezza, possa essere interpretata non come una proprietá di "ottimalitá" della rete stessa, bensì come un riflesso della non stazionarietà delle forzanti.

Table of Contents

Introduction

In this thesis we propose a model that we conjecture represents a new and original formulation of the Optimal Transport Problem. This type of problems naturally arises in several real life applications in which one seeks least-cost strategies to reallocate "resources" from one place to another. Many biological transport systems (for example blood vessels in animal bodies, plant roots, river networks, etc.) grow as the result of a complex evolutionary process that promotes "optimality" in response to natural selection principles. Also many human-built systems (such as road and communication networks) are designed to minimize construction costs and, at the same time, to guarantee optimal transport of resources. The study of the leading principles shaping "optimal transportation" structures is the fundamental question addressed by OT theory.

Optimal Transport Problem (OTP). The first mathematical formulation of the OTP was introduced by Gaspard Monge in 1781 in *"Mémoire sur la théorie des déblais et des remblais"* [52] as a problem of military fortification construction, where we have to move optimally some material from a starting to a final configuration. The mathematical formulation of the Monge problem, pictorially

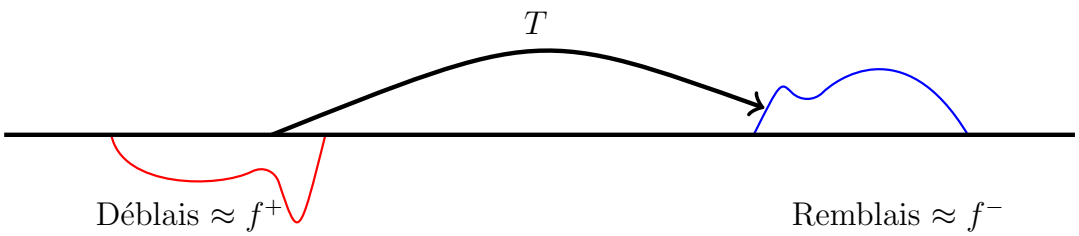

Figure 1: The Monge OTP formulation: find the least effort map T moving the soil from one excavation (Déblais=f^+) to an embankment (Remblais=f^-) of equal volume.

represented in Figure 1, reads as follows. Consider $\Omega \subset \mathbb{R}^d$ and take two density functions $f^+, f^- : \Omega \mapsto [0, \infty[$ with $\int_\Omega f^+ = \int_\Omega f^-$, describing respectively the initial and the final configuration of the mass to be transported. Given a transport cost $c : \Omega \times \Omega \mapsto \mathbb{R}$, we want to find among all measurable maps T from Ω to Ω the optimal solution T^* solving

$$\inf_{T:\Omega \mapsto \Omega} \left\{ \int_\Omega c(x, T(x)) df^+(x) \quad \text{s.t. : } \begin{array}{l} T_\# f^+(A) = f^-(A) \\ \forall A \text{ measurable set in } \Omega \end{array} \right\}$$

where $T_\# f^+(A) := f^+(T^{-1}(A))$. Monge Problem was reformulated by Leonid Kantorovich who introduced a relaxed version, inspired by the Linear Programming, nowadays called Monge-Kantorovich Problem. This topic has received a lot of attention in the last years during which many different formulations and problems have been introduced (see [73, 64]).

Divergence constrained Problem. One typical formulation of the OTPtries to find the optimal mass movement in the case where mass concentration along the transport is penalized. like, for example, in the study of urban traffic or crowd motion. On the contrary, mass concentration is a common strategy adopted by different "transport infrastructure" such as tree branches and roots, blood vessels, river networks, etc. These two problems have a common formulation: given $0 < q < 2$, find among all vector-valued functions $v : \Omega \mapsto \mathbb{R}^d$, orthogonal to the boundary of Ω, the optimal v^* that solves

$$\inf_v \left\{ \int_\Omega |v|^q \ : \ \mathrm{div}(v) = f^+ - f^- \right\} \tag{1}$$

In the case $q \in]1, 2[$, this problem is called *Congested Transport Problem* since the convex power penalizes concentration ($a^q + b^q < (a + b)^q$). When $q = 1$ it is called *Beckmann Problem* and is equivalent to the Monge-Kantorovich OTP with cost equal to the Euclidean distance. In the case $q \in]0, 1[$, mass concentration is encouraged ($(a + b)^q < a^q + b^q$) and the problem is called *Branched or Ramified Transport Problem* (BTP). Note that in this latter case the singularity of the resulting structures requires a careful definition of the above integral.

Dynamic Monge-Kantorovich equations. The new formulation of the OTP proposed in this thesis is an infinite dimensional extension of a discrete model introduced in [68] in order to describe the behavior of a slime mold called *Physarum*

Polycephalum (PP), that recently captured the interest of mathematicians and biologists for its optimization abilities (this explain the words "Biologically Inspired" in the title). The simple structure of PP allows the encoding in a simple but effective mathematical model of its surviving strategies that can be considered as optimal, since they have been tuned during thousands of years of natural selection. Optimal transport policies can be deduced by studying how PP reallocate nutrients in its body. In [68] the authors note that "Experimental observation shows that tubes (that form the body of PP) with larger fluxes are reinforced (i.e. they expand), while those with smaller fluxes degenerate (i.e. they shrink)". In order to describe such adaptation the authors introduce a time-varying conductivity satisfying an evolution equation that, in our infinite-dimensional framework, is transposed into the second equations of the following system:

$$-\operatorname{div}\Big(\mu(t,x)\,\nabla\,u(t,x)\Big) = f(x) = f^{+}(x) - f^{-}(x) \tag{2a}$$

$$\partial_t \mu(t,x) = \left[\mu(t,x)|\,\nabla\,u(t,x)|\right]^{\beta} - \mu(t,x) \tag{2b}$$

$$\mu(0,x) = \mu_0(x) > 0 \qquad \mu(t,x)\,\nabla\,u(t,x)\cdot n_{\partial\Omega} = 0 \tag{2c}$$

where $\mu : ([0,+\infty[\times\Omega) \mapsto]0,+\infty[$ is an isotropic conductivity coefficient, and $u : ([0,+\infty[\times\Omega) \mapsto \mathbb{R}$ is a potential function. The functions f^+ and f^- represent, respectively, the mass injected and absorbed, and thus they have to be balanced, which means $\int_\Omega f^+ = \int_\Omega f^-$. The adaptive dynamics described by Equation (2b) has two components: the first one is the increasing part given by the flux magnitude modulated by a power $\beta > 0$, while the second is a decay term.

Conjectures. Inspired by analogous results developed for the discrete setting in [13], we conjecture that the solution $(\mu(t,\cdot), u(t,\cdot))$ of Equation (2) converges at large-times to an equilibrium configuration $(\mu^*_\beta(\cdot), u^*_\beta(\cdot))$, and that the vector field $v^*_\beta(\cdot) = -\mu^*_\beta(\cdot)\,\nabla\,\mu^*_\beta(\cdot)$ is related to the solution of the problem in Equation (1).

Existence and Uniqueness. Our results are still in the form of conjectures for the main reason that the problem of showing existence and uniqueness for the solution pair $(\mu(t), u(t))$ is still open. We are able to prove a partial result, namely a local-in-time existence and uniqueness theorem in the case $\beta = 1$ and under the assumption of $f^+, f^- \in L^\infty(\Omega)$ and $\mu_0 \in \mathcal{C}^\delta(\Omega)$ with $0 < \delta < 1$. The proof requires an original extension of classical results of regularity theory of elliptic equations in Hölder spaces.

Lyapunov-candidate functional. An additional step in support of our conjectures is the fact that we are able to identify a Lyapunov-candidate functional $\mathcal{L}_\beta$ that decreases in time along $(\mu(t), u(t))$ and that reads as:

$$\mathcal{L}_\beta(\mu, u) := \frac{1}{2}\int_\Omega \mu |\nabla u|^2 \, dx + \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{2-\beta}{\beta}} \, dx \tag{3}$$

In the case $0 < \beta \leq 1$, we are able to show that the minimization of the above functional is equivalent to the problem in Equation (1) with $q = 2 - \beta$. In the case $1 < \beta < 2$, convincing numerical results and formal calculations persuade us to conjecture the existence of an extension of the above equivalence, whereby the relation $q = 2 - \beta$ (i.e., $q \in]0, 1[$ in Equation (1)) relates the steady state solution $(\mu^*_\beta(\cdot), u^*_\beta(\cdot))$ of our DMK and the solution of the BTP (always given by Equation (1)). To the writer's best knowledge, this equivalence represents a new formulation of the Congested, Beckmann, and Branched Problems.

NewP In addition, we can empirically re-interpret these problems as the search for the "transport infrastructure" μ^* that gives the "optimal" trade off between the energy dissipated in transporting the mass f^+ towards f^- via potential flow (the first term of $\mathcal{L}_\beta$) and the cost of building the "optimal infrastructure". The cost is assumed to be proportional to the total mass of μ weighted with the power $P(\beta) = \frac{2-\beta}{\beta}$. According to this re-interpretation, for $\beta = 1$, the two cost terms in $\mathcal{L}_\beta$ are balanced and it is convenient to build a transport infrastructure that contained within the convex envelop of the supports of f^+ and f^-. For $0 < \beta < 1$, since the exponent $P(\beta)$ is greater than 1, the convexity of the function $(\cdot)^{P(\beta)}$ encourages the spreading of the support of μ beyond the convex envelop (i.e., we can afford to build a spatially distributed transport infrastructure). On the contrary, for $1 < \beta < 2$ the formation of concentrated transport patterns is determined by the concavity of $P(\beta)$ that favors an infrastructure with a hierarchically increasing transport capacity. In all cases the optimal vector field is $-\mu^*_\beta \nabla u^*_\beta$ where ∇u^*_β gives the transport direction. We should remark here that the case $\beta > 1$ does not possess a unique solution, but rather the time-asymptotic configuration depends upon the given initial solution μ_0. Another important characteristic of our formulation is that it provides an empirical dynamics with which the transport density adapts towards the optimal solution, which can be then reinterpreted as an asymptotic equilibrium of an infinite-dimensional system. We would like to note that, while our partial theoretical results supply abundant information in

support of our claims for the case $0 < \beta < 2$, convincing numerical results show that the model could be extended beyond the limit and distinctive fractal-like structures consistently emerge also for $\beta > 2$.

Numerical solution of DMK. An important element of the proposed formulation, actually the distinctive property at the origin this study, is that its numerical solution is feasible and efficient using "simple", albeit rigorous, discretization approaches. In fact, the time-dynamic of the DMK model can be reinterpreted as pseudo-transient giving rise to approximates the solution of the minimization problem in Equation (1).

The equations of the DMK model can be easily projected into standard finite dimensional spaces. The most stable discretization proceeds following the ensuing steps. First, the ordinary differential equation for the transport density in Equation (2b) is projected onto a piecewise constant FEM space defined on a triangulation of the domain. Then the elliptic equation in Equation (2a) is discretized using a linear Galerkin FEM method defined on uniformly refined triangles. Finally, the resulting non-linear differential-algebraic system of equations is solved by means of a first order Euler method (forward or backward) and is coupled with a simple Picard iteration to resolve the non-linearity. The procedure is iterated in time until relative differences on the spatial norm of the transport density are smaller than a predefined tolerance.

After a careful verification step aimed at determining experimentally the convergence characteristics of the numerical approach, we solve several numerical test cases in $\mathbb{R}^2$. Some of these tests are taken from the relevant literature, while some were specifically designed to support our conjectures. The developed solver turned out to be surprisingly robust and efficient, being able to find convincing solutions even in the most extreme cases with highly irregular patterns. We strongly presume that these characteristics are inherited by the numerical scheme from the original DMK model structure. In other words, from the numerical point of view, time acts as a relaxation parameter that drastically smooths out the difficulties in finding an approximation to the problem solution.

In all cases, the numerical tests support the conjecture that indeed the dynamic model possesses a time-asymptotic equilibrium point, and the asymptotic solutions confirm our formal results. In the case $\beta > 1$ we consistently obtain an approximation of μ^*_β that displays a branching structure resembling the

solution of the BTP. In this case, additional numerical difficulties arise in the solution of the linear systems stemming from the FEM discretization of the elliptic equation. We have studied appropriate (and novel) modifications to the Preconditioned Conjugate Gradient method that allow us to successfully tackle these highly ill-conditioned systems.

We would like to remark that the numerical efficiency of our solver can be easily improved by using slightly more complicated but standard numerical algorithms. Currently under study is the development of a Newton method for the solution of the non-linear systems in the case of implicit Euler time-stepping, to completely exploit the geometric convergence towards steady state only hinted at in the present thesis. Still, more theoretical work is needed to determine the exact relationships between the spatial discretization spaces used for u_h and μ_h that guarantee stability of the approach.

Gradient Flow structure. Looking to obtain a gradient flow structure, we modify the extended DMK equations maintaining the power β only on the transport density term, while fixing $\beta = 2$ for the magnitude of the gradient of the transport potential. We then obtain the following:

$$-\operatorname{div}(\mu(t,x)\nabla u(t,x)) = f^+(x) - f^-(x) \tag{4a}$$

$$\partial_t \mu(t,x) = \mu^\beta(t,x)|\nabla u(t,x)|^2 - \mu(t,x) \tag{4b}$$

$$\mu(0,x) = \mu_0(x) \quad \mu(t,x)\nabla u(t,x)\cdot n_{\partial\Omega} = 0 \tag{4c}$$

The corresponding Lyapunov-candidate functionalcan be evaluated as:

$$\Phi_\beta(\mu,u) := \frac{1}{2}\int_\Omega \mu|\nabla u|^2\,dx + \frac{1}{2}\int_\Omega \frac{\mu^{2-\beta}}{2-\beta}\,dx$$

The conjecture above can be reproposed for Equation (4) simply changing the relation between the exponents β and q, that now read as $q = 2\frac{2-\beta}{3-\beta}$. The principal interest of the above alternative version is that formal calculations show that the system in Equation (4) can be recast within the framework of a *Gradient Flow* by introducing an appropriate metric in the ambient space of the variable μ. The Gradient Flow approach would give the proof of existence and uniqueness, and it would also give a reinterpretation of the model as a steepest decedent algorithm for the minimization of Φ_β.

The complexity contained in the proofs of local existence under the assumption $f^+, f^- \in L^\infty(\Omega)$ and $\mu_0 \in \mathcal{C}^\delta(\Omega)$, as well the difficult technical issues we are facing

in the Gradient Flow approach show that the proper tuning of the spaces where the solution of Equations (2) and (4) lives is a very delicate question. Nevertheless we are convinced that this avenue is worth pursuing and that our model can represent a new, unified, formulation of transport problem as in Equation (1).

Thesis Structure. The thesis is organized as follows. First, in Chapter 1 we present common definitions and well known results related to Optimal Transport Problems , starting from the L^1 Monge-Kantorovich problem and concluding with the Branched and Congested Transport Problems . Then, we expose the original contributions of this work. In Chapter 2 we describe the deduction of our dynamic model and we prove local existence and uniqueness of the solution for $\beta = 1$. Moreover, we introduce the Lyapunov-candidate functional $\mathcal{L}$, and prove the equivalence between its minimization and Beckmann Problem. The chapter concludes with the derivation of the numerical method used to discretize the model, and with the experiments conducted to support our conjecture In Chapter 3 we describe the extended DMK model, the corresponding Lyapunov-candidate functional $\mathcal{L}_\beta$, and the connections between its minimization and the solution of the congested and branched transport problems. Subsequently we present the numerical experiments supporting our conjecture, for the cases $0 < \beta < 1$ and $\beta > 1$. Chapter 4 is dedicated to the development and testing of the numerical strategies adopted to solve the extremely ill-posed linear system arising from the discretization of the elliptic equation for $\beta > 1$.

In Chapter 5 we present the variant of the DMK model described in Equation (4) and show the formal calculations that lead to the conjecture that this variant can be interpreted as a Gradient Flow in metric spaces.

Finally, Chapter 6 discusses a few application examples that show the applicability of the proposed model and its numerical solver to tackle real-world problems in the field of geomorphology, plant-root dynamics, complex networks, etc.

Chapter 1

Optimal Transport Problem

In this chapter we present an overview of the Optimal Transport Problem (OTP), that studies how to find the least-cost strategy to reallocate a mass from an initial configuration to another. This chapter contains a series of fundamental definitions and well known results of the OTP theory, highlighting those arguments more related to our contribution in this field. We start from the original formulation given in 1781 by Gaspard Monge, then we describe the relaxed Kantorovich formulation. We focus on the case with the transport cost equal to the Euclidean distance, called L^1-OTP, and discuss in different equivalent formulations. Then we define of the so-called Monge-Kantorovich partial differential equations that play a central rule in the L^1-OTP theory and in this thesis. In the last two sections we present the *Branched and Congested Transport Problems*, that consider transports where mass concentration along the transport is either penalized or favored.

1.1 Monge Formulation

The first formulation of the OTP was introduced by Gaspard Monge in 1781 in *"Mémoire sur la théorie des déblais et des remblais"* [52] as a problem of military fortification construction. He studied the least work strategy to move a certain amount of earth from the original place (the "déblais") to an embankment of equal volume (the "remblais"), assuming that the transport cost is given by to the product of mass to be moved times and the distance to be covered.

In modern mathematical terms the "Déblais" and the "Remblais" are repre-

sented as two non-negative measures f^+ and f^-, with equal volume (hereafter, we will denote with $\mathcal{M}_+(X)$ the set of the non-negative measures defined on a measure space X). In great generality, the ambient spaces for the measures f^+, f^- are two complete and separable spaces X and Y, but in most of the cases we will assume $X = Y = \Omega$ where Ω is an open, bounded, convex, and connected domain in $\mathbb{R}^d$ and with smooth boundary. The mass movements of the Monge problem are called *Transport Maps* and they belong to the set

$$\mathcal{T}(f^+, f^-) := \left\{ \begin{array}{c} \text{Measurable map } T : X \mapsto Y \\ \text{s.t. : } T_\# f^+ = f^- \end{array} \right\}$$

where the image measure $T_\# f^+$ is defined as

$$T_\#(f^+)(A) := f^+(T^{-1}(A)) \quad \forall\ A \text{ measurable set in } X$$

The Monge problem now reads:

Problem 1 (Monge Problem)**.** *Given two non-negative finite measures f^+ and f^- on X and Y satisfying $f^+(X) = f^-(Y)$, a cost functional $c : X \times Y \mapsto \mathbb{R}$, find $T^* \in \mathcal{T}(f^+, f^-)$ solving*

$$\min_{T \in \mathcal{T}(f^+, f^-)} I(T) := \int_X c(x, T(x)) df^+(x) \tag{1.1}$$

The problem described by Monge is a particular case with $c(x, y) = |x - y|$. In general Problem 1 can be ill-posed and the optimal map may not exist. For example when f^+ is a Dirac measure and f^- is not, the class $\mathcal{T}(f^+, f^-)$ is empty since the image measure $T_\# f^+$ is atomic (see [64, Section 1.4] and [72] for more examples and counterexamples). Even assuming that the the measures f^+ and f^- have smooth densities, the idea of using direct method of the calculus of variation in order to find an optimal plan as the limit of a minimizing sequence in $T_n \in \mathcal{T}(f^+, f^-)$ for the functional $I(\cdot)$ in Equation (1.1) may not always work (see [30, p. 5-6] or [64, Exercise 1]).

NewP The mathematical difficulties arising from the attempt of solving directly the Monge Problem were overcome by Leonid Kantorovich who introduced a relaxed version of the Monge Problem, that we present in the next section.

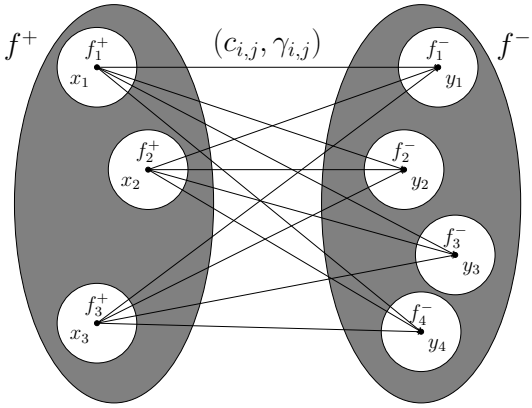

Figure 1.1: The gray ellipses represent the supports of f^+ and f^- of the Monge Problem. Consider the points $x_i \in \mathbb{R}^d$ ($y_i \in \mathbb{R}^d$) in the support of f^+ (f^-), and associate to them the mass f_i^+ (f_j^-) concentrated in a small ball centered in x_i, (y_i). For any pair of points x_i, y_j a transport cost $c_{i,j}$ is associated. Then the optimization problem described in Equation (1.2) is a natural discretization of the Monge Problem.

1.2 Kantorovich Relaxation

The Monge Problem was reformulated by Leonid Kantorovich in [44] in a relaxed version inspired by the following discrete formulation of the Monge Problem. Consider n points $(x_i)_{i=1,n} \in \mathbb{R}^d$ with associated masses $(f_i^+)_{i=1,n}$ and m points $(y_j)_{j=1,m} \in \mathbb{R}^d$ with masses $(f_j^-)_{j=1,m}$, and the additional requirement that $\sum_i^n f_i^+ = \sum_j^m f_j^-$ (we may think of them as atomic discretizations of f^+ and f^-, as described in Figure 1.1). Given the real numbers $(c_{i,j})$ for $(i = 1, \dots, n, j = 1, \dots, n)$ representing the cost of moving one unit of mass from point x_i to point y_j, we look for $\gamma_{i,j}^*$ solution of the following minimization problem

$$\min_{\gamma_{i,j}} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{i,j}\gamma_{i,j} \tag{1.2a}$$

$$\text{s.t. } : \sum_{j=1}^{n} \gamma_{i,j} = f_i^+ \quad \sum_{i=1}^{n} \gamma_{i,j} = f_j^- \quad \gamma_{i,j} \geq 0 \tag{1.2b}$$

The discrete problem defined above rewrites the OTP formulated by Monge in the form of a Linear Programming Problem [1] (see in [30, Appendix] for the

[1] Kantorovich is consider one of the fathers of the Linear Programming theory and for this reason he was awarded in 1975 the Nobel Prize in Economics (see [71])

details). This reformulation suggests that instead of searching the solution of the OTP among the transport maps, we should search on the set of *Transport Plans* defined as follows

$$\Pi(f^+, f^-) := \left\{\gamma \in \mathcal{M}_+(X \times Y) \text{ s.t } : \ (\pi_x)_\# \gamma = f^+ \ , \ (\pi_y)_\# \gamma = f^-\right\}$$

where π_x and π_y are projection maps $(x, y) \mapsto x$ and $(x, y) \mapsto y$. The transport plan γ and the relative constraints are clearly the continuous version of $\gamma_{i,j}$ and the constraints described in Equation (1.2b) , respectively.

The infinite dimensional version of the problem in Equation (1.2), nowadays known as *Kantorovich Primal Problem*, reads as follow:

Problem 2 (Kantorovich Primal Problem)**.** *Given two non-negative finite measures f^+ and f^- on X and Y satisfying $f^+(X) = f^-(Y)$, and given a cost function $c : X \times Y \mapsto \mathbb{R}^+$, find the optimal transport plan $\gamma^* \in \Pi(f^+, f^-)$ that solves*

$$\min_{\gamma \in \Pi(f^+, f^-)} \mathcal{K}_c(\gamma) := \int_{X \times Y} c(x, y) d\gamma(x, y)$$

The Kantorovich formulation is weaker than the Monge's, since for any map $T \in \mathcal{T}(X, Y)$ we can define the plan $\gamma_T = (Id, T)_\# f^+$ that belongs to $\Pi(X, Y)$, and if T^* solves Problem 1 then γ_{T^*} solves Problem 2. The first advantage of the Kantorovich relaxed formulation is that under very mild assumptions on the cost c it is easy to prove the existence of a minimizer for Problem 2, as stated in the following theorem.

Theorem 3. *For any $c : X \times Y \mapsto \mathbb{R}$ lower semi-continuous, Problem 2 admits a solution $\gamma^* \in \Pi(f^+, f^-)$*

The proof is a based on the direct method of the calculus of variations and the proof can be found in [64, Theorem 1.4 and 1.5] for the case with X, Y compact and c lower semi-continuous and bounded from below. The general proof is then given in [64, Theorem 1.7].

The second main advantage of the Kantorovich formulation is that Problem 2 admits a dual problem. Denoting with $\mathcal{C}_b(X)$ the space of continuous and bounded functions on a metric space X, we have the following

Problem 4 (Kantorivich Dual Problem)**.** *Given two non-negative finite measures f^+ and f^- on X and Y satisfying $f^+(X) = f^-(Y)$, and given a cost function*

$c : X \times Y \mapsto \mathbb{R}$. *Let* $\mathcal{L}_c$ *be the set*

$$\mathcal{L}_c := \left\{ \begin{array}{c} (u,v) \in \mathcal{C}_b(X) \times \mathcal{C}_b(Y) \quad s.t. \ : \\ u(x) + v(y) \leq c(x,y) \ \forall (x,y) \in X \times Y \end{array} \right\}$$

Find $(u^*, v^*) \in \mathcal{L}_c$ *solving the maximization problem*

$$\sup_{(u,v)\in\mathcal{L}_c} \mathcal{I}_{(f^+,f^-)}[u,v] := \int_X u(x) df^+(x) + \int_Y v(y) df^-(y) \tag{1.3}$$

Problem 4 is called *Kantorovich Dual Problem* and we have the following theorem:

Theorem 5 (Kantorovich Duality)**.** *Given two non-negative finite measures* f^+ *and* f^- *on* X *and* Y *satisfying* $f^+(X) = f^-(Y)$*, and a cost function* $c : X \times Y \mapsto \mathbb{R}$ *lower semi-continuous, the following equality holds*

$$\min_{\gamma\in\Pi(f^+,f^-)} \mathcal{K}_c(\gamma) = \max_{(u,v)\in\mathcal{L}_c} \mathcal{I}_{(f^+,f^-)}(u,v)$$

This last result can be proved by applying Theorem 55, under the assumption X, Y compact and c continuous. Actually, as already mentioned in [72, Remark 1.4], according to the definitions given in appendix A.1, the Kantorovich Primal Problem is the "real" dual of Problem 2. The complete proof of these statements can be found in [72, Theorem 1.3]. In the same book can be found an extension that considers general assumptions on X, Y and c of Theorem 5.

Remark 1. *The dual problem of the Kantorovich formulation can be seen as the extension to the continuum of the following problem*

$$\max \sum_{i=1}^{n} u_i f_i^+ + \sum_{j=1}^{m} v_j f_j^- \tag{1.4}$$
$$u_i + v_j \leq c_{i,j} \ (i = 1,.,n; \ j = 1,.,m)$$

which is the dual of the discrete problem defined in Equation (1.2)*. As well known, Theorem 5 can be viewed as the extension of duality result of linear programming (see [66]) which says*

$$\min_x \left\{ \boldsymbol{c} \cdot \boldsymbol{x} \ s.t. \ : \ \begin{array}{c} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ \boldsymbol{x} \geq 0 \end{array} \right\} = \max_y \left\{ \boldsymbol{b} \cdot \boldsymbol{y} \ s.t. \ : \ \boldsymbol{A}^T \boldsymbol{y} \leq \boldsymbol{c} \right\} \tag{1.5}$$
$$\boldsymbol{A} \in \mathbb{R}^{m,n} \quad \boldsymbol{x}, \boldsymbol{c} \in \mathbb{R}^n \quad \boldsymbol{y}, \boldsymbol{b} \in \mathbb{R}^m$$

In [30, Appendix] the discrete and dual problems in Equations (1.2) *and* (1.4) *are written in the form of Equation* (1.5)*. The analogies between the continuous and the discrete problem are summarized in Figure 1.2.*

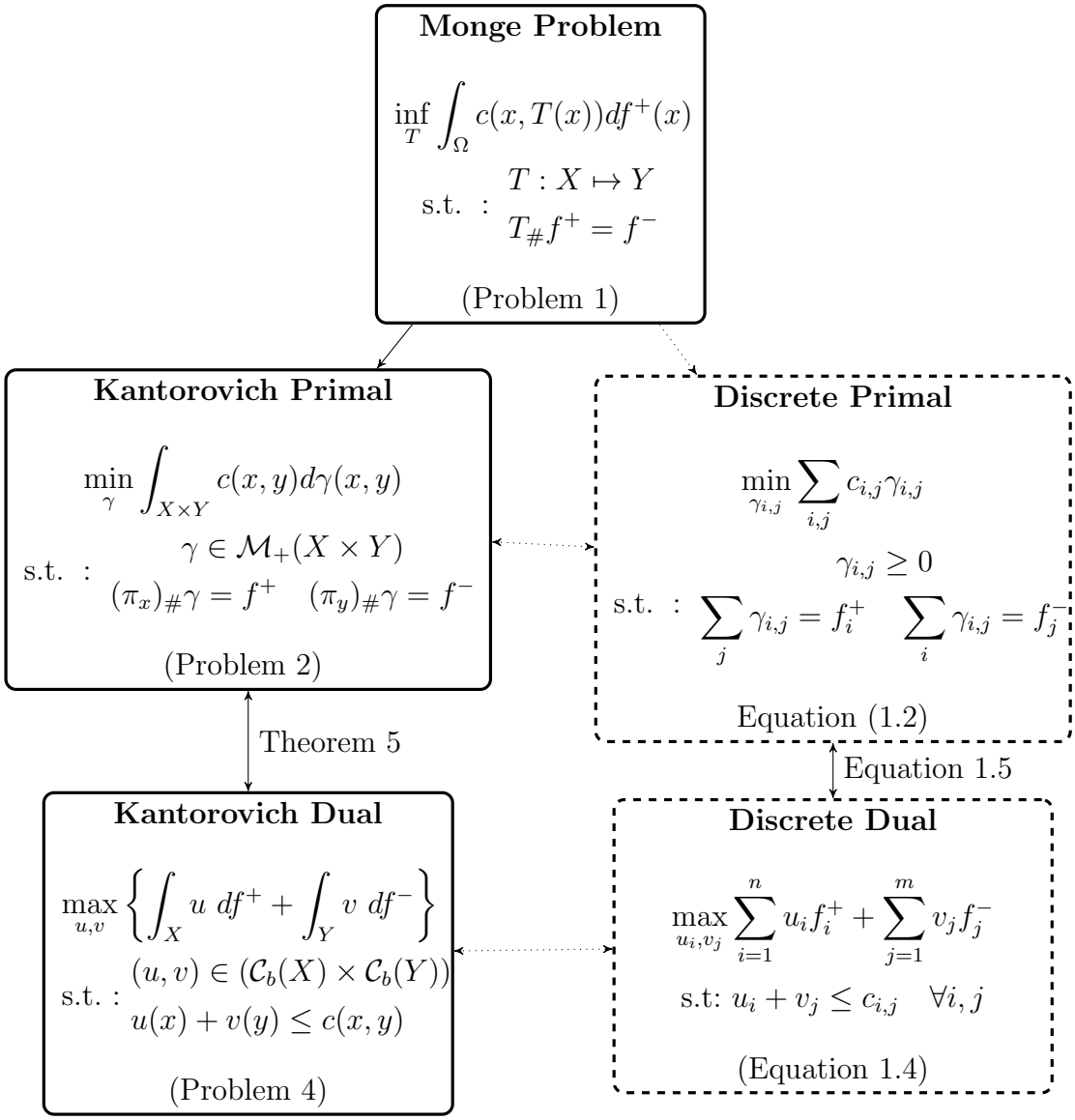

Figure 1.2: Schematic representation of the connections among the different formulations of the OTP discussed in this section. We highlight the analogy between the continuous (left blocks) and the discrete (right dashed blocks) formulations, as well as the analogy between the Kantorovich Duality in Theorem 55 and the duality result in Equation (1.5). The deduction of the discrete primal problem from the Monge Problem is discussed in Figure 1.1.

1.2.1 c-concave function and c-transform

We now introduce the definition of $c/\bar{c}$-transform and $c/\bar{c}$-concave functions.

Definition 6 (c and $\bar{c}$-Transform)**.** *Consider a cost function $c : X \times Y \mapsto \bar{\mathbb{R}}$. Given $u : X \mapsto \bar{\mathbb{R}}$, the c-transform $u^c : Y \mapsto \bar{\mathbb{R}}$ of u is defined by*

$$u^c(y) := \inf_{x \in X} (c(x, y) - u(x))$$

Given $v : Y \mapsto \bar{\mathbb{R}}$, the $\bar{c}$-transform $v^{\bar{c}} : X \mapsto \bar{\mathbb{R}}$ of v is defined by

$$v^{\bar{c}}(x) := \inf_{x \in Y} (c(x, y) - v(y))$$

Definition 7 (c-concave and $\bar{c}$-concave functions)**.** *Consider a cost function $c : X \times Y \mapsto \bar{\mathbb{R}}$. A function $v : Y \mapsto \bar{\mathbb{R}}$ such that exists $u : X \mapsto \bar{\mathbb{R}}$ with $v = u^c$ is called $\bar{c}$-concave. We denote with $\bar{c} - conc(X)$ the set of $\bar{c}$-concave functions defined on Y. A function $u : X \mapsto \bar{\mathbb{R}}$ such that exists $v : Y \mapsto \bar{\mathbb{R}}$ with $u = v^{\bar{c}}$ is called c-concave. We denote with $c - conc(X)$ the set of c-concave functions defined on X.*

We introduced the above definitions in order to state the following proposition

Proposition 8. *Assume that X, Y are compact and the function c is continuous then Problem 4 admits a solution pair $(u^*, v^*) \in \mathcal{L}_c$ with $v^* = (u^*)^c \in \bar{c} - conc(Y)$. This means that Problem 4 can be rewritten as:*

$$\sup_{u \in c - conc(X)} \int_X u(x) df^+(x) + \int_Y (u^c)(y) df^-(y)$$

The function u^* is called Kantorovich Potential.

Proof. See [64, Proposition 1.11] □

1.3 Existence of Optimal Plan and Map

The Kantorovich formulation of the OTP described in previous section is known as Monge-Kantorovich (MK) Transport Problem, and has been studied by several authors in the recent years. Many different formulations of the OTP have been introduced, with connections with areas that are not a priori related to OTP. For the purposes of this thesis we do not need to present all these results, and

we only the reader to [73, 64] for a complete overview of the recent advances in the OTP theory. However, we want to report below a general result, taken from [64], that ensures uniqueness of an optimal transport plan that is solutions of the Kantorovich Primal Problem and, more remarkably, the existence of an optimal transport map for the Monge Problem.

Proposition 9. *Consider a compact domain $\Omega \subset \mathbb{R}^d$, two balanced measures $f^+, f^- \in \mathcal{M}_+(\Omega)$, such that $\partial\Omega$ is f^+-negligible, and f^+ is absolutely continuous with respect to the Lebesgue measure. Assume that the transport cost is of the form $c(x,y) = h(|x-y|)$ with h a strictly convex function, then there exists a unique transport plan $\gamma^* \in \Pi(f^+, f^-)$ of the form $\gamma^* = (Id, T^*)_\# f^+$, with $T^* \in \mathcal{T}(f^+, f^-)$. Moreover, there exists a Kantorovich potential u and T^* satisfies the following relation:*

$$T^*(x) = x - (\nabla h)^{-1}(\nabla(u^*(x)))$$

The above result is fundamental for the so called L^p-OTP in which the cost reds as $c(x,y) = |x-y|^p$ with $p > 1$, that is one of the most studied. Moreover it reconciles the Monge and the Kantorovich formulations. Unfortunately the above proposition can not be applied when we consider the $c(x,y) = |x-y|$, which is the cost studied by in the original Monge Problem, since the strict convexity of function h plays a crucial rule in the proof of the above theorem.

1.4 The L^1-OTP: $c(x,y) = |x-y|$

In this sections we present some results for the OTP with cost $c(x,y) = |x-y|$, called L^1-OTP. This case displays more pathological behavior than those described in Proposition 9. First, the uniqueness of an optimal plan is not ensured. The classical counterexample is the book shifting problem ([2]) Given $n \geq 1$, we consider $f^+ = \chi[0,n] \cdot \mathcal{L}^1$ and $f^- = \chi[1,n+1] \cdot \mathcal{L}^1$ ($\chi[a,b]$ is the indicator function of an interval $[a,b] \subset \mathbb{R}$) . For these measures there exist two optimal maps with equal total cost n. One map is $T_1^*(t) = t + 1$, the other map is

$$T_2^*(t) = \begin{cases} t+n & \text{on } [0,1] \\ t & \text{on } [1,n] \end{cases}$$

thus, also the two plans, $\gamma_1 = (Id, T_1^*)_\# f^+$ and $\gamma_2 = (Id, T_2^*)_\# f^+$ are optimal for the Kantorovich Primal Problem. However, the minimal assumptions on

Ω, f^+, f^- that ensure the existence of optimal transport map solution of the Monge Problem are still a matter of research. Despite these difficulties and peculiarities, the L^1-OTP has a rich mathematical theory, with different formulations that are the main topic of this section.

We will restrict to the case $X = Y = \Omega$, **where $\Omega \subset \mathbb{R}^d$ is an open, bounded, connected, and convex domain with smooth boundary**. Note that most of the following results can be extended to $\mathbb{R}^d$ and to more general framework of Riemanian manifold (see [72]), but for simplicity we consider Ω as above. The first result on the L^1-OTP is the following:

Theorem 10 (Kantorovich-Rubinstein Theorem)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. The Kantorovich Dual Problem in Equation (1.3) with cost function $c(x, y) = |x - y|$ can be rewritten as find $u \in Lip_1(\Omega)$ that solves*

$$\sup_{u \in Lip_1(\Omega)} \int_\Omega u \; df \tag{1.6}$$

with $f = f^+ - f^-$. $Lip_1(\Omega)$ denotes the set of the Lipschitz continuous functions of Ω, with Lipschitz constant equal to 1.

The proof can be found in [72, Theorem 1.14], where it is proved the equivalence between the sets $\mathrm{Lip}_1(\Omega)$ and $c - \mathrm{conc}(\Omega)$ when the cost function c is equal to the Euclidean distance, and that $u^c = -u$. Then Proposition 8 applies.

We now first present a minimization problem that, as stated in the immediately following proposition, turns out to be equivalent to the problem in Equation (1.6)

Problem 11 (Beckmann Problem)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Find $v^* \in [\mathcal{M}(\Omega)]^d$ solving*

$$\inf_{v \in [\mathcal{M}(\Omega)]^d} \Big\{ |v|(\Omega) \;\; : \; \mathrm{div}(v) = f \quad \Big\}$$

with $f = f^+ - f^-$. The divergence constraint on v is the sense of distributions, i.e.

$$\int_\Omega \nabla \varphi \cdot dv = -\int_\Omega \varphi df \quad \forall \varphi \in \mathcal{C}^1(\bar{\Omega})$$

Proposition 12. *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$, then Problem 11 and problem in Equation* (1.6) *are equivalent which means*

$$\sup_{u \in Lip_1(\Omega)} \int_\Omega u \; df = \inf_{v \in [\mathcal{M}(\Omega)]^d} \Big\{ |v|(\Omega) \;\; : \; \mathrm{div}(v) = f \quad \Big\}$$

with $f = f^+ - f^-$

The proof of the above results is, once again an application of Theorem 55. The proof can be found in [17]. The proper characterization of the equivalence between the Beckmann Problem and the the Dual Kantorovich problem, requires the introduction of the quantity called Optimal Transport Density, described in the following section.

1.4.1 Optimal Transport Density and Monge Kantorovich equations

We now give the definition of *Optimal Transport Density* (OT density) as given in [14], which is deeply related to all the formulations of the L^1-OTP presented so far.

Definition 13 (Optimal Transport Density)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Given $\gamma^* \in \Pi(f^+, f^-)$ a minimizer for the Kantorovich Primal Problem (Problem 2) with cost function $c(x,y) = |x - y|$, the* Optimal Transport Density *$\mu^* \in \mathcal{M}_+(\Omega)$ associated to f^+, f^- is defined as:*

$$\langle \mu^*, \varphi \rangle := \int_{\Omega \times \Omega} \int_0^1 |w'_{x,y}(t)| \varphi(w_{x,y}(t)) dt \; d\gamma(x,y) \quad \forall \varphi \in \mathcal{C}(\Omega) \tag{1.7}$$

where

$$w_{x,y}(t) = (1-t)x + ty$$

(The requirement on Ω to be convex is fundamental to consider the curve $w_{x,y}(t)$).

The following theorem summarized a series of the results in [2, 35, 24, 25, 62] obtained by different authors on uniqueness and summability of the OT density.

Theorem 14. *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. If f^+ (or f^-) admits L^1-density with respect to the Lebesgue measure, then the OT density μ^* associated to f^+, f^- is uniquely defined and admits L^1-density with respect to the Lebesgue measure, thus we will indicate it with $\mu^*(f^+, f^-)$ (or, alternatively, $\mu^*(f)$ with $f = f^+ - f^-$). Moreover, if f^+ and f^- admit L^p densities for $1 \leq p \leq +\infty$ then the same holds for μ^*.*

The OT density μ^* plays a crucial rule in the L^1-OTP as the following proposition states.

Proposition 15. *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$, If f^+ and f^- admit L^p-density with $1 \leq p \leq +\infty$ a solution v^* of Problem 11 belongs to $[L^p(\Omega)]^d$ and is can be written as*

$$v^* = -\mu^* \nabla u^* \tag{1.8}$$

where μ^* is OT density $\mu^*(f^+, f^-)$ and u^* is solution of the Dual Kantorovich Problem.

A straightforward consequence of the previous proposition and of Theorem 14 it is the following corollary.

Corollary 16. *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. If f^+ and f^- admit L^1-densities with respect to the Lebesgue measure, then Problem 11 rewrites as the following minimization problem*

$$\min_{v \in [L^1(\Omega)]^d} \left\{ \int_\Omega |v| \, dx \ : \operatorname{div}(v) = f \right\} \tag{1.9}$$

with $f = f^+ - f^-$, and it admits a unique solution v^* given by Equation (1.8)

We finally approach the most important result of L^1-OTP for this thesis, which states that the OT density is described by the following system of equations, called *Monge Kantorovich equations* (MK equations), introduced with different approaches and goals in [16, 32].

Proposition 17 (Monge-Kantorovich Equations)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Assume that f^+ and f^- admit L^1-densities. The OT density $\mu^*(f^+, f^-)$ (defined in Equation (1.7)) and the Kantorovich potential u^* (solution of Equation (1.6)) solve the following equations*

$$-\operatorname{div}(\mu^* \nabla u^*) = f \ \text{ in } \Omega \tag{1.10a}$$

$$|\nabla u^*| \leq 1 \qquad \text{in } \Omega \tag{1.10b}$$

$$|\nabla u^*| = 1 \qquad \text{a.e. in } \mu^* > 0 \tag{1.10c}$$

with $f = f^+ - f^-$.

By the results reported in Proposition 15 the OT density μ^* is uniquely defined, while the Kantorovich Potential u^* solution of the MK equations is not uniquely defined outside the support of μ^*.

Remark 2. *The MK equations and Proposition 15 are still valid for $f^+, f^- \in \mathcal{M}_+(\Omega)$, but as shown in [15] special tools are needed to define the term ∇u^*. We prefer not to consider this generalization for sake of simplicity.*

1.4.2 MK equations via Mass Optimization Problem

The MK equations were studied in [16] as the following Mass Optimization Problem (MOP):

Problem 18 (Mass Optimization Problem)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Find $\nu^* \in \mathcal{M}_+(\Omega)$ that solves*

$$\min_{\nu \in \mathcal{M}_+(\Omega)} \{\mathcal{E}_f(\nu) \quad : \ \nu(\Omega) = 1\}$$

where

$$\mathcal{E}_f(\nu) := \sup_{\varphi \in \mathcal{C}^1(\bar{\Omega})} \Gamma_f(\nu, \varphi) \tag{1.11}$$

$$\Gamma_f(\nu, \varphi) := \int_\Omega \left(\varphi df - \frac{|\nabla \varphi|^2}{2} d\nu \right) \tag{1.12}$$

with $f = f^+ - f^-$.

In [14] the quantity $\mathcal{E}_f(\mu)$ is called *Compliance* and represents the dissipated energy for a given forcing term $f = f^+ - f^-$ and a given conductor distribution ν.

The equivalence between the MOP and MK equations is given by the following proposition:

Proposition 19. *Given the OT density $\mu^*(f^+, f^-)$, a solution of problem Problem 18 is give by*

$$\nu^* = \frac{\mu^*}{\int_\Omega d\mu^*}$$

The proof of the above proposition was first presented in [16].

1.4.3 MK equations via ∞-Poisson Equation

System 1.10 was introduced in [32] in the form of ∞-Poisson equation with forcing term $f = f^+ - f^-$, described by the following proposition.

Proposition 20 (∞-Poisson)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Assume that f^+ and f^- admit Lipschitz continuous densities with respect to the Lebesgue measure. The solution pair (μ^*, u^*) of system 1.10 is the limit for $p \to \infty$ of $(|\nabla u_p|^{p-2}, u_p)$ in where u_p solves the p-Poisson equations*

$$-\operatorname{div}(|\nabla u_p|^{p-2} \nabla u_p) = f$$

with $f = f^+ - f^-$. (the limit for $|\nabla u_p|^{p-2}$ must be understood in the sense of the weak-topology of $L^\infty(\Omega)$, while for u_p the limit is in the topology induced by the uniform norm).*

The assumptions on the Lipschitz continuity of the densities is mainly introduced to prove the existence of an optimal map T^* moving f^+ into f^-, as stated in the following proposition:

Proposition 21 (Solution of the Monge Problem)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Assume that f^+*

and f^- admit Lipschitz continuous densities with respect to the Lebesgue measure. Take $x \in supp(f^+)$ and consider $z(t,x)$ solution of the ODE

$$z'(t) = Z(t, z(t)) \quad Z(t,z) = \frac{-\mu^*(z)\,\nabla\, u^*(z)}{(1-t)f^+(z) + tf^-(z)}$$

with initial data $z(0) = x$. The map T^* defined as

$$T^*(x) := z(1,x)$$

goes from $supp(f^+)$ into $supp(f^-)$, and is the solution of the Monge Problem, with cost equal to Euclidean distance.

The proof of last Proposition can be found in [32] it requires approximation arguments to deal with the term $1/\left((1-t)f^+(z) + tf^-(z)\right))$ outside of the support of f^+ and f^-.

Remark 3. *The result in Proposition 21 was historically the first solution of the OTP two centuries after the original Monge formulation .*

We summarized in Figure 1.3 a schematic representation of the OTP, with particular focus on the different formulations for the L^1-case.

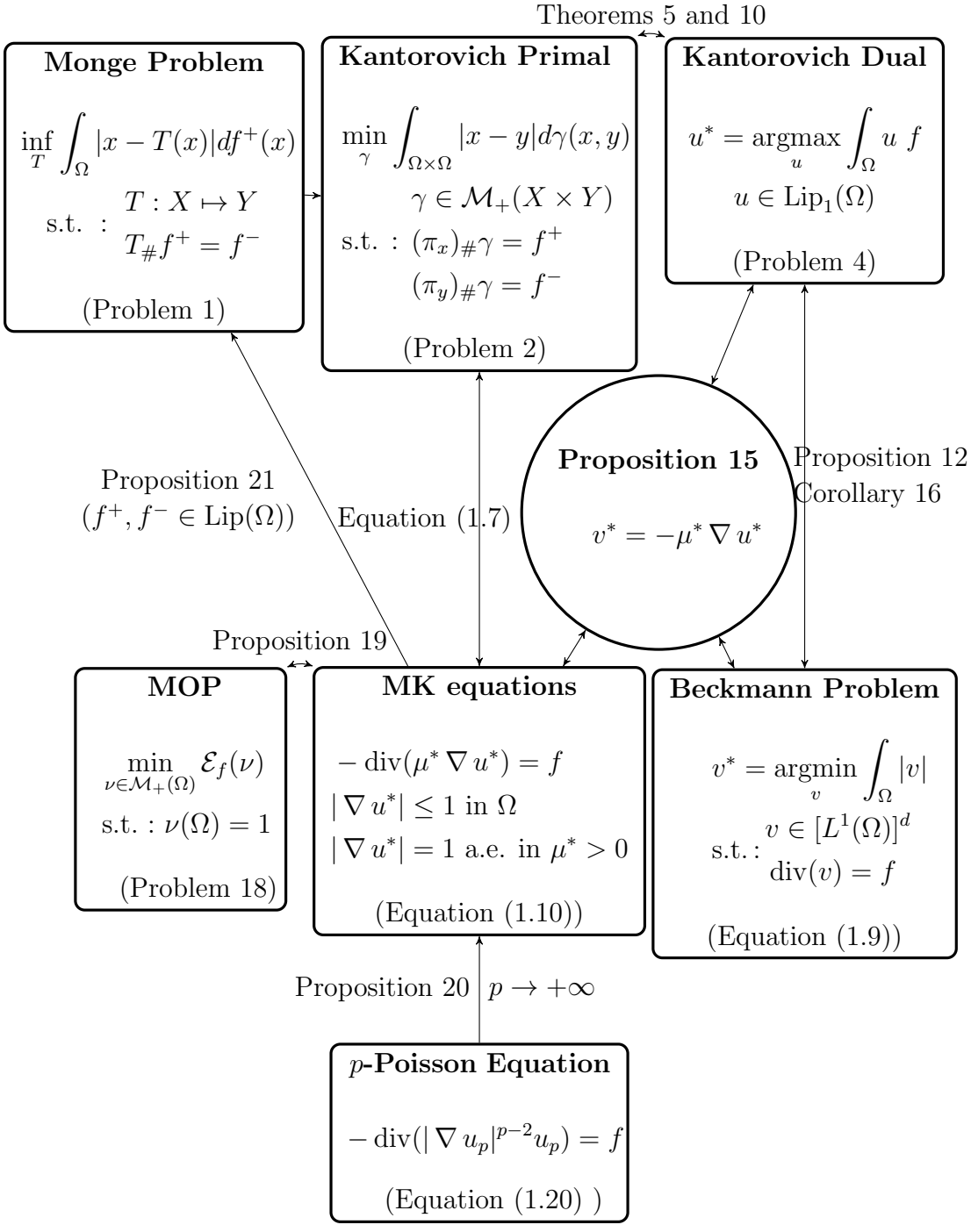

Figure 1.3: Map of the OTP formulations and results for the case $c(x, y) = |x-y|$ in a convex, bounded, connected domain $\Omega \subset \mathbb{R}^d$. The starting measure f^+ and final measure f^- are assumed to have densities with respect to the Lebesgue measure, even if many definition and results do not require this assumption. f denotes $f^+ - f^-$.

1.4.4 Cost in the case of non-uniform distance

In this section we introduce a generalized version of the L^1-OTP in which the Euclidean distance is replaced by a geodetic distance. To this aim we introduce a positive and smooth function $k : \Omega \mapsto \mathbb{R}^+$, so that the distance induced by k can be defined as

$$d_k(x,y) := \inf_{\gamma \in \mathcal{C}^1([0,1],\Omega)} \left\{ \int_0^1 k(\gamma(t))|\gamma'(t)|dt \quad \gamma(0) = x \ , \ \gamma(1) = y \right\} \tag{1.13}$$

It is clear that, when $k \equiv 1$, $d_k(x,y) = |x - y|$, thus the we recover the L^1-OTP. But for a general $k(x)$, using the distance d_k as cost of the OTP, we can recover all the formulations and results presented in Section 1.4 for the case $c(x,y) = |x - y|$. This is standard extension of the L^1-OTP, studied in in [64, 5], in which the positive function $k(x)$ describes the spatial pattern of the resistance to flow.

The Kantorovich dual problem in Equation (1.6) for the cost $c = d_k$ reads as

$$\sup \left\{ \int_\Omega u \ df \quad : \ u \in \mathrm{Lip}_k(\Omega) \right\} \tag{1.14}$$

where

$$\mathrm{Lip}_k(\Omega) = \left\{ \begin{array}{c} u : \Omega \mapsto \mathbb{R} \quad \text{s.t.} \ : \\ \sup\limits_{x \neq y} \dfrac{|u(x) - u(y)|}{d_k(x,y)} < +\infty \end{array} \right\}$$

The divergence constrained problem in Equation (1.9) rewrites as

$$\inf_{v \in [L^1(\Omega)]^d} \left\{ \int_\Omega k(x)|v(x)| \ : \ \mathrm{div}(v) = f \right\} \tag{1.15}$$

Problems in Equations (1.14) and (1.15) are equivalent to the same problems previously defined for the for the Euclidean distance case. The analogous of the MK equations reads:

$$-\mathrm{div}(\mu^* \nabla u^*) = f \ \text{ in } \Omega \tag{1.16a}$$

$$|\nabla u^*| \leq k(x) \text{ in } \Omega \tag{1.16b}$$

$$|\nabla u^*| = k(x) \text{ a.e. in } \mu^* > 0 \tag{1.16c}$$

with $f = f^+ - f^-$

1.5 Branched Transport Problems

In the Monge-Kantorovich formulation the total transport cost does not depend on the intermediate phases between the starting and final configuration of the mass transported. However in many real-life transport problems one may be interested in penalizing or favoring mass concentration along the transport. In this section we present the so called, *Branched Transport Problem*(BTP) an area of the OTP theory that studies the case in which we prefer to move the mass together, (the opposite case where mass concentration is penalized is discussed in the next session). Gilbert in [38] gave the first formulation of the BTP as a problem of finding the minimal cost communication network. In the recent years the Gilbert problem has been reformulated and extended by several authors.

One of the biggest issues in the BTPis that its numerical solution, even in the discrete case of the Gilbert Problem, is known to be NP hard. In this section we give a short overview of the main characteristics and formulations of BTP, following [64, 74].

1.5.1 The Gilbert-Steiner Problem

We now present a simple problem, visualized in Figure 1.4, that exemplifies the main characteristics of the BTP. Consider a courier that has to deliver some boxes from a delivery center to different destinations. In terms of the OTP, the final configuration of the boxes is represented by m Dirac masses $(f_j^- = \delta_{x_j})_{j=1,\dots,n}$ where x_j is the destination location, while f^+ is a single Dirac source with mass m, located at the delivery center. The answer of L^1-OTP is to send each box to the corresponding recipient along straight lines, but it can be more convenient to first move the boxes together and then split them when we get closer to the destination. With two delivery destinations the transport path "drawn" by the L^1-OTP is "V" shaped, while it is "Y" shaped when transport aggregation is encouraged. When we consider multiple destinations we see "star" shaped and "branched" paths.

In the L^1-OTP the "V" or the "star" shaped paths are optimal since the transport cost per unit length is exactly proportional to the amount of goods transported (this is clear looking at the OTP described in Sections 1.1 and 1.2 both in the discrete and continuous cases). The requirement that mass concen-

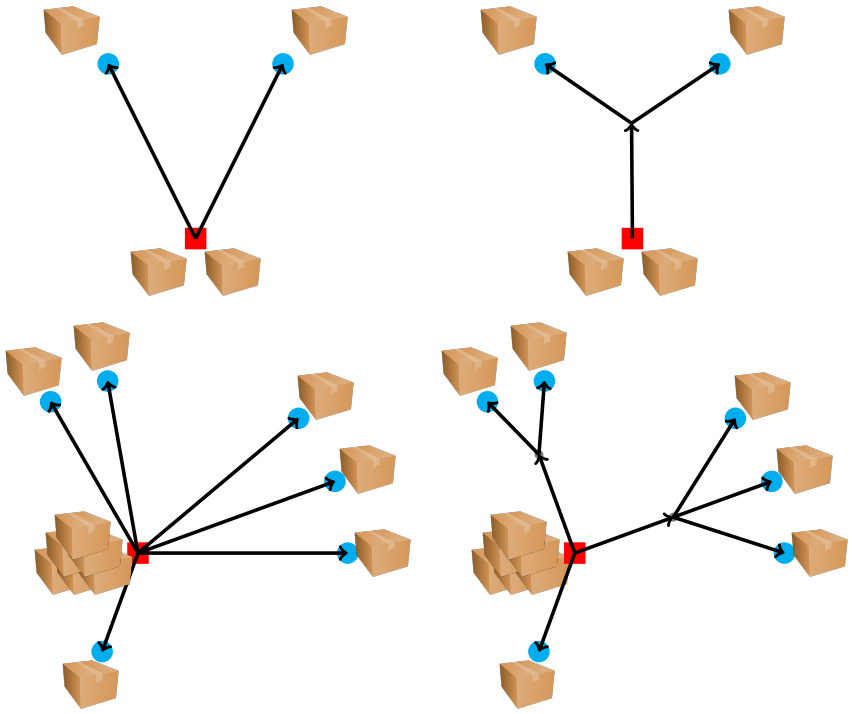

Figure 1.4: Schematic approximate representation of the problem of delivering some boxes from one location (red squares) to different destinations (blue circles). The upper panel reports the paths "drawn" by the the solutions of L^1-OTP (left panel) and BTP(right panel) for the case with one starting point e two destinations, while the lower panel reports the case with several destinations.

tration is encouraged is encoded introducing a functions φ that describes the transport cost for unit length and that has to satisfy the following properties:

$$\varphi(\max(m_1, m_2)) \leq \varphi(\ m_1 + m_2) \leq \varphi(m_1) + \varphi(m_2)$$
$$\varphi(m_1 + c) - \varphi(m_1) \leq \varphi(m_2 + c) - \varphi(m_2) \quad \forall c > 0 \quad m_1 > m_2$$

The first property says that the cost increases with mass transported but it is sub-additive. The second property says that the marginal cost generated by adding some mass to a given background quantity is smaller for bigger backgrounds. The typical choice is the concave function $\varphi(m) = m^\alpha$ with $0 < \alpha < 1$, which satisfies both properties.

The first formulation of this type of transport cost was introduced by Gilbert in [38]. He was studying how to build the minimal cost network connecting the cities located in some point $x_1, \ldots, x_n$ on the plane. To this aim, he introduced a generalization the Steiner problem (which looks for the minimal length network connecting all points $x_1, \ldots, x_n$) in which the cost of construction for unit length described by a sub-additive function φ. In the case $\varphi = m^\alpha$ this problem is called the Gilbert-Steiner Problem. To properly describe the Gilbert-Steiner Problem, we need to give the definition of *Transport Path* as in [74]:

Definition 22 (Transport Path)**.** *Consider two atomic measures $f^+ = \sum_i^n f_i^+ \delta_{x_i}$, $f^- = \sum_j^m f_j^- \delta_{y_j}$ with $\sum_j^m f_j^- = \sum_i^n f_i^+$ (x_i, y_j are points in $\Omega \subset \mathbb{R}^d$ and δ_x is the Dirac measure centered at x). An admissible* Transport Path *from f^+ to f^- is a pair composed by an oriented graph $G = (V, E)$ (V and E denote respectively the set of nodes and the set of edges of the graph G) and a flow function $q : E \mapsto [0, \infty[$ satisfying Kirchhoff law*

$$\sum_{e \in \sigma(v)} q_e = \begin{cases} f_i^+ & \text{if } v = x_i \text{ for some } i \\ -f_j^- & \text{if } v = y_j \text{ for some } j \\ 0 & \text{otherwise} \end{cases} \tag{1.17}$$

where $\sigma(v)$ is the "star" of v, i.e., the set of edges having vertex v in common. The set of all admissible Transport Path from f^+ to f^- is denote by $\mathcal{P}(f^+, f^-)$.

Problem 23 (Gilbert-Steiner Problem)**.** *Given f^+, f^- two balanced atomic masses as in definition 22, and $0 \leq \alpha \leq 1$, we want to find find the Transport Path in $(G, q) \in \mathcal{P}(f^+, f^-)$ minimizing the* Gilbert-Steiner energy

$$E_\alpha(G, q) = \sum_{e \in E(G)} (q_e)^\alpha L_e \tag{1.18}$$

where L_e indicates the length of an edge $e \in E$.

When $\alpha = 0$ we recover Steiner Problem, while the case $\alpha = 1$ can be seen as the discrete version of the L^1-OTP. The problem with $0 < \alpha < 1$ produces the branching structure described in the introduction, and is the starting point of the so called *Branched Transport Problem* (BTP).

It is not trivial to show that the Problem 23 admits a solution. We give here just a sketch of the proof that can be found in [74, Propositions 2.1 and 2.2]. The first step is to restrict the search into the set of acyclic graphs (i.e. graphs with no loops or sequences of connected edges which starts and ends at the same vertex). In fact, given an path (G, q) with G containing loops, we can always obtain another path $(\tilde{G}, \tilde{q})$ with smaller energy E_α, by removing these loops. This assumption uniformly bounds the number of branching vertices by $m + n - 2$. Thus there are finite topologically equivalent Transport Paths and thus there is enough compactness to prove existence of a minimizer.

This discussions suggests that it can be difficult identify a minimizer since we have to explore all possible topological configurations, whose number increases dramatically with the number of source/sink points. Indeed, the solution this problem is known to be NP-hard.

1.5.2 Extension to the continuum

The extension of Problem 23 to general mass densities $f^+, f^- \in \mathcal{M}_+(\Omega)$ with $\Omega \subset \mathbb{R}^d$ was introduced by Xia in [74]. The original idea is to consider two sequence of atomic measures f_n^+ and f_n^- such that

$$f_n^+ \rightharpoonup f^+ \quad f_n^- \rightharpoonup f^-$$

and define the optimal path from f^+ to f^- as the limit of the optimal Transport Paths $(\mathcal{P}_n^*, q_n^*) \in \mathcal{P}(f_n^+, f_n^+)$. In order to make sense to such limit procedure, we need to reformulate Problem 23 in term of Measure Theory. To this aim we need to introduce the notion of 1-rectifiable set in $\mathbb{R}^d$ that, without entering into the details of a proper definition, can be seen as a countable union of Lipschitz curves. Now, consider a triple composed by a 1-rectifiable set $K \subset \Omega$, a vector field $\tau : K \mapsto S^{d-1}$, and a function $q : K \mapsto \mathbb{R}^+$ integrable with respect to the 1-dimensional Hausdorff measure $\mathcal{H}^1$: we can define a vector measure $v =$

$[K, \tau, q] \in [\mathcal{M}(\Omega)]^d$ through the following equation

$$\langle [K, \tau, q], \zeta \rangle := \int_K q(x)\tau(x) \cdot \zeta(x) d\mathcal{H}^1(x) \quad \zeta \in [\mathcal{C}(\Omega)]^d$$

Given a path $(G, q) \in \mathcal{P}(f_n^+, f_n^-)$ we associate to it the vector measure $v_{G,q}$ composed by the triple $[E, \tau_E, q]$ where E is the union of the graph edges, τ_E is the vector measure with unit value defined by the edge direction, and q the weight satisfying Equation (1.17). Kirchhoff law Equation (1.17) rewrites as $\mathrm{div}(v_{(G,q)}) = f_n^+ - f_n^-$, in the sense of distribution that we recall means

$$\int_\Omega \nabla \varphi \cdot dv_{(G,q)} = -\int_\Omega \varphi(df_n^+ - df_n^-) \quad \forall \varphi \in \mathcal{C}^1(\bar{\Omega})$$

The Gilbert-Steiner energy in Equation (1.18) becomes

$$E_\alpha(G, q) = \sum_{e \in E} (q_e)^\alpha L_e = \int_E |q(x)|^\alpha d\mathcal{H}^1(x)$$

We can now extend the definition of E_α to general $v \in [\mathcal{M}(\Omega)]^d$ as follows

$$E_\alpha(v) := \inf \left\{ \liminf_n E_\alpha(G_n, q_n) \ : \ \begin{array}{c} v_{(G_n,q_n)} \rightharpoonup v \\ f_n^+ - f_n^- \rightharpoonup f^+ - f^- \\ \mathrm{div}(v_{(G_n,q_n)}) = f_n^+ - f_n^- \end{array} \right\} \tag{1.19}$$

It can be proved that those $v \in [\mathcal{M}(\Omega)]^d$ with $E_\alpha(v) < +\infty$ are 1-rectifiable and that the energy functional in Equation (1.19) can be rewritten as:

$$E_\alpha(v) = \begin{cases} \displaystyle\int_E |v|^\alpha d\mathcal{H}^1(x) & \text{if} \quad v = [E, \tau, q] \\ +\infty & \text{otherwise} \end{cases}$$

Finally we can give the definition of the BTP for general $f^+, f^- \in \mathcal{M}_+(\Omega)$ as in [74].

Problem 24. *Find $v^* \in [\mathcal{M}(\Omega)]^d$ solving*

$$\inf_{v \in [\mathcal{M}(\Omega)]^d} \left\{ E_\alpha(v) \quad : \quad \mathrm{div}(v) = f^+ - f^- \quad \right\}$$

In general this problem may have no solution. For example if $\Omega = [-a, a]^d$, given $f^+ \in \mathcal{M}_+(\Omega)$ with total mass equal to 1 and f^- equal the unitary Dirac mass centered at 0, then there exists v with $\mathrm{div}(v) = f^+ - f^-$ such that $E_\alpha(v) < +\infty$ if and only if $\alpha > 1 - \frac{1}{d}$. There exist counterexamples that show that the threshold $\alpha^* = 1 - \frac{1}{d}$ is sharp (we refer to [64] for more details on these arguments).

1.6 Congested Transport Problem and p-Poisson Equations

In this section we give a short presentation of that area of OTP in which we want to penalize mass-concentration when we move f^+ into f^-, called *Congested Transport Problem*(CTP). These type of problems has many real-life applications, for example, in the study of urban traffic or crowd motion. One formulation of these type problem reads as follows: we want to find the optimal vector field $v^*\Omega \mapsto \mathbb{R}^d$ that solves

$$\min_v \left\{ \int_\Omega H(|v|)\, dx \ : \ \mathrm{div}(v) = f^+ - f^- \right\}$$

where H is a real super-linear function (see [65] for more details). When we consider $H(|v|) = |v|^q$ with $1 < q < 2$ the above minimization problem is equivalent to solve a well known non-linear elliptic equation, the p-Poisson that reads:

Problem 25 (p-Poisson Equation)**.** *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Assume that the forcing terms f^+ and f^- admit L^q-densities, with $q > 1$, and let p to be the conjugate exponent of q, i.e.*

$$\frac{1}{p} + \frac{1}{q} = 1$$

We want to find the solution of the following non-linear equation

$$-\mathrm{div}(|\nabla u_p|^{p-2} \nabla u_p) = f^+ - f^- = f \tag{1.20}$$

complemented with zero Neumann boundary condition. The above equation is called p-Poisson equation.

The p-Poisson equation (in the weak form) are the Euler-Lagrange of the following minimization problem:

$$\min_{u \in W^{1,p}(\Omega)} \int_\Omega \left(\frac{1}{p} |\nabla u|^p - fu \right) dx$$

with $f = f^+ - f^-$. This formulation of p-Poisson equation, together with the following proposition, explains the relations between the Congested Transport Problem and the p-Poisson equations.

Proposition 26. *Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative measures f^+ and f^- on Ω such that $df^+(\Omega) = df^-(\Omega)$. Assume that forcing term f^+ and f^- admits L^q-densities with $1 < q < +\infty$. Then the following equivalence holds*

$$\min_{u \in W^{1,p}(\Omega)} \int_\Omega \left(\frac{1}{p} |\nabla u|^p - fu \right) dx = \max_{v \in [L^q(\Omega)]^d} \left\{ - \int_\Omega \frac{|v|^q}{q} \, dx \ : \ \mathrm{div}(v) = f \right\}$$

where $f = f^+ - f^-$ and p conjugate exponent of q. The solution u_p of the left-hand side problem and solution $\bar{v}$ the right-hand side problem $\bar{v}$ satisfy the following relation

$$\bar{v} = -|\nabla u_p|^{p-2} \nabla u_p$$

The proof is based on Theorem 55 and can be easily derived from the results in [29, Example 2.2, Chapter 4] for the case with homogeneous Dirichlet boundary condition.

In Figure 1.5 we summarize the optimal transport formulations in the form of minimization problem defined for vector field $v : \Omega \mapsto \mathbb{R}^d$ with divergence equal to $f^+ - f^-$.

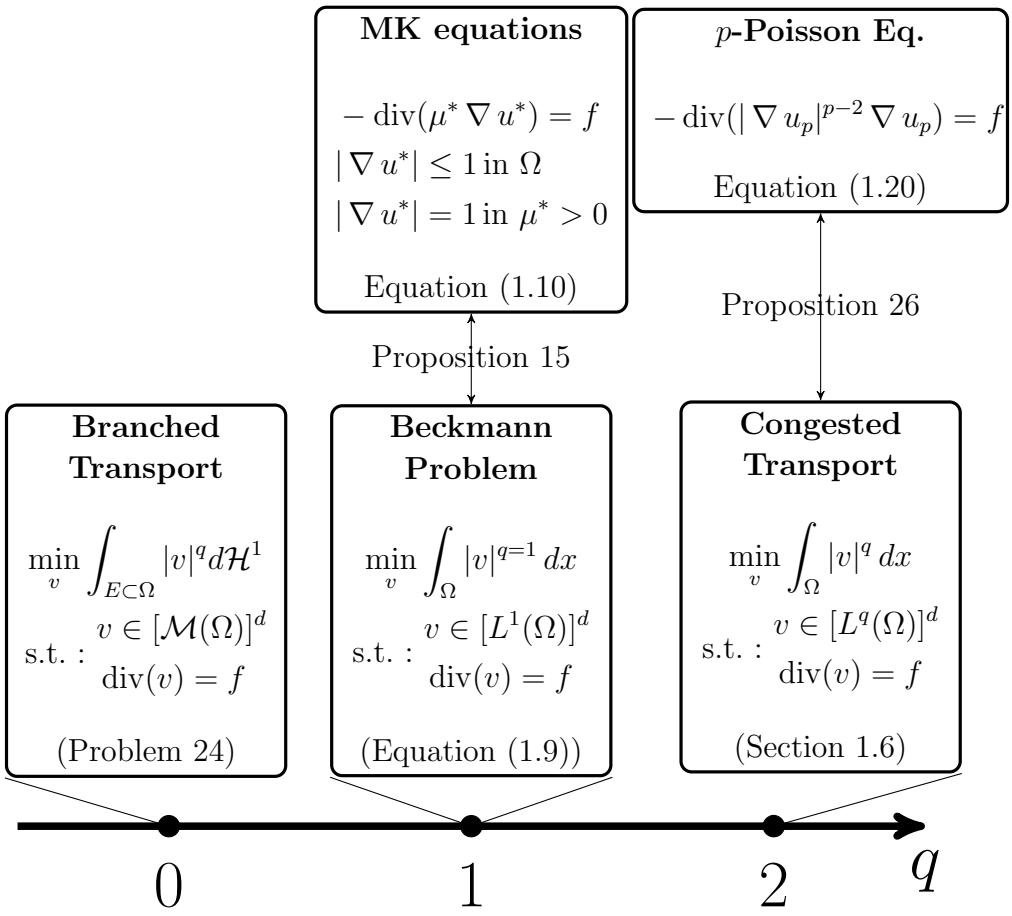

Figure 1.5: Schematic representation of different transport problems that move the mass f^+ into f^- casted in the form of minimization with divergence constrained, or as PDE formulations. The exponents q and p satisfy the relation $1/q+1/p = 1$ and $f = f^+ - f^-$. The exponent $q \in [0, 2]$ modulates how want the mass to move between the initial and the final configurations. For $q \in]0, 1[$ mass concentration is encouraged, while the opposite holds for $q \in]1, 2[$. In the case $q = 1$ the problem is equivalent to the L^1-OTP. The extremal values $q = 0$ and $q = 2$ correspond to the Steiner Problem and the Poisson equation, respectively.

Chapter 2

Dynamic Monge-Kantorovich

In this chapter we develop an original a dynamical formulation of the L^1-OTP and we detail our main conjecture of equivalence between the solution of the MK equations described in Equation (1.10) and the larger-time equilibrium solution of the model proposed in this chapter. . We restrict to the case of f^+ and f^- continuous with respect to the Lebesgue measure, using f^+ and f^- to denote their densities. Our problem is the following: find the pair $(\mu, u) : ([0, +\infty[, \Omega) \mapsto (\mathbb{R}^+, \mathbb{R})$ that solves

$$-\operatorname{div}\Big(\mu(t,x)\,\nabla\, u(t,x)\Big) = f(x) = f^+(x) - f^-(x) \tag{2.1a}$$

$$\partial_t \mu(t,x) = \mu(t,x)|\,\nabla\, u(t,x)| - \mu(t,x) \tag{2.1b}$$

$$\mu(0,x) = \mu_0(x) > 0 \tag{2.1c}$$

complemented with zero-Neumann boundary conditions. Even if a complete proof of our conjecture is still missing, many theoretical and numerical indications support our thesis. In this chapter we first derive the model and analyze its theoretical properties. Next we show how the model can be efficiently solved numerically, and we argue that the proposed approach can be used to approximate the solution of the L^1-OTP. The analysis of existence and uniqueness of the solution pair $(\mu(t,x), u(t,x))$ of system relies on the transformation of the coupled system into an ODE in Banach spaces by defining the operator $u(\mu)$ as the weak solution of Equation (2.1a) given μ. Under the hypothesis of $\mu_0 \in \mathcal{C}^\delta(\Omega)$ and $f \in L^\infty(\Omega)$ we prove existence and uniqueness for $t \in [0, \tau_0[$ with $\tau_0 > 0$ depending on the initial data. Moreover, we identify a Lyapunov-candidate functional $\mathcal{L}$, i.e. a function that decreases along the $\mu(t)$ trajectory, These results are collected in

the paper [33, 34]. Finally we also prove that the minimization of this functional is equivalent to the Beckmann Problem and thus it is equivalent to solving the MK equations.

A schematic summary of the connections between the proposed model and the different L^1-OTP formulations is reported in Figure 2.2.

2.1 From Physarum Polycephalum to Dynamic Monge-Kantorovich

2.1.1 Modeling the dynamics of Physarum Polycephalum

In a recent paper, [68] proposed a mathematical model proposed describing the dynamics of *Physarum Polycephalum* (PP) that, on the basis of experimental evidence [54], is able to find the most efficient network path between food sources. The experiments suggest that in a maze colonized by PP the slime reorganizes itself concentrating on the shortest path connecting the two food sources as reported in Figure 2.1. The abilities of PP of shortest path has have been used effectively for the experimental analysis of transportation networks, with many researchers suggesting that this slime mold is capable of identifying the optimal many-site connecting transportation network, such as the railroads of Tokyo and Spain [69, 1]. Many further surprising properties of PP have been experimentally identified, but we now focus on the mathematical model proposed by [68]. The PP in the channels of the maze is schematized as a undirected planar graph $G = (V, E)$, reported in figure Figure 2.1, with positive edge length $\{L_e\}_{e\in E}$, and two nodes $v = 1, n$ indices where two unitary food sources are located. To each edge $e \in E$ is associated a "conductivity" function D_e and to each node $v \in V$ is associated a "potential"(or pressure) function p_v. The problem is then to find

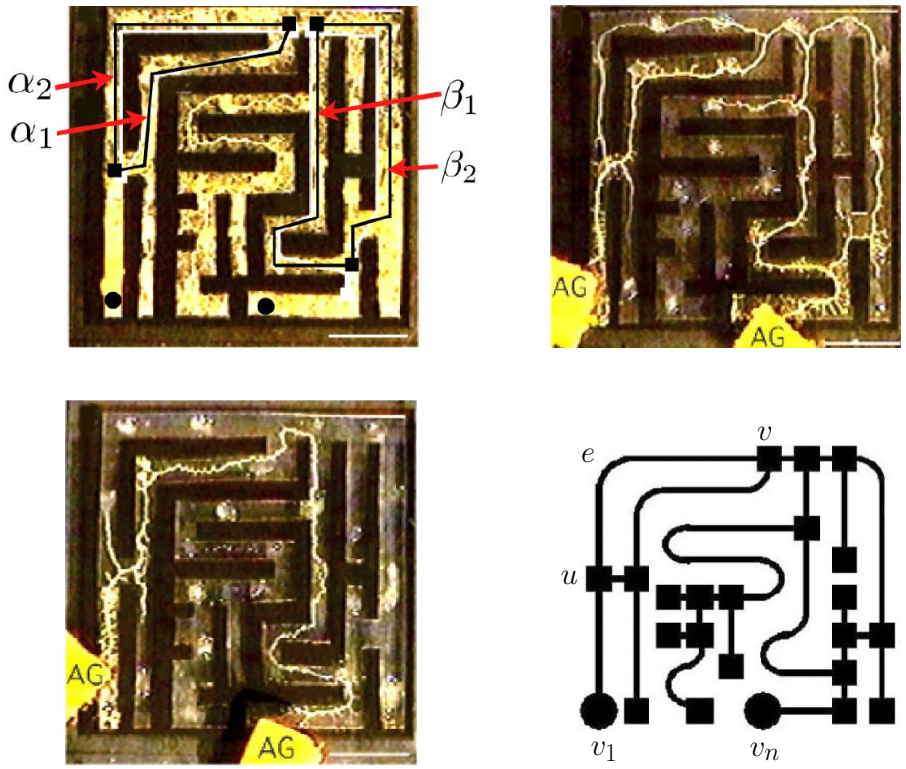

Figure 2.1: Setup of the Physarum Polycephalum experiment described in [54]. The top left panel shows the experimental maze initially filled with PP. The top tight panel shows how, after introduction of two food sources, PP starts retiring from the dead ends of the maze. The bottom left panel displays the final configuration of PP, which concentrates only on the shortest path connecting the two food sources. The bottom right panel shows the graph describing the topology of the maze channels as used in [68]. (Figures reprinted from [68] Copyright (2018), with permission from Elsevier)

the optimal distribution of the pair (D_e, p_v) that satisfies

$$\sum_{e\in\sigma(v)} Q_e(t) = f_v = \begin{cases} +1 & v = 1 \\ -1 & v = n \\ 0 & v \neq 1, n \end{cases} \qquad \forall v \in V, \quad (\text{“Kirchhoff-law”}) \tag{2.2a}$$

$$Q_e(t) = D_e(t)\frac{(p_u(t) - p_v(t))}{L_e} \qquad \forall e \in E, \quad (\text{“Fick-Poiseuille”}) \tag{2.2b}$$

$$D'_e(t) = g\left(|Q_e(t)|\right) - D_e(t) \qquad \forall e \in E, \quad (D_e \text{ dynamics}) \tag{2.2c}$$

$$D_e(0) = \hat{D}_e(0) > 0 \qquad \forall e \in E, \quad (\text{initial data}) \tag{2.2d}$$

where $e = (u, v)$ denotes the edge of G connecting vertices u and v where $\sigma(v)$ is the “star” of v, i.e., the set of edges having vertex v in common, and $g : \mathbb{R}^+ \mapsto \mathbb{R}^+$ is a non-decreasing function with $g(0) = 0$. This model can be explained heuristically using a classical hydraulic analogy, eventually motivating the above introduced terms “balance law-Kirchhoff” and “Fick-Poiseuille”. We think of the graph G as representing the set of pipes where the flow of a fluid driven by the vertex source function f_v occurs. Then, the first Equation (2.2a) can be identified as the enforcement of the fluid mass balance, while Equation (2.2b) is the momentum balance stating that the flux in each graph edge is proportional to the discrete gradient of the vertex potential function p_v via a conductance coefficient D_e (inverse of a resistance). Hydraulic resistance to flow is known to be proportional to the pipe perimeter, and hence to its diameter. Thus, the evolutive Equation (2.2c), which forms the innovative core of the model, asserts the intuitive behavior that to optimally (with minimal energy loss) accommodate larger fluxes the pipe diameter must increase, although it needs to remain bounded. From this observation it can be concluded that the function $g(x)$ must be non-decreasing. Moreover, to avoid unboundedness, the growth of the hydraulic conductivity needs to be compensated by introducing the balancing decay term $-D_e(t)$. In [68] several numerical results using this model were presented in the graph describing the geometry of the maze in Figure 2.1. The authors show that when $g(x) = x$ the conductivity D_e at large times tends to localize (have a local support) on the edges of the shortest path between the two external sources.

More recently Bonifaci et al. proved in [13] that in the case $g(x) = x$, for $t \to \infty$, indeed the distribution of D_e converges to the shortest path for a general planar graph G. Moreover, the same authors prove that the above model is

equivalent to an optimal transport problem on the graph G when we consider a balanced forcing term f satisfying

$$\sum_{v \in V} f_v = 0 \tag{2.3}$$

Such problem can be recasted as finding $Q = \{Q_e\}_{e \in E}$ such that:

$$\begin{aligned} &\min_{Q \in \{Q_e\}_{e \in E}} \sum_{e \in E} Q_e L_e \quad \text{s.t.:} \\ &\sum_{e \in \sigma(v)} Q_e = f_v \qquad \text{for all } v \in V. \end{aligned} \tag{2.4}$$

In fact, under some general assumptions on the graph structure, the solution of system Equation (2.2) converges to a stationary solution Q^* that is also solution of the above optimal transport problem in G.

2.1.2 Dynamic Monge-Kantorovich (DMK) Model

In this section we generalize the model given in eq. Equation (2.2) by removing the graph structure and defining the problem on an open bounded domain $\Omega \subset \mathbb{R}^d$. We restrict this study to the case of $g(x) = x$. Then, given a forcing function $f : \Omega \to \mathbb{R}$, a continuous analogue of Equation (2.3) tries to find the pair of functions $(\mu, u) : [0, +\infty[\times \Omega \mapsto \mathbb{R}^+ \times \mathbb{R}^d$ that satisfies:

$$-\operatorname{div}\Big(\mu(t,x)\,\nabla u(t,x)\Big) = f(x) \tag{2.5a}$$

$$\partial_t \mu(t,x) = \mu(t,x)|\,\nabla u(t,x)| - \mu(t,x) \tag{2.5b}$$

$$\mu(0,x) = \mu_0(x) > 0 \tag{2.5c}$$

complemented by zero Neumann boundary conditions. Here, $\partial_t \mu$ indicates partial differentiation with respect to time, and $\nabla = \nabla_x$. This generalization is intuitively justified by comparing Equation (2.2) with Equation (2.5). In fact, Equation (2.5a) states the spatial balance of a (continuum) Fick-Poiseuille flux $q = -\mu \nabla u$ with potential function u, while Equation (2.5b) is the analogue in the continuous setting of the dynamics in the original discrete model described by Equation (2.2c).

In analogy with the discrete model we conjecture that system Equation (2.5) converges to an equilibrium point as $t \to +\infty$. At equilibrium, the time derivatives should vanish ($\partial_t \mu \to 0$), and thus Equation (2.5b) becomes the constraint

stating that the norm of the gradient of u must be unitary where μ is strictly greater than zero. Note that outside the support of μ no constrain is imposed. In particular, the bound on $|\nabla u|$ can not be deduced. These observations are crucial to the development of our conjecture, which reads as:

Conjecture 1. *The solution pair $(\mu(t), u(t))$ of Equation (2.5) with $f = f^+ - f^-$ converges for $t \to +\infty$ to the pair (μ^*, u^*) where $\mu^* = \mu^*(f^+, f^-)$ is the OT density and u^* is a Kantorovich potential u^*, solution of the L^1-OTP.*

2.2 Existence and Uniqueness

We would like to introduce the discussion on existence and uniqueness of the pair $(\mu(t), u(t))$ solution of Equation (2.5) for all $t \geq 0$, by posing three fundamental questions related to the proposed model:

1. which kind of solution pair for system 2.5, we expect to find?
2. What are the necessary assumptions on Ω, μ_0, f^+, and f^-?
3. In which function space does the solution pair (μ, u) live?

In [33] we approach the first question requiring the PDE in Equation (2.5a) to be in weak form, and the dynamic equation Equation (2.5b) to be in mild form. In this setting we give in [33] the following result

Theorem 27. *Given Ω an open, bounded, convex, and connected domain in $\mathbb{R}^d$ with smooth boundary, $f \in L^\infty(\Omega)$ with zero mean and $\mu_0 \in \mathcal{C}^\delta(\Omega)$ with $\mu_0 > 0$ and $0 < \delta < 1$ there exists $\tau_0 > 0$ depending on f and μ_0, such that the system*

$$\int_\Omega \mu(t,x)\,\nabla u(t,x)\,\nabla \varphi(x)\,dx = \int_\Omega f(x)\varphi(x)\,dx \quad \forall \varphi \in H^1(\Omega) \tag{2.6a}$$

$$\partial_t \mu(t,x) = \mu(t,x)|\nabla u(t,x)| - \mu(t,x) \tag{2.6b}$$

$$\mu(0,x) = \mu_0(x) > 0 \tag{2.6c}$$

$$\int_\Omega u(t,x)\,dx = 0 \tag{2.6d}$$

admits a unique solution pair

$$(\mu, u) \in \mathcal{C}^1([0,\tau_0[, \mathcal{C}^\delta(\Omega)) \times \mathcal{C}^1\left([0,\tau_0[, \mathcal{C}^{1,\delta}(\Omega)\right)$$

The proof of Theorem 27 is based on the idea of rewriting Equation (2.6) in the form of an Ordinary Differential Equation (ODE) in the variable μ. In fact observe that Equation (2.6a) uniquely defines u for a fixed μ, and thus uniquely defining the right hand of Equation (2.6b). Thanks to results on the regularity of the solution of elliptic equation, properly extended to be applied to our problem, we can proof that the right hand of Equation (2.6b) contains only functionals that are Lipschitz continuous in μ, at least locally. Standard arguments in the theory of ODEs in Banach Spaces ensure local existence and uniqueness of the solution $\mu(t)$. In other words, there exists a sufficiently small $\tau(\mu_0) > 0$ such that the fix point problem Equation (2.32) admits a solution $\mu \in \mathcal{C}^1([0, \tau(\mu_0); \mathcal{C}^\delta(\Omega))$. In order to present the complete of Theorem 27, we first need to introduce some definitions and results to describes properly such ideas.

2.2.1 Elliptic Equations: weak solutions and regularity

As already mentioned above, the proof of Theorem 27 is based on regularity results of solution of elliptic equations in weak form, thus in this preliminary part we introduced this type of PDEs. We focus on equations in the form of Equation (2.6a), in the general case of a forcing term in the form $f + \mathrm{div}(G)$ with $G : \Omega \mapsto \mathbb{R}^n$, for a reason that will be clear later.

Problem 28. *Given Ω an open, bounded, convex, and connected domain in $\mathbb{R}^d$ with smooth boundary, consider $f \in L^2(\Omega)$ such that $\int_\Omega f\,dx = 0$, $G \in [L^2(\Omega)]^d$, and $\mu \in L^\infty(\Omega)$ bounded from below i.e.:*

$$\exists \lambda > 0 \ : \ \mu(x) \geq \lambda \quad \forall x \in \Omega$$

Find $u \in H^1(\Omega)$ with $\int_\Omega u\,dx = 0$ that satisfies the following equation

$$\int_\Omega \mu \nabla u \nabla \varphi\,dx = \int_\Omega f\varphi\,dx + \int_\Omega G \cdot \nabla \varphi\,dx \quad \forall \varphi \in H^1(\Omega)$$

The above problem is well posed, as stated by the following:

Proposition 29. *Problem 28 admits a unique solution, denoted by*

$$u_{f,G}(\mu)$$

When $G = 0$ we write simply $u_f(\mu)$.

The proof of the previous proposition is based on the Lax-Milgram Theorem and on the Poincare-Wirtinger Inequality ([31]).

Now let us define two sets: the first one is given by

$$\mathcal{F} := \left\{ f \in L^\infty(\Omega) : supp(f) \subsetneq \Omega \text{ and } \int_\Omega f\, dx = 0 \right\}.$$

The second set is given by:

$$\mathcal{D} := \left\{ \mu \in \mathcal{C}^\delta(\Omega) \text{ such that } \lambda(\mu) := \min_{x \in \bar{\Omega}} \mu(x) \geq \alpha > 0 \right\},$$

where $0 < \delta < 1$. Here we denote with $\mathcal{C}^\delta(\Omega)$ the set of the Hölder continuous functions in Ω with Hölder exponent δ:

$$\mathcal{C}^\delta(\Omega) = \left\{ v : \Omega \mapsto \mathbb{R} \ : \ v_{[\delta,\Omega]} := \sup_{x \neq y} \frac{|v(x) - v(y)|}{|x - y|^\delta} < +\infty \right\}$$

with the norm

$$\|v\|_{\mathcal{C}^\delta} := \sup_\Omega v + v_{[\delta,\Omega]}$$

We now prove a fundamental lemma, whose long proof is given later in proof of 33, that extends classical results of regularity theory of elliptic equations with Hölder continuous coefficients taken from [70, 37]. Our contribution to the statement of the lemma is a detailed description of the dependence upon $\|\mu\|_{\mathcal{C}^\delta(\Omega)}$ and $\lambda(\mu)$ of the constants the appearing in these estimates.

Lemma 30. *Given Ω an open, bounded, convex, and connected domain in $\mathbb{R}^d$ with smooth boundary, Consider f, G, and μ as in Problem 28 with the additional assumptions that $f \in \mathcal{F}$, $G \in [\mathcal{C}^\delta(\Omega)]^d$, and $\mu \in \mathcal{D}$. Then the solution $u_{f,G}(\mu)$ of Problem 28 belongs to $\mathcal{C}^{1,\delta}(\Omega)$ and the following estimate holds:*

$$\| \nabla u_{f,G}(\mu)\|_{\mathcal{C}^\delta(\Omega)} \leq K(d, \Omega, \delta) K_\mu(\mu) \left(\|f\|_{L^\infty(\Omega)} + \|G\|_{\mathcal{C}^\delta(\Omega)} \right) \tag{2.7}$$

where $K(d, \Omega, \delta)$ is a constant depending on the dimension d, the domain Ω, and the Hölder regularity δ of μ, and:

$$K_\mu(\mu) = K_\mu\left(\lambda(\mu), \|\mu\|_{\mathcal{C}^\delta(\Omega)}\right) = \frac{1}{\lambda(\mu)} \left(\frac{\|\mu\|_{\mathcal{C}^\delta(\Omega)}}{\lambda(\mu)} \right)^{\frac{d+\delta}{2\delta}}. \tag{2.8}$$

This Lemma is analogous to Theorem 5.19 of [37] simplified to a scalar elliptic equation but extended to explicitly determine the dependence of the inequality

constants upon μ. We will denote with C or c generic constants that may depend upon d, Ω, and the Hölder continuity exponent δ but are always independent of μ. Before detailing the proof of Lemma 30, we recall and adapt some classical results of regularity theory of elliptic PDEs

Lemma 31 (Elliptic Decay)**.** *Let $v \in H^1(\Omega)$ be any solution of*

$$\int_\Omega \nabla v \nabla \varphi \, dx = 0 \quad \forall \varphi \in H_0^1(\Omega) \tag{2.9}$$

then there exists a constant $c(d)$ such that:

$$\int_{B(x_0,\rho)} |\nabla v|^2 \, dx \leq c(d) \left(\frac{\rho}{R}\right)^d \int_{B(x_0,R)} |\nabla v|^2 \, dx \tag{2.10}$$

$$\int_{B(x_0,\rho)} |\nabla v - (\nabla v)_{x_0,\rho}|^2 \, dx \leq c(d) \left(\frac{\rho}{R}\right)^{d+2} \int_{B(x_0,R)} |\nabla v - (\nabla v)_{x_0,R}|^2 \, dx \tag{2.11}$$

for arbitrary balls $B(x_0,\rho) \Subset B(x_0,R) \Subset \Omega$.

The above lemma is a revisited version of from Proposition 5.8 in [37]. The proof follows from the observation that the derivatives of v satisfy the weak form of Laplace equation (see also [3], page 61). Note that the constant $c(d)$ depends only on the problem dimension d as we are considering Laplace equation.

We also use the following result from Lemma 5.13 in [37] and Lemma 9.2 in [2]:

Lemma 32 (Iteration lemma)**.** *Let $\phi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ be a non-negative and non increasing function satisfying*

$$\phi(\rho) \leq A \left[\left(\frac{\rho}{R}\right)^\alpha + \epsilon \right] \phi(R) + B \, R^\beta \tag{2.12}$$

for some $A, \alpha, \beta > 0$, with $\alpha > \beta$ and for all $0 < \rho \leq R \leq R_0$, where $R_0 > 0$ is given.

Then there exist constants $\epsilon_0 = \epsilon_0(A,\alpha,\beta)$ and $C = C(A,\alpha,\beta)$ such that

$$\text{if} \quad \epsilon \leq \epsilon_0 = \left(\frac{1}{2A}\right)^{\frac{2\alpha}{\alpha-\beta}} \qquad \text{then} \quad \phi(\rho) \leq C \left[\frac{\phi(R)}{R^\beta} + B\right] \rho^\beta. \tag{2.13}$$

We will be using thebootstrap technique introduced by [53, 19] and used more recently by [22] to show the regularity of local minimizers of double phase variational integrals. The technique can be described by the following steps.

First we consider a compact set $K \Subset \Omega$ and prove that $u \in L^{2,\nu}(K)$ for a suitable regularity exponent ν with $0 < \nu < d$. Then, $u \in \mathcal{L}^{2,d+2\delta}(K)$ where $L^{2,\nu}(K)$ and $\mathcal{L}^{2,d+2\delta}(K)$ are the Morrey and Campanato spaces, respectively. The results are extended to the entire domain by assuming enough regularity of $\partial\Omega$. This latter step is not reported in the following proof for brevity. Finally, the equivalence between the Campanato spaces $\mathcal{L}^{2,d+2\delta}(\Omega)$ and $\mathcal{C}^{\delta}(\bar{\Omega})$ is used to prove estimate Equation (2.7) and to derive the expression of the constant K_μ given in Equation (2.8).

We recall that the norm of a function $u : \Omega \to \mathbb{R}^m$ (in our case we have either $m = 1$ or $m = d$) belonging to a Morrey space is given by:

$$\|u\|_{L^{2,\gamma}(\Omega)} = \Big(\sup_{\substack{x_0 \in \Omega \\ \rho > 0}} \rho^{-\gamma} \int_{\Omega(x_0,\rho)} |u|^2 \, dx \Big)^{\frac{1}{2}}$$

where $\Omega(x_0, \rho) = \Omega \cap B(x_0, \rho)$ and $0 \leq \gamma < d$. For $0 \leq \gamma < d + 2$, the norm of u belonging to a Campanato space is given by:

$$\|u\|_{\mathcal{L}^{2,\gamma}(\Omega)} = \|u\|_{L^2(\Omega)} + \Big(\sup_{\substack{x_0 \in \Omega \\ \rho > 0}} \rho^{-\gamma} \int_{\Omega(x_0,\rho)} |u - (u)_{x_0,\rho}|^2 \, dx \Big)^{\frac{1}{2}}$$

where $(u)_{x_0,\rho} = \int_{\Omega(x_0,\rho)} u \, dx / |\Omega(x_0, \rho)|$ is the average integral.

Proof of 33. *The first step of the bootstrap proceeds as follows. Consider $x_0 \in K$ and the ball $B_R := B(x_0, R) \Subset \Omega$. In this ball we use Korn's technique (freezing the coefficients) to decompose the solution as $u = v + w$ where $v \in H^1(B_R)$ satisfies the equations:*

$$\int_{B_R} \mu(x_0) \, \nabla v \, \nabla \varphi \, dx = 0 \quad \forall \varphi \in H^1_0(B_R) \tag{2.14}$$

with $v = u$ in ∂B_R and the second equation is to be interpreted in the sense that $v - u \in H^1_0(B_R)$. The second function $w \in H^1_0(B_R)$ satisfies the equation:

$$\int_{B_R} \mu(x_0) \, \nabla w \, \nabla \varphi \, dx = \int_{B_R} \Big[f\varphi + G \cdot \nabla \varphi - (\mu(x) - \mu(x_0)) \, \nabla u \cdot \nabla \varphi \Big] \, dx \qquad \forall \varphi \in H^1_0(B_R) \tag{2.15}$$

with $w = 0$ in ∂B_R. Since $\mu(x_0)$ in Equation (2.14) is a strictly positive and bounded scalar number it can be eliminated from the equation, hence w simply solves the weak form of Laplace equation:

$$\int_{B_R} \nabla v \, \nabla \varphi \, dx = 0 \qquad \forall \varphi \in H^1_0(\Omega) \tag{2.16}$$

with $v = u$ in $\partial\Omega$. Thus we can use Lemma 31 to obtain:

$$\int_{B_\rho} |\nabla v|^2\,dx \le c(n)\left(\frac{\rho}{R}\right)^n \int_{B_R} |\nabla v|^2\,dx \tag{2.17}$$

Recall that at this point our goal is to estimate the Morrey norm $\|\nabla u\|_{L^{2,\nu}(K)}$ with $\nu < d$. We use the above decomposition of u to estimate $\phi(\rho) := \int_{B_\rho} |\nabla u|^2\,dx$, $0 < \rho \le R$. Thus we can write:

$$\begin{aligned}
\int_{B_\rho} |\nabla u|^2\,dx &= \int_{B_\rho} |\nabla v + \nabla w|^2\,dx \le 2\int_{B_\rho} |\nabla v|^2\,dx + 2\int_{B_\rho} |\nabla w|^2\,dx \\
&\le c(d)\left(\frac{\rho}{R}\right)^d \int_{B_R} |\nabla v|^2\,dx + 2\int_{B_\rho} |\nabla w|^2\,dx \\
&= c(d)\left(\frac{\rho}{R}\right)^d \int_{B_R} |\nabla u - \nabla w|^2\,dx + 2\int_{B_\rho} |\nabla w|^2\,dx \\
&\le c(d)\left(\frac{\rho}{R}\right)^d \int_{B_R} |\nabla u|^2\,dx + c(d)\left(\frac{\rho}{R}\right)^d \int_{B_\rho} |\nabla w|^2\,dx \\
&\quad + 2\int_{B_\rho} |\nabla w|^2\,dx \\
&\le c(d)\left(\frac{\rho}{R}\right)^d \int_{B_R} |\nabla u|^2\,dx + c(d)\int_{B_R} |\nabla w|^2\,dx
\end{aligned}$$

Note that, somewhat improperly, we always use the symbol $c(d)$ to indicate a constant depending on d only and that may assume different meaning even within the same equation. To estimate $\int_{B_R} |\nabla w|^2\,dx$ we use $\varphi = w$ in Equation (2.15) to get:

$$\begin{aligned}
\lambda(\mu)\int_{B_R} |\nabla w|^2\,dx &\le \int_{B_R} \mu(x_0)|\nabla w|^2\,dx \\
&= \int_{B_R} \left[fw + G\cdot\nabla w - (\mu(x) - \mu(x_0))\,\nabla u\,\nabla w\right]\,dx
\end{aligned} \tag{2.18}$$

Using Hölder continuity of μ, and Poincaré and Cauchy-Schwarz inequalities, we can bound the right-hand-side of the previous equation to obtain:

$$\int_{B(x_0,R)} fw\,dx \le \|f\|_{L^2(B(x_0,R))}\; c(d)\,\|\nabla w\|_{L^2(B(x_0,R))}$$

$$\int_{B(x_0,R)} G\cdot\nabla w\,dx \le \|G\|_{L^2(B(x_0,R))}\;\|\nabla w\|_{L^2(B(x_0,R))}$$

$$\int_{B(x_0,R)} (\mu(x) - \mu(x_0))\,\nabla u\,\nabla w\,dx \le R^\delta \|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}\|\nabla u\|_{L^2(B(x_0,R))}\,\|\nabla w\|_{L^2(B(x_0,R))}$$

In the end, using Minkowski inequality to remove the double products, we can write:

$$\int_{B_R} |\nabla w|^2\,dx \leq 2\frac{1}{(\lambda(\mu))^2}\Big[(c(d))^2\|f\|^2_{L^2(B_R)} + \|G\|^2_{L^2(B_R)} + R^{2\delta}\|\mu\|^2_{\mathcal{C}^\delta(\bar{\Omega})}\|\nabla u\|^2_{L^2(B_R)}\Big] \tag{2.19}$$

Since $f \in L^\infty(\Omega)$*, implying that* $f \in L^{2,\nu}(\Omega)$ *and* $\|f\|^2_{L^{2,\nu}(\Omega)} \leq c(d)\|f\|^2_{L^\infty(\Omega)}$ *for* $0 \leq \nu < d$*, we obtain:*

$$\|f\|^2_{L^2(B_R)} \leq c(d)\|f\|^2_{L^\infty(B_R)}R^\nu \leq c(d)\|f\|^2_{L^\infty(\Omega)}R^\nu \tag{2.20}$$

Since $G \in \mathcal{C}^\delta(\bar{\Omega})$ *implies that (each component of)* $G \in \mathcal{L}^{2,\gamma}(\Omega)$ *for all* $0 \leq \gamma \leq d + 2\delta$*, noting that we require* $0 \leq \nu < d$ *and in this case* $L^{2,\nu}(\Omega) \equiv \mathcal{L}^{2,\nu}(\Omega)$*, we obtain:*

$$\|G\|^2_{L^2(B_R)} \leq \|G\|^2_{\mathcal{L}^{2,\gamma}(B_R)}R^\gamma \leq c(d)\|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}R^\gamma \tag{2.21}$$

Taking $\nu < d$ *in Equation* (2.20) *and* $\gamma = \nu$ *in Equation* (2.21) *we get:*

$$\int_{B_\rho} |\nabla u|^2\,dx \leq c(d)\left[\left(\frac{\rho}{R}\right)^d + R^{2\delta}\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)}\right)^2\right]\int_{B_R} |\nabla u|^2\,dx + c(d)\left(\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2}\right)R^\nu \tag{2.22}$$

Now we rewrite inequality Equation (2.22) *in the form of the hypotheses of Lemma 32, i.e.:*

$$\phi(\rho) := \int_{B_\rho} |\nabla u|^2\,dx, \qquad \alpha = d,\ \beta = \nu,$$

$$\epsilon = R^{2\delta}\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)}\right)^2,\ A = c(d),\ B = c(d)\left(\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2}\right)$$

for $\rho \leq R$*. Considering* R *such that:*

$$R^{2\delta}\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)}\right)^2 \leq \left(\frac{1}{2A}\right)^{\frac{2n}{d-\nu}} = A_0$$

we have that:

$$R \leq R_0 = A_0^{\frac{1}{2\delta}}\left(\frac{\lambda(\mu)}{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}\right)^{\frac{1}{\delta}} \tag{2.23}$$

We can now apply Lemma 32 to arrive at the following estimate valid for $0 < \rho \leq R \leq R_0$*:*

$$\int_{B_\rho} |\nabla u|^2\,dx \leq C(A,d,\nu)\rho^\nu\left(\frac{\int_{B_R}|\nabla u|^2\,dx}{R^\nu} + B\right) \tag{2.24}$$

Incorporating all the constants into one single constant $C(d,\nu)$ we obtain:

$$\int_{B_\rho} |\nabla u|^2\,dx \le C(d,\nu)\rho^\nu\left(\frac{\int_{B_R}|\nabla u|^2\,dx}{R^\nu}+\frac{\|f\|^2_{L^\infty(\Omega)}+\|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{(\lambda(\mu))^2}\right) \tag{2.25}$$

The previous estimate is valid for every $B_R \Subset \Omega$. Varying $x_0 \in K$ and using the continuity inequality in the Lax-Milgram Lemma we obtain the desired estimate of this first step of the bootstrap procedure, i.e.:

$$\begin{aligned}\|\nabla u\|^2_{L^{2,\nu}(K)} &\le C(d,\nu)\left(\frac{\int_{B_{R_0}}|\nabla u|^2\,dx}{R_0^\nu}+\frac{\|f\|^2_{L^\infty(\Omega)}+\|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{(\lambda(\mu))^2}\right)\\ &\le C(d,\nu)\left(\frac{C(\Omega)\|f\|^2_{L^\infty(\Omega)}+\|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{(\lambda(\mu))^2}\frac{1}{R_0^\mu}+\frac{\|f\|^2_{L^\infty(\Omega)}+\|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{(\lambda(\mu))^2}\right)\\ &\le C(d,\nu)C(\Omega)\frac{\|f\|^2_{L^\infty(\Omega)}+\|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{(\lambda(\mu))^2}\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{\nu}{\delta}}\end{aligned} \tag{2.26}$$

where $C(d,\nu)$ is bounded for all $\nu < d$.

The second step of the bootstrap procedure starts by noting that Equation Equation (2.15) can be rewritten using $R = R_0$ as defined above:

$$\begin{aligned}\int_{B_R}\mu(x_0)\nabla w\nabla\varphi\,dx = \int_{B_R}\Big[f\varphi+(G-(G)_R)\cdot\nabla\varphi-\\ (\mu(x)-\mu(x_0))\nabla u\cdot\nabla\varphi\Big]\,dx \qquad \forall\varphi\in H^1_0(B_R)\\ w=0 \text{ in } \partial B_R\end{aligned} \tag{2.27}$$

We continue by using again the decomposition $u = v+w$ and Lemma 31 to obtain:

$$\begin{aligned}\int_{B_\rho}|\nabla u-(\nabla u)_\rho|^2\,dx &= \int_{B_\rho}|\nabla v+(\nabla v)_\rho+\nabla w+(\nabla w)_\rho|^2\,dx\\ &\le c(d)\left(\frac{\rho}{R}\right)^{d+2}\int_{B_R}|\nabla v-(\nabla v)_R|^2\,dx+2\int_{B_\rho}|\nabla w-(\nabla w)_\rho|^2\,dx\\ &\le c(d)\left(\frac{\rho}{R}\right)^{d+2}\int_{B_R}|\nabla u-(\nabla u)_R|^2\,dx+c(d)\int_{B_R}|\nabla w|^2\,dx\end{aligned}$$

where the last inequality arises from the minimality of the mean. We follow the same developments as before, but now we explicitly include the factor R in the constant of Poincaré inequality to obtain:

$$\begin{aligned}\int_{B_\rho}|\nabla u-(\nabla u)_\rho|^2\,dx \le c(d)\left(\frac{\rho}{R}\right)^{d+2}\int_{B_R}|\nabla u-(\nabla u)_R|^2\,dx\\ +2c(d)\frac{R^2\|f\|^2_{L^2(B_R)}+\|G-(G)_R\|^2_{L^2(B_R)}+R^{2\delta}\|\mu\|^2_{\mathcal{C}^\delta(\bar\Omega)}\|\nabla u\|^2_{L^2(B_R)}}{(\lambda(\mu))^2}\end{aligned} \tag{2.28}$$

Since $\nabla u \in L^{2,\nu}(K)$ *for* $0 < \nu < d$ *we can take* $\nu = d - \delta$ *in Equation* (2.26) *to get:*

$$\| \nabla u\|^2_{L^2(B_R)} = \frac{\int_{B_R} |\nabla u|^2 \, dx}{R^{n-\delta}} R^{d-\delta} \leq \| \nabla u\|_{L^{2,d-\delta}} R^{d-\delta}$$

Using $\nu = d-2+\delta$ *in Equation* (2.20) *we obtain* $\|f\|^2_{L^2(B_R)} \leq c(d)\|G\|^2_{L^\infty(\Omega)} R^{d-2+\delta}$, *while using* $\gamma = d+\delta$ *in Equation* (2.21) *we have* $\|G-(G)_R\|^2_{L^2(B_R)} \leq \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})} R^{d+\delta}$. *Substitution of these inequalities in Equation* (2.28) *yields:*

$$\begin{aligned}
&\int_{B_\rho} |\nabla u - (\nabla u)_\rho|^2 \, dx \leq c(d) \left(\frac{\rho}{R}\right)^{d+2} \int_{B_R} |\nabla u - (\nabla u)_R|^2 \, dx \\
&\quad + c(d) \frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2} R^{d+\delta} \\
&\quad + R^{2\delta} \frac{\|\mu\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2} C(d, d-\delta) C(\Omega) \frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2} \left(\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)} \right)^{\frac{d-\delta}{\delta}} \right) R^{d-\delta} \\
&\leq c(d) \left(\frac{\rho}{R}\right)^{d+2} \int_{B_R} |\nabla u - (\nabla u)_R|^2 \, dx \\
&\quad + C(d, \Omega, \delta) \frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2} \left(1 + \frac{\|\mu\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2} \left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)} \right)^{\frac{d-\delta}{\delta}} \right) R^{d+\delta} \\
&\leq c(d) \left(\frac{\rho}{R}\right)^{d+2} \int_{B_R} |\nabla u - (\nabla u)_R|^2 \, dx \\
&\quad + C(d, \Omega, \delta) \frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{(\lambda(\mu))^2} \left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)} \right)^{\frac{d+\delta}{\delta}} R^{d+\delta}
\end{aligned}$$

Application of Lemma 32 with $\phi(\rho) := \int_{B_\rho} |\nabla u - (\nabla u)_\rho|^2 \, dx$ *yields for* $0 < \rho \leq R \leq R_0$*:*

$$\begin{aligned}
&\int_{B_\rho} |\nabla u - (\nabla u)_\rho|^2 \, dx \leq \rho^{d+\delta} C(d, \Omega, \delta) \cdot \\
&\cdot \left[\frac{\int_{B_R} |\nabla u - (\nabla u)_R|^2 \, dx}{R^{d+\delta}} + \left(\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)^2} \right) \left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu)} \right)^{\frac{d}{\delta}+1} \right]
\end{aligned}$$

from which, using again the minimality of the mean and the estimate of R_0 *given*

in Equation (2.23)*, we can evaluate:*

$$\begin{aligned}
&\frac{\int_{B_\rho} |\nabla u - (\nabla u)_\rho|^2\,dx}{\rho^{d+\delta}} \leq C(d,\delta,\Omega)\cdot \\
&\cdot\left[\frac{\int_{B_{R_0}} |\nabla u - (\nabla u)_{R_0}|^2\,dx}{R_0^{d+\delta}} + \left(\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\right)\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1}\right] \\
&\leq C(d,\delta,\Omega)\left(\int_{B_{R_0}} |\nabla u - (\nabla u)_{R_0}|^2\,dx + \frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\right)\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1} \\
&\leq C(d,\delta,\Omega)\left(\int_{\Omega} |\nabla u|^2\,dx + \frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\right)\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1} \\
&\leq C(d,\delta,\Omega)\left(\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\right)\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1}
\end{aligned}$$

Hence $\nabla u \in \mathcal{L}^{2,d+\delta}(K)$ *and we can write:*

$$\|\nabla u\|^2_{\mathcal{L}^{2,d+\delta}(K)} \leq C(d,\delta,\Omega)\left(\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\right)\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1} \tag{2.29}$$

The bootstrap procedure is restarted from Equation (2.27) *using* $\nu = d-2+2\delta$ *in Equation* (2.20) *and* $\gamma = d+2\delta$ *in Equation* (2.21)*, and estimate Equation* (2.29) *in Equation* (2.28) *so that a term* $R^{d+2\delta}$ *can be factored. Thus we can write:*

$$\begin{aligned}
\int_{B_\rho} |\nabla u - (\nabla u)_\rho|^2\,dx \leq\ & c(d)\left(\frac{\rho}{R}\right)^{d+2}\int_{B_R} |\nabla u - (\nabla u)_R|^2\,dx \\
&+ c(d)\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|_{\mathcal{C}^\delta(\bar\Omega)}}{(\lambda(\mu))^2}R^{d+2\delta} \\
&+ C(d,\delta,\Omega)\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1}R^{d+2\delta}
\end{aligned}$$

and finally, applying once again Lemma 32, we have the final result:

$$\|\nabla u\|^2_{\mathcal{L}^{2,d+2\delta}(K)} \leq C(d,\delta,\Omega)\frac{\|f\|^2_{L^\infty(\Omega)} + \|G\|^2_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)^2}\left(\frac{\|\mu\|_{\mathcal{C}^\delta(\bar\Omega)}}{\lambda(\mu)}\right)^{\frac{d}{\delta}+1} \tag{2.30}$$

Extension of the previous estimate to the entire domain Ω *can be obtained following the same bootstrap procedure starting from the analogue of the elliptic decay Lemma 31 on hemispheres (similarly to what is proposed in [37], Theorem 5.21). Such process introduces a dependence on the regularity of the boundary* $\partial\Omega$ *in the constant* $C(d,\delta,\Omega)$ *in Equation* (2.30)*, but we do not explicitly write such*

dependence. By the equivalence between $\mathcal{L}^{2,d+2\delta}(\Omega)$ *and* $\mathcal{C}^{\delta}(\bar{\Omega})$ *we get:*

$$\| \nabla u\|_{\mathcal{C}^{\delta}(\bar{\Omega})} \leq C(d,\delta,\Omega)\frac{\|f\|_{L^{\infty}(\Omega)} + \|G\|_{\mathcal{C}^{\delta}(\bar{\Omega})}}{\lambda(\mu)} \left(\frac{\|\mu\|_{\mathcal{C}^{\delta}(\bar{\Omega})}}{\lambda(\mu)}\right)^{\frac{d+\delta}{2\delta}}$$

which proves Equation (2.7) *and Equation* (2.8)*. From this, using Theorem 1.40 of [70], we directly obtain that* $u \in \mathcal{C}^{1,\delta}(\Omega)$*.*

2.2.2 Proof of Theorem 27

We can now resume the proof Theorem 27. First we can give the following definitions, that are well defined thanks to Proposition 29 and Lemma 30 and the fact the $\mathcal{C}^{\delta}(\Omega)$-norm is sub-multiplicative

Definition 34 (Potential)**.** *Let* $\mu \in \mathcal{D}$ *and* $f \in \mathcal{F}$*. The Potential Operator* $\mathcal{U} : \mathcal{D} \mapsto \mathcal{C}^{1,\delta}(\Omega)$*, that maps* μ *into* $\mathcal{U}(\mu)$*, is defined as follows*

$$\mu \mapsto \mathcal{U}(\mu) := u_{f,0}(\mu)$$

where $u_{f,0}(\mu)$ *is defined in Proposition 29*

Definition 35 (Flux)**.** *Let* $\mu \in \mathcal{D}$ *and* $f \in \mathcal{F}$*. The operator* $\mathcal{Q} : \mathcal{D} \mapsto \mathcal{C}^{\delta}(\Omega)$ *is defined as:*

$$\mu \mapsto \mathcal{Q}(\mu) := \mu|\nabla \mathcal{U}(\mu)|.$$

Thus we can recast Equation (2.6) in ODE-form in the variable μ:

$$\partial_t \mu(t) = \mathcal{Q}\left(\mu(t)\right) - \mu(t) \tag{2.31a}$$

$$\mu(0) = \mu_0 \tag{2.31b}$$

The solution pair $(\mu(t), u(t))$ of system Equation (2.6) introduced in Theorem 27 is defined as $(\mu(t), \mathcal{U}(\mu(t))$. The mild ($\mathcal{C}^0$-semigroup) formulation of Equation (2.31) reads as:

$$\mu(t) = e^{-t}\mu_0 + \int_0^t e^{s-t}\mathcal{Q}\left(\mu(s)\right) ds \tag{2.32}$$

This shows immediately that $\mu(t) \geq e^{-t}\min(\mu_0)$ for all $t \geq 0$, ensuring that $\mathcal{U}(\mu(t))$ is well-posed, i.e., the associated bilinear form is coercive. Moreover we are then able to show that the operator $\mathcal{Q}(\mu)$ is locally Lipschitz continuous, and we can then invoke Banach-Caccioppoli fixed-point theorems to show local existence and uniqueness of the solution μ. However, the fact that Lipschitz

continuity is only local in μ, prevents the extension of this result to larger times. The proof of the local-Lipschitz continuity of the operator $\mathcal{Q}$ starts by re-defining the subspace $\mathcal{D}$ the union of open and convex subsets:

$$\mathcal{D} = \bigcup_{0<a<b<+\infty} \mathcal{D}(a,b)$$
$$\mathcal{D}(a,b) := \left\{\mu \in \mathcal{C}^\delta(\Omega) \text{ such that } a < \lambda(\mu) \leq \|\mu\|_{\mathcal{C}^\delta(\Omega)} < b\right\}$$

for $0 < a < b < \infty$.

2.2.3 Local Lipschitz Continuity

We can now state the following proposition that establishes the sufficient hypothesis for the application of standard fix point theorem for existence and uniqueness of the solution of ODEs of the type:

$$\partial_t \mu(t) = \mathcal{Q}\left(\mu(t)\right) - \mu(t)$$

Proposition 36. *The Potential and Flux operators $\mathcal{U}$ and $\mathcal{Q}$ are Lipschitz continuous and bounded in $\mathcal{D}(a,b)$ for all a and b, $0 < a < b < \infty$. In other words we have that for every $\mu \in \mathcal{D}(a,b)$,*

$$\|\mathcal{U}(\mu)\|_{\mathcal{C}^{1,\delta}(\Omega)} \leq C_1(a,b)$$
$$\|\mathcal{Q}(\mu)\|_{\mathcal{C}^{\delta}(\Omega)} \leq C_2(a,b)$$

and there exist constants $L_{\mathcal{U}}(a,b)$ and $L_{\mathcal{Q}}(a,b)$ such that, for every $\mu_1, \mu_2 \in \mathcal{D}(a,b)$:

$$\|\mathcal{U}(\mu_1) - \mathcal{U}(\mu_2)\|_{\mathcal{C}^{,1}(\Omega)} \leq L_{\mathcal{U}}(a,b)\|\mu_1 - \mu_2\|_{\mathcal{C}^\delta(\Omega)}$$
$$\|\mathcal{Q}(\mu_1) - \mathcal{Q}(\mu_2)\|_{\mathcal{C}^\delta(\Omega)} \leq L_{\mathcal{Q}}(a,b)\|\mu_1 - \mu_2\|_{\mathcal{C}^\delta(\Omega)}$$

Proof. From Lemma 30 the boundedness of $\mathcal{U}(\mu)$ for $\mu \in \mathcal{D}(a,b)$ follows immediately with $G = 0$ in Equation (2.7). The local Lipschitz continuity of $\mathcal{U}$ derives from the following considerations. Given $\mu_1, \mu_2 \in \mathcal{D}(a,b)$ and $u_k = \mathcal{U}(\mu_k)$ with $k = 1, 2$, we note that

$$\int_\Omega \mu_1 \nabla u_1 \nabla \varphi \, dx = \int_\Omega f\varphi \, dx = \int_\Omega \mu_2 \nabla u_2 \nabla \varphi \, dx \quad \forall \varphi \in H^1(\Omega)$$
$$\int_\Omega \mu_1 \nabla (u_1 - u_2) \nabla \varphi \, dx = \int_\Omega (\mu_2 - \mu_1) \nabla u_2 \nabla \varphi \, dx \quad \forall \varphi \in H^1(\Omega)$$

Application of Lemma 30 with $f = 0$ and $G = -(\mu_1 - \mu_2)\,\nabla\,\mathcal{U}(\mu_2))$, which belongs to $[\mathcal{C}^\delta]^d$, yields:

$$\begin{aligned}
\|\,\nabla(u_1 - u_2)\|_{\mathcal{C}^\delta(\Omega)} &\leq K(\Omega, d, \delta) K_\mu(\mu_1) \|(\mu_1 - \mu_2)\,\nabla\, u_2\|_{\mathcal{C}^\delta(\Omega)} \\
&\leq K(\Omega, d, \delta) K_\mu(\mu_1) \|\mu_1 - \mu_2\|_{\mathcal{C}^\delta(\Omega)} \|\,\nabla\, u_2\|_{\mathcal{C}^\delta(\Omega)} \\
&= K(\Omega, d, \delta)^2 K_\mu(a, b)^2 \|f\|_{L^\infty(\Omega)} \|\mu_1 - \mu_2\|_{\mathcal{C}^\delta(\Omega)} \qquad (2.33)
\end{aligned}$$

We can also prove that the flux operator $\mathcal{Q}$ is bounded in $\mathcal{D}(a, b)$. In fact, since the Hölder norm is sub-multiplicative, we can write:

$$\|\mathcal{Q}(\mu)\|_{\mathcal{C}^\delta(\Omega)} = \|\,\mu|\,\nabla\,\mathcal{U}(\mu)|\,\|_{\mathcal{C}^\delta(\Omega)} \leq K(\Omega, d, \delta)\; b\, K_\mu(a, b) \|f\|_{L^\infty(\Omega)}$$

Lipschitz continuity of $\mathcal{Q}$ derives from Equation (2.33) as follows:

$$\begin{aligned}
\|\mathcal{Q}(\mu_1) - \mathcal{Q}(\mu_2)\|_{\mathcal{C}^\delta(\Omega)} &= \|\mu_1|\,\nabla\, u_1| - \mu_2|\,\nabla\, u_2|\,\|_{\mathcal{C}^\delta(\Omega)} \\
&= \|\mu_1\,(|\,\nabla\, u_1| - |\,\nabla\, u_2|) - (\mu_2 - \mu_1)\,|\,\nabla\, u_2|\,\|_{\mathcal{C}^\delta(\Omega)} \\
&\leq \|\mu_1\|_{\mathcal{C}^\delta(\Omega)} \|\,|\,\nabla\,[u_1 - u_2]\,|\,\|_{\mathcal{C}^\delta(\Omega)} \\
&\qquad + K(\Omega, d, \delta) K_\mu(a, b) \|f\|_\infty \|\mu_1 - \mu_2\|_{\mathcal{C}^\delta(\Omega)} \\
&\leq L_q(a, b) \|\mu_1 - \mu\|_{\mathcal{C}^\delta(\Omega)}
\end{aligned}$$

□

The $\mathcal{C}^1$-regularity in time of $u(t)$ is stated in the proposition that follows. Heuristically, the proof is based on the observation that, assuming that both $\partial_t\mu(t)$ and $\partial_t u(t)$ exist, we can take the derivative in time equation $-\operatorname{div}(\mu(t)\,\nabla\, u(t)) = f$ and use the fact that the source function is independent on time, thus obtaining the following equation for $u(t)$

$$-\operatorname{div}(\mu(t)\,\nabla\,\partial_t u(t)) - \operatorname{div}(\partial_t\mu(t)\,\nabla\, u(t)) = 0$$

.

Proposition 37. *The solution $u(t)$ of Equation (2.6) belongs to $\mathcal{C}^1([0, \tau(\mu_0)[; \mathcal{C}^{1,\delta}(\Omega))$. For each $t \in [0, \tau(\mu_0)[$ its time derivative $\partial_t u(t)$ solves the following equation:*

$$\begin{aligned}
\int_\Omega \mu(t)\,\nabla\,\partial_t u(t) \cdot \nabla\,\varphi\, dx &= -\int_\Omega \partial_t\mu(t)\,\nabla\, u(t) \cdot \nabla\,\varphi\, dx \quad \forall \varphi \in H^1(\Omega) \\
\int_\Omega \partial_t u(t)\, dx &= 0
\end{aligned} \qquad (2.34)$$

Proof. Let $t \in [0, \tau(\mu_0)[$ and choose $h > 0$ such that $t + h < \tau(\mu_0)$. The solution $\mu(t)$ of Equation (2.31) belongs to a ball $B(\mu_0, R)$ centered in μ_0 and with appropriate radius R. From Equation (2.32), it is possible to find two constants $a(\mu_0), b(\mu_0)$ such that $\mu(t) \in \mathcal{D}\left(a(\mu_0), b(\mu_0)\right)$. This allows us to write:

$$K_\mu(\mu(t)) = \frac{1}{\lambda(\mu(t))}\left(\frac{\|\mu(t)\|_{\mathcal{C}^\delta(\bar{\Omega})}}{\lambda(\mu(t))}\right)^{\frac{n+\delta}{2\delta}} \leq \frac{1}{a(\mu_0)}\left(\frac{b(\mu_0)}{a(\mu_0)}\right)^{\frac{n+\delta}{2\delta}} \quad \forall t \in [0, \tau(\mu)[$$

which shows that both the Potential and Flux operators are bounded and Lipschitz-continuous in $\mathcal{D}\left(a(\mu_0), b(\mu_0)\right)$. We first note that $u(t) = \mathcal{U}(\mu(t))$ is Lipschitz-continuous in time, since u is locally Lipschitz-continuous and $\mu \in \mathcal{C}^1\left([0, \tau(\mu_0)[\,; \mathcal{C}^\delta(\Omega)\right)$. Next, we can write the following equation, that holds $\forall \varphi \in H^1(\Omega)$

$$\int_\Omega \mu(t)\,\nabla u(t) \cdot \nabla \varphi\, dx = \int_\Omega f \varphi\, dx = \int_\Omega \mu(t+h)\,\nabla u(t+h) \cdot \nabla \varphi\, dx$$

Changing sign and adding to both sides the term $\int_\Omega \mu(t)\,\nabla u(t+h) \cdot \nabla \varphi\, dx$, yields:

$$\begin{aligned}\int_\Omega \mu(t)\,\nabla[u(t+h) - u(t)] \cdot \nabla \varphi\, dx = \\ - \int_\Omega [\mu(t+h) - \mu(t)]\,\nabla u(t+h) \cdot \nabla \varphi\, dx\end{aligned} \tag{2.35}$$

Now, at each time $t \in [0, \tau(\mu_0)[$, we define $w(t)$, that heuristically should be $\partial_t u(t)$, as the unique solution of:

$$\begin{cases} \int_\Omega \mu(t)\,\nabla w(t) \cdot \nabla \varphi\, dx = -\int_\Omega \partial_t \mu(t)\,\nabla u(t) \cdot \nabla \varphi\, dx \quad \forall \varphi \in H^1(\Omega) \\ \int_\Omega w(t)\, dx = 0 \end{cases} \tag{2.36}$$

Thanks to Lemma 30 it is easy to verify that $w(t) \in \mathcal{C}^{1,\delta}(\bar{\Omega})$. Now we multiply Equation (2.36) by $-h$ with and Equation (2.35) to obtain:

$$\begin{aligned}&\int_\Omega \mu(t)\,\nabla[u(t+h) - u(t) - hw(t)] \cdot \nabla \varphi\, dx \\ &\quad = -\int_\Omega [\mu(t+h) - \mu(t)]\,\nabla u(t+h) \cdot \nabla \varphi\, dx + h\int_\Omega \partial_t \mu(t)\,\nabla u(t) \cdot \nabla \varphi\, dx \\ &\quad = -\int_\Omega \{[\mu(t+h) - \mu(t) - h\partial_t \mu(t)]\,\nabla u(t+h) \\ &\qquad\qquad + h\partial_t \mu(t)\,(\nabla u(t+h) - \nabla u(t))\} \cdot \nabla \varphi\, dx \\ &\quad = -\int_\Omega [G_1(t,h) + h\, G_2(t,h)] \cdot \nabla \varphi\, dx\end{aligned}$$

with

$$G_1(t,h) = [\mu(t+h) - \mu(t) - h\partial_t \mu(t)]\,\nabla u(t+h); \quad G_2(t,h) = \partial_t \mu(t)\,[\nabla u(t+h) - \nabla u(t)]$$

Since $\mu \in \mathcal{C}^1(0,\tau;\mathcal{C}^\delta(\bar\Omega))$, we can estimate the above functions G_1 and G_2 as:

$$\begin{aligned}\|G_1(t,h)\|_{\mathcal{C}^\delta(\bar\Omega)} &\le \|\mu(t+h)-\mu(t)-h\partial_t\mu(t)\|_{\mathcal{C}^\delta(\bar\Omega)}\|\nabla u(t+h)\|_{\mathcal{C}^\delta(\bar\Omega)}\\ &\le K(n,\Omega,\delta)K_\mu(\mu(t))\|f\|_{L^\infty(\Omega)}\cdot o(h)\\ &\le K(n,\Omega,\delta)K(\mu_0)\|f\|_{L^\infty(\Omega)}\cdot o(h)\end{aligned}$$

and, since the Potential operator is Lipschitz-continuous, we have also:

$$\begin{aligned}\|G_2(t,h)\|_{\mathcal{C}^\delta(\bar\Omega)} &= \|\partial_t\mu(t)\left[\nabla u(t+h)-\nabla u(t)\right]\|_{\mathcal{C}^\delta(\bar\Omega)}\\ &\le \|\partial_t\mu(t)\|_{\mathcal{C}^\delta(\bar\Omega)}\|\nabla u(t+h)-\nabla u(t)\|_{\mathcal{C}^\delta(\bar\Omega)}\\ &\le L(\mu_0)h\end{aligned}$$

where $L(\mu_0)$ is a function of f, K. Thus we can write:

$$\|G_1+hG_2\|_{\mathcal{C}^\delta(\bar\Omega)} \le \|G_1\|_{\mathcal{C}^\delta(\bar\Omega)} + \|G_2\|_{\mathcal{C}^\delta(\bar\Omega)} = o(h)$$

and, for Lemma 30 using $G=-(G_1+hG_2)$, we obtain:

$$\lim_{h\to 0}\frac{\|\nabla[u(t+h)-u(t)-hw(t)]\|_{\mathcal{C}^\delta(\bar\Omega)}}{h}=0$$

that shows that $\partial_t u \in \mathcal{C}^1(0,\tau;\mathcal{C}^\delta(\bar\Omega))$ with $\partial_t u = w$. □

These results are collected in [33].

2.3 The Lyapunov-candidate functional

The existence and uniqueness result obtained in the previous section is only local in time and it does not allow us to pass to the limit with $t\to+\infty$ in Equation (2.6). Nevertheless we are able to identify a Lyapunov-candidate functional, i.e., a function that decreases along the $\mu(t)$-trajectories.

NewP The Lyapunov-candidate functional, that we introduced for the first time in [34], is defined for general $\mu \in L^1(\Omega)$ and is given by

$$\mathcal{L}(\mu) := \mathcal{E}_f(\mu) + \mathcal{M}(\mu) \tag{2.37}$$

$$\mathcal{E}_f(\mu) = \sup_{\varphi\in\mathcal{C}^1(\bar\Omega)}\int_\Omega\left(f\varphi-\mu\frac{|\nabla\varphi|^2}{2}\right)dx \qquad \mathcal{M}(\mu) := \frac{1}{2}\int_\Omega \mu\,dx \tag{2.38}$$

Note that $\mathcal{E}_f(\mu)$ was already defined in Equation (1.11) for general $\mu\in\mathcal{M}_+(\Omega)$. When we restrict $f\in\mathcal{F}$ and $\mu\in\mathcal{D}$ the functional rewrites as

$$\mathcal{L}(\mu) := \frac{1}{2}\int_\Omega \mu|\nabla u(\mu)|^2\,dx + \frac{1}{2}\int_\Omega \mu\,dx \tag{2.39}$$

We can state the following:

Proposition 38. *The function $\mathcal{L} : \mathcal{D} \longmapsto \mathbb{R}^+$ defined above is strictly decreasing in time along the solution $\mu(t)$ of Equation (2.6) for $t \in [0, \tau(\mu_0)[$. Its time derivative is given by*

$$\frac{d\mathcal{L}(\mu(t))}{dt} = -\frac{1}{2}\int_\Omega \mu(t)\left(|\nabla u(t)| - 1\right)^2\left(|\nabla u(t)| + 1\right) dx$$

Proof. Thanks to the $\mathcal{C}^1$-regularity in time of the solution pair $(\mu(t), u(t))$ given by Theorem 27, we can compute the time derivative of $\mathcal{L}(\mu(t))$ and prove it is strictly negative. In fact we have:

$$\frac{d\mathcal{L}(\mu(t))}{dt} = \frac{1}{2}\int_\Omega \left(\partial_t\mu(t)|\nabla u(t)|^2 + 2\mu(t)\nabla\partial_t u(t)\nabla u(t)\right) dx + \frac{1}{2}\int_\Omega \partial_t\mu\, dx$$

Substituting $\varphi = u(t)$ in Equation (2.34) we obtain

$$\int_\Omega \mu(t)\nabla\partial_t u(t)\cdot\nabla u(t)\, dx = -\int_\Omega \partial_t\mu(t)|\nabla u(t)|^2\, dx$$

Thus

$$\begin{aligned}
\frac{d\mathcal{L}(\mu(t))}{dt} &= \frac{1}{2}\int_\Omega -\partial_t\mu(t)|\nabla u(t)|^2 + \partial_t\mu(t)\, dx \\
&= -\frac{1}{2}\int_\Omega \partial_t\mu(t)\left(|\nabla u(t)|^2 - 1\right) dx \\
&= -\frac{1}{2}\int_\Omega \mu(t)\left(|\nabla u(t)| - 1\right)\left(|\nabla u(t)|^2 - 1\right) dx \\
&= -\frac{1}{2}\int_\Omega \mu(t)\left(|\nabla u(t)| - 1\right)^2\left(|\nabla u(t)| + 1\right) dx < 0
\end{aligned}$$

□

It is clear from previous equation that the time-derivative of $\mathcal{L}$ along $\mu(t)$-trajectory is equal to zero only if $|\nabla u(t)| = 1$ on the support of $\mu(t)$. Again, this result gives only one of the constrains of the MK equations, whit the bound on the norm of the gradient in the whole domain not imposed.

Since $\mathcal{L}(\mu(t))$ decreases in time, it is natural to investigate if $\mathcal{L}$ admits a minimum. The following proposition, shows the equivalence between the minimization of Lyapunov-candidate functional $\mathcal{L}$ and the Beckmann Problem which is equivalent to solve the MK equations by in Proposition 15. This original result provides further support to Conjecture 1.

Proposition 39. *Given Ω an open, bounded, convex, and connected domain in $\mathbb{R}^d$ with smooth boundary. Consider $f \in L^1(\Omega)$ with zero mean, then Beckmann*

Problem (as given in Corollary 16) and the minimization of $\mathcal{L}$ are equivalent which means

$$\min_{v\in[L^1(\Omega)]^d}\left\{\int_\Omega |v|\,dx \;:\; \mathrm{div}(v)=f \right\} = \min_{\mu\in L^1_+(\Omega)} \mathcal{L}(\mu) \tag{2.40}$$

where $L^1_+(\Omega)$ indicates the space of the non-negative function in $L^1(\Omega)$.

Moreover, the OT density $\mu^*(f)$ is a point of minimum for $\mathcal{L}$.

The proof makes use of the following duality result, already used in [15], that holds for general $\mu \in \mathcal{M}_+(\Omega)$ and $f \in \mathcal{M}(\Omega)$.

Lemma 40. *Given Ω an open, bounded, convex, and connected domain in $\mathbb{R}^d$ with smooth boundary. Consider $\mu \in \mathcal{M}_+(\Omega)$, $f \in \mathcal{M}(\Omega)$ with zero mean, then the following equalities hold*

$$\inf_{\varphi\in\mathcal{C}^1(\bar\Omega)} \mathcal{I}_{f,\mu}(\varphi) = \sup_{\xi\in[L^2_\mu(\Omega)]^d}\left\{-\int_\Omega |\xi|^2 d\mu \;:\; \mathrm{div}(\xi\mu)=f \right\} = -\mathcal{E}_f(\mu) \tag{2.41}$$

where

$$\mathcal{I}_{f,\mu}(\varphi)\int_\Omega \frac{1}{2}\,|\nabla\varphi|^2 d\mu - df\varphi$$

Proof. The proof is based on Theorem 55. In fact functional $\mathcal{I}_{f,\mu}$ rewrites in the following form

$$\mathcal{I}_{f,\mu}(u) = F(u) + G(\Lambda(u))$$

with

$$\begin{aligned} F : \mathcal{C}^1(\bar\Omega)\mapsto\mathbb{R} \quad &: \quad F(u) = -\int_\Omega u df \\ G : (L^2_\mu(\Omega))^d \mapsto \mathbb{R} \quad &: \quad G(p) = \int_\Omega \frac{|p|^2}{2} d\mu \\ \Lambda : V \mapsto (L^2_\mu(\Omega))^d \quad &: \quad \Lambda(u) = \nabla u \end{aligned}$$

The Legendre transform of F and G, and the conjugate operator of Λ are

$$\begin{aligned} F^* : V^* \mapsto \mathbb{R} \quad &: \; F^*(u^*) = \sup_{u\in V}\langle u^*-f, u\rangle = \begin{cases} 0 & \text{if } u^*-f=0 \\ +\infty & \text{otherwise} \end{cases} \\ G^* : (L^2_\mu(\Omega))^d \mapsto \mathbb{R} \quad &: \; G^*(p^*) = \frac{1}{2}\int_\Omega |p^*|^2 d\mu \\ \Lambda^* : (L^2_\mu(\Omega))^d \mapsto V^* \; &: \; \Lambda^*(p^*) = \mathrm{div}_\mu\, p^* \end{aligned}$$

where $f = \mathrm{div}_\mu(p^*)$ means

$$\int_\Omega \varphi df = -\int_\Omega p^* \cdot \nabla \varphi \; d\mu \tag{2.42}$$

The dual problem $\mathcal{P}^*$ in Equation (A.18) reads as

$$\sup_{p^* \in (L^2_\mu(\Omega))^d} \left\{ -F^*(\mathrm{div}(p^*)) - \int_\Omega \frac{|p^*|^2}{2} \; d\mu \right\} =$$
$$\sup_{p^* \in (L^2_\mu(\Omega))^d} \left\{ -\int_\Omega \frac{|p^*|^2}{2} d\mu \; : \; \mathrm{div}_\mu(p^*) = f \right\}$$

proving the equivalence in Equation (2.41) The extremality condition in Equation (A.20) says that the following equality holds

$$\int_\Omega \frac{|\nabla \bar{u}|^2}{2} \; d\mu + \int_\Omega \frac{|\bar{p}^*|^2}{2} \; d\mu = -\int_\Omega \bar{p}^* \cdot \nabla \bar{u} \; d\mu \tag{2.43}$$

which implies, by Young inequality, that

$$\bar{p}^* = -\mu \nabla \bar{u} \quad \mu - \text{a.e on} \tag{2.44}$$

This completes the proof. □

We can now proceed withe the proof of Proposition 39.

Proof. We begin rewriting $\mathcal{E}_f(\mu)$ for $\mu \in L^1_+(\Omega)$ in the following variational form:

$$\mathcal{E}_f(\mu) = \sup_{\varphi \in \mathcal{C}^1(\bar{\Omega})} \int_\Omega \left(f\varphi - \mu \frac{|\nabla \varphi|^2}{2} \right) dx$$

By using the duality result in Lemma 40, we can write $\forall \mu \in L^1_+(\Omega)$ the following equalities

$$\mathcal{L}(\mu) = \mathcal{E}_f(\mu) + \mathcal{M}(\mu) = \inf_{\xi \in L^2_\mu(\Omega))^d} \left\{ \Upsilon(\mu, \xi) \; : \; \mathrm{div}(\xi\mu) = f \quad \right\}$$

where

$$\Upsilon(\mu, \xi) := \frac{1}{2} \int_\Omega |\xi|^2 \mu \, dx + \frac{1}{2} \int_\Omega \mu \, dx$$

Now for any $\mu \in L^1_+(\Omega)$ and for any $\xi \in (L^2_\mu(\Omega))^d$, by Young inequality we obtain:

$$\int_\Omega |\xi| \mu \, dx \leq \frac{1}{2} \int_\Omega |\xi|^2 \mu \, dx + \frac{1}{2} \int_\Omega \mu \, dx = \Upsilon(\mu, \xi) \quad \forall \xi \in (L^2_\mu(\Omega))^d \tag{2.45}$$

By taking the infimum on $\xi \in (L^2_\mu(\Omega))^d$ with $\mathrm{div}(\xi\mu) = f$ in the last inequality we obtain

$$\inf_{\xi\in(L^2_\mu(\Omega))^d} \left\{ \int_\Omega |\xi|\mu \, dx \ : \ \mathrm{div}(\xi\mu) = f \quad \right\} \leq \mathcal{L}(\mu) \quad \forall \mu \in L^1_+(\Omega)$$

Using these last inequalities and the fact that, by Proposition 15, the integral of OT density $\mu^*(f)$ is equal to value of the Beckmann we can write the following equalities and inequalities

$$\begin{aligned}
\int_\Omega \mu^* \, dx &= \inf_{v\in[L^1(\Omega)]^d} \left\{ \int_\Omega |v| \, dx \ : \ \mathrm{div}(v) = f \quad \right\} \\
&\leq \inf_{\mu,\xi} \left\{ \int_\Omega |\xi|\mu \, dx \ : \ \begin{array}{c} (\mu,\xi) \in L^1_+(\Omega) \times [L^2_\mu(\Omega)]^d \\ \mathrm{div}(\xi\mu) = f \end{array} \right\} \\
&\leq \mathcal{L}(\mu)
\end{aligned}$$

that hold for any $\mu \in L^1_+(\Omega)$. Thus the last inequality holds also for $\inf_{\mu\in L^1_+(\Omega)} \mathcal{L}(\mu)$. But now

$$\inf_{\mu\in L^1_+(\Omega)} \mathcal{L}(\mu) \leq \mathcal{L}(\mu^*) = \int_\Omega \mu^* \, dx$$

All the above inequalities are equalities and the proof is complete. □

2.3.1 Deduction of the Lyapunov-candidate functional

In this section we describe an empirical rationale corroborating the deduction of the Lyapunov-candidate functional $\mathcal{L}$. No real proofs are reported here, but this, together with Proposition 39 and the numerical experiments reported in the next chapter, contributes to forming the idea that the DMK is a valid alternative for the approximate quantification of the MK equations. Given the necessarily informal character of this paragraph, we restrict our analysis to absolutely continuous measure (with respect to Lebesgue) with L^1- density. thus the same holds for the OT density associated to f. In the second part of this paragraph we report a (again empirical) derivation starting form the Kantorovich Dual Problem in Equation (1.6).

From Mass Optimization Problem. We rewrite the Mass Optimization Problem (MOP) by definition of $\mathcal{E}_f$ in Equation (1.11):

$$\min_{\nu\in L^1_+(\Omega)} \left\{ \mathcal{E}_f(\nu) \ : \int_\Omega \nu \, dx = 1 \right\} = \min_{\nu\in L^1_+(\Omega)} \left\{ \max_{\varphi\in \mathrm{Lip}(\Omega)} \Gamma_f(\nu,\varphi) \ : \int_\Omega \nu \, dx = 1 \right\} \tag{2.46}$$

Instead of studying the above constrained minimization problem we can introduce a Lagrange multiplier $\lambda \in \mathbb{R}$ and write following equivalent problem

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}} \min_{\nu\in L^1(\Omega)} \left\{ \mathcal{E}_f(\nu) + \lambda \int_\Omega \left(\nu - \frac{1}{|\Omega|} \right) dx \right\} = \\
&\max_{\lambda\in\mathbb{R}} \min_{\nu\in L^1(\Omega)} \left\{ \max_{\varphi\in \mathrm{Lip}(\Omega)} \Gamma_f(\nu,\varphi) + \lambda \int_\Omega \left(\nu - \frac{1}{|\Omega|} \right) dx \right\} = \\
&\max_{\lambda\in\mathbb{R}} \min_{\nu\in L^1(\Omega)} \max_{\varphi\in \mathrm{Lip}(\Omega)} \Phi(\nu,\varphi,\lambda)
\end{aligned}$$

where

$$\Phi(\nu,\varphi,\lambda) := \int_\Omega f\varphi - \nu \frac{|\nabla \varphi|^2}{2}\, dx + \lambda \int_\Omega \left(\nu - \frac{1}{|\Omega|} \right) dx$$

Since the pair $(\nu^*, \varphi^*) = (\frac{\mu^*}{c^*}, c^* \cdot u^*)$ where $c^* := \int_\Omega d\mu^*$ solves the saddle point problem in Equation (2.46), and satisfies the constrain $\int_\Omega \nu = 1$, for any $\lambda \in \mathbb{R}$. Computing the first variation of Φ with respect to ν we can determine the value of optimal Lagrange Multiplier, which is $\lambda^* = (c^*)^2/2)$. By removing additive constants from the minimization problem of $\Phi(\nu, \varphi, (c^*)^2/2)$ we obtain

$$\operatorname*{argmin}_{\nu\in L^1_+(\Omega)} \left\{ \mathcal{E}_f(\nu) \ : \int_\Omega \nu\, dx = 1 \right\} = \operatorname*{argmin}_{\nu\in L^1_+(\Omega)} \left\{ C(\nu) \ : \ \nu \in L^1(\Omega) \right\}$$

where

$$C(\nu) := \mathcal{E}_f(\nu) + \frac{(c^*)^2}{2} \int_\Omega \nu\, dx$$

We now introduce the change of variable $\Psi : L^1(\Omega) \mapsto L^1(\Omega)$ defined as $\Psi(\mu) = \mu/c^*$, and let $\tilde{C}(\mu) := C(\Phi(\mu))$ such that

$$\tilde{C}(\mu) = C(\Phi(\mu)) = c^* \mathcal{E}_f(\mu) + c^*/2 \int_\Omega d\mu = c^* \mathcal{L}(\mu) \tag{2.47}$$

from which we readily obtain

$$\begin{aligned}
\operatorname*{argmin}_{\mu\in L^1(\Omega)} \mathcal{L}(\mu) &= \operatorname*{argmin}_{\mu\in L^1(\Omega)} \tilde{C}(\mu) = \Psi^{-1} \left(\operatorname*{argmin}_{\nu\in L^1(\Omega)} C(\nu) \right) \\
&= \Psi^{-1} \left(\operatorname*{argmin}_{\nu\in L^1_+(\Omega)} \left\{ \mathcal{E}_f(\nu) \ : \int_\Omega \nu\, dx = 1 \right\} \right) \\
&= \Psi^{-1} \left(\frac{\mu^*}{\int_\Omega \mu^*\, dx} \right) \\
&= \mu^*
\end{aligned}$$

From Kantorovich Dual Problem. We now show an informal proof of the equivalence between the minimization of $\mathcal{L}$ and the Kantorovich Dual Problem, that we recall, as given in Theorem 10, can be stated as the following optimization problem:

$$\sup_{u \in \mathrm{Lip}_1(\Omega)} \int_\Omega uf \, dx$$

The constraint $u \in \mathrm{Lip}_1(\Omega)$ can be written as $|\nabla u| \leq 1$, or, equivalently, as $\frac{|\nabla u|^2 - 1}{2} \leq 0$. Then we can introduce a Lagrange multiplier $\lambda \in L^1_+(\Omega)$ and study the unconstrained problem

$$\inf_{\lambda \in L^1_+(\Omega)} \sup_{u \in \mathrm{Lip}(\Omega)} \int_\Omega uf \, dx - \int_\Omega \lambda \left(\frac{|\nabla u|^2 - 1}{2} \right) dx = \inf_{\lambda \in L^1_+(\Omega)} \mathcal{E}_f(\lambda) + \mathcal{M}(\lambda)$$

recalling that

$$\mathcal{E}_f(\lambda) = \sup_{\varphi \in \mathrm{Lip}(\Omega)} \int_\Omega f\varphi - \lambda \frac{|\nabla \varphi|^2}{2} \, dx \quad \mathcal{M}(\lambda) = 1/2 \int_\Omega \lambda \, dx$$

thus the equivalence is proved.

Remark 4. *These simple deduction of the functional $\mathcal{L}$, which reinterprets the OT density as a Lagrange multiplier of problem in Equation* (1.6)*, was already present as comment in the first lines in [30, p. 36] but the computation were not developed.*

2.4 Extension to non-uniform metric

We now introduce a generalized version of Equation (2.6) that extends the results obtained in the previous section to the case of the OTP with cost equal to a geodetic distance described in Section 1.4.4. To this aim the function $k(x)$, that defines the metric, can be used to describe, for example, the spatial pattern of the resistance to flow, whereby large values of k imply large energy losses and hence large gradients of the potential u. Thus Equation (2.31) is replaced by

$$\partial_t \mu(t,x) = \mu(t,x) |\nabla u(t,x)| - k(x)\mu(t,x) \tag{2.48}$$

Existence and uniqueness of above ODE can be obtained as a straight forward extension of Theorem 27, while the Lyapunov-candidate functionalin Equa-

tion (2.37) rewrites as

$$\mathcal{L}_{k(x)}(\mu) := \mathcal{E}_f(\mu) + \mathcal{M}_{k(x)}(\mu) \tag{2.49}$$
$$\mathcal{E}_f(\mu) = \frac{1}{2}\int_\Omega \mu|\nabla u(\mu)|^2\,dx \qquad \mathcal{M}_{k(x)}(\mu) := \frac{1}{2}\int_\Omega k^2(x)\mu\,dx$$

We can easily obtain the generalization of Propositions 38 and 39, summarized in the followings

Proposition 41. *The functional $\mathcal{L}_{k(x)}$ defined in Equation* (2.49) *is decreasing along the μ-trajectories of Equation* (2.48)*. Its time derivative is given by*

$$\frac{d\mathcal{L}_{k(x)}(\mu(t))}{dt} = -\frac{1}{2}\int_\Omega \mu(t)\left(|\nabla u(t)| - k(x)\right)^2\left(|\nabla u(t)| + k(x)\right)dx < 0$$

Moreover the minimization of $\mathcal{L}_{k(x)}$ over the field $\mu \geq 0$ is equivalent to the generalized Beckmann Problem in Equation (1.15)*, i.e.:*

$$\min_{v\in[L^1(\Omega)]^d}\left\{\int_\Omega k(x)|v|\,dx \;\; : \mathrm{div}(v) = f\right\} = \min_{\mu\in L^1_+(\Omega)} \;\mathcal{L}_{k(x)}(\mu)$$

The proof of this last result is a straightforward adaptation of the arguments used in the proof of Equation (2.40).

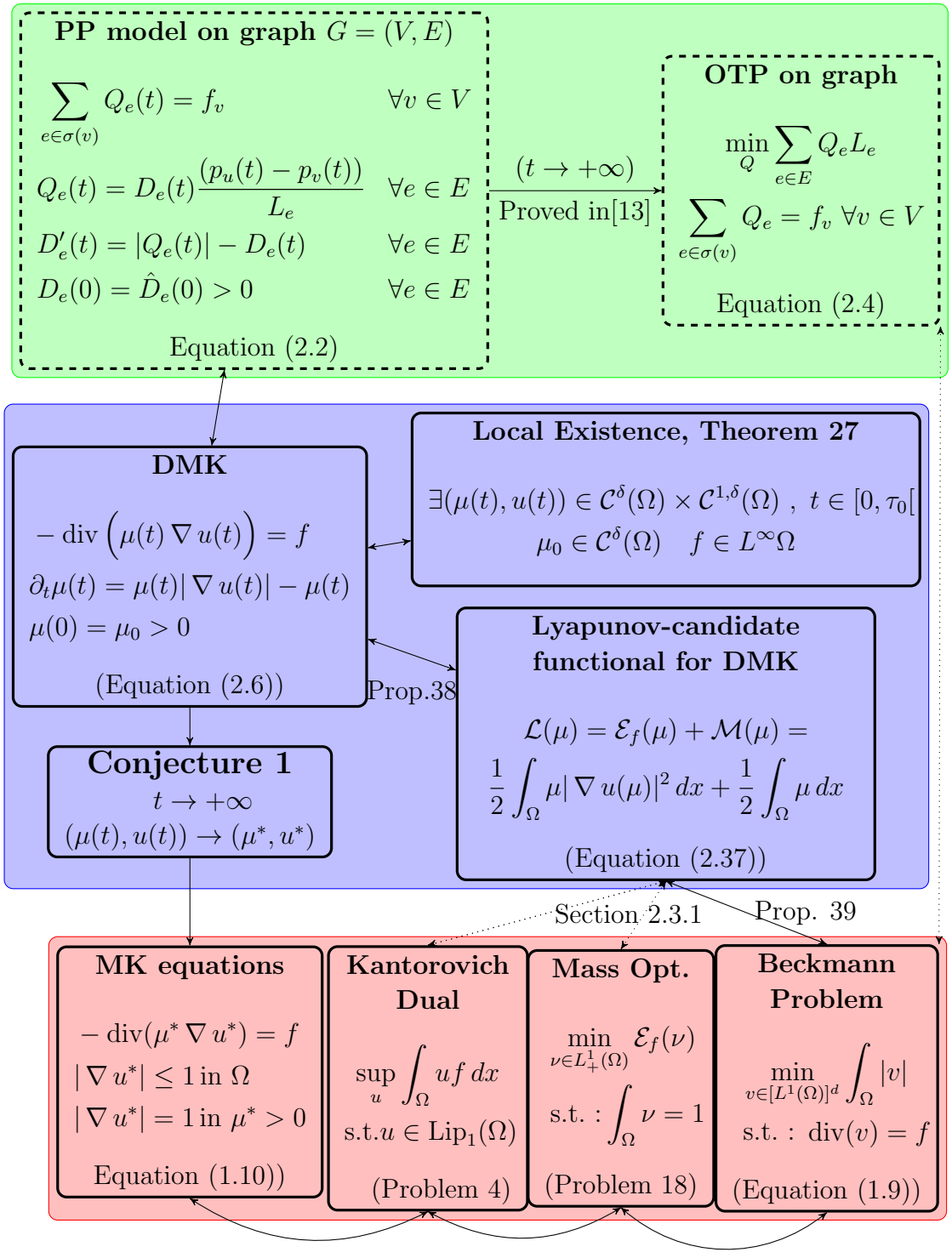

Figure 2.2: Map of the results of this chapter. We highlight the connections among the discrete models (green block), the DMK model(blue block), and the different formulations of the L^1-OTP(red block).

2.5 Numerical Solution of MK equations by using DMK

In this section we analyze how the DMK model presented in Chapter 2 can be used as new method for the numerical solution of the L^1-PDE based optimal transportation model. We describe and test different numerical approaches for the solution of our problem based on the finite element method. Because of the lack of global control on $|\nabla u|$. proof of the convergence of the FEM is beyond reach. Thus we resort to numerical experimentation. the Choices of the discretization spaces is guided by the expected regularity of the OT density and potential (μ^*, u^*) suggested by Theorem 14. Thus it seems reasonable to look fora continuous approximation of the Kantorovich potential and a more flexible and less regular approximation of the OT density. It is then intuitive to use linear conforming Finite Elements for the elliptic equation coupled with piecewise constant. Independently of the spatial discretization, the ensuing nonlinear differential-algebraic equation is discretized by means of a first order Euler method (forward or backward) and a simple Picard iteration is used to resolve the nonlinearity when necessary. The procedure is iterated in time until relative differences of spatial norm of the transport density are smaller than a predefined tolerance.

We study the experimental convergence rate of the proposed solution approaches and discuss limitations and advantages of these formulations. An extensive set of test cases, including problems that admit an explicit solution to the MK equations so that error norms can be accurately evaluated, are appropriately designed to verify and test the expected numerical properties of the solution methods. The results show optimal convergence toward the asymptotic equilibrium point is achieved for sufficiently regular forcing function. All the obtained numerical solutions support Conjecture 1 that indeed the dynamic model possess a time-asymptotic equilibrium point that coincides with the solution of the MK equations . Also the existence and coherence of the Lyapunov-candidate functional is confirmed.

We obtain similar results also for the case $k \not\equiv 1$ described in Section 2.4, and we were able to reproduced the PP experiment who inspired the the discrete model in [68].

2.5.1 Numerical discretization

2.5.1.1 Projection spaces

The numerical approach at the solution of Equation (2.6) is based on the method of lines. Spatial discretization is achieved by projecting the weak formulation of the system of equations onto a pair of finite dimensional spaces $(\mathcal{V}_h, \mathcal{W}_h)$. We denote with $\mathcal{T}_h(\Omega)$ a regular triangulation of the (assumed polygonal) domain Ω, characterized by n nodes and m triangles, where h indicates the characteristic length of the elements. We also denote with $\mathcal{P}_0(\mathcal{T}_h(\Omega)) = \text{span}\{\psi_1(x), \ldots, \psi_M(x)\}$ the space of element-wise constant functions on $\mathcal{T}_h(\Omega)$, i.e., $\psi_i(x)$ is the characteristic function of triangle T_i, and with $\mathcal{P}_1(\mathcal{T}_h(\Omega)) = \text{span}\{\varphi_1(x), \ldots, \varphi_N(x)\}$ the space of element-wise linear Lagrangian basis functions defined on $\mathcal{T}_h(\Omega)$. We consider two different choices of the space $\mathcal{V}_h$ used in the projection of the elliptic Equation (2.6a), namely $\mathcal{V}_h = \mathcal{P}_{1,h} = \mathcal{P}_1(\mathcal{T}_h(\Omega))$ and $\mathcal{V}_h = \mathcal{P}_{1,h/2} = \mathcal{P}_1(\mathcal{T}_{h/2}(\Omega))$. Here $\mathcal{T}_{h/2}(\Omega)$ is the triangulation generated by conformally refining each triangle $T_k \in \mathcal{T}_h(\Omega)$ (.i.e. each element T_k is divided in 2^d sub-elements having as nodes the gravity centers of the 2^{d-1}-faces contained in T_k) . Again we consider different choices of spaces also for the projection of the dynamic equation Equation (2.6b) by using alternatively $\mathcal{W}_h = \mathcal{P}_{1,h}$ and $\mathcal{W}_h = \mathcal{P}_{0,h} = \mathcal{P}_0(\mathcal{T}_h(\Omega))$, when the projection is done on the same mesh used for the elliptic equations, or $\mathcal{W}_h = \mathcal{P}_{1,h/2}$ $\mathcal{W}_h = \mathcal{P}_{0,h} = \mathcal{P}_0(\mathcal{T}_h(\Omega))$, when we use the sub-grid.

NewP Following this approach and separating the temporal and spatial variables, the discrete potential $u_h(t,x)$ and diffusion coefficient $\mu_h(t,x)$ are written as:

$$u_h(t,x) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \quad \varphi_i \in \mathcal{V}_h \qquad \mu_h(t,x) = \sum_{k=1}^{M} \mu_k(t)\psi_k(x) \quad \psi_k \in \mathcal{W}_h$$

where N and M are the dimensions of $\mathcal{V}_h$ and $\mathcal{W}_h$, respectively. The finite element discretization yields the following problem: for $t \geq 0$ find $(u_h(t,\cdot), \mu_h(t,\cdot)) \in \mathcal{V}_h \times \mathcal{W}_h$ such that

$$a_{\mu_h(u_h,\varphi_j)} = \int_\Omega \mu_h \nabla u_h \cdot \nabla \varphi_j \, dx = (f, \varphi_j) = \int_\Omega f\varphi_j \, dx \quad j = 1, \ldots, N \quad (2.51a)$$

$$\int_\Omega \partial_t \mu_h \psi_l \, dx = \int_\Omega (|\mu_h \ \nabla u_h| - \mu_h)\psi_l \, dx \quad l = 1, \ldots, M \quad (2.51b)$$

$$\int_\Omega \mu_h(0,\cdot)\psi_j \, dx = \int_\Omega \mu_0 \psi_l \, dx \quad l = 1, \ldots, M \quad (2.51c)$$

where we add to Equation (2.51a) the zero-mean constraint $\int_\Omega u_h\,dx = 0$ to enforce well-posedness. In matrix form, indicating with $\boldsymbol{u}(t) = \{u_i(t)\}$, $i = 1,\dots,N$, and $\boldsymbol{\mu}(t) = \{\mu_k(t)\}$, $k = 1,\dots,M$, the vectors that describe the time evolution of the projected system, we can write the following index-1 nonlinear system of differential algebraic equations (DAE):

$$\boldsymbol{A}[\boldsymbol{\mu}(t)]\ \boldsymbol{u}(t) = \boldsymbol{b}, \tag{2.52a}$$

$$\boldsymbol{M}\ \partial_t\boldsymbol{\mu}(t) = \boldsymbol{B}(\boldsymbol{u}(t))\ \boldsymbol{\mu}(t), \qquad \boldsymbol{M}\ \boldsymbol{\mu}(0) = \boldsymbol{\mu}_0 \tag{2.52b}$$

The $N \times N$ stiffness matrix $\boldsymbol{A}[\boldsymbol{\mu}(t)]$ is given by:

$$A_{ij}[\boldsymbol{\mu}(t)] = \sum_{k=1}^{M} \mu_k(t) \int_\Omega \psi_k\,\nabla\,\varphi_i \cdot \nabla\,\varphi_j\,dx$$

The N-dimensional source vector $\boldsymbol{b}$ whose components are $b_i = \int_\Omega f\ \varphi_i\,dx$. The singularity due to the homogeneous Neumann boundary conditions is removed by forcing the solution $\boldsymbol{u}$ to remain orthogonal to vector $a_i = \int_\Omega \varphi_i\,dx$ [11]. The $M \times M$ mass matrix $\boldsymbol{M}$ is expressed by:

$$M_{k,l} = \int_\Omega \psi_k\psi_l\,dx,$$

The $M \times M$ matrix $\boldsymbol{B}$ has the same structure of $\boldsymbol{M}$ and is defined as

$$B_{k,l}[\boldsymbol{u}(t)] = \int_\Omega \left(|\sum_{i=1}^{N} u_i(t)\,\nabla\,\varphi_i| - 1 \right) \psi_k\psi_l\,dx$$

and the M-dimensional vector $\boldsymbol{\mu}_0$ contains the projected initial condition $\mu_{0,l} = \int_\Omega \mu_0\psi_l\,dx$. When we consider $\mathcal{W}_h = \mathcal{P}_{0,h}$, matrices $\boldsymbol{M}$ and $\boldsymbol{B}$ are diagonal and Equation (2.52) simplifies to:

$$\boldsymbol{A}[\boldsymbol{\mu}(t)]\boldsymbol{u}(t) = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}(t) = 0 \tag{2.53a}$$

$$\partial_t\boldsymbol{\mu}(t) = \boldsymbol{D}[\boldsymbol{u}(t)]\boldsymbol{\mu}(t), \qquad \boldsymbol{\mu}(0) = \tilde{\boldsymbol{\mu}}_0 \tag{2.53b}$$

where $\boldsymbol{D}$ is the $M \times M$ diagonal matrix given by:

$$D_{k,k}[\boldsymbol{u}(t)] = \frac{1}{|T_k|} \int_{T_k} \left(|\sum_{i=1}^{N} u_i(t)\,\nabla\,\varphi_i| - 1 \right) dx$$

where $|T_k|$ is the measure of the element T_k, and $\tilde{\boldsymbol{\mu}}_0$ is the M-dimensional vector with components given by $\tilde{\mu}_{0_k} = \frac{1}{|T_k|}\int_{T_k} \mu_0\,dx$,i.e., the L^2-projection of μ_0 on the triangles of $\mathcal{T}_h$.

2.5.1.2 Time discretization

In order to solve the DAE Equation (2.52) or Equation (2.53) we define a discretization in time using either a forward or a backward Euler scheme. Denoting with Δt_k the time-step size so that $t_{k+1} = t_k + \Delta t_k$ and $(\boldsymbol{u}^k, \boldsymbol{\mu}^k) = (\boldsymbol{u}(t_k), \boldsymbol{\mu}(t_k))$, the approximate solutions at time t_k can be written as $u_h^k(x) = \sum_i^N u_i^k \varphi_i(x)$ and $\mu_h^k(x) = \sum_{l=1}^M \mu_l^k \psi_l(x)$.

Case $\mu_h(t,\cdot) \in \mathcal{P}_{1,h}$. In this case, the forward Euler scheme yields:

$$\begin{aligned} &\boldsymbol{A}[\boldsymbol{\mu}^k]\ \boldsymbol{u}^k = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}^k = 0 \\ &\boldsymbol{\mu}^{k+1} = (I + \Delta t_k \boldsymbol{M}^{-1}\boldsymbol{B}[\boldsymbol{u}^k])\boldsymbol{\mu}^k, \qquad \boldsymbol{\mu}^0 = \boldsymbol{M}^{-1}\mu_0 \end{aligned}$$

When backward Euler is employed, the time-stepping scheme becomes:

$$\begin{aligned} &\boldsymbol{A}[\boldsymbol{\mu}^{k+1}]\boldsymbol{u}^{k+1} = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}^{k+1} = 0 \\ &\boldsymbol{M}\boldsymbol{\mu}^{k+1} = \boldsymbol{M}\boldsymbol{\mu}^k + \Delta t_k \boldsymbol{B}[\boldsymbol{u}^{k+1}]\boldsymbol{\mu}^{k+1}, \qquad \boldsymbol{\mu}^0 = \boldsymbol{M}^{-1}\mu_0 \end{aligned}$$

and the nonlinearity is resolved by means of the following successive iteration (Picard) scheme:

$$\begin{aligned} &\boldsymbol{\mu}^{0,k+1} = \boldsymbol{\mu}^k \\ &\text{for } m = 0, 1, 2, \ldots \\ &\qquad \boldsymbol{A}[\boldsymbol{\mu}^{m,k+1}]\boldsymbol{u}^{m,k+1} = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}^{m,k+1} = 0 \\ &\qquad \boldsymbol{\mu}^{m+1,k+1} = (\boldsymbol{M} - \Delta t_k \boldsymbol{B}[\boldsymbol{u}^{m,k+1}])^{-1}\ \left(\boldsymbol{M}\boldsymbol{\mu}^k\right) \end{aligned} \tag{2.54}$$

which can be repeated until the relative difference $\rho(\mu_h^{m+1,k+1}, \mu_h^{m,k+1}) \leq \tau_{\mathrm{NL}}$, where

$$\rho(\mu_h^{m+1,k+1}, \mu_h^{m,k+1}) = \frac{\|\mu_h^{m+1,k+1} - \mu_h^{m,k+1}\|_{L^2(\Omega)}}{\|\mu_h^{m,k+1}\|_{L^2(\Omega)}}$$

or the number of Picard iterations m reaches a prefixed maximum m_{MAX}.

Case $\mu_h(t,\cdot) \in \mathcal{P}_{0,h}$. In this case the forward Euler scheme reads:

$$\begin{aligned} &\boldsymbol{A}[\boldsymbol{\mu}^k]\ \boldsymbol{u}^k = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}^k = 0 \\ &\boldsymbol{\mu}^{k+1} = (I + \Delta t_k \boldsymbol{D}[\boldsymbol{u}^k])\boldsymbol{\mu}^k, \qquad \boldsymbol{\mu}^0 = \tilde{\mu}_0 \end{aligned}$$

while the backward Euler discretization yields the non-linear system of equations:

$$\begin{aligned}
&\boldsymbol{A}[\boldsymbol{\mu}^{k+1}]\ \boldsymbol{u}^{k+1} = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}^{k+1} = 0 \\
&\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^{k} + \Delta t_k \left(\boldsymbol{D}[\boldsymbol{u}^{k+1}]\boldsymbol{\mu}^{k+1}\right), \qquad \boldsymbol{\mu}^{0} = \tilde{\mu}_0
\end{aligned}$$

solved again by Picard iteration:

$$\begin{aligned}
&\boldsymbol{\mu}^{0,k+1} = \boldsymbol{\mu}^{k} \\
&\\
&\boldsymbol{A}[\boldsymbol{\mu}^{m,k+1}]\boldsymbol{u}^{m,k+1} = \boldsymbol{b}, \qquad \boldsymbol{a}\cdot\boldsymbol{u}^{m,k+1} = 0 \\
&\boldsymbol{\mu}^{m+1,k+1} = (\boldsymbol{I} - \Delta t_k \boldsymbol{D}[\boldsymbol{u}^{m,k+1}])^{-1}\boldsymbol{\mu}^{k}
\end{aligned} \tag{2.55}$$

repeated until $\rho(\mu_h^{m+1,k+1}, \mu_h^{m,k+1}) \leq \tau_{\mathrm{NL}}$ or $m > m_{\mathrm{MAX}}$. Note again that the matrix $(I - \Delta t_k \boldsymbol{D})$ in Equation (2.55) is trivially invertible being diagonal.

2.5.1.3 Solution of the linear system

At each Picard iteration we need to solve a large sparse symmetric linear system that is positive semi-definite because homogeneous Neumann boundary conditions are used. We solve the system by a preconditioned conjugate gradient (PCG) method and employ the approach suggested by [11] to construct the Krylov subspace orthogonal to the null space of the system matrix. We choose this iterative linear solver for two reasons: first, we are solving a sequence of slightly varying linear systems, thus at each system solution we can obtain additional data, like the initial solution or spectral informations (a strategy exploited in Chapter 4) to solve more efficiently the next linear system. Second, we can cope with the near singularity of the stiffness matrix more easily than by using a direct solver.

Convergence of the PCG iteration is considered achieved when the Euclidean norm of the residual relative to the initial residual norm is smaller than the tolerance τ_{CG}. We start the iteration from u_h^k, i.e., solution at the previous time step and use an incomplete Choleski factorization with no fill-in as preconditioner. Since the system dynamics drives the transport density μ_h to zero in large portions of the domain Ω, we set a lower limit to it and impose that $\mu_h \geq 10^{-10}$ everywhere. This is sufficient to guarantee the ellipticity of the FEM bilinear form and allows the system condition number to remain bounded so that PCG converges within a limited number of iterations.

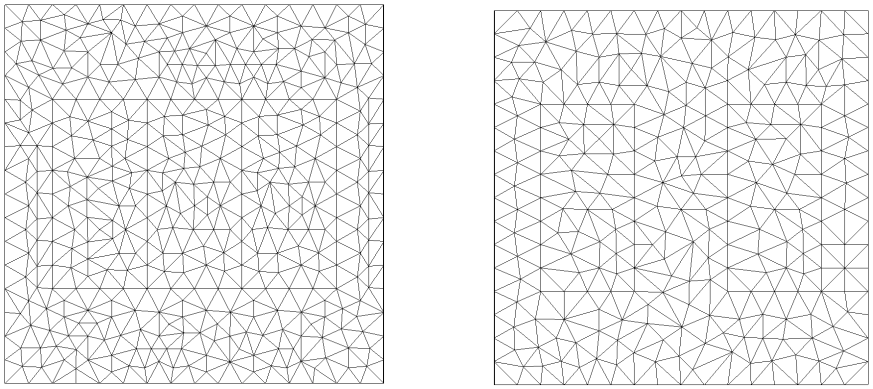

Figure 2.3: Domain Ω and supports Q^+, Q^c, and Q^- for Test Case 1. Unrefined initial meshes. From the left to the right: Mesh 1 (constrained Delaunay, 438 nodes and 810 elements), and Mesh 2 (constrained Delaunay, 297 nodes and 528 elements). The edges of Mesh 1 are aligned with the supports of f and $\mu^*(f)$. Mesh 2 is aligned only with the supports of f.

2.5.2 Numerical experiments

We study the numerical scheme described in Section 2.5.1 with two test-cases. In the first test-case we compare the large-time solution of the projected system against the closed form solution proposed by [18] for sufficiently regular forcing functions. We also verify the convergence toward steady-state for increasingly refined grids. In the second test-case, we extend the comparison against those reported in [5]. We also analyze the stability of the spatial discretization method used.

2.5.2.1 Test Case 1: comparison with closed-form solutions

In this first set of tests we experiment the convergence of the different numerical schemes toward the OT density by comparing the numerical solution with the closed-form solution of the MK equations discussed in [18]. To this aim, we consider a square domain in $\mathbb{R}^2$, $\Omega = [0,1] \times [0,1]$, and a zero-mean forcing function f supported in two rectangles Q^+ and Q^- contained in Ω, where f

assumes opposite signs (Figure 2.3). The different supports are given by:

$$Q^+ = \left\{(x,y) \in \Omega : (x,y) \in \left[\frac{1}{8},\frac{3}{8}\right] \times \left[\frac{1}{4},\frac{3}{4}\right]\right\}$$
$$Q^c = \left\{(x,y) \in \Omega : (x,y) \in \left[\frac{3}{8},\frac{5}{8}\right] \times \left[\frac{1}{4},\frac{3}{4}\right]\right\}$$
$$Q^- = \left\{(x,y) \in \Omega : (x,y) \in \left[\frac{5}{8},\frac{7}{8}\right] \times \left[\frac{1}{4},\frac{3}{4}\right]\right\}$$

We are optimally transporting $f^+ = f(Q^+)$ into $f^- = f(Q^-)$ and look for the density $\mu(t,x)$ and the potential $u(t,x)$ that satisfy Equation (2.6) as $t \to \infty$ and approximate μ^* solution of Equation (1.10). We consider that a time-equilibrium condition has been achieved when the relative variation in μ_h ($\mathrm{var}(\mu_h)$) is smaller than a tolerance, i.e.,

$$\mathrm{var}(\mu_h) := \rho(\mu_h^{k+1}, \mu_h^{tstep})/\Delta t_k < \tau_{\mathrm{T}}.$$

We indicate with t^* the time when time equilibrium is numerically reached and with μ_h^* the corresponding μ_h^k.

To test our numerical schemes we set up two different problems that differentiate by the specific choice of f. In particular, the first test case considers a continuous forcing function f_1 with opposite sign in Q^+ and Q^-, while the second case considers a piecewise constant function f_2:

$$f_1(x,y) = \begin{cases} 2\sin\left(4\pi\left(x-\frac{1}{8}\right)\right)\sin\left(2\pi\left(y-\frac{1}{4}\right)\right) & \text{in } Q^+ \\ -2\sin\left(4\pi\left(x-\frac{5}{8}\right)\right)\sin\left(2\pi\left(y-\frac{1}{4}\right)\right) & \text{in } Q^- \\ 0 \quad \text{elsewhere} \end{cases}$$

and

$$f_2(x,y) = \begin{cases} 2 & \text{in } Q^+ \\ -2 & \text{in } Q^- \\ 0 & \text{elsewhere} \end{cases}$$

Correspondingly, the OT density $\mu^*(f)$ is given by [18]:

$$\mu^*(f_1)(x,y) = \begin{cases} \frac{1}{2\pi}\left(1-\cos\left(4\pi\left(x-\frac{1}{8}\right)\right)\right)\sin\left(2\pi\left(y-\frac{1}{4}\right)\right) & \text{in } Q^+ \\ \frac{1}{\pi}\sin\left(2\pi\left(y-\frac{1}{4}\right)\right) & \text{in } Q^c \\ \frac{1}{2\pi}\left(1-\cos\left(4\pi\left(\frac{7}{8}-x\right)\right)\right)\sin\left(2\pi\left(y-\frac{1}{4}\right)\right) & \text{in } Q^- \\ 0 \quad \text{elsewhere} \end{cases}$$

and

$$\mu^*(f_2)(x,y) = \begin{cases} 2\left(x-\frac{1}{8}\right) & \text{in } Q^+ \\ \frac{1}{2} & \text{in } Q^c \\ 2\left(\frac{7}{8}-x\right) & \text{in } Q^- \\ 0 \quad \text{elsewhere} \end{cases}$$

Note that the support of $\mu^*(f)$ is given by $Q^\mu = Q^+ \cup Q^c \cup Q^-$. With this explicit solution, we can verify the experimental convergence rates in space at infinite times for the different proposed schemes. We use two different initial triangulation settings (Mesh 1 and Mesh 2, Figure 2.3) that are uniformly refined four times to assess FEM convergence. Both meshes are constrained to be aligned with the exact supports of f^+ and f^-, so that the condition $\sum_i \int_\Omega f(x)\varphi_i\,dx = 0$ can be imposed exactly, and have approximately the same number of nodes and elements. Mesh 1 (Figure 2.3, left) is a constrained Delaunay triangulations with edges aligned with the boundary of Q^μ. Mesh 2 is also a constrained Delaunay triangulation but is not aligned with Q^μ in the area between Q^+ and Q^-. Hence, in the latter case, we expect convergence to be influenced also by the geometric convergence of the mesh element boundaries toward the support Q^μ of μ^*. Sensitivity to initial conditions is tested by employing the following different initial data $\mu_0^{(i)}$:

$$\begin{gathered} \mu_0^{(1)}(x,y) = 1; \quad \mu_0^{(2)}(x,y) = 0.1 + 4\left((x-0.5)^2 + (y-0.5)^2\right); \\ \mu_0^{(3)}(x,y) = 3 + 2\sin(8\pi x)\sin(8\pi y). \end{gathered} \tag{2.56}$$

Note that in these tests we are not interested in computational speed, but only on the numerical behavior of the schemes. Thus we do not limit the minimum

time step size and the maximum number of iterations (in both time-stepping and the PCG algorithm used to solve the linear system of algebraic equations), and use a very tight tolerance to determine when time equilibrium is reached. We employ very tight tolerances, i.e., $\tau_{\mathrm{NL}} = 10^{-11}$ and $\tau_{\mathrm{T}} = 5 \times 10^{-9}$, and a PCG exit tolerance on the relative residual $\tau_{\mathrm{CG}} = 10^{-13}$ in order to test the actual limits of the numerical scheme. We would like to remark that much more efficient simulations are obtained with much more relaxed tolerances ($\approx 10^{-3}$) maintaining meaningful quantitative results, for which the spatial distribution of μ_h^* is approximated with an accuracy that can be compared with the discretization error. In the simulations presented here we adopted, both for the forward and back time-stepping, a varying Δt_k by setting $\Delta t_{k+1} = \min(1.05 \times \Delta t_k, \Delta t_{\max})$, where $\Delta t_{\max} = 0.5$ was found experimentally to ensure the stability of the forward Euler scheme, or equivalently, the convergence of the Picard iteration, (more detailed discussion on the Picard scheme is discussed later). Convergence as $h \to 0$ is explored by looking at the time behavior of the $L^2(\Omega)$ relative error defined as:

$$\mathrm{err}(\mu_h(t), f) := \frac{\|\mu_h(t) - \mu^*(f)\|_{L^2(\Omega)}}{\|\mu^*(f)\|_{L^2(\Omega)}}$$

Convergence toward steady-state equilibrium. We analyze the experimental convergence toward an equilibrium point (μ_h^*, u^*) by looking at the time evolution of $\mathrm{var}(\mu_h(t))$ and $\mathrm{err}(\mu_h(t))$. Figure 2.4 reports the log-log scale plots of these two quantities calculated in the case of continuous forcing function. The different curves in each sub-plot display the behavior obtained on successive uniform refinements of the two different initial meshes of Figure 2.3, while the columns identify the different combinations of spatial discretizations. Only results of the Explicit Euler time-stepping scheme are shown, the results of the Implicit Euler method being identical. The first set of plots (first two rows) are relative to the Q^μ-aligned mesh set, while the lower set reports the results for the Q^f-aligned meshes.

The μ_h variation, $\mathrm{var}(\mu_h(t))$, displays a monotone behavior for all schemes, with an expected geometric convergence rate toward steady-state, as evidenced by the slope of the rectilinear portions of the curves which coincides for all mesh-levels and for both mesh types. At increasing refinement levels the plots deviate as a consequence of the higher accuracy of the spatial discretization. This is testified

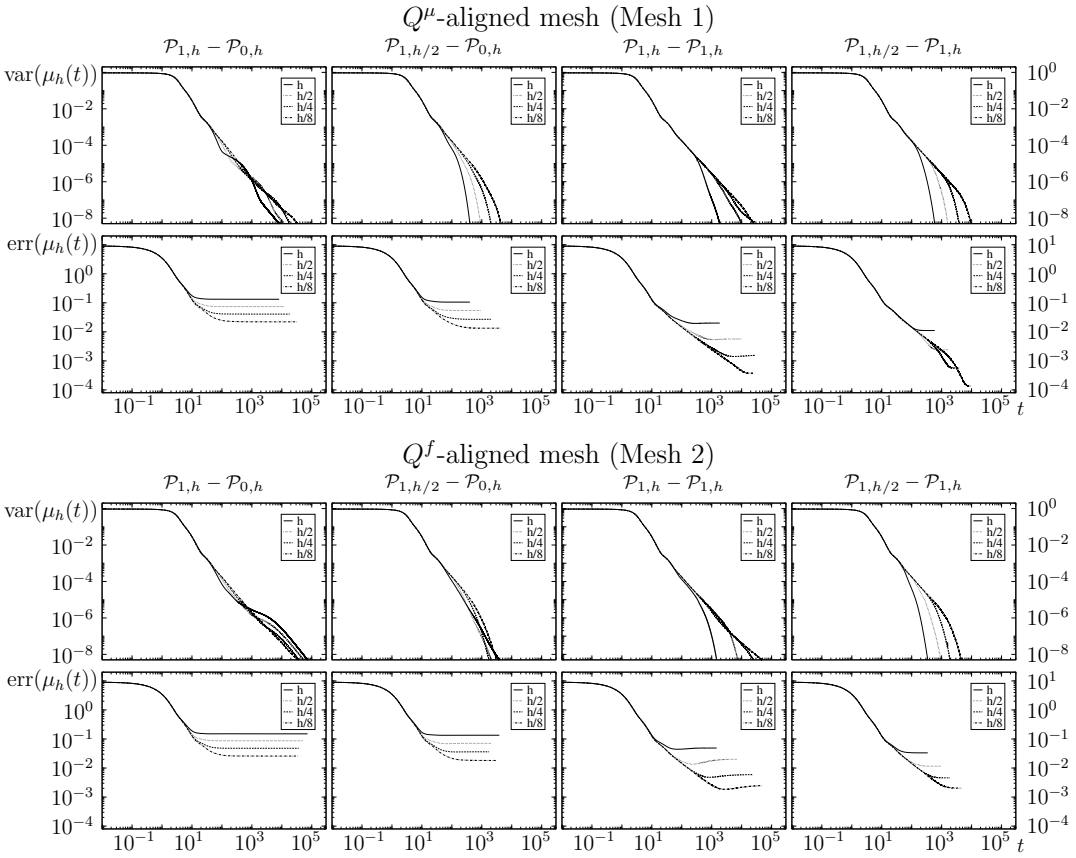

Figure 2.4: Convergence toward equilibrium in the case of continuous forcing ($f = f_1$). The log-log plots of $\mathrm{var}(\mu_h(t,\cdot))$ and $\mathrm{err}(\mu_h(t,\cdot))$ vs. time are reported for Mesh 1 (Q^{μ}-aligned, top block) and for Mesh 2 (Q^{f}-aligned, bottom block). The columns refer from left to right to the results obtained with $\mathcal{P}_{1,h} - \mathcal{P}_{0,h}$, $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$, $\mathcal{P}_{1,h} - \mathcal{P}_{1,h}$, $\mathcal{P}_{1,h/2} - \mathcal{P}_{1,h}$, respectively.

also by the fact that the convergence curves for all schemes initially coincide, starting to diverge approximately when the corresponding spatial accuracy limit is attained. Accuracy saturation in the error plots ($\mathrm{err}(\mu_h(t))$ vs. t) occurs at the same time at which $\mathrm{var}(\mu_h(t))$ start diverging. More uncertain profiles are obtained when spatial discretization is performed on the same mesh for the pair (μ_h, u_h) for both $\mathcal{P}_1 - \mathcal{P}_0$ and $\mathcal{P}_1 - \mathcal{P}_1$ discretization spaces. The reason for the loss of regularity is to be attributed to oscillations in the cell gradients that cause amplified oscillations in the corresponding transport density. Spatial average of the gradient magnitudes, leading to the $\mathcal{T}_h - \mathcal{T}_{h/2}$, shows a much smoother behavior with a faster convergence towards equilibrium. We postpone a more detailed discussion of this phenomenon to Section 2.5.2.2, where a more challenging test case is approached.

Looking at the bottom half of Figure 2.4, we see the effect of using meshes that are not aligned with the support of the OT density . Because of the discontinuity in μ_h occurring across the boundary of Q^μ, convergence is limited by the geometric convergence of the triangular shapes towards this boundary, and the global attainable accuracy is bounded by this error. We observe a consistent behavior of the error for both mesh-types at different h levels. The accuracy levels at which the error saturates decrease consistently with the expected order of spatial convergence of the different schemes, when the geometric error is negligible. This is clearly observable by looking at the plots of $\mathrm{var}(\mu_h(t))$ for the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$, and the $\mathcal{P}_{1,h/2} - \mathcal{P}_{1,h}$, cases, where the optimal second order convergence of the latter approach is observable from the fact that difference in the attained accuracy levels are doubled with respect to the first order approach. In the case of the Q^f-aligned mesh, the higher order approach displays more accurate results with respect to the first order method, but the geometric error prevents the realization of optimal convergence rates.

Convergence of the spatial discretization. We would like to recall that continuity of the transport density $\mu^*(f)$ when the forcing term f is continuous was proved in $\mathbb{R}^2$ in [36] under some assumptions over f. However, except for partial regularity results along transport rays [18], the general case seems to be open question. In our test cases, for both f_1 and f_2 forcings, strong variations in μ_h are present in a direction orthogonal to the boundary of Q^μ in the central portion of the domain (outside Q^f). Because of these variations, which in

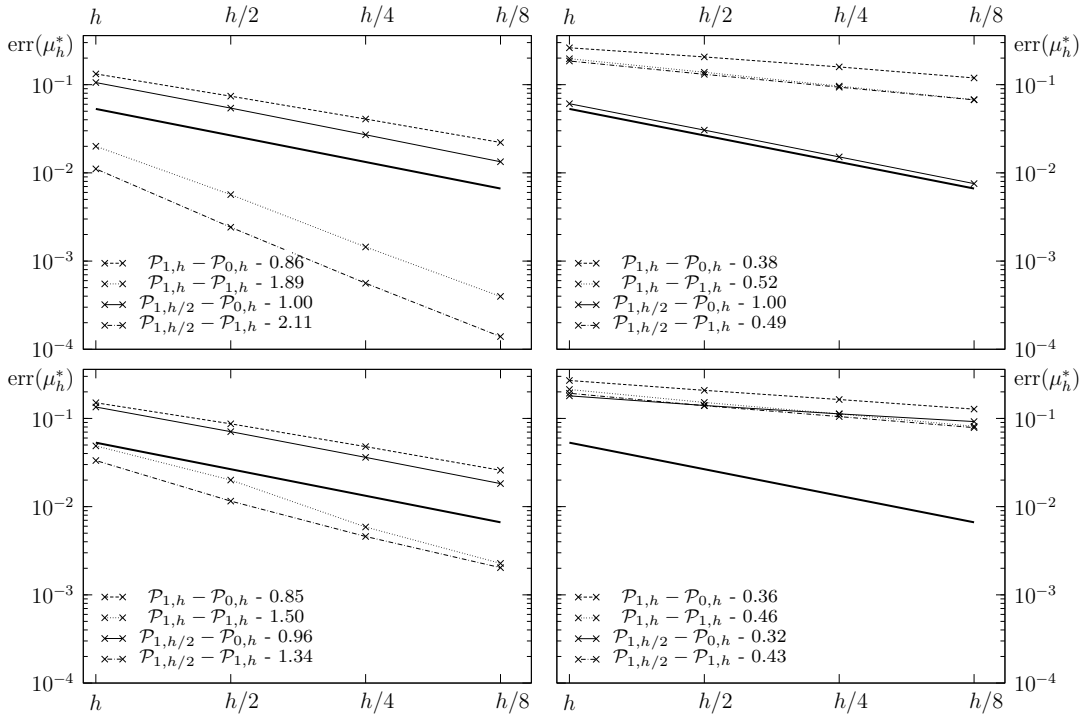

Figure 2.5: Behavior of $\text{err}(\mu_h^*)$ vs. h for the different discretization methods. The results for the continuous forcing function f_1 are shown in the left column, while the right column reports the results for f_2. The top row is relative to the Mesh-1 sequence (aligned with Q^μ), while the bottom row corresponds to the Mesh-2 sequence (aligned only with Q^f). For visual reference, the first order convergence line is also plotted with a thick solid trait. The average experimental convergence rates are reported in the legends of each plot next to the discretization method.

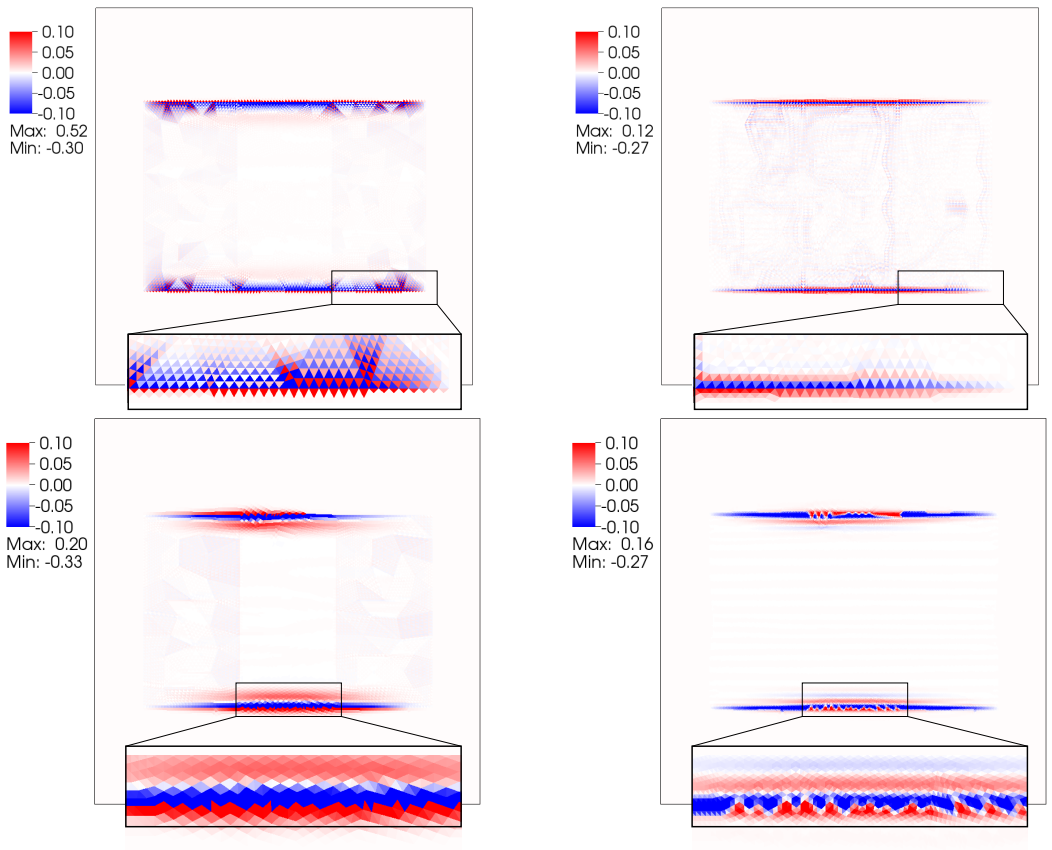

Figure 2.6: Spatial distribution of the error $\mu_h^* - \mu^*(f_2)$ at steady state for the piecewise constant forcing function f_2. The upper row reports the results on the finest level of Mesh-1 obtained with the $\mathcal{P}_{1,h} - \mathcal{P}_{0,h}$, (left) and $\mathcal{P}_{1,h} - \mathcal{P}_{1,h}$, (right). The lower row shows the results on the finest level of Mesh-2 from the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$, (left) and $\mathcal{P}_{1,h/2} - \mathcal{P}_{1,h}$, (right) approaches. Note that classical checkerboard oscillations clearly appear for the scheme that use a single grid, while no oscillations are produced by the two-grid algorithm.

the discontinuous forcing case are actual μ^*-discontinuities, we expect a loss of convergence in the FE solution. We note, however, that convergence towards steady-state is not affected by spatial errors, as shown in the previous discussion.

The experimental convergence profiles for the different methods are reported in Figure 2.5. The column on the left groups the results relative to the more regular case raised by the continuous forcing function f_1 (left). The right column reports the results obtained in the case with the piecewise constant forcing f_2. The top and bottom rows identify the mesh sequences aligned with the boundary of Q^μ or with the boundary of Q^f, respectively. From the two plots on the left, we can argue that all methods attain optimal convergence when the Mesh-1 sequence is used. The $\mathcal{T}_h - \mathcal{T}_{h/2}$ combination is characterized by a smoother behavior. The use of the Mesh-2 sequence, which we recall is aligned only with the boundaries of Q^f and not those of Q^μ, triggers the emergence of geometrical errors that cause a sizeable reduction on the convergence rates of both $\mathcal{P}_1 - \mathcal{P}_1$ and $\mathcal{P}_1 - \mathcal{P}_0$ schemes.

As expected, the results for the discontinuous forcing function (right column) are characterized by an important loss of convergence rate for all schemes, except the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ in combination with the Q^μ-aligned meshes. The loss of regularity in the solution due the lower regularity of f, together with the eventual geometric error in the Mesh-2 sequence, reduces the convergence rates of the schemes. The use of a $\mathcal{T}_h - \mathcal{T}_{h/2}$ combination seems to alleviate somehow the rate loss. This is confirmed by the spatial distribution of the error $\mu_h^* - \mu^*(f_2)$ shown in Figure 2.6. In this figure we report the results obtained with the $\mathcal{P}_{1,h} - \mathcal{P}_{0,h}$ (upper left panel) and the $\mathcal{P}_{1,h} - \mathcal{P}_{1,h}$ (upper right panel) for Mesh 1, and the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ (lower left panel) the $\mathcal{P}_{1,h/2} - \mathcal{P}_{1,h}$ (lower right panel) for Mesh 2. The plots suggest that the $\mathcal{P}_1 - \mathcal{P}_1$ approach localizes the error on the north and south boundaries of Q^μ, where the jump in μ^* from 0 to 0.5 localizes. The $\mathcal{P}_1 - \mathcal{P}_0$ approach, on the other hand, displays an additional small but non negligible error on the support of the forcing function Q^f. The zooms on the pictures show clear oscillations for the one-mesh methods (upper row) in both directions orthogonal and parallel to the μ_h-discontinuity. On the contrary, the methods based on two-meshes (lower row) exhibit a monotone error behavior along the boundary of Q^μ, but the misalignment of the triangle edges causes an increased error as compared to the Mesh-1 results. The error slightly oscillates in the direction normal to the μ_h-jump due to the gradient reconstruction. It is evident that the smoothing due to

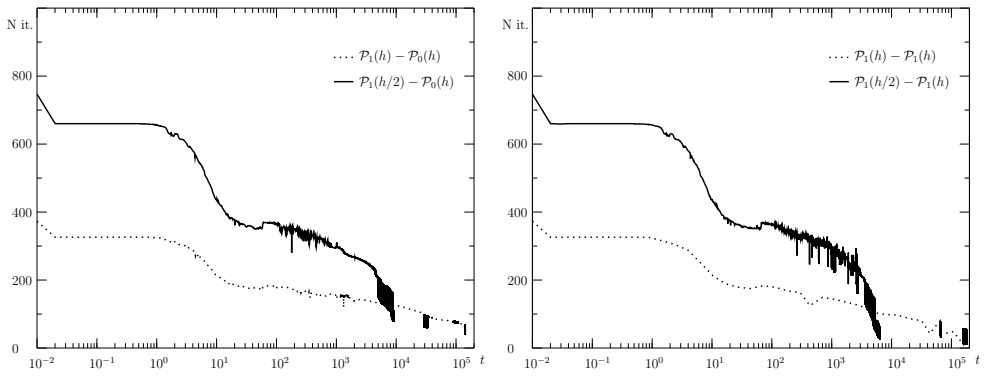

Figure 2.7: Number of iterations required by the PCG scheme to solve the linear system arising from the discretization of the elliptic equation, when $\mu_h \in \mathcal{P}_{0,h}$ (left panel) and $\mu_h \in \mathcal{P}_{1h}$ (right panel). Both panels include the results obtained with and without the use of the sub-grid.

the averaging of the gradient magnitude on the larger triangles helps in reducing overall oscillations. This will become more evident when we will discuss these oscillations in connection with a more challenging test case in Section 2.5.2.2.

Computational cost. We would like to remark here that our numerical approach is not optimized for computational cost. In fact, we are mainly interested in exploring the accuracy and feasibility of our solution approach. Nonetheless, it is interesting to include a discussion on computational cost in the worst-case scenarios described here. A number of cost-saving strategies could be envisaged, including using coarse-mesh solutions to extrapolate initial guesses, re-use of stiffness and preconditioning matrices, developing a Newton-Raphson strategy to improve stability and employ larger time-step sizes, etc. On the other hand, here we are interested in showing that, although far from optimal, our approach is potentially very effective and competitive with literature approaches in the solution of the Monge-Kantorovich equations.

Before initiating this discussion, we note that our transient simulations show an initial phase (approximately until $\mathrm{var}(\mu_h^k) \approx 10^{-3}$) showing profiles of $\mathrm{var}(\mu_h^k)$ and $\mathrm{err}(\mu_h^k)$ that are superimposed for all mesh-refinements. This initial phase is characterized by strong variations of μ_h, during which the support Q^μ starts to delineate with a fast decay of μ_h in the regions where the OT density is zero. After this initial phase, μ_h varies more slowly and stabilizes within Q^μ to its

final value. At the same time, in $\Omega \setminus Q^{\mu}$, the decay continues towards the final preset limit. This phase is characterized by larger time-step sizes and faster PCG convergence.

We measure computational cost by looking at the total number of iterations needed by the PCG method to achieve convergence at each linear system solve, which is obtained when the relative residual is smaller than τ_{CG}. The linear solve phase is prevalent with respect to the matrix assembly phase, being approximately between 60 and 80% of the total cost. Moreover, it is the only phase whose cost varies with time, the assembly phase having a constant computational effort. We remark here that the nonlinearity of the problem and the time-variability of μ_h forces the reassembly of the stiffness matrix at each time-step for Explicit Euler, and at each Picard iteration for Implicit Euler.

Figure 2.7 shows the number of PCG iterations to convergence vs. time in the solution of the linear system arising from the FEM discretization of the elliptic equation Equation (2.6a) using either the $\mathcal{P}_{1,h/p} - \mathcal{P}_{0,h}$ (left panel) or $\mathcal{P}_{1,h/p} - \mathcal{P}_{1,h}$ (right panel) approaches on the finest Mesh-1 refinement level ($p = 1$ or 2). Both methods based on a single mesh ($p = 1$) and on the use of the sub-mesh ($p = 2$) are reported in each panel. For both approaches the initial transient is the most expensive, with $p = 2$ characterized by a doubled number of linear iterations, reflecting the fact that the meshes are parametrized by the mesh parameter h used for the discretization of μ_h. After the initial transient the number of linear iterations starts decreasing, and tends to zero as time progresses, with a faster reduction for the two-mesh methods.

Implicit Euler and convergence of the Picard scheme. In the case of implicit Euler time-stepping, the nonlinear system is solved by Picard iteration as described in Equation (2.54) or in Equation (2.55). All numerical experiments show that the number of iterations of the Picard scheme increases linearly with the time step Δt_k, suggesting a fixed rate of contraction. We estimated it by computing the relative μ_h-variation defined as:

$$C(k) := \frac{\|\mu_h^{m^*,k} - \mu_h^{m^*-1,k}\|_{L^2(\Omega)}}{\|\mu_h^{m^*-1,k} - \mu_h^{m^*-2,k}\|_{L^2(\Omega)}} \tag{2.57}$$

where m^* is the Picard iteration number when convergence occurs. Independently of the spatial discretization method used we obtain the following estimate

$$C(k) \approx K\Delta t_k \qquad K = 1$$

This suggests that Δt_k can be used as a proxy to control the time-step evolution in the case of implicit Euler. Unfortunately, since there is no uniform bound on $|\nabla u(T)| \forall t \geq 0$ we can not deduced the experimental estimate. $C(k) \approx K\Delta t_k$ by the numerical analysis of the convergence scheme. In fact, at the first Picard iteration the constant $C(k)$ can be fixed equal to $\max_k |\nabla u_h^{m,k}|_{T_k}$, but this quantity is useless already at the next Picard iteration. In fact, in other numerical experiments, not reported in here, we noted that, using $\Delta t_k \geq 1$, the Picard scheme fails. While adopting the maximum time step $\Delta t_{\max} = 0.5$, we ensure convergence of the Picard scheme with a number of Picard iterations much smaller than pre-fixed maximum number $m_{\text{MAX}} = 30$. In fact, in all simulations, the maximum number of iterations experimented is 23, while the average number of iterations, computed as the (total number of Picard iterations)/(total number of time iterations), is between 2.19 and 7.19. The choice of $\Delta t_{\max} = 0.5$ offers a good trade-off between minimizing the number of Picard iterations and maximizing the time step. With fixed time step Δt_k, the number of iterations required by the Picard scheme monotonically decreases as we approach the equilibrium, and at the end of the time-iterations very few fix-point iterations are required by the Picard scheme to converge (between 2 and 5). This is clearly due to the exponential decay of the solution as time progresses, as predicted by the mild solution of Equation (2.31).

Dynamics of $\mathcal{L}(\mu(t))$. As a further verification to test the convergence towards steady state, we look at the time behavior of the Lyapunov-candidate functional $\mathcal{L}(\mu_h(t))$ for the different initial conditions described in Equation (2.56). Figure 2.8 reports this behavior for the finest mesh of set 2 of the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ method using a log scale in time. The results for other methods and mesh sets are practically indistinguishable, and are not reported here. We see that $\mathcal{L}$ decreases monotonically and always attains the same minimum value in time independently on the initial conditions. After $t \approx 100$ the value of $\mathcal{L}(\mu_h(t))$ becomes apparently stationary.

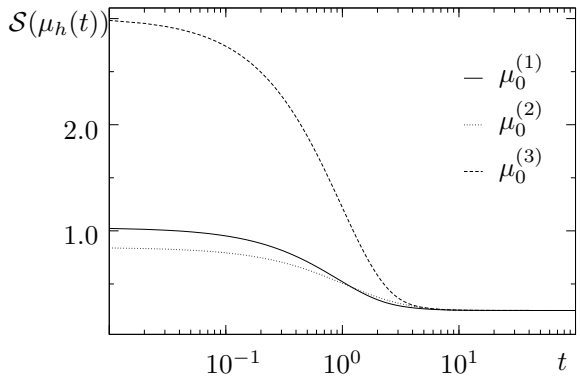

Figure 2.8: Time behavior of the Lyapunov-candidate functional $\mathcal{L}(\mu_h^k)$.

2.5.2.2 Test Case 2: comparison with literature

Homogeneous case. In this section we compare our algorithm to the results obtained by [5] using a finite element method to solve a regularized mixed formulation of the MK equations. Their discretization approach yields a highly nonlinear algebraic system that is solved using a modified successive over-relaxation method. They propose a set of test problems that we use here to practically show the characteristics of our method for the solution of the MK equations, and, at the same time, we highlight how the use of different spatial discretization methods affects the results. We address the first test case proposed by [5], which considers the transport of a uniform density supported on a circle towards a disjoint ellipses . Figure 2.9 (left) shows the domain Ω where problem Equation (1.10) is defined and the supports Q^+ and Q^- of the forcing term $f = f^+ - f^-$, with $f^+(x) = 2$ for $x \in Q^+$ and zero otherwise, and $f^-(y)$, appropriately rescaled for $y \in Q^-$ to ensure mass balance. The coarse initial mesh is also shown in light blue lines, and its uniform refinement is shown in thin dashed lines. This mesh, characterized by 820 nodes and 1531 triangles, is a constrained Delaunay triangulation that follows the boundaries of both Q^+ and Q^-. The forcing function is adapted to this triangulation to enforce the compatibility condition of zero mean. The same Figure shows in the right panel the time-converged spatial distribution of the transport density numerically evaluated with the most stable discretization method, $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$, on the finest mesh, which is used as a reference solution. The spatial distribution of μ_h is in good agreement with the results obtained by [5], achieving its maximum value (0.482) on the boundary of the left circle,

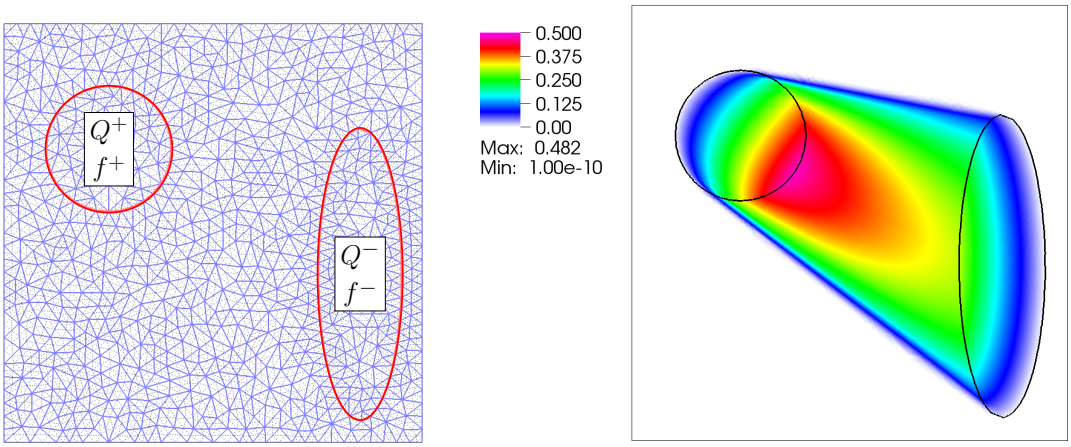

Figure 2.9: Domain and supports of the forcing function used in the discretization of the problem in Equation (1.10) for the solution of Example 1 of [5]. The two triangulations $\mathcal{T}_h$ and $\mathcal{T}_{h/2}$ are also shown with blue and dashed black lines, respectively.

and its minimum value 10^{-10} set by the prescribed lower bound as discussed in Section 2.5.1.3. In this section we use this sample test to experimentally discuss the need to use different FEM spaces for the discretization of the transport density and of the transport potential.

We start this discussion by presenting the results obtained using the $\mathcal{P}_{1,h} - \mathcal{P}_{0,h}$ approach on the coarsest grid and look at three different times during the evolution. The times are selected so that $\mathrm{var}(\mu_h^k)$ reaches the values 10^{-3}, 10^{-4}, 5×10^{-8}, namely $t_1 = 2.73 \times 10^2$, $t_2 = 1.36 \times 10^3$, and $t_3 = 2.5 \times 10^5$, the latter time corresponding to the time-converged solution. We plot in Figure 2.10 both μ_h (upper panels) and $|\nabla u_h|$ (lower panels).

NewP At the first time the solution clearly resembles the reference solution shown in Figure 2.9 (right), although at a much coarser resolution. Correspondingly, the gradient in the second row displays some slight but acceptable overshoots in a region that resembles Q^μ. Already at this early time, which occurs after 1630 time steps, some oscillations are visible.

NewP At time t_2 these oscillations are much more pronounced with a checkerboard pattern that suggests an intrinsic instability of the scheme. We should note that the color scale in the plots are limited above and below by suitable values that emphasize the oscillations. The maximum and minimum values for both μ_h and $|\nabla u_h|$ are reported right below each legend. We observe that there

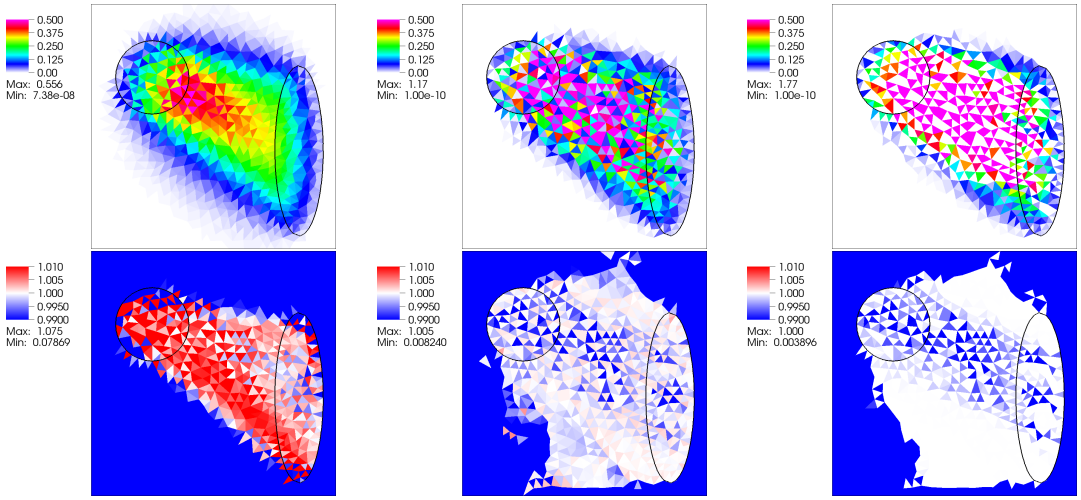

Figure 2.10: Solution of Test Case 2 at $t_1 = 2.73 \times 10^2$ (left), $t_2 = 1.36 \times 10^3$(center) , and $t_3 = 2.5 \times 10^5$(right), using the $\mathcal{P}_{1,h} - \mathcal{P}_{0,h}$ approach. Top row: spatial distribution of $\mu_h \in \mathcal{P}_{0,h}$. Bottom row: spatial distribution of $|\nabla u_h| \in \mathcal{P}_{0,h}$ as calculated from $u_h \in \mathcal{P}_{1,h}$.

is no overshoot in $|\nabla u_h|$, which at the final time is never greater than one. Still, checkerboard-like oscillations are visible. These fluctuations cause the dynamic equation to drive μ_h to zero quickly, causing a drastic deterioration of the solution accuracy.

The situation does not improve by using higher order spaces for μ_h. Figure 2.11 shows the results obtained by using a $\mathcal{P}_{1,h} - \mathcal{P}_{1,h}$ approach. We still observe oscillations, albeit appearing at a later time and with a different pattern. Once oscillations in $|\nabla u_h|$ around the unit value start developing the dynamic equation determines a decay of μ_h within the elements where $|\nabla u_h| < 1$ even if located within Q^μ. This decay quickly reinforces in time leading to the observed checkerboard pattern. This behavior resembles the classical lack of stability due to a violation of an inf-sup-like constraint, but at this point we are not able to identify this condition.

On the other hand, oscillations completely disappear if we employ a two-mesh approach. Looking at the checkerboard oscillations displayed in Figure 2.10, it is intuitive to think that averaging the gradient magnitude between neighboring triangles should compensate the fluctuations. This observation led us to employ the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ discretization described in Section 2.5.1. Indeed, with this approach the gradients calculated from $u_h \in \mathcal{P}_1(\mathcal{T}_{h/2})$ are projected onto the space

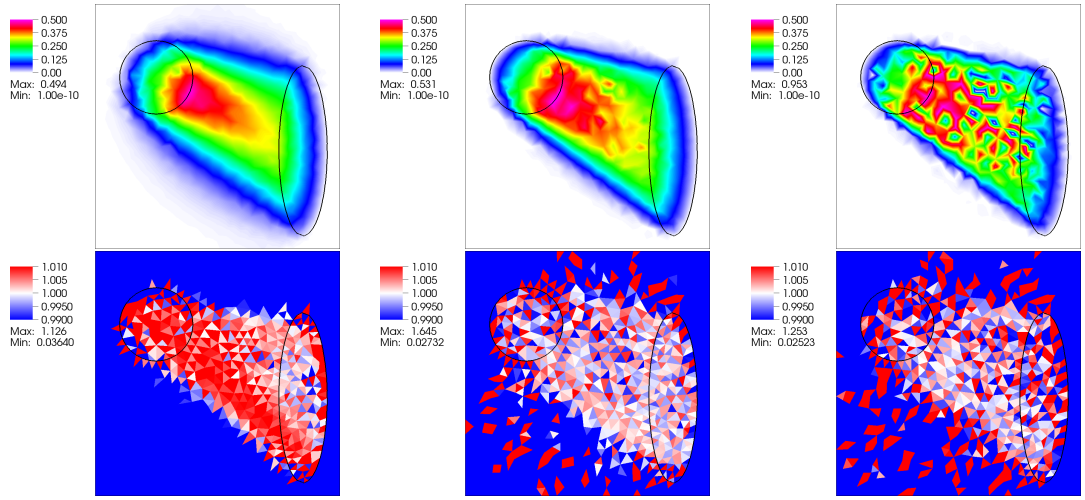

Figure 2.11: Solution of Test Case 2 at $t_1 = 6.82^1$ (left), $t_2 = 5.36 \times 10^2$ (center), and $t_3 = 9.65 \times 10^3$, (left), using the $\mathcal{P}_{1,h} - \mathcal{P}_{1,h}$ discretization. Top row: spatial distribution of $\mu_h \in \mathcal{P}_{1,h}$. Bottom row: spatial distribution of $|\nabla u_h| \in \mathcal{P}_{0,h}$ calculated from $u_h \in \mathcal{P}_{1,h}$.

$\mathcal{P}_0(\mathcal{T}_h)$ for insertion into the dynamic equation Equation (2.51b). This projection is equivalent to averaging the piecewise constant gradients over the four triangles of $\mathcal{T}_{h/2}$ that form one triangle of $\mathcal{T}_h$. This results in a oscillation free μ_h field, as shown in Figure 2.12. It is evident that no μ_h oscillations form even at the coarsest mesh level used in this test. Note that the spatial discretization of the elliptic equation does not guarantee monotonicity [57]. In fact, the gradient magnitudes arising from $u_h \in \mathcal{P}_{1,h/2}$ still show the classical checkerboard fluctuations (Figure 2.12, middle row). However, the projection of $|\nabla u_h|$ onto $\mathcal{P}_{0,h}$ (Figure 2.12, bottom row) does not show oscillations, albeit small overshooting occurs especially at the earlier times. We should emphasize that $|\nabla u_h|$ is plotted here using an extremely narrow color scale ranging within $[0.9999, 1.0001]$.

Looking at the final time-converged solution, the value $|\nabla u_h^*|$ within the support of μ_h^* and neighboring regions is exactly unitary, and remains bounded by 1 almost everywhere, in compliance with the constraint of the MK equations. Only one small region with $|\nabla u_h^*| > 1$ develop with a maximum value approaching 1.00009, considered consistent with the tolerance used in the PCG linear solve and the lower bound set on μ_h that cannot be smaller than 10^{-10}. Indeed, oscillations of the order of 10^{-5} in the gradient magnitude may be indistinguishable by the linear solver of the u_h equation.

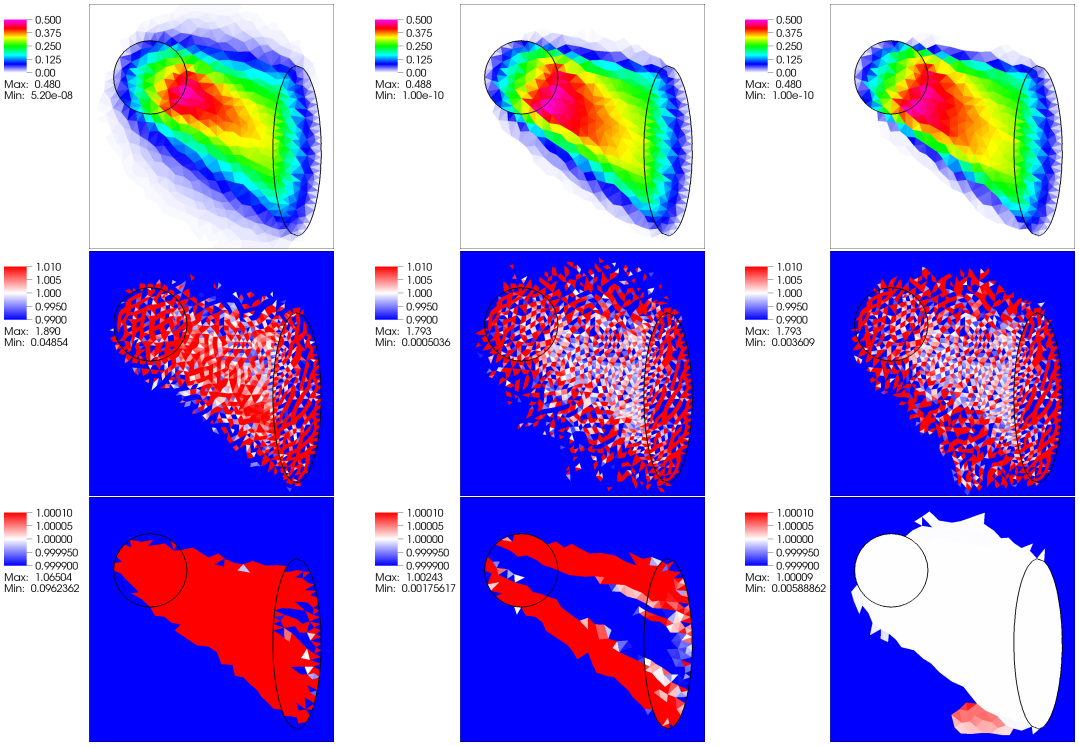

Figure 2.12: Solution of Test Case 2 at $t_1 = 6.75 \times 10^1$ (left), $t_2 = 2.04 \times 10^2$ (center), and $t_3 = 1.54 \times 10^3$ (right) using the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ discretization. Top row: spatial distribution of $\mu_h \in \mathcal{P}_{0,h}$. Middle row: spatial distribution of $|\nabla u_h| \in \mathcal{P}_{0,h/2}$ calculated from $u_h \in \mathcal{P}_{1,h/2}$. Bottom row: spatial distribution of $|\nabla u_h| \in \mathcal{P}_{0,h}$.

Similar considerations can be done in the case $\mu_h \in \mathcal{P}_{1,h}$ (not shown here) but in this case some oscillations in $|\nabla u_h|$ persist even when $u_h \in \mathcal{P}_{1,h/2}$. This reinforces the conjecture that some sort of inf-sup stability condition exists that couples the discretization spaces for u_h and μ_h, and will be the subject of further studies.

One final observation for this test case concerns the computational cost of our approach. In comparison with the technique proposed by [5], our method seems to be computationally advantageous. In fact, as already mentioned, the simulations reported in [5] where obtained using a mixed FEM approach in combination with adaptive mesh refinement, leading to nonlinear systems of dimension approaching 60000. In our case, the dimensions for the smallest test case are 1531 (number of triangles in $\mathcal{T}_h$) and 3170 (number of nodes in $\mathcal{T}_{h/2}$) for the diagonal dynamic algebraic system and the elliptic system, respectively, leading to a total of 4701 degrees of freedom. Note that the finest solution of Figure 2.9 was obtained with a total of 73917 degrees of freedom. Our confidence that the approach we propose is superior to that [5] is is reinforced by the observation that effective simulations can be obtained at intermediate mesh levels. Moreover, time-convergence can be considered achieved at much earlier times then the ones employed in this section if we look at the stationarity of the Lyapunov-candidate functional. Obviously, adding simple adaptive mesh refinement strategies would greatly enhance the performance of the studied methodology.

2.5.3 Heterogeneous $k(x)$

The numerical method described in Section 2.5.1 can be easily adapted to approximate Equation (2.48) with $k(x) \not\equiv 1$. The numerical results for the heterogeneous case are shown in Fig. 2.13 where the steady-state spatial distribution of the flux magnitude $|q_h| = \mu_h|\nabla u|$ is plotted in the case of $k_e = 0.01$ (top panels) and $k_e = 100$ (bottom panels). We first note that in this heterogeneous case the gradient is bounded by $k(x)$ and not by one as in the previous test case. For this reason we chose to to plot the flux magnitude $|q_h| = \mu_h|\nabla u_h|$ instead of μ_h. Two successively refined triangulations are used, leading to linear systems of dimensions $N_h + M_{2h} = 3603 + 1738$ and $N_h + M_{2h} = 14157 + 6952$ for the coarser and the finer meshes, respectively. The results are qualitatively comparable with those of [5], although obtained with a much coarser discretization. It is interest-

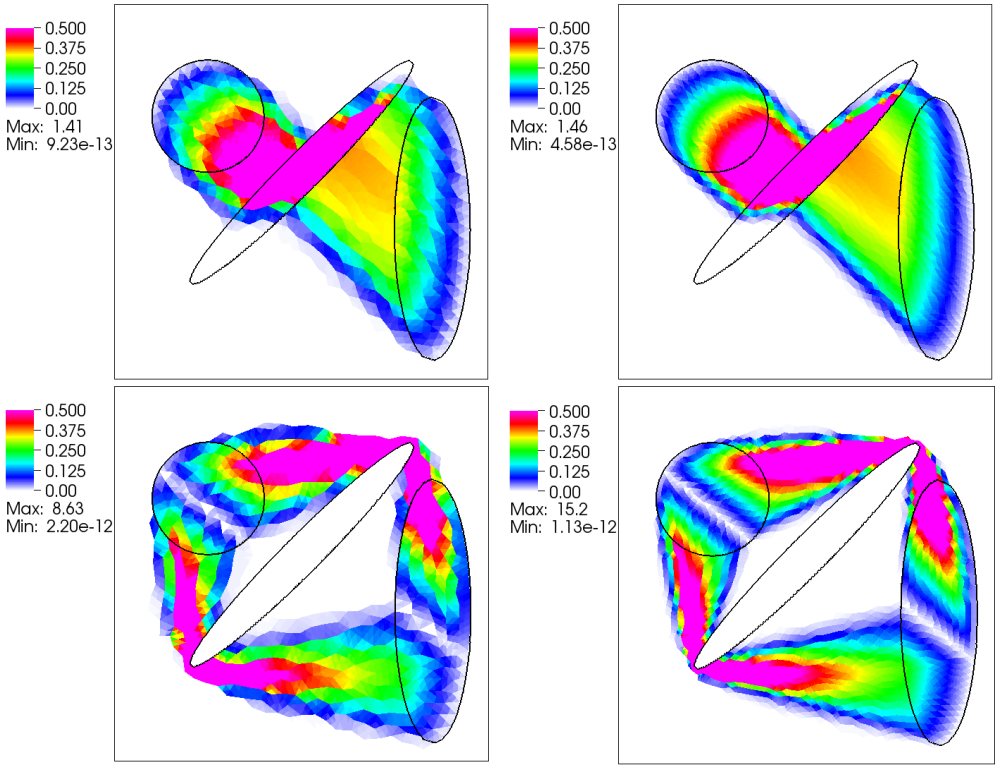

Figure 2.13: Numerical solution of the heterogeneous test case for $k_e = 0.01$ (top) and $k_e = 100$ (bottom) in the central ellipse. The figures show the optimal flux magnitude $|q_h| = \mu_h|\nabla u_h|$ for the mesh with 1738(6952) triangles and 933(3603) (left) and the once-refined mesh 6952(27808) triangles and 3603(14157) nodes (right).

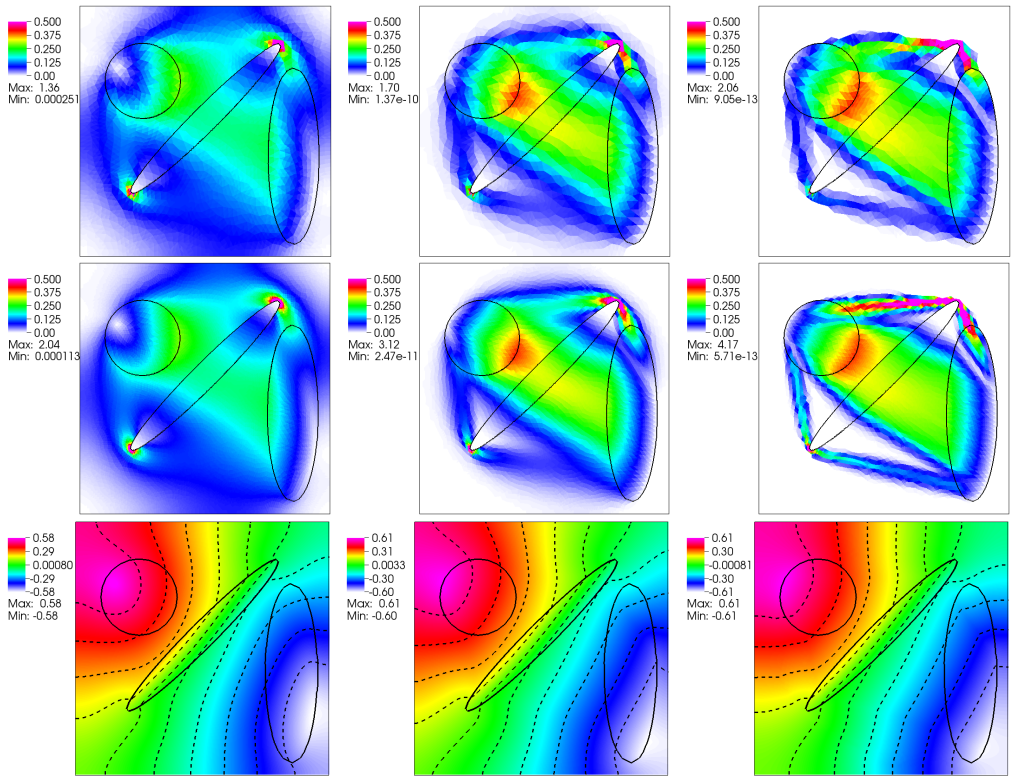

Figure 2.14: Numerical solution of the heterogeneous test case with $k_e = 3$ in the central ellipse in terms of OT flux magnitude $|q_h| = \mu_h |\nabla u_h|$ for 3 different times ($\mathrm{var}(\mu_h) = 0.1, 0.01, 5 \times 10^{-9}$ from left to right) and 2 refinement levels (from top to bottom). On the lower panels we report u_h at three times considered for the finest mesh. The left circle and the right ellipse represent the support of f^+ and f^-, respectively. The central ellipse represents the portion of the domain where $k(x) = 3$.

ing to note that the qualitative features of the solution are obtained already at the coarser mesh, with no visible numerical artifacts barring mesh roughness. We would like to stress here the fact that, notwithstanding the fact that the mesh nodes are not aligned with the support of μ_h, the geometrical features of the solution are well captured at all mesh resolution levels. From the spatial distribution of the flux magnitude, we see that values of k_e lower than one promotes larger fluxes across the central ellipses. On the contrary, values substantially larger than one restricts through-flow, and promotes the circumnavigation of the high low conductivity areas.

Finally, Fig. 2.14 shows the distribution of $|q_h| = \mu_h |\nabla u_h|$ for the case of

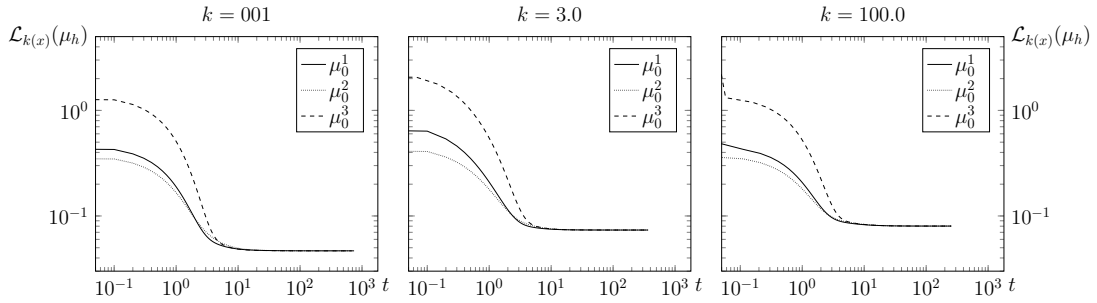

Figure 2.15: Time evolution for $\mathcal{L}_{k(x)}(\mu_h)$ for with different initial data μ_0. The values $k(x) = 0.01, 3, 100$ are imposed inside the central ellipse reported in Figure 2.13.

$k_e = 3$ at three different times (left to right) and two successive refinement levels (top to bottom). The three times are chosen so that the μ_h variation reaches the thresholds $\text{var}(\mu_h(\hat{t}_1)) = 0.1$, $\text{var}(\mu_h(\hat{t}_2)) = 0.01$, and $\text{var}(\mu_h(\hat{t}_3)) = 5 \times 10^{-9}$. Correspondingly, we have $\hat{t}_1 \approx 5.2$ and $\hat{t}_2 \approx 21$, remaining the same for both tested triangulations, and $\hat{t}_3 = 1600$ for the coarser level and $\hat{t}_3 = 2200$ for the finer mesh as steady state is achieved at a later time for the finer mesh, reflecting the fact that the overall error is smaller. In fact, the converged steady state solution occurs after 6616 and 8955 time steps for the coarse and fine triangulations, respectively. Note that, for this last heterogeneous test case, the time-stepping sequence employed an upper bound on Δt_k equal to 0.25. Also in this case the steady-state numerical solution is similar to the results reported by [5]. We see from the time sequence that our model constructs the transport map gradually. Starting from the uniformly distributed initial condition, it first identifies the larger flow paths and then refines them to arrive at the final configuration. The last numerical results we report is the time behavior of the Lyapunov-candidate functional $\mathcal{L}_{k(x)}$, defined in Equation (2.49). As reported in Figure 2.15 the numerical simulations confirm that $\mathcal{L}_{k(x)}(\mu_h(t))$ is the strict decreasing and it converges to the same asymptotic value for different initial configurations of μ_0. The overall results are consistently pointing towards the veracity of the conjecture that the infinite-time solution to our problem indeed coincides with the solution of the Monge-Kantorovich equations in the support of the OTP path.

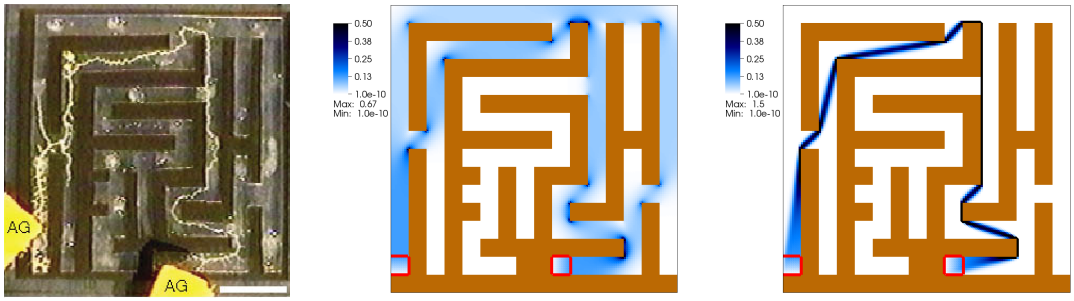

Figure 2.16: Simulation of the dynamics of PP mass reorganization in the maze experiment of [68, 54]: distribution of PP transport density at dimensionless times $t = 60.3$ (central panel), and $t = 9.6 \times 10^3$ (right panel), compared to the experimental distribution shown on the left (from [54], reprinted with permission). The simulation was done on a triangulation $\mathcal{T}_h$ with 32768 triangles and 16641 nodes. At the last time step of the simulation the μ_h variation was smaller than $\tau_{\mathrm{T}} = 5 \times 10^{-9}$.

2.5.3.1 Numerical simulation of PP dynamics

The proposed model and its numerical discretization just described can deal with function $k(x)$ describing very complex geometry. We applied the method to the simulation of the dynamics of the experiment described in [68, 54]. The domain encompassing the entire maze setup is discretized by means of a uniform triangulation obtained by subdividing each edge of the square-shaped maze into 128 subdivisions yielding the coarser mesh $\mathcal{T}_h$ comprised of 32768 elements and 16641 nodes. This high resolution is required to follow accurately the walls of the maze, which are described by setting $\kappa(x) = 1000$ (brown colors in the upper left panel of Fig. 2.16) while the maze paths are characterized by $\kappa(x) = 1$. These values are calibrated experimentally to enforce that no measurableflux through the maze walls occurs. The initial condition μ_0 is set to 10^{-10} on the maze walls and one elsewhere. The two food sources $f^+ = 1$ and $f^- = -1$ are shown as red squares in the figure. We employ a variable time step size starting from $\Delta t_0 = 10^{-2}$ and with $\Delta t_{j+1} = \min(1.01\Delta t_k, 0.5)$, to ensure stability of the Forward Euler scheme is verified for all times with an ample safety margin. The simulation is stopped when $\mathrm{var}(\mu_h) \leq 5 \times 10^{-9}$. Again, this tolerance is exceedingly small, and could be replaced with no noticeable changes by a tolerance several orders of magnitude larger, We use low tolerance to verify the robustness of

our approach. Fig. 2.16 shows the distribution of μ_h at two different times, chosen in agreement with the simulations reported in [68, 54]. These times are useful to identify the intermediate phase when the P.Polycephalum starts retreating from the dead ends ($t = 60.3$) and the final steady state configuration achieved at $t = 9.6\times10^3$. Note that the same numerical solution is obtained, albeit at different dimensionless times, starting from different initial conditions μ_0. The central panel shows the intermediate phase when PP completely retreated from the dead end paths of the maze but persists on all the possible paths connecting the two food sources. We note a stronger concentration of μ_h at the edges of the maze walls, indicating that PP starts accumulating around a narrow band along the shortest route. At the final time (right panel), μ_h is distributed along the optimal route displaying varying approximation levels depending on the alignment of the mesh triangles with the support of μ_h. In fact, the vertical portion is one element thick, while the oblique routes encompass more than one triangle. This is a common feature that is reproduced experimentally at all mesh refinement levels. All these observations are in line with the results proposed by [68], although in our case the graph structure is not imposed a-priori but it is mimicked through the appropriate definition of $k(x)$. Note that in our continuous setting the presence of $k(x)$ is related to the cost of through-flow, while the original graph-based formulation allows only flow through the graph edges.

Chapter 3

Extension of the DMK equations. An unexpected branching source

In this chapter we discuss a simple extension of the DMK model described in Chapter 2 that leads to solutions that resembles the congested and the the branched transport discussed in Section 1.5. This extension, already suggested in [68], considers a dynamics where by the time derivative of the transport density grows non-linearly with the transport flux. A sub-linear growth penalizes the flux intensity (i.e. the transport density) and promotes distributed transport. Corresponding solutions are reminiscent of the CTP described in Section 1.6. Indeed, for this case we are able to relate our proposed dynamics with the p-Poisson equations, typical of CTP. On the contrary, a super-linear growth favors flux intensity and promotes a transport that occurs preferentially in a concentrated manner, leading to "singular" and "fractal-like" solutions that resemble those of the BTP (Problem 24). In this case we are able to formally derive a Lyapunov-candidate functional that resembles the functional $\mathcal{L}$ defined in Chapter 2, albeit we remain in Lebesgue-measure setting and we are unable to formally address concentrated measures. Hence, using the necessary caution, we conjecture that the proposed extend DMK is a version of branched transport, and again its numerical is highly efficient and, using ad-hoc linear solver strategies, we are able to obtain highly promising results.

The extended DMK modifies the dynamical equation for $\mu(t)$ raising the flux

$|\mu(t)\,\nabla\, u(t)|$ to the power $\beta > 0$, thus leading to the following system

$$-\operatorname{div}(\mu(t)\,\nabla\, u(t)) = f(x) = f^{+}(x) - f^{-}(x) \tag{3.1a}$$

$$\partial_t \mu(t) = [\mu(t)|\,\nabla\, u(t)|]^{\beta} - \mu(t) \tag{3.1b}$$

$$\mu(0) = \mu_0(x) > 0 \tag{3.1c}$$

complemented with zero Neumann boundary conditions. Assuming existence and uniqueness of a solution of the above equations, we claim that the solution pair $(\mu(t), u(t))$ converges toward a steady state configuration (μ^*, u^*), as in the case $\beta = 1$. We find that the behavior of our model changes drastically for the case for the case $0 < \beta < 1$ and the case $\beta > 1$. For the case $0 < \beta < 1$, we claim that for $0 < \beta < 1$ the pair $(\mu(t), u(t))$ tends to $(|\,\nabla\, u_p|^{p-2}, u_p)$, with

$$p = \frac{2-\beta}{1-\beta}$$

where u_p is the solution of the p-Poisson equation with forcing term f. Note that analogously to the case $\beta = 1$, this new formulation of the p-Poisson equation leads to very efficient and accurate numerical solutions schemes.

In the case $\beta > 1$ the numerical evidence suggests a connection between the steady state (μ^*, u^*) and the BTP solution described in Section 1.5.Although, in this case we are still not able to exactly identify the relations, formal calculations supported by several numerical results, are used to derive these relationships that suggests possible connections to a far-fetching "negative p-Laplacian". In the next sections we will described these formal arguments and the implications and the intuitions that are drawn by these calculations. Then we report the numerical results that back-up our claims, and some future research suggested by the numerical indications. For this case, the numerical simulations show a strong dependence of the steady state configuration on the initial data μ_0. We conjecture that this is due to the non-convexity of the Lyapunov-candidate functional, which causes the numerical algorithm to stall in local minima configurations.

3.1 Lyapunov-candidate functional

In this section we present the formal derivation of the Lyapunov-candidate functional for all $\beta > 0$. The local existence result presented in Section 2.2 can be, a

priori, extended to the case $\beta > 1$, but numerical results and theoretical considerations suggest that the assumption of Hölder-continuous μ is too strong. Nevertheless, since we are mostly concerned on the asymptotic behavior of $(\mu(t), u(t))$ solution of Equation (3.1), we assume existence and uniqueness of a solution pair for all $t \geq 0$.

The most important result is that, similarly to the case $\beta = 1$, we can identify a Lyapunov-candidate functional given by

$$\mathcal{L}_\beta(\mu) := \mathcal{E}_f(\mu) + \mathcal{M}_\beta(\mu) \tag{3.2}$$

$$\mathcal{E}_f(\mu) := \frac{1}{2}\int_\Omega \mu |\nabla u(\mu)|^2 \, dx \qquad \mathcal{M}_\beta(\mu) := \begin{cases} \frac{1}{2}\int_\Omega \ln(\mu) & \text{if } \beta = 2 \\ \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{2-\beta}{\beta}} & \text{otherwise} \end{cases}$$

Assuming that the formal computations done to obtain the time derivative of the term $\mathcal{E}_f(\mu(t))$ can be reproposed also for the case $\beta \neq 1$, we can state the following proposition

Proposition 42. *Assume that a solution pair $(\mu(t), u(t))$ of Equation (3.1) exists and is $\mathcal{C}^1$-regular in time. Then the derivative along the $\mu(t)$ trajectory of functional $\mathcal{L}_\beta$ is given by*

$$\frac{d}{dt}\left(\mathcal{L}_\beta(\mu(t))\right) = \tag{3.3}$$
$$-\frac{1}{2}\int_\Omega \mu(t)^\beta \left(|\nabla u(\mu(t))|^\beta - \mu^{(\frac{1-\beta}{\beta})\beta}(t)\right)\left(|\nabla u(\mu(t))|^2 - \left(\mu^{\frac{1-\beta}{\beta}}(t)\right)^2\right) dx$$

Proof.

$$\begin{aligned}
\frac{d}{dt}\left(\mathcal{L}(\mu(t))\right) &= -\frac{1}{2}\int_\Omega \partial_t\mu(t)\left(|\nabla u(t)|^2 - \mu^{\frac{2-\beta}{\beta}-1}\right) dx \\
&= -\frac{1}{2}\int_\Omega \left([\mu(t)|\nabla u(t)|]^\beta - \mu(t)\right)\left(|\nabla u(t)|^2 - \mu^{\frac{2-\beta}{\beta}-1}\right) dx \\
&= -\frac{1}{2}\int_\Omega \mu(t)^\beta \left(|\nabla u(t)|^\beta - \mu^{1-\beta}(t)\right)\left(|\nabla u(t)|^2 - \mu^{2\frac{1-\beta}{\beta}}\right) dx \\
&= -\frac{1}{2}\int_\Omega \mu(t)^\beta \left(|\nabla u(t)|^\beta - \mu^{\frac{1-\beta}{\beta}\beta}(t)\right)\left(|\nabla u(t)|^2 - \left(\mu^{\frac{1-\beta}{\beta}}\right)^2\right) dx
\end{aligned}$$

Setting

$$\left(g_1(t,x) = |\nabla u(t,x)| \quad g_2(t,x) = \mu^{\frac{1-\beta}{\beta}}(t,x)\right)$$

we can write:

$$\frac{d}{dt}\left(\mathcal{L}_\beta(\mu(t))\right) = -\frac{1}{2}\int_\Omega \mu(t)^\beta \left(g_1(t,x)^\beta - g_2(t,x)^\beta\right)\left(g_1(t,x)^2 - g_2(t,x)^2\right) dx < 0$$

where we introduced the functions g_1 and g_2 to make clear that the last equation is strictly negative, since $\left(g_1(t,x)^\beta - g_2(t,x)^\beta\right)$ and $(g_1(t,x)^2 - g_2(t,x)^2)$ have the same sign.

□

Equation (3.3) shows that $\mathcal{L}_\beta(\mu(t))$ is strictly decreasing in time and we can also deduce, at least formally, that its derivative is equal to zero if the following equations are satisfied

$$\begin{cases} -\operatorname{div}\left(\mu^* \nabla u^*\right) = f^+ - f^- \\ \mu^* = |\nabla u^*|^{\frac{\beta}{1-\beta}} \quad \text{on} \quad \{\mu^* > 0\} \end{cases} \tag{3.4}$$

It is clear that Equation (3.4) is equivalent to $\partial_t \mu(t) = 0$ in Equation (3.1b). Moreover Equation (3.4) immediately suggests a link between the large-time equilibrium state of Equation (3.1) and the p-Poisson equation

$$-\operatorname{div}\left(|\nabla u_p|^{p-2}\nabla u_p\right) = f^+ - f^-$$

if the following relation between the exponents β and p holds

$$p - 2 = \frac{\beta}{1-\beta}$$

3.2 Case $0 < \beta < 1$

The informal equivalence between the steady state version of Equation (3.1) and the p-Poisson equation, together with the decrease in time of the functional $\mathcal{L}_\beta$, suggests to investigate if the Lyapunov-candidate functional $\mathcal{L}_\beta$ admits a minimum and if this minimum is related to the p-Poisson equations. This intuitive idea is confirmed by the following generalization of Proposition 39 to the case $0 < \beta < 1$.

Proposition 43. *Let* $0 < \beta < 1$, $q = 2 - \beta$ *and* $P(\beta) = \frac{2-\beta}{\beta}$. *Then the following equality holds*

$$\inf_{\mu \in L_+^{P(\beta)}(\Omega)} \mathcal{L}_\beta(\mu) = \inf_{v \in [L^q(\Omega)]^d} \left\{ \int_\Omega \frac{|v|^q}{q}\, dx \ : \ \operatorname{div}(v) = f \right\} \tag{3.5}$$

Moreover, the functional $\mathcal{L}_\beta$ admits a unique minimizer $\mu_\beta^* \in L_+^{\frac{2-\beta}{\beta}}(\Omega)$ given by

$$\mu_\beta^* = |\nabla u_p|^{p-2}$$

where u_p is the solution of p-Poisson equation

$$-\operatorname{div}(|\nabla u_p|^{p-2}\nabla u_p) = f$$

with p the conjugate exponent of q:

$$p = \frac{2-\beta}{1-\beta}$$

Proof. The proof is an adaptation of the arguments used to prove Proposition 39. Equation (3.5) can be shown as follows. First, by Lemma 40, we can rewrite $\mathcal{L}_\beta(\mu)$ as follows

$$\mathcal{L}_\beta(\mu) = \inf_{\xi\in[L^2_\mu(\Omega)]^d}\{\Theta(\mu,\xi) \ : \operatorname{div}(\xi\mu) = f\} \quad \forall \mu \in L_+^{P(\beta)}(\Omega) \tag{3.6}$$

where

$$\Theta_\beta(\mu,\xi) := \frac{1}{2}\int_\Omega |\xi|^2\mu\,dx + \frac{1}{2}\int_\Omega \mu^{\frac{2-\beta}{\beta}}\,dx.$$

Now, by using Young/Hölder inequality with conjugated exponents $2/q$ and $\frac{2}{2-q}$, we obtain the following inequality

$$\int_\Omega |\xi\mu|^q\,dx = \int_\Omega |\xi|^q\mu^{q/2}\mu^{q/2}\,dx \le \frac{q}{2}\int_\Omega |\xi|^2\mu\,dx + \frac{2-q}{2}\int_\Omega (\mu^{\frac{q}{2}})^{\frac{2}{2-q}}\,dx,$$

which holds for any pair $(\mu,\xi) \in \left(L_+^{P(\beta)}(\Omega), [L^2_\mu(\Omega)]^d\right)$. Imposing $\frac{q}{2-q} = \frac{2-\beta}{\beta}$ (which gives the relation $q = 2-\beta$) and dividing by q, yields

$$\int_\Omega \frac{|\xi\mu|^{(2-\beta)}}{(2-\beta)}\,dx \le \frac{1}{2}\int_\Omega |\xi|^2\mu\,dx + \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{2-\beta}{\beta}}\,dx = \Theta_\beta(\mu,\xi),$$

which holds for all $\mu \in L_+^{P(\beta)}(\Omega)$ and all $\xi \in [L^2_\mu(\Omega)]^d$. Now we take the infimum over the μ-divergence constrained $\xi \in [L^2_\mu(\Omega)]^d$ on both side of the previous inequality, and use Equation (3.6) to obtain :

$$\inf_{\xi\in[L^2_\mu(\Omega)]^d}\left\{\int_\Omega \frac{|\xi\mu|^{(2-\beta)}}{(2-\beta)}\,dx \ : \operatorname{div}(\xi\mu) = f\right\} \le \mathcal{L}_\beta(\mu) \quad \forall \mu \in L_+^{P(\beta)}(\Omega).$$

Now, taking the infimum over all $\mu \in L_+^{P(\beta)}(\Omega)$ yields:

$$\inf_{\mu\in L_+^{P(\beta)}(\Omega)}\left\{\inf_{\xi\in[L^2_\mu(\Omega)]^d}\left\{\int_\Omega \frac{|\xi\mu|^{(2-\beta)}}{(2-\beta)}\,dx \ : \operatorname{div}(\xi\mu) = f\right\}\right\} \le \inf_{\mu\in L_+^{P(\beta)}(\Omega)}\mathcal{L}_\beta(\mu). \tag{3.7}$$

According to Proposition 26 we have that for any $q > 1$

$$\operatorname*{argmin}_{v\in[L^q(\Omega)]^d}\left\{\int_\Omega \frac{|v|^q}{q}\,dx \ : \operatorname{div}(v) = f\right\} = v^* = -|\nabla u_p|^{p-2}\nabla u_p,$$

where p is the conjugate exponent of q. Now considering $q = 2-\beta$ (and thus $p = (2-\beta)(1-\beta)$) the following chain of equalities and inequalities holds

$$\begin{aligned}
&\int_\Omega \frac{|\nabla u_p|^p}{2-\beta}\,dx = \int_\Omega \frac{|v^*|^{(2-\beta)}}{(2-\beta)}\,dx \\
&= \inf_{v\in[L^{(2-\beta)}(\Omega)]^d}\left\{\int_\Omega \frac{|v|^{(2-\beta)}}{(2-\beta)}\,dx \ : \operatorname{div}(v) = f\right\} \qquad (3.8)\\
&\le \inf_{\mu\in L_+^{P(\beta)}(\Omega)}\left\{\inf_{\xi\in[L^2_\mu(\Omega)]^d}\left\{\int_\Omega \frac{|\xi\mu|^{(2-\beta)}}{(2-\beta)}\,dx \ : \operatorname{div}(\xi\mu) = f\right\}\right\}
\end{aligned}$$

by Equation (3.7) we obtain

$$\begin{aligned}
&\le \inf\left\{\mathcal{L}_\beta(\mu) \ : \ \mu \in L_+^{P(\beta)}(\Omega)\right\} \\
&\le \mathcal{L}(|\nabla u_p|^{p-2}) = \mathcal{E}_f(|\nabla u_p|^{p-2}) + \mathcal{M}_\beta(|\nabla u_p|^{p-2}) \\
&= \int_\Omega \frac{|\nabla u_p|^p}{2}\,dx + \frac{1}{2}\int_\Omega \frac{|\nabla u_p|^{(p-2)\frac{2-\beta}{\beta}}}{\frac{2-\beta}{\beta}}\,dx \qquad (3.9)\\
&= \int_\Omega \frac{|\nabla u_p|^p}{(2-\beta)}\,dx
\end{aligned}$$

where in the last equality we used

$$\begin{aligned}
\int_\Omega \frac{|\nabla u_p|^p}{2}\,dx &= \int_\Omega f u_p - |\nabla u_p|^{p-2}\frac{|\nabla u_p|^2}{2}\,dx \\
&\le \sup_{V\in\mathcal{C}_1(\Omega)}\int_\Omega f\varphi - |\nabla u_p|^{p-2}\frac{|\nabla \varphi|^2}{2}\,dx \\
&= \mathcal{E}_f(\mu_p = |\nabla u_p|^{p-2}) \\
&= \inf_{\xi\in[L^2_{\mu_p}(\Omega)]^d}\left\{\int_\Omega \frac{|\xi|^2}{2}\mu_p\,dx \ : \operatorname{div}(\xi\mu_p) = f\right\} \\
&\le \int_\Omega \frac{|\nabla u_p|^p}{2}\,dx
\end{aligned}$$

Thus, all the inequalities in Equations (3.8) and (3.9) are actually equalities, which shows that μ_β^* is a unique minimum, with uniqueness following from the strict convexity of $\mathcal{L}_\beta$. In fact, $\mathcal{E}_f$ is convex, being the sup of functionals that are linear with respect to μ, and $\mathcal{M}_\beta$ is strictly convex for $0 < \beta < 1$. Thus the sum is strictly convex. This completes the proof. □

Propositions 43 and 45 suggest the following conjecture for the case $0 < \beta < 1$

Conjecture 2. *For $0 < \beta < 1$ the pair $(\mu(t), u(t))$ solution of Equation (3.1) converges to the pair $(|\nabla u_p|^{p-2}, u_p)$ where u_p is the solution of the p-Poisson equation with*

$$p = \frac{2-\beta}{1-\beta} \tag{3.10}$$

This holds for any initial data μ_0.

We want to highlight two remarkable facts regarding Proposition 43 and Conjecture 2 when $\beta \to 1$ and $\beta \to 0$. In the first case, according to Equation (3.10), when β tends to 1 we have that $p \to +\infty$, in good agreement with the fact that the MK equations are the limit of the p-Poisson equation as reported in Proposition 20. Similarly, the exponent $q = 2 - \beta$ tends to 1, which is coherent with the equivalent formulation of the MK equations described in Equation (1.9). Note that it is possible to include the case $\beta = 0$ in Conjecture 2 since in this case $\mu(t) \to 1$ and the system of equations converges to the classical Poisson equation ($p = 2$).

3.3 Case $\beta > 1$

In this section we discuss our attempts to extend the arguments presented in previous section to the case $\beta > 1$. We are particularly interested in understanding if functional $\mathcal{L}_\beta$ admits a minimizer and, if it does, if $\mu(t)$ is converging to this minimizer as $t \to \infty$. The first part of the proof of Proposition 43 partially answers these questions and it can be reproposed, at least for the case $1 < \beta < 2$, where, by the relations $q = 2 - \beta$, the exponent q remains positive. From Equations (3.8) and (3.9) we obtain:

$$\begin{aligned} &\inf_{v \in [L^q_+(\Omega)]^d} \left\{ \int_\Omega \frac{|v|^q}{q}\, dx \ : \operatorname{div}(v) = f \right\} \\ \leq &\inf_{\mu,\xi} \left\{ \int_\Omega \frac{|\xi\mu|^q}{p}\, dx \ : \begin{array}{l} (\mu,\xi) \in L^{P(\beta)}_+(\Omega) \times [L^2_\mu(\Omega)]^d \\ \operatorname{div}(\xi\mu) = f \end{array} \right\} \\ \leq &\inf_{\mu \in L^{P(\beta)}_+(\Omega)} \mathcal{L}_\beta(\mu) \end{aligned} \tag{3.11}$$

Unfortunately we are not able to identify a candidate minimum for $\mathcal{L}_\beta$ and state the analogous of Proposition 43 for $1 < \beta < 2$. Nevertheless, the minimization

problem in Equation (3.11) resembles the BTP formulated by Xia, described in Problem 24, with the exponent q playing the rule of the branch exponent α. However, we have to highlight an important difference between Equation (3.11) and Problem 24. In fact, in Xia's formulation the integrals are computed with respect to 1-dimensional Hausdorff measure, while in our computations we always used the Lebesgue measure.

Despite these differences, the numerical simulations presented in Chapter 3 suggest that system 3.1 admits a steady state solution (μ^*, u^*) where the supports of the numerical solutions μ_h^* seem to approximate the 1-dimensional structures related to the BTP.

Thus, we can state the following:

Conjecture 3. *For $\beta > 1$, the solution $(\mu(t), u(t)$ of Equation (3.1) admits an equilibrium point (μ_β^*, u_β^*), which depends on the initial condition μ_0. This solution is a minimum of the Lyapunov-candidate functional $\mathcal{L}_\beta$.*

Another informal argument supporting our claims is to look at Equation (3.4) describing the steady state of system Equation (3.1). Relation $p = (2-\beta)/(1-\beta)$ suggests that the steady state should solve a p-Poisson equation with a negative p exponent. The only reference found in the literature regarding the existence of solutions of this equation is contained in the BTP theory of Xia in[75], even if there the results are defined only on graphs. Note how the exponents $q = 2 - \beta$ and $p = (2 - \beta)/(1 - \beta)$ are one the conjugate of the other. In Figure 3.1 we summarize the behavior of exponents p and q with respect to β.

Remark 5. *One possible strategy that can be adopted to reconcile the incompatibility in the measure used for the integration, is inspired by the Modica-Mortola approach described in [63]. The main idea is to introduce a parameter $\varepsilon > 0$ in Equation (2.45) using the Young's inequality with ε*

$$ab \leq \frac{\varepsilon^p}{p} a^p + \frac{1}{\varepsilon^q q} a^q$$

in order to weight differently the terms $\mathcal{E}_f(\mu)$ and $\mathcal{M}_\beta(\mu)$ that form $\mathcal{L}_\beta(\mu)$. In Section 3.4.3.1 we will present some preliminary numerical results that go into this direction. We do not know at this moment if such procedure will introduce a new relation between the exponent β and α of the BTP.

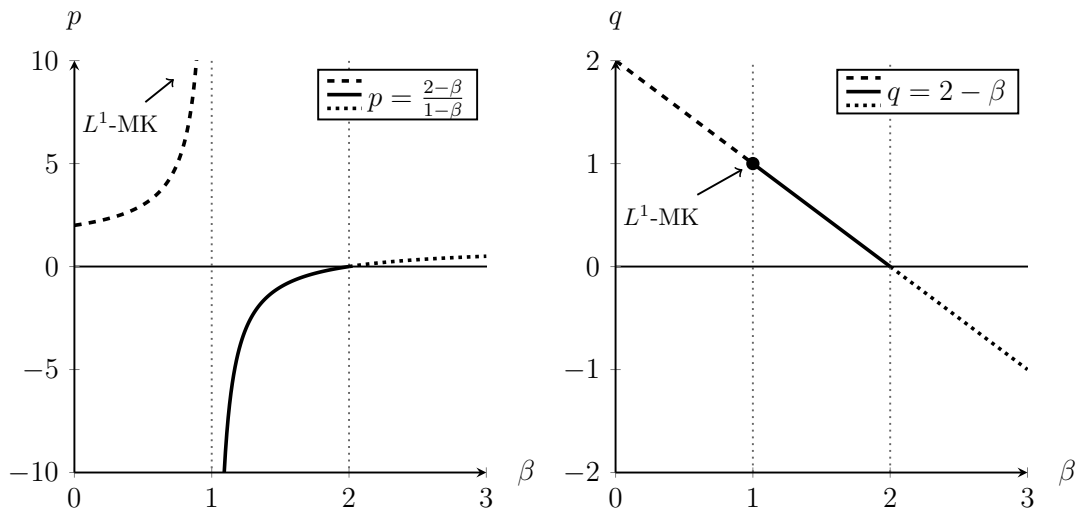

Figure 3.1: Values of p (left) given by Equation (3.10) and its conjugate $q = p'$ (right) with $0 \leq \beta < 3$. We divided each graph in three portions. The first portion (dashed line) restricted to the case $0 < \beta < 1$ represents the p and q values for which we claim (Conjecture 2) the equivalence with the p-Poisson equations. The second portion (sold line) corresponds to the interval $1 < \beta < 2$, for which we conjecture (and see numerically) a connection with the BTP, where the exponent q belongs to $]0, 1[$. The last part (dotted line), $\beta > 2$, is where even the heuristic analysis of the minimization problem associate to the Lyapunov-candidate functionalfails, with $q < 0$, even if in our numerical experiments we see branching structures appear. We also remark the case $\beta = 1$ which corresponds to the MK equations and to the p-Poisson equations for $p \to \infty$

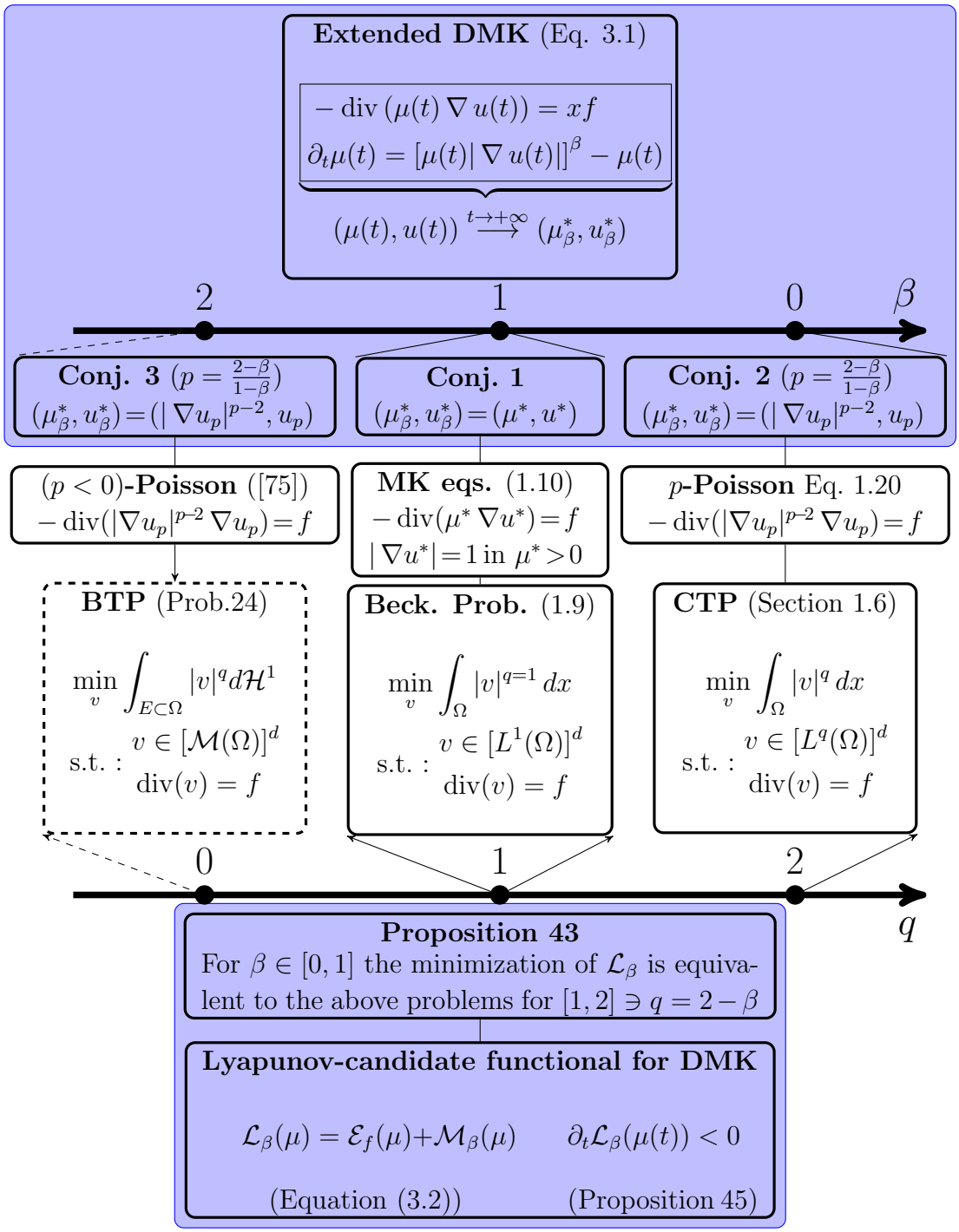

Figure 3.2: Schematic representation of the results and statements in this section. It is a revisited version of Figure 1.5 (blue box) where we highlight the connections between the extended DMK model and the transport problems written as minimization problems with divergence constraint.

3.4 Simulations of the Extended DMK equations

In this section we present numerical evidence in support of the conjectures that system 3.1 admits a steady state configuration for $t \to +\infty$ connected with the congested transport problem for $0 < \beta < 1$ and with the Branched Transport Problem for $\beta > 1$. The discretization method used is the combination of the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ spatial discretization with forward Euler time stepping described in Section 2.5.1. By using this scheme the necessary adjustments for $\beta \neq 1$ are minimal.

For $0 < \beta < 1$ we test Conjecture 2 in a 2-dimensional example, comparing our result with an exact solution u_p of the p-Poisson equation. All the simulations considered confirm our conjecture, showing that the numerical solution via the discretization of Equation (3.1) can be proposed also for the numerical solution of the p-Poisson equations. For the case $\beta > 1$ we present some examples of the equilibrium configurations for different powers β and different types of f^+ and f^-.

3.4.1 Numerical approach

We discretize Equation (3.1) with the $\mathcal{P}_{1,h/2} - \mathcal{P}_{0,h}$ strategy for the spatial discretization of (u, μ), combined with forward Euler scheme. Accordingly, the approximate pair $(\mu_h(t,x), u_h(t,x)$ can be written as

$$u_h(t,x) = \sum_{i=1}^{N} u_i(t)\varphi_i(x) \quad \varphi_i \in \mathcal{P}_1(\mathcal{T}_{h/2})$$
$$\mu_h(t,x) = \sum_{k=1}^{M} \mu_k(t)\psi_k(x) \quad \psi_k \in \mathcal{P}_0(\mathcal{T}_h)$$

Following the notation adopted in Section 2.5, the discretization scheme described above leads to the following sequence of linear systems

$$\boldsymbol{A}[\boldsymbol{\mu}^k]\boldsymbol{u}^k = \boldsymbol{b} \tag{3.12}$$
$$\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \Delta t_k \left[\boldsymbol{B}_\beta[\boldsymbol{u}^k] \left(\boldsymbol{\mu}^k\right)^\beta - \boldsymbol{\mu}^k\right] \tag{3.13}$$

where $\boldsymbol{A}[\boldsymbol{\mu}^k]$ is the stiffness matrix associate to $\boldsymbol{\mu}^k$ and $\boldsymbol{B}_\beta[\boldsymbol{u}^k]$ is matrix defining the norm of the gradient of $u_h(t^k, x)$ raised to the power β. Starting from the

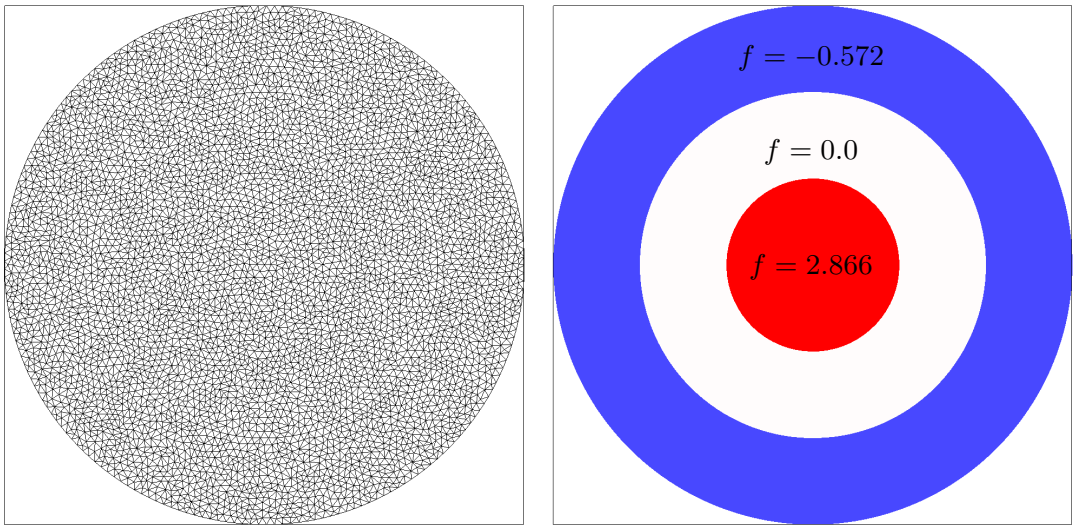

Figure 3.3: Triangulation $\mathcal{T}_h$ (with 5191 nodes and 10179 triangles) and the forcing term $f(x, y) = F(r)$. The mesh points follow the concentric circles that compose the boundary of the supports of f^+ and f^-. The right panel reports spatial distribution of f^+ and f^- on their supports.

projected initial data μ_h^0, we iterate until

$$\text{var}(\mu_h^k) := \frac{\|\mu_h^{k+1} - \mu_h^k\|_{L^2(\Omega)}}{\Delta t_k^k \|\mu_h^k\|_{L^2(\Omega)}})$$

is below a fixed threshold τ_{T}, in which case we assume a steady state configuration is reached. We impose a lower bound of 10^{-10} to $\mu_h(t, x)$ to avoid singularity of the linear system arising from the elliptic equation in Equation (3.1a).

The linear system in Equation (3.12) is solved via Preconditioned Conjugate Gradient (PCG). In Chapter 4 we will discuss this problem and the strategies devised for the efficient solution of the sequence of linear systems given in Equations (3.12) and (3.13), in particular in the case $\beta > 1$.

3.4.2 Numerical Experiments for $0 < \beta \leq 1$

We now present a series of numerical experiments relative to the case $0 < \beta \leq 1$. We compare the long-time limit of $\mu_h(t, x)$, denoted as μ_h^*, with $\mu_\beta^* := |\nabla u_p|^{p-2}$, where u_p is the solution of the p-Poisson equation for which an explicit formula is known. The test case considered is a two dimensional example taken from [4], where the forcing term is radially symmetric, which means that $f(x, y) = F(r)$

with $r = \sqrt{x^2+y^2}$, and $F :]0,1[\mapsto \mathbb{R}$. Under these assumptions, the exact solution of the p-Poisson equation reads:

$$u_p(x,y) = U(r) = -\int_r^1 \text{sign}(Z(t))|Z(t)|^{\frac{1}{p-1}}dt \quad Z(r) = -\frac{1}{r}\int_0^r tF(t)dt$$

According to the relation $p = (2-\beta)/(1-\beta)$ in Equation (3.10), we can write an explicit formula for μ_β^* that reads as

$$\mu_\beta^*(x,y) = |Z(r)|^{\frac{p-2}{p-1}} = |Z(r)|^\beta \tag{3.14}$$

In our numerical experiment we take F as a piecewise constant function, positive on the interval $]0,1/3[$, zero $[1/3,2/3]$, and negative on $]2/3,1[$. The value of F on the positive and the negative parts are two constants c_1, c_2 calculated such that the right hand side of the linear system arising from the elliptic equation is orthogonal to the constant vector, up to machine precision. The tuning of the constants corrects quadrature errors c_1, c_2 is necessary to provide accurate approximation of the integrals and correct errors introduced during the triangulation of the supports of f^+ and f^-. The mesh $\mathcal{T}_h$ and the forcing term f are plotted in Figure 3.3.

Our numerical experiments consist in testing the existence of a steady state μ_h^* for different values of β and evaluating the the error with respect to the candidate exact solution μ_β^*, error defined as

$$\text{err}(\mu_h^k) := \frac{\|\mu_h^k - \mu_\beta\|_{L^2(\Omega)}}{\|\mu_\beta\|_{L^2(\Omega)}}$$

We repeat the procedure for a sequence of conformally refined grid and we evaluate the experimental rate of convergence $\text{err}(\mu_h^*)$. We consider the exponents $\beta = 0.25, 0.5, 0.75$ and the limit value 1.0, which corresponds to the case $p = +\infty$, exploiting the fact that expression in Equation (3.14) is well defined also for $\beta = 1$ and represents the optimal transport density of the MK equations for the considered forcing term.

In Figure 3.4 we report the time evolution of $\text{var}(\mu_h(t))$ and $\text{err}(\mu_h(t))$ (that are the linear interpolations of the $(\text{var}(\mu_h^k), \text{err}(\mu_h^k))$, as in Section 2.5), for the values of β considered, and the four levels of mesh refinement. The tolerance τ_T used to fix the achievement of the steady state configuration is 5×10^{-7}. As shown in Figure 3.4, in all simulations the equilibrium configuration is achieved, with rate of convergence that increase at lower values of β. For practical purposes, the

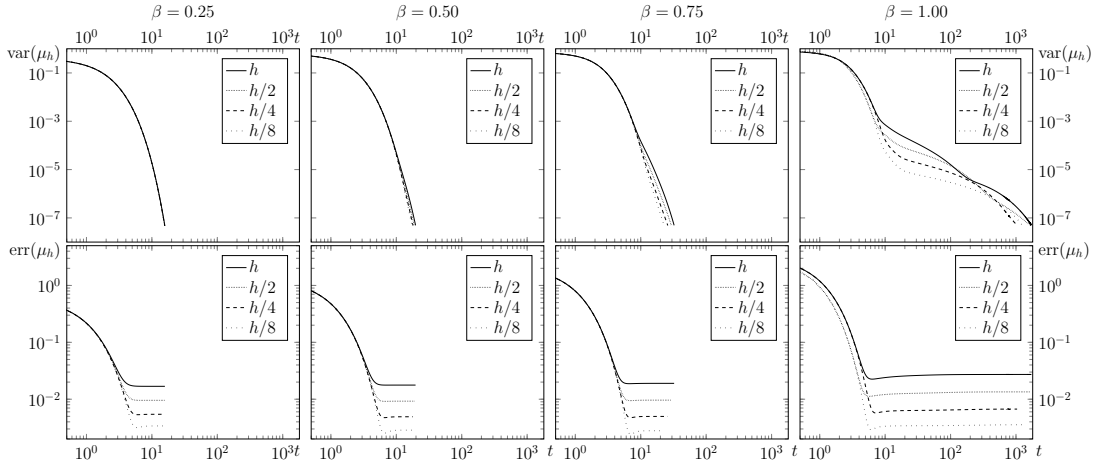

Figure 3.4: Log-log plots of $\mathrm{var}(\mu_h(t,\cdot))$ (upper panels) and $\mathrm{err}(\mu_h(t,\cdot))$ (lower panels) vs. time. The columns refer, from left to right, to the results obtained with $\beta = 0.25, 0.5, 0.75, 1.0$.

value of τ_{T} could be raised to much bigger tolerance without affecting the $\mathrm{err}(\mu_h^*)$, since in all simulations $\mathrm{err}(\mu_h(t))$ is practically stationary after $t = 10$, when $\mathrm{var}(\mu_h(t) \in [10^{-3}, 10^{-4}]$. Instead, for reason of numerical testing we continua the simulations until the indicated threshold is achieved.

In Figure 3.5 we report the behavior of $\mathrm{err}(\mu_h^*)$ for successive refinements of the mesh $\mathcal{T}_h$, for the four values of β considered. The experimental rate of convergence of the scheme (reported in the legend in Figure 3.5) is proportional to h^m where the power m increases with β, passing from $m = 0.775$ for $\beta = 0.25$ to from $m = 0.981$ for $\beta = 1$.

We report in Figure 3.6 the time evolution of the functional $\mathcal{L}_\beta(\mu_h(t))$, starting from three different initial data μ_0^i $(i = 1, 2, 3)$ with formulas given in Equation (2.56). In all simulations $\mathcal{L}_\beta(\mu_h(t))$ decreases monotonically and always attains the same minimum value in time independently on the initial conditions. After $t \approx 10$ the value of $\mathcal{L}_\beta(\mu_h(t))$ becomes apparently stationary, even if continues to decrease. This result is further confirmation of the correctness of Conjecture 2 and of Proposition 45.

3.4.3 Numerical Experiments for $\beta > 1$

In this section we present the numerical results obtained when $\beta > 1$ is imposed in Equation (3.1). We adopt the same discretization method adopted for the

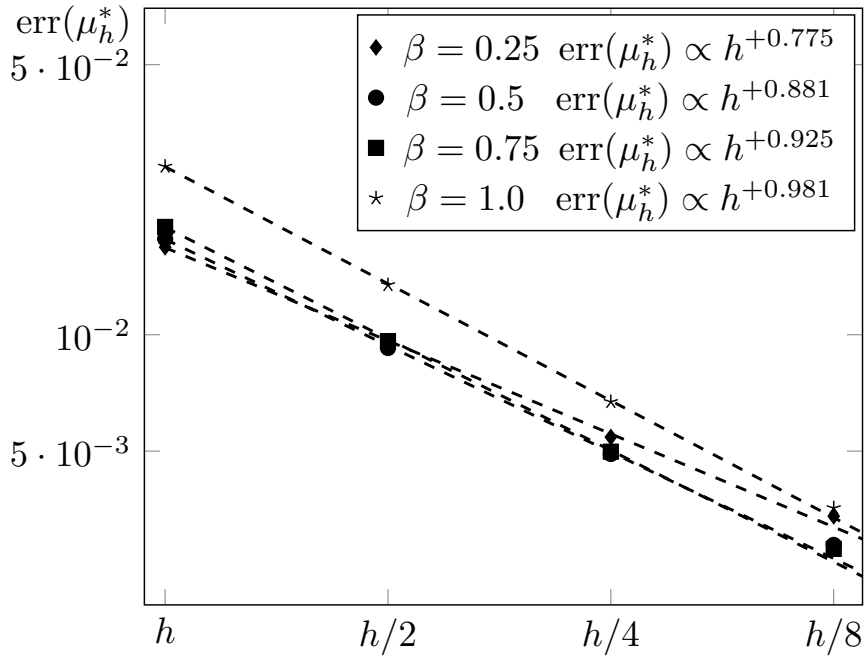

Figure 3.5: Log-log plot of $\mathrm{err}(\mu_h^*)$ vs. the mesh parameter h for $\beta = 0.25, 0.5, 0.75, 1.0$. In the legend we report the average experimental convergence rate for each power β.

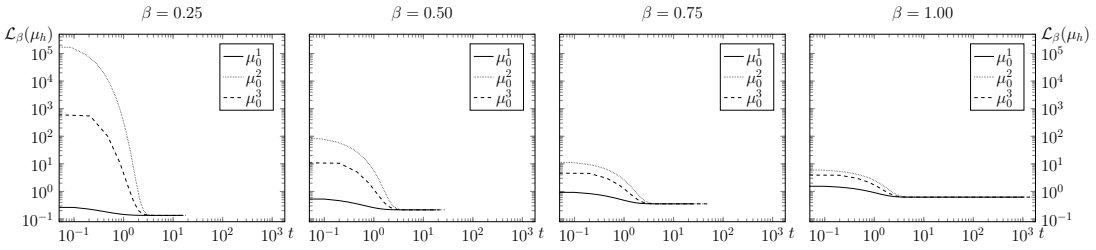

Figure 3.6: Time behavior of the Lyapunov-candidate functional $\mathcal{L}_\beta(\mu_h(t))$, for $\beta = 0.25, 0.5, 0.75, 1.0$ (from left to right) starting from three different initial data μ_0. We report the results of the coarser mesh, without refinement, since for the other cases they are practically indistinguishable.

the case $0 < \beta \leq 1$. In our numerical experiments we consider two test cases. In the first case we take f equal to the piecewise constant function f_2 defined in Section 2.5.2.1. We will refer to this study test as (TC1).

(NewP) The second test case (TC2) considers a forcing term $f = f^+ - f^-$ with f^+ is formed by 50 Dirac sources with unit mass randomly distributed in the region $\Omega = [0.1, 0.9] \times [0.1, 0.9]$, while f^- is a single Dirac located at $(0.05, 0.05)$ that balances f^+. We also consider different initial data μ_0, using a uniformly unitary transport density and the initial data $\mu_0^{2,3}$ defined in Equation (2.56). The simulations are conducted for different values of β (in particular $\beta = 1.1$,1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9, 2.0, 2.2, 2.5, 3.0). Higher values of β are not considered since the linear systems arising from the discretization of the elliptic equations become extremely ill-conditioned and their solution with standard schemes unfeasible. In Chapter 4 we will discuss about these problems and the strategies adopted to cope with it.

We proceed with successive refinements of an initial grid $\mathcal{T}_h$. For (TC1) we use a grid of 1615 nodes and 3100 triangles, aligned with the support of f, while for the (TC2) we use an initial triangulation of the domain $[0, 1] \times [0, 1]$ with 1661 nodes and 3192 triangles. The 51 points where the f is concentrated coincide with some nodes of the grid. The magnitude of the atomic forcing terms is calculated so that the elements of the right-hand side vector $\boldsymbol{b}$ in Equation (3.12) have the same values for different refinement. Again, $\tau_{\mathrm{T}} = 5 \times 10^{-7}$. We discuss the behavior of the model solution by looking at the results obtained for $\beta = 1.5$ as representative examples for the other values of β. In fact, most of considerations that now we will present still hold for true any other value of β, grid or initial data μ_0 considered. We will discuss later the main differences when different values of β are employed. In all the numerical simulations, we experimented strong and sudden variations of μ_h, since the term $\Delta\mu_h^k = \boldsymbol{B}_\beta[\boldsymbol{u}^k] \left(\boldsymbol{\mu}^k\right)^\beta - \boldsymbol{\mu}^k$ in Equation (3.13) can rapidly increases by several orders of magnitude. This effect is amplified for greater values of β. To preserve the stability of the forward Euler scheme we use a time step Δt_k whose size is tuned according to term $\Delta\mu_h^k$.

In Figure 3.7 we report the time evolution of $\mathrm{var}(\mu_h(t))$ with initial data $\mu_0 \equiv 1$, for both test cases considered. Unlike the numerical experiments for $\beta \leq 1$ we see that the $\mathrm{var}(\mu_h(t))$ is not monotonically decreasing in time, and oscillations are present for different levels of refinement.

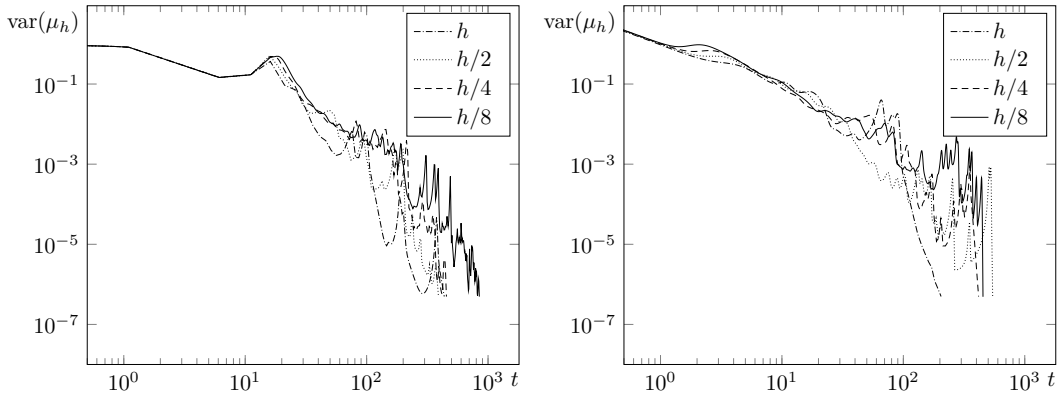

Figure 3.7: Time evolution of $\mathrm{var}(\mu_h(t))$ on successive refinement of grid $\mathcal{T}_h$ for $\beta = 1.5$ and $\mu_0 \equiv 1$. The results for TC1 are reported on the left, while those obtained for TC2 on the right.

Despite the worrying and irregular behavior of $\mathrm{var}(\mu_h(t))$, the average variation is decreasing steadily and for both test cases all the simulations show convergence toward an equilibrium configuration (μ_h^*, u_h^*) for all values of β and for every grid and initial data μ_0 we have considered.

Looking at the results for both test cases, we see that the support of μ_h^* (i.e. the union of all triangles in me $\mathcal{T}_h$ where μ_h is above the minimal imposed threshold) tends to create a network. This network is made of narrow channels connecting the supports of f^+ and f^-. We denoted these supports by the symbols Q^+ and Q^-, respectively. The created network presents a hierarchical structure in which the channels with higher flow capacity, determined by the values of μ_h, repetitively branch into sub-channels until the whole support of f is covered. The networks described above are shown in Figures 3.8 and 3.9, where we report the spatial distribution behavior of μ_h^* at successive grid refinements for $\beta = 1.5$, and $\mu_0 \equiv 1$. Even if it is difficult to compare μ_h^* for different refinement levels, an underlying limit network is clearly appearing. This result is confirmed for any other values of β, grid or initial data μ_0 considered. In particular for TC1, we see in Figure 3.8 that inside Q^+ and Q^- μ_h^* forms a branching structures with allegedly fractal features. In the region outside Q^+ and Q^-, the support μ_h^* concentrates on a series of connected triangles, creating a tight channel with high conductivity. These effects persists at each refinement level, and the support of μ_h^* seems to approximate a 1-dimensional structure.

In TC2 we perceive another phenomenon, severe several branches in the μ_h^*

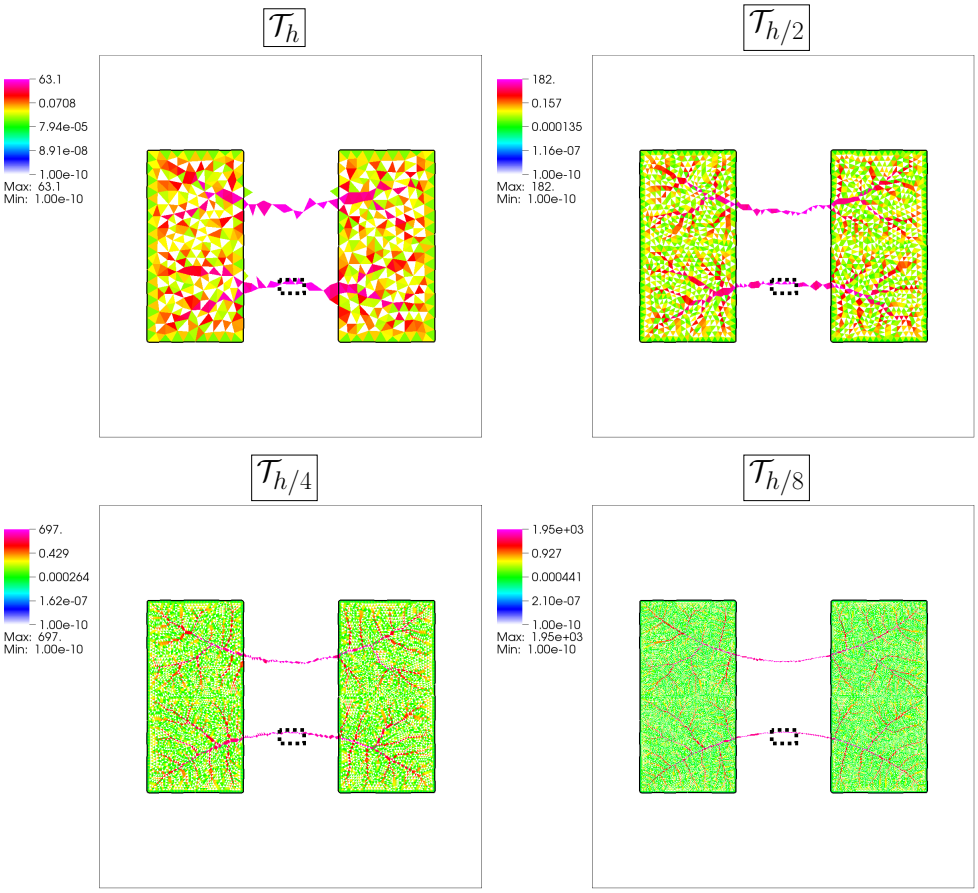

Figure 3.8: Numerical approximation μ_h^* for $\beta = 1.5$ for the piecewise constant forcing term The initial data μ_0 is uniformly equal to 1 on the entire domain. We show the results obtained on different mesh refinements levels. The supports of f^+ and f^- are contoured in black. A constant unit mass is transported from the left to the right rectangles.

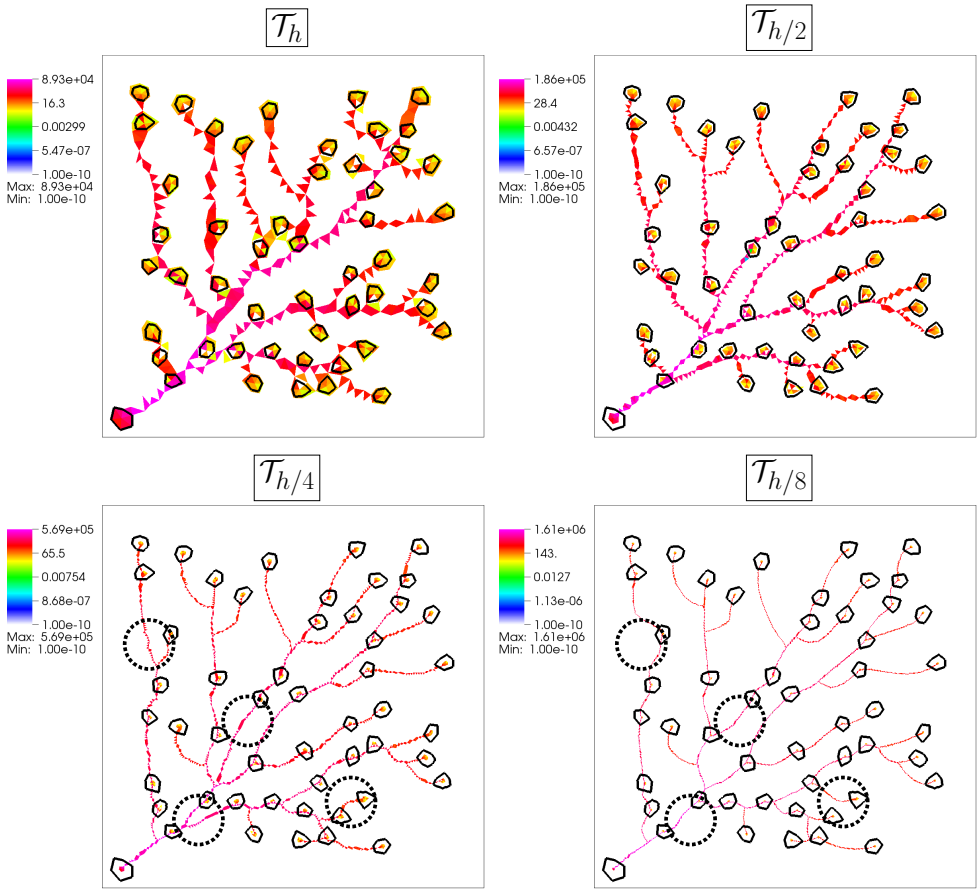

Figure 3.9: Numerical approximation μ_h^* for $\beta = 1.5$ for TC2. The initial data μ_0 is uniformly equal to 1 on the entire domain. We show the results obtained on different levels of mesh refinements. The small black circles indicate the approximate position of the Dirac masses. In the bottom panels we have indicated with dashed circles the area where topological changes on the network structures occur.

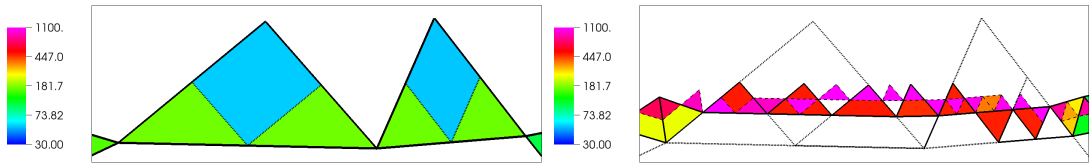

Figure 3.10: Zoom of the small dotted rectangles located on the central channels indicated in in Figure 3.8. The left panel reports the zoom for $\mathcal{T}_h$ and $\mathcal{T}_{h/2}$, while the right panel those for $\mathcal{T}_{h/4}$ and $\mathcal{T}_{h/8}$. We overlap the support of μ_h^* for a finer grid above the support of μ_h^* for the coarser one. We show only the triangles where μ_h^* is above the threshold 10^{-10}. We see that the first two triangles (blue) of $\mathcal{T}_h$, those of $\mathcal{T}_{h/2}$ (green), those of $\mathcal{T}_{h/4}$ (red), and finally those of $\mathcal{T}_{h/8}$ (pink).

tree are not straight lines. We attribute this occurrence to a problem of grid-alignment of the numerical solution, presumably attributable to the extreme spatial irregularity of μ_h^*. Indeed, we are surprised by the capabilities of our numerical scheme to reproduce, albeit with inaccuracies, these singular structures.

In Figure 3.10 we show a small region inside the dashed rectangles in Figure 3.8. The "section" of these channel decreases with h, while the crossing flux remains constant (since the leading area as the same), thus the values of μ_h^* increase using finer grids. This means that for finer grids we obtain μ_h^* with more irregular support and with increasing maximal values reported in Figures 3.8 and 3.9. As consequence of such effects we have a rapid increase of the conditioning number of the matrix in the linear system arising from the elliptic equation, not shown here, with extreme cases in which the PCG scheme does not converges. We cope to this problem with the strategy described in Chapter 4. Nevertheless the method requires the construction of Incomplete Cholesky (IC) factorization with partial fill-in of the SPD $\boldsymbol{A}[\boldsymbol{\mu_h}]$ in Equation (3.12), that in the cases of very refined meshes and high values of β can not always be computed.

Unlike the case $\beta \leq 1$, for $\beta > 1$ there is a dependence with respect to the initial data μ_0. In Figure 3.11 we report the results obtained for μ_h^* on the finest grid starting with μ_0^2, μ_0^3 (left) as described in Equation (2.56). In the case $\mu_0 = \mu_0^2$, in which the initial data attain the minimum value of 0.01 at the center of the square, the support of the equilibrium configuration seems to avoid altogether regions with lower values of μ_0. We conjecture that the converged

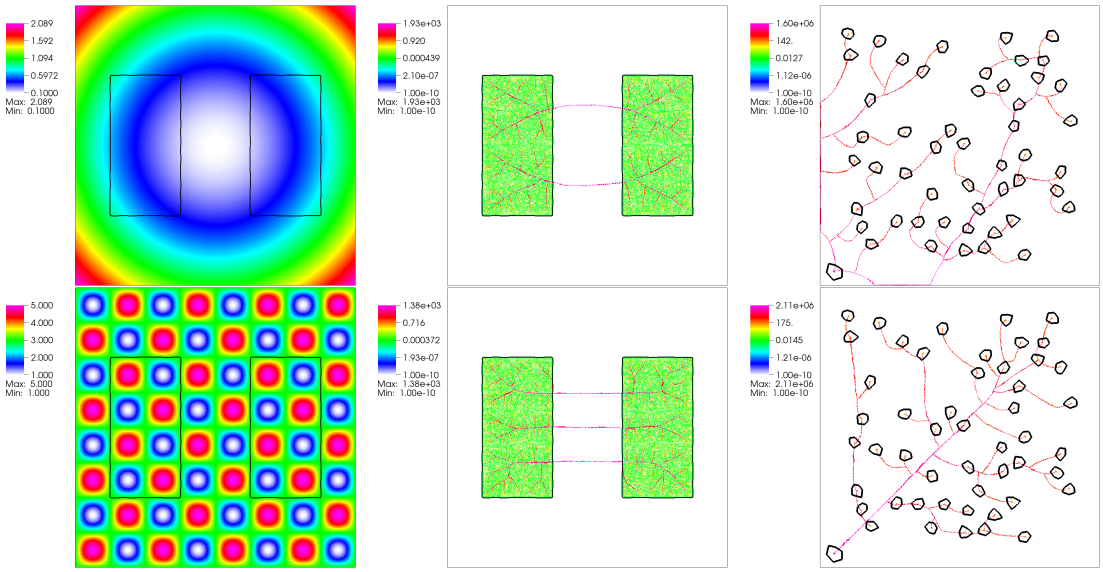

Figure 3.11: Spatial distribution of initial data $(\mu_h)_i$ for $i = 2, 3$ defined in Equation (2.56) (left panel), and the corresponding asymptotic state μ_β^* for TC1 (central panel) and TC2 (right panel). We report only the results obtained using the finest grid and $\beta = 1.5$.

numerical solutions correspond to local minima for $\mathcal{L}_\beta$. As a consequence, the patterns reported in Figure 3.9 display topological changes (highlighted with circles) between different mesh levels. For example, from the figure we can see that going from the third to the fourth level four topological changes are found.

We would like to highlight that, independently of initial data, power β, or grid, the support of μ_h^* in TC2 has the structure of an acyclic graph connecting all the sink/source points. The absence of loops is a fundamental characteristic of the solution of the BTP, that is never imposed a priori in our model, and provides a further confirmation of the strong connection of the extend DMKmodel and the BTP.

Finally, Figures 3.12 and 3.13 report a comparison among the approximations μ_h^* obtained for the different values of β, using the finest grid and the initial data $\mu_0 \equiv 1$. The panels show the results at increasing values of β from left to right and top to bottom. For all test cases the tendency for creating more concentrated networks with high conductivity channels increases with the power β. In fact, we note that the number of central channels created at the equilibrium varies from three for $\beta = 1.1$ to one for $\beta = 3$. The final configurations shows

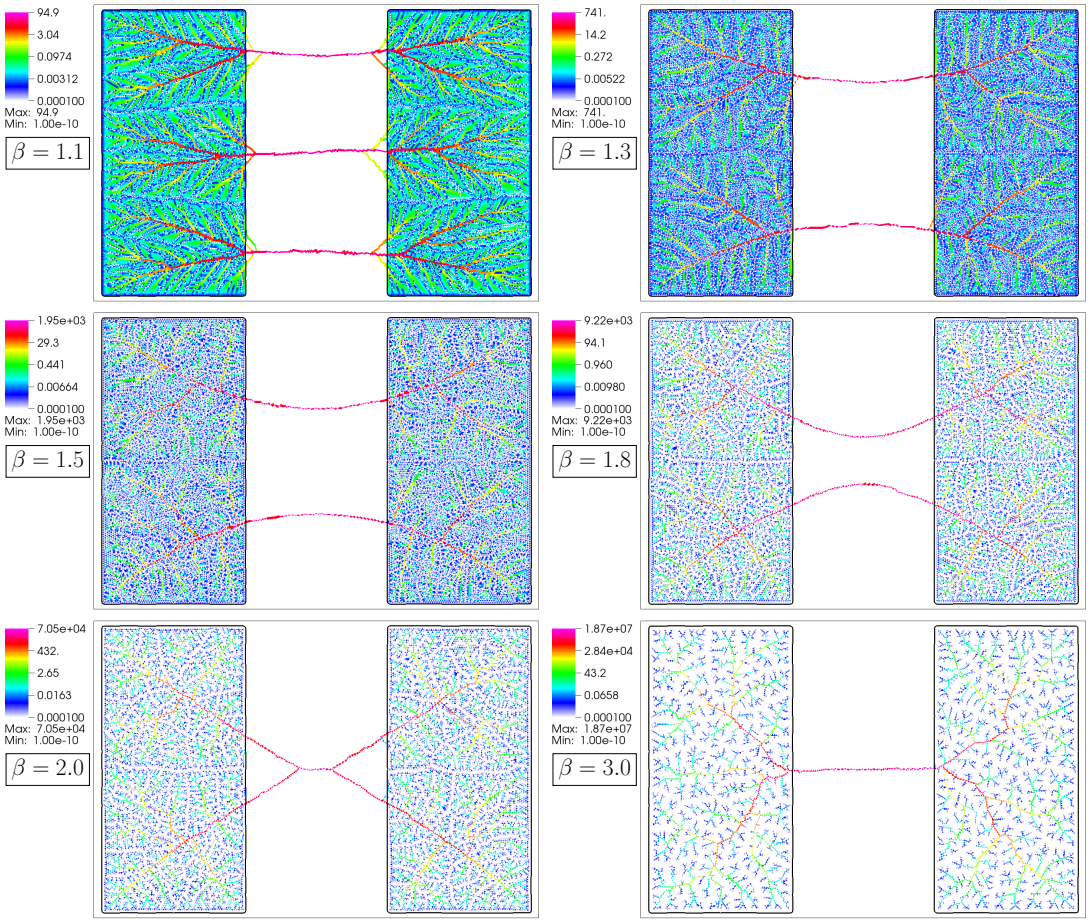

Figure 3.12: Behavior of the spatial distribution of μ_β^* for TC1 for different values of β. We remark that, although the color scale starts from the value 10^{-4}, the white regions indicate where μ_β^* attains the minimal value 10^{-10}.

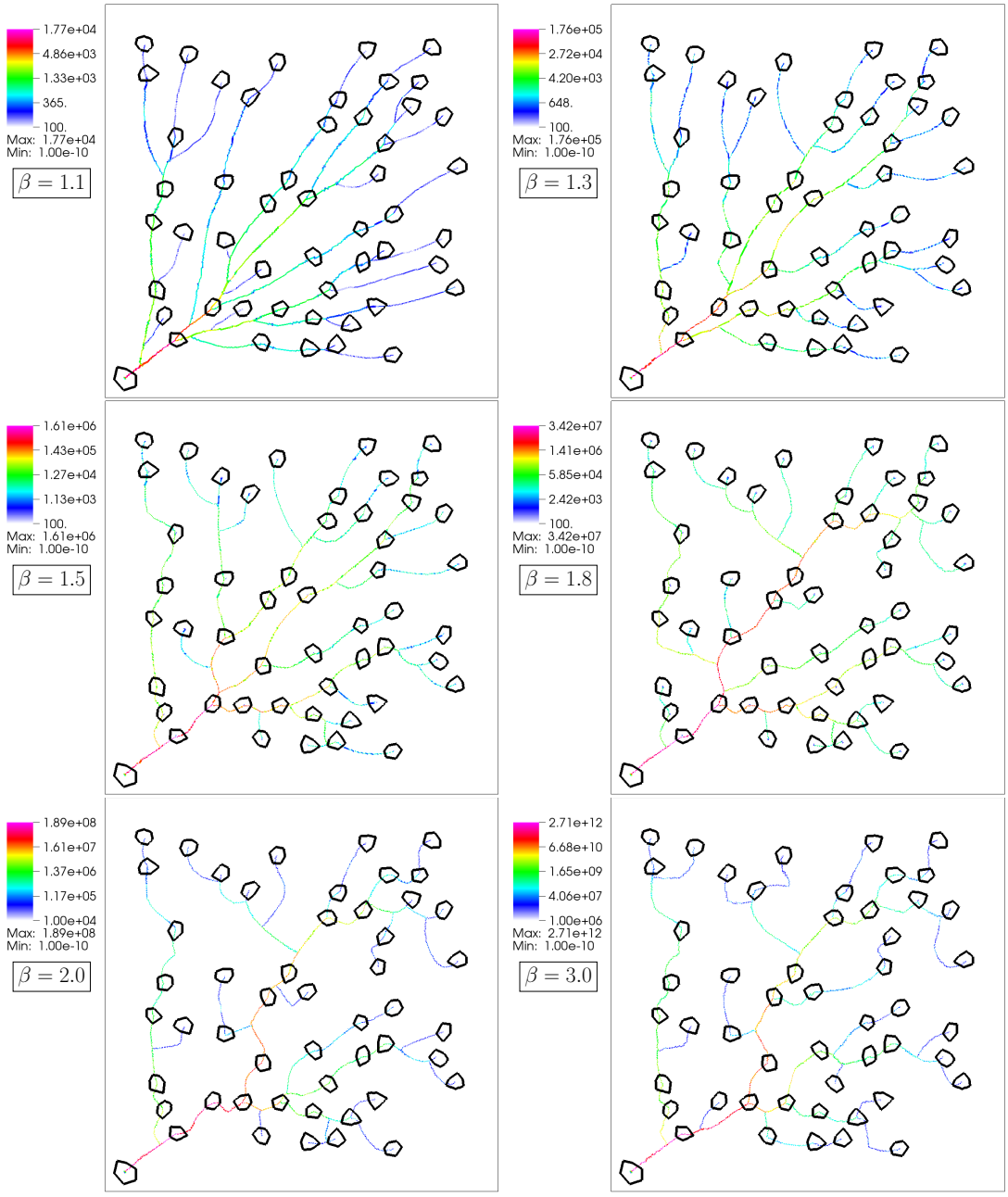

Figure 3.13: Behavior of the spatial distribution of μ^*_β for TC2 for different values of β. We remark that, although the color scale starts from different initial values, the white regions indicate where μ^*_β attains the minimal value 10^{-10}.

no branching points in the channel and seems to be the most affected by grid alignment problems.

3.4.3.1 Lyapunov, Energy, and Mass Functionals

In this section we analyze numerically the time evolution of the Lyapunov-candidate functional $\mathcal{L}_\beta(\mu_h(t))$, its constituents $\mathcal{E}_f(\mu_h(t))$ and $\mathcal{M}_\beta(\mu_h(t))$, and $\int_\Omega(\mu_h(t))$. In addition, we look for a power-law scaling of these quantities as the mesh parameter h is refined. This latter exercise is an attempt to find a proper scaling in the search for a mean to approximate singular measure using the Lebesgue integrals typical of the FEM method (see also remark 5).

Figure 3.14 shows the time evolution of the different components of of $\mathcal{L}_\beta(\mu_h(t))$ at the different refinement levels, using $\beta = 1.5$, and the initial data $\mu_0 \equiv 1$. The numerical simulations support the statements in Proposition 45 on the decrease in time of the $\mathcal{L}_\beta(\mu_h(t))$. These results are confirmed for all the powers β, initial data μ_0, and for both forcing terms considered. The dependence on the initial data μ_0 clearly influences the asymptotic value $\mathcal{L}_\beta(\mu_h^*)$, that unlike for the cases $0 < \beta < 1$, is not the same.

As already anticipated, we note that the value of $\mathcal{L}_\beta(\mu_h^*)$ scales with respect to the mesh parameter h, in a form that resembles a power law:

$$\mathcal{L}_\beta(\mu_h^*) \propto h^m \tag{3.15}$$

Also functionals $\mathcal{E}_f(\mu_h^*), \mathcal{M}_\beta(\mu_h^*)$ and total mass $\int_\Omega \mu_h^* \, dx$ present similar behaviors, as shown in Figure 3.15, where we report the log-log plots of the different quantities vs the mesh parameter h for $\beta = 1.5$. We have also calculated the least-square lines whose slope gives the value of the power m in Equation (3.15) (the values of m for the different functionals are reported in the legend).

We note that the total mass always decreases with h. This is due to the tendency of μ_h to concentrate in progressively narrower channels. Intuitively, this limit structure should be measured via a singular Hausdorff-type measure, while the FEM method used in our calculations employees the Lebesgue measure. The components $\mathcal{E}_f$ and $\mathcal{M}_\beta$ of the Lyapunov-candidate functional increase at a practically constant rate that depends on the forcing, as seen from the values of the power m that approaches 0.3 or TC1 and 0.5 for TC2. The value of m changes with the power β of the dynamics, as seen in Figure 3.16 that reports

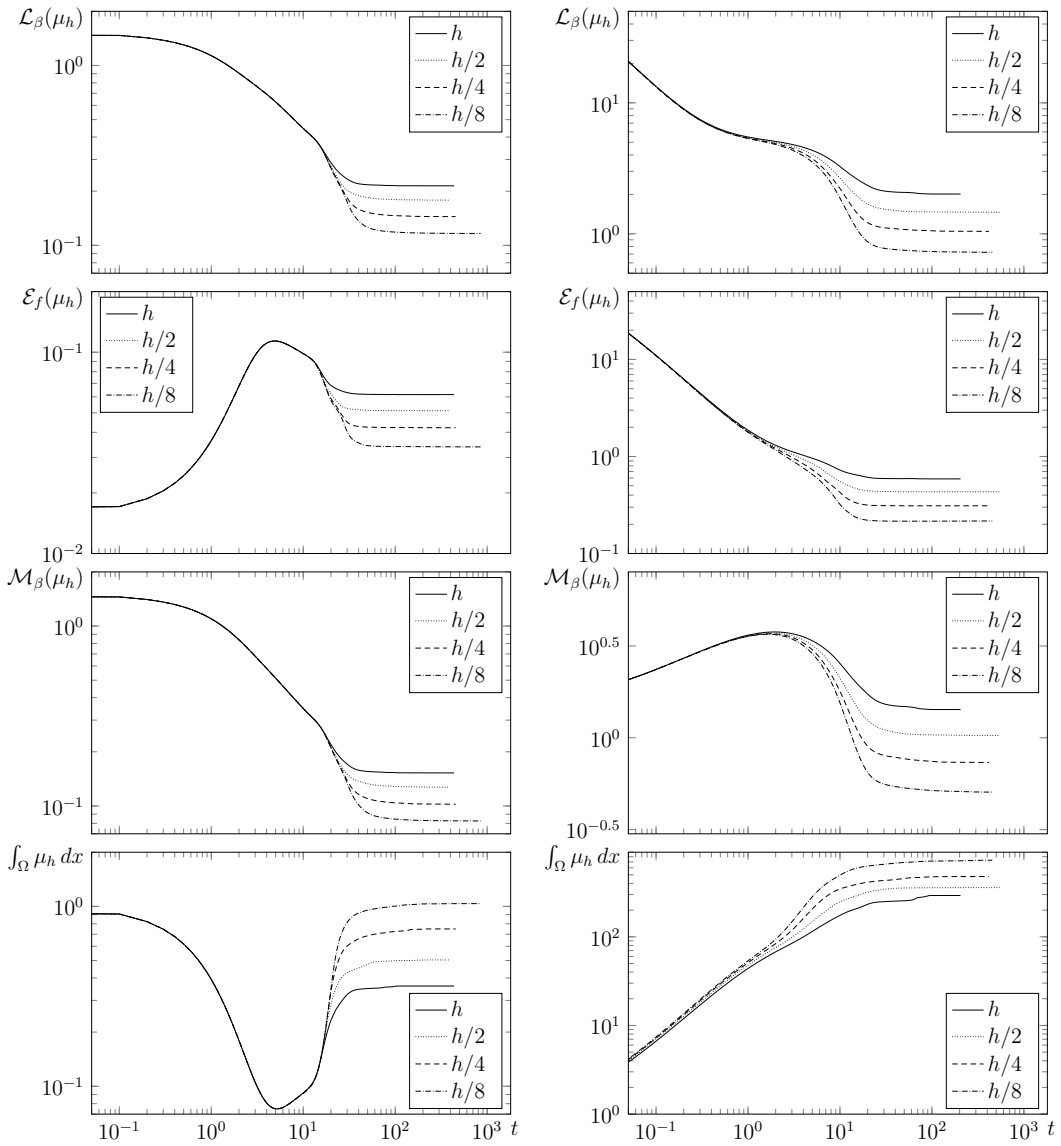

Figure 3.14: Time evolution of $\mathcal{L}_\beta(\mu_h(t)), \mathcal{E}_f(\mu_h(t)), \mathcal{M}_\beta(\mu_h(t))$, and $\int_\Omega \mu_h(t)\,dx$ (top to bottom) for the test case 1 (left) and 2 (right). Each figure reports the time evolution for each grid refinement level. We used the exponent $\beta = 1.5$ and the initial data $\mu_0 \equiv 1$.

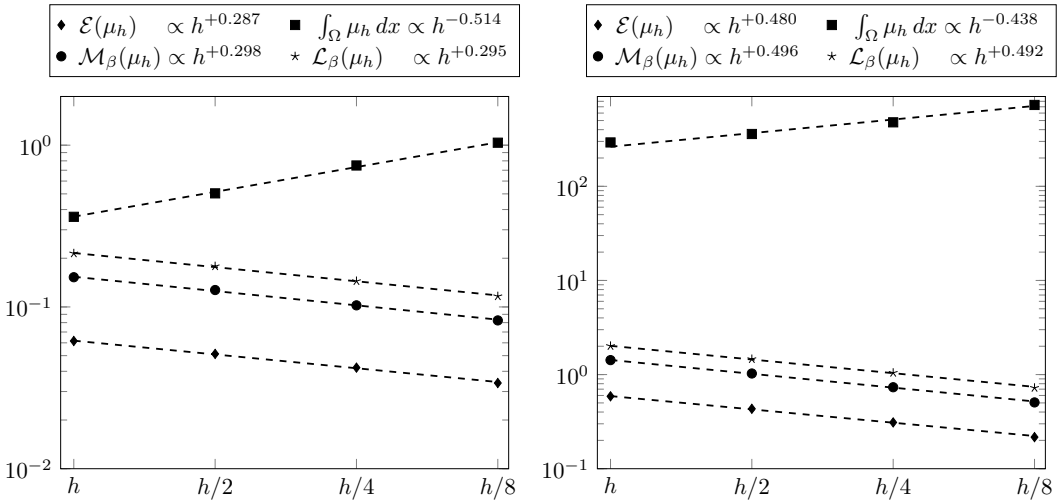

Figure 3.15: Log-log plot values of $\mathcal{L}_\beta(\mu_h^*)$, $\mathcal{E}_f(\mu_h^*)$, $\mathcal{M}_\beta(\mu_h^*)$, and $\int_\Omega \mu_h^*\,dx$ vs the mesh parameter h. The left panels show the results obtained for the piecewise constant forcing term, the right panel those for the atomic forcing term. We used the exponent $\beta = 1.5$ and the initial data $\mu_0 \equiv 1$. Note that the scale of the abscissa is reversed.

m vs. β. It is interesting to note that while the power m for the total mass varies linearly with β, the behavior for the Lyapunov-candidate functional is different and varies with the forcing term. The results of this section are an attempt at finding a proper scaling for the two components of the Lyapunov-candidate functional function as an effort to approximate singular measures using the Lebesgue integrals characteristic of the finite element method. While this essay has been unsuccessful, we are reassured by these results that show the power of an accurate and robust numerical tool to suggest possible strategies for the deeper understanding of the considered phenomenon.

In conclusion of this chapter, we would like to remark that there exist sparse examples in the literature addressing the numerical solution of the BTP both in the discrete and in the continuous settings. For example, for the case of L^1-OTP, in addition to the work of [5], we can cite the work of [26] who propose a numerical approach based on finite difference discretization and an original application of Newton method. In the case of BTP [55] address the minimization problem via finite difference discretization and a conjugate gradient minimization method, but they report problems similar to those described in this section (grid alignment and convergence to local minima). Despite all the criticalities that are still present in

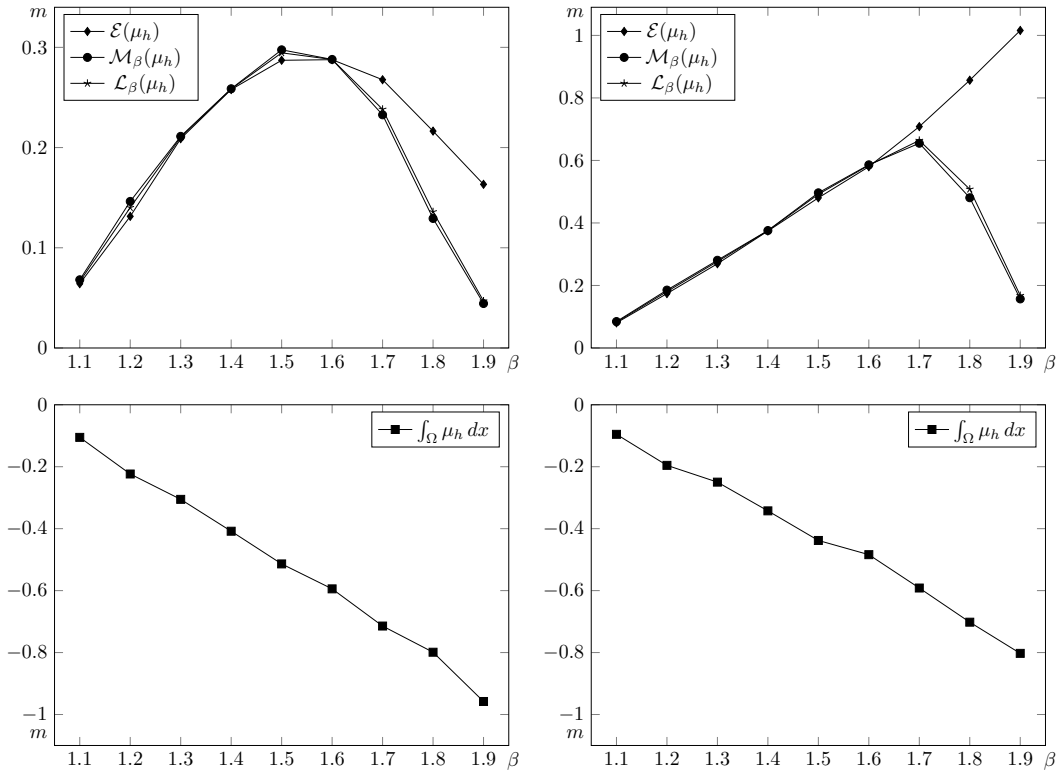

Figure 3.16: The left panels show the results obtained for the piecewise constant forcing term, the right panel those for the atomic forcing term. Approximate exponent m for $\mathcal{L}_\beta(\mu_h^*)$, $\mathcal{E}_f(\mu_h^*)$, $\mathcal{M}_\beta(\mu_h^*)$, and $\int_\Omega \mu_h^*\, dx$ for $\beta = 1.1$,1.2, 1.3, 1.4, 1.5, 1.6, 1.7, 1.8, 1.9. The initial data considered is $\mu_0 \equiv 1$.

our approach, starting from the fact that we are not able to formulate an exact relationship between the equilibrium configurations (μ^*_β, u^*_β) and the solutions of BTP, we are confident that our conjectures, or some appropriate variants, are true. Then our numerical formulation would represent an original and highly efficient approach for the numerical solution of OTP and BTP.

Chapter 4

Spectral Preconditioner for Extended DMK with $\beta > 1$

In this chapter we present the strategy we have developed to efficiently solve the sequences of linear systems arising form FEM discretization of Equation (3.1) described in Section 3.4. As already mentioned in Section 3.4.3, when $\beta > 1$ the linear systems to be solved at each time step are characterized by a large, sparse, ill conditioned symmetric positive definite (SPD) matrix $\boldsymbol{A}$. Extreme cases in same instances prevent the convergence of PCG with standard preconditioners such as the Incomplete Cholesky (with partial fill-in) factorization of $\boldsymbol{A}$. We investigate several preconditioning strategies that incorporate partial approximated spectral information. We present numerical evidence that the proposed techniques are efficient in reducing the condition number of the preconditioned systems, not only decreasing the number of PCG iterations and the overall CPU time, but also enabling the correct solution of linear system in the cases where standard preconditioners fail. These results are collected in the manuscript [6].

4.1 Spectral Method

The numerical discretization by finite elements in space and explicit Euler in time described in Sections 3.4 and 3.4.3 requires the solution of the large, sparse SPD linear system. Following the notation adopted in Section 2.5 and Equation (3.12),

this sequence of linear systems reads as

$$\boldsymbol{A}[\boldsymbol{\mu}^k]\boldsymbol{u}^k = \boldsymbol{b} \tag{4.1}$$

$$\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \Delta t_k \left[\boldsymbol{B}_\beta[\boldsymbol{u}^k] \left(\boldsymbol{\mu}^k\right)^\beta - \boldsymbol{\mu}^k\right] \tag{4.2}$$

where k is the time step index, $\boldsymbol{A}[\boldsymbol{\mu}^k]$ is the stiffness matrix associated to $\boldsymbol{\mu}^k$, and $\boldsymbol{B}_\beta[\boldsymbol{u}^k]$ is the matrix defining the norm of the approximate gradient of $u_h(t^k, x)$ raised to the power β. Because of typically large dimensions, we use the Preconditioned Conjugate Gradient (PCG) method solver to solve the linear system in Equation (4.1). PCG convergence becomes increasingly difficult as time progresses since the condition number of the system matrix grows with β. In fact, the dynamics of the model is such that μ_h tends to zero in large portions of Ω. To avoid non-coerciveness of the elliptic partial differential equation, we impose a minimum threshold for μ_h equal to 10^{-10}. However, the maximum value of μ_h increases as its support concentrates along thinner paths appearing for increasing values of β, as shown in Section 3.4.3.

According to [58] we can estimate the minimum and maximum eigenvalue of the stiffness metric $\boldsymbol{A}[\mu_h]$ as follows

$$\lambda_{\min}(\boldsymbol{A}[\mu_h]) \leq C_1 h^2 \mu_{\min} \quad \lambda_{\max}(\boldsymbol{A}[\mu_h]) \geq C_2 \mu_{\max}$$

with the constants depending on the domain Ω and the triangulations $\mathcal{T}_h$ and $\mathcal{T}_{h/2}$. Thus the condition number of $\boldsymbol{A}[\mu_h^k]$ increases with time, possibly leading to non-convergence of the PCG iteration with a standard preconditioner such as an Incomplete Cholesky (IC) factorization with partial fill-in, which can not always be computed.

The strategy to develop efficient preconditioners adopted in [6] explicitly takes into consideration the spatial and temporal variability of the transport density by calculated spectral information of the involved stiffness matrices. The idea of using partial spectral knowledge to accelerate linear system solvers has been described in several papers such as [20, 27, 39] and, more recently in [50]. In all these papers the authors start with an initial preconditioner $\boldsymbol{P}_0$ and use an approximation of a few eigenvectors of the preconditioned matrix to update $\boldsymbol{P}_0$ with a low-rank matrix. Another characteristic shared by all these previous papers is that the coefficient matrix of the linear systems to be solved $\boldsymbol{A}\boldsymbol{x}_k = \boldsymbol{b}_k$ remains unchanged throughout the whole sequence. This allow the incremental

refinement of the set of eigenvectors used to update the low-rank correction matrix. In this model we consider instead sequences of sparse linear systems with changing coefficient matrices dynamically depending on the transport density. We present numerical evidence that the proposed techniques are efficient in reducing the condition number of the preconditioned systems, thus decreasing the number of PCG iterations and the CPU time. In the section we investigate several strategies that incorporate incomplete spectral information [8] on previous matrices to update the preconditioner for the current and future system.

4.2 The spectral preconditioner

Consider the sequence of linear systems of the form

$$\boldsymbol{A}_k \boldsymbol{x}_k = \boldsymbol{b}, \tag{4.3}$$

where $\boldsymbol{A}_k \in \mathbb{R}^{n\times n}$ is an SPD matrix, $\boldsymbol{x}_k,\ \boldsymbol{b} \in \mathbb{R}^n$. For a given linear system $\boldsymbol{A}_k \boldsymbol{x}_k = \boldsymbol{b}$ we study the acceleration of the PCG solver provided by the following spectral preconditioner:

$$\boldsymbol{P} = \boldsymbol{P}_0 + \boldsymbol{V}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{V}_p^T, \tag{4.4}$$

where $\boldsymbol{V}_p = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p]$ and $\boldsymbol{v}_j, j = 1, \ldots, p$ are approximate eigenvectors either of $\boldsymbol{P}_0\boldsymbol{A}_k$ or of $\boldsymbol{A}_k$; $\boldsymbol{\Lambda}_p = \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$, and $\lambda_j,\ j = 1, \ldots, p$ are the corresponding smallest eigenvalues. When $\boldsymbol{V}_p$ contains eigenvectors of $\boldsymbol{P}_0\boldsymbol{A}_k$, the effect of the low-rank correction is easily shown to be:

$$\boldsymbol{P}\boldsymbol{A}_k \boldsymbol{v}_j = (\lambda_j + 1)\boldsymbol{v}_j, \qquad j = 1, \ldots, m.$$

so that some of the eigenvalues of the new preconditioned matrix are incremented by 1 with an obvious reduction of the condition number.

We propose two different ways to obtain the approximated eigenvectors needed to construct the spectral preconditioner: evaluating the sought eigenpairs with an external enslave (Deflation-Accelerated Conjugate Gradient, DACG) or approximating them directly from the PCG iterations at previous time-steps. For simplicity, from now on we will write $\boldsymbol{A}$ for $\boldsymbol{A}_k$ when no confusion arises.

4.2.1 Approximating the smallest eigenpairs by DAGC

Following [8], we propose to approximate some of the leftmost eigenvectors of a given coefficient matrix $\boldsymbol{A}_k$ by performing some preliminary iterations of an eigen-

Algorithm 1 DACG method

- INPUT: tolerance $\tau_{\text{DACG}}, \boldsymbol{P}_0, p$. Set $\boldsymbol{V}_p = 0$.
- FOR $j = 1$ TO p
  1. Choose a unit $2-$norm $\boldsymbol{x}_0$ such that $\boldsymbol{V}_p^T \boldsymbol{x}_0 = 0$;
  2. Find the minimum of the RQ over all $\boldsymbol{x}$ such that $\boldsymbol{V}_p^T \boldsymbol{x} = 0$ by a nonlinear PCG procedure, with starting point $\boldsymbol{x}_0$ and preconditioner $\boldsymbol{P}_0$. Stop whenever the following test is satisfied:

$$\frac{\|\boldsymbol{A}\boldsymbol{x}_{\text{DACG}} - q(\boldsymbol{x}_{\text{DACG}})\boldsymbol{x}_{\text{DACG}}\|}{q(\boldsymbol{x}_{\text{DACG}})\|\boldsymbol{x}_{\text{DACG}}\|} \leq \tau_{\text{DACG}} \tag{4.5}$$

  3. Set $\lambda_j = q(\boldsymbol{x}_{\text{DACG}}), \quad \boldsymbol{v}_j = \dfrac{\boldsymbol{x}_{\text{DACG}}}{\|\boldsymbol{x}_{\text{DACG}}\|}, \quad \boldsymbol{V}_p = [\boldsymbol{V}_p,\ \boldsymbol{v}_j]$.
- END FOR

value solver. We chose the Deflation-Accelerated Conjugate Gradient (DACG) eigensolver [7, 10], which is based on the preconditioned conjugate gradient (nonlinear) minimization of the Rayleigh Quotient (RQ) $q(\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x}/\boldsymbol{x}^T\boldsymbol{x}$. The leftmost eigenpairs are computed sequentially, by minimizing RQ over a subspace orthogonal to the previously computed eigenvectors. This method, which applies only to symmetric positive definite matrices, has been proven very efficient in the solution of eigenproblems arising from discretization of PDEs in [10]. DACG also proved very suited to parallel implementation as documented in [9] where an efficient parallel matrix vector product has been employed. Our implementation of DACG is shown in Alg. Section 4.2.1. The main computational cost of one DACG iteration is given by:

1. One matrix-vector product.

2. One application of the preconditioner.

3. Orthogonalization of the search direction against the previously computed eigenpairs (columns of matrix $\boldsymbol{V}_p$). The cost of this step is increasing with the number of eigenpairs begin sought.

Convergence of DACG is strictly related to the relative separation between consecutive eigenvalues, namely

$$\xi_j = \frac{\lambda_j}{\lambda_{j+1} - \lambda_j}. \tag{4.6}$$

When two eigenvalues are relatively close, DACG convergence may be very slow. Also DACG takes advantage of preconditioning, which in our case is chosen to be the IC factorization of matrix $\boldsymbol{A}$.

Once a small number of leftmost eigenvectors has been computed and stored as columns of $\boldsymbol{V}_p$, different low-rank corrections of a given preconditioner $\boldsymbol{P}_0$ can be defined as e.g. described in [49]. For example a BFGS-style preconditioner can be written as

$$\begin{aligned} \boldsymbol{P} &= \boldsymbol{V}_p(\boldsymbol{V}_p^T\boldsymbol{A}\boldsymbol{V}_p)^{-1}\boldsymbol{V}_p^T \\ &+ \quad \left(\boldsymbol{I} - \boldsymbol{V}_p(\boldsymbol{V}_p^T\boldsymbol{A}\boldsymbol{V}_p)^{-1}\boldsymbol{V}_p^T\boldsymbol{A}\right)\boldsymbol{P}_0\left(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{V}_p(\boldsymbol{V}_p^T\boldsymbol{A}\boldsymbol{V}_p)^{-1}\boldsymbol{V}_p^T\right) \\ &\approx \boldsymbol{V}_p\boldsymbol{\Lambda}_p^{-1}\boldsymbol{V}_p^T + \left(\boldsymbol{I} - \boldsymbol{V}_p\boldsymbol{V}_p^T\right)\boldsymbol{P}_0\left(\boldsymbol{I} - \boldsymbol{V}_p\boldsymbol{V}_p^T\right) \end{aligned} \tag{4.7}$$

A simplified version of this BFGS preconditioner neglects the left and right projectors on $\boldsymbol{P}_0$, and thus takes the same form as in (Equation (4.4)):

$$\boldsymbol{P} = \boldsymbol{V}_p\boldsymbol{\Lambda}_p^{-1}\boldsymbol{V}_p^T + \boldsymbol{P}_0.$$

It can be shown [49] that the preconditioned matrix $\boldsymbol{P}\boldsymbol{A}$ has a better spectral distribution than $\boldsymbol{P}_0\boldsymbol{A}$.

4.2.2 Recovering spectral information by the Lanczos process

Another strategy we are going to use is to recover the partial eigenspectrum of $\boldsymbol{A}$ from the Krylov subspace built by the linear solver, using the Lanzcos process embedded within the PCG algorithm. Denoting again by $\boldsymbol{P}_0$ an initial preconditioner for matrix $\boldsymbol{A}$, during the PCG method we save the first m preconditioned residuals as columns of a matrix $\boldsymbol{W}_m$:

$$\boldsymbol{W}_m = \left[\frac{\boldsymbol{P}_0\boldsymbol{r}_0}{\sqrt{\boldsymbol{r}_0^T\boldsymbol{P}_0\boldsymbol{r}_0}}, \frac{\boldsymbol{P}_0\boldsymbol{r}_1}{\sqrt{\boldsymbol{r}_1^T\boldsymbol{P}_0\boldsymbol{r}_1}}, \ldots, \frac{\boldsymbol{P}_0\boldsymbol{r}_{m-1}}{\sqrt{\boldsymbol{r}_{m-1}^T\boldsymbol{P}_0\boldsymbol{r}_{m-1}.}}\right]$$

Matrix $\boldsymbol{W}_m$ is such that $\boldsymbol{W}_m^T\boldsymbol{P}_0^{-1}\boldsymbol{W}_m = \boldsymbol{I}_m$, in view of the $\boldsymbol{P}_0-$orthogonality of the residuals generated by the PCG method. Moreover, we can form the Lanczos

tridiagonal matrix using the PCG coefficients α_k, β_k as follows:

$$\boldsymbol{T}_m = \begin{bmatrix} \frac{1}{\alpha_0} & -\frac{\sqrt{\beta_1}}{\alpha_0} & & & \\ -\frac{\sqrt{\beta_1}}{\alpha_0} & \frac{1}{\alpha_1}+\frac{\beta_1}{\alpha_0} & -\frac{\sqrt{\beta_2}}{\alpha_1} & & \\ & & \ddots & & \\ & & & -\frac{\sqrt{\beta_{m-1}}}{\alpha_{m-2}} & \\ & & -\frac{\sqrt{\beta_{m-1}}}{\alpha_{m-2}} & \frac{1}{\alpha_{m-1}}+\frac{\beta_{m-1}}{\alpha_{m-2}} & \end{bmatrix}$$

Matrices $\boldsymbol{W}_m$ and $\boldsymbol{T}_m$ obey to the classical Lanczos relation i.e.:

$$\boldsymbol{W}_m^T \boldsymbol{A} \boldsymbol{W}_m = \boldsymbol{T}_m.$$

After eigensolving $\boldsymbol{T}_m$ we obtain $\boldsymbol{T}_m = \boldsymbol{Q}\boldsymbol{\Lambda}_m\boldsymbol{Q}^T$, where the coefficients of the diagonal matrix $\boldsymbol{\Lambda}_m$ approximate the eigenvalues of $\boldsymbol{P}_0\boldsymbol{A}$ while the columns of $\boldsymbol{V}_p = \boldsymbol{W}_m\boldsymbol{Q}_p$ (where $\boldsymbol{Q}_p$ contains the first p columns of $\boldsymbol{Q}$) are approximations of the p leftmost eigenvectors of $\boldsymbol{P}_0\boldsymbol{A}$. In fact, first note that $\boldsymbol{V}_p^T\boldsymbol{A}\boldsymbol{V}_p = \boldsymbol{Q}_p^T\boldsymbol{W}_m^T\boldsymbol{A}\boldsymbol{W}_m\boldsymbol{Q}_p = \boldsymbol{Q}_p^T\boldsymbol{T}_m\boldsymbol{Q}_p = \boldsymbol{\Lambda}_p \equiv \mathrm{diag}(\lambda_1, \ldots, \lambda_p)$. Then, let $\boldsymbol{U} = \boldsymbol{P}_0^{-1/2}\boldsymbol{V}_p$ we obtain:

$$\boldsymbol{U}^T\boldsymbol{U} = \boldsymbol{V}_p^T\boldsymbol{P}_0^{-1}\boldsymbol{V}_p = \boldsymbol{I}_m \tag{4.8}$$

$$\boldsymbol{\Lambda}_p = \boldsymbol{V}_p^T\boldsymbol{A}\boldsymbol{V}_p = \boldsymbol{U}^T\boldsymbol{P}_0^{1/2}\boldsymbol{A}\boldsymbol{P}_0^{1/2}\boldsymbol{U} \tag{4.9}$$

corresponding to the Lanczos process applied to matrix $\boldsymbol{P}_0^{1/2}\boldsymbol{A}\boldsymbol{P}_0^{1/2}$. Hence the columns of $\boldsymbol{U}$ approximate the eigenvectors of $\boldsymbol{P}_0^{1/2}\boldsymbol{A}\boldsymbol{P}_0^{1/2}$ and the columns of $\boldsymbol{V}_p$ approximate the eigenvectors of $\boldsymbol{P}_0\boldsymbol{A}$, as can be seen from the following relationships:

$$\begin{aligned} \boldsymbol{P}_0^{1/2}\boldsymbol{A}\boldsymbol{P}_0^{1/2}\boldsymbol{U} \approx \boldsymbol{U}\boldsymbol{\Lambda}_p &\iff \boldsymbol{P}_0\boldsymbol{A}\boldsymbol{P}_0^{1/2}\boldsymbol{U} \approx \boldsymbol{P}_0^{1/2}\boldsymbol{U}\boldsymbol{\Lambda}_p \\ &\iff \boldsymbol{P}_0\boldsymbol{A}\boldsymbol{V}_p \approx \boldsymbol{V}_p\boldsymbol{\Lambda}_p, \end{aligned}$$

4.3 Implementation

Approximation of a number of leftmost eigenpairs is a costly task and cannot be performed at each linear system solution. To reduce the impact of this cost on the overall process we devise different strategies depending on how we obtain the spectral information.

4.3.1 Initial preconditioner $\boldsymbol{P}_0$

For all the experiments the initial preconditioner is an IC preconditioner obtained by setting the maximum number of nonzero elements per row LFIL= 30 and a drop tolerance $\tau_{IC} = 10^{-4}$. The use of a smaller LFIL and/or a larger τ_{IC} does not guarantee the existence of the IC factorization for all systems leading to a breakdown of the simulations. This choice of parameters produced a rather dense Cholesky factor with a number of nonzero elements roughly 8 times that of the triangular part of $\boldsymbol{A}$. For this reason, the computation of this preconditioner for each linear system of the sequence was not effective. We decided to compute the IC preconditioner for a given matrix $\boldsymbol{A}_k$ if $k = 1$ or the number of PCG iterations in the previous linear system was above a fixed value, it_{chol}. We used the previously computed IC preconditioner, otherwise.

4.3.2 Eigenpairs of $\boldsymbol{A}$ obtained by DACG

The computation of a number of the leftmost eigenpairs by DACG is a preprocessing stage that in principle should be executed prior to every system solution. However, in view of the slow variability of the system matrices $\boldsymbol{A}_k$ at increasing k, we propose to evaluate selectively the eigenpairs, whenever the PCG solution of a generic linear system $\boldsymbol{A}_k\boldsymbol{x}_k = \boldsymbol{b}$ takes more than a fixed number of iterations ($it_k \geq it_{\text{prec}}$). In this case, except for $k = 1$, it is effective to use as initial DACG guess the previously computed eigenvectors. The final algorithm is reported in Alg. algorithm 2.

4.3.3 Eigenpairs of $\boldsymbol{P}_0\boldsymbol{A}$ obtained by Lanczos-PCG

Computation of matrices $\boldsymbol{T}_m$ and $\boldsymbol{W}_m$ is carried out during the PCG process and adds negligible computational costs due to the saving of the PCG residual vectors. The main computational burden in this strategy is given by the matrix-matrix product $\boldsymbol{V}_m = \boldsymbol{W}_m\boldsymbol{Q}_p$ implemented via BLAS-3 subroutines, with a consequent optimal use of memory accesses. Due to the slow convergence of the Lanczos process to the smallest eigenvalues, and also for memory reasons, it is convenient to recover a relatively small number of eigenpairs (independently of the size m of $\boldsymbol{V}_m$, which nonetheless should be taken sufficiently large to ensure the completeness of the calculated leftmost eigenspectrum). In the Lanczos process

Algorithm 2 PCG with spectral DACG preconditioner

- INPUT: $it_{\text{prec}}, it_{\text{chol}}, p, \tau_{\text{DACG}}$.
- Set chol_switch = TRUE; switch = TRUE;
- FOR $k = 1$ TO n_sys
  - IF chol_switch THEN
    compute $\boldsymbol{P}_0 = IC(\boldsymbol{A}_k)$; set chol_switch = FALSE;
  - IF switch THEN
    1. Compute the p leftmost eigenpairs by the DACG procedure with preconditioner $\boldsymbol{P}_0$ and accuracy τ_{DACG}.
    2. Form matrices $\boldsymbol{V}_p, \boldsymbol{\Lambda}_p$.
    3. Solve the $k-$th linear system by PCG preconditioned by $\boldsymbol{P}_0 + \boldsymbol{V}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{V}_p^T$.
    4. switch= FALSE.
  - IF $it_k > it_{\text{prec}}$ switch = TRUE
  - IF $it_k > it_{\text{chol}}$ chol_switch = TRUE

  END FOR

Algorithm 3 PCG with spectral Lanczos preconditioner

- INPUT: $it_{\text{prec}}, it_{\text{chol}}, m_{\max}, p$.
- Set chol_switch = TRUE; switch = TRUE;
- FOR $k = 1$ TO n_sys
  - IF chol_switch THEN
    compute $\boldsymbol{P}_0 = IC(\boldsymbol{A}_k)$; set chol_switch = FALSE;
  - IF switch THEN
    1. Solve the $k-$th linear system by the PCG method preconditioned by $\boldsymbol{P}_0$.
    2. Construct the tridiagonal Lanczos matrix $\boldsymbol{T}_m$, with $m = \min\{m_{\max}, it_k\}$.
    3. Extract from $\boldsymbol{T}_m$ and $\boldsymbol{W}_m$ the p smallest eigenpairs and form matrices $\boldsymbol{V}_p, \boldsymbol{\Lambda}_p$.
    4. switch= FALSE.
  - ELSE
    1. Solve the $k-$th linear system by PCG preconditioned by $\boldsymbol{P}_0 + \boldsymbol{V}_p \boldsymbol{\Lambda}_p^{-1} \boldsymbol{V}_p^T$.
  - IF $it_k > it_{\text{prec}}$ switch = TRUE
  - IF $it_k > it_{\text{chol}}$ chol_switch = TRUE

END FOR

we use only the $p = \{10, 20\}$ smallest eigenvalues and corresponding eigenvectors thus obtaining a $n \times p$ matrix $\boldsymbol{V}_p$ and a $p \times p$ diagonal matrix $\boldsymbol{\Lambda}_p$. The final algorithm is reported in Alg. algorithm 3.

4.4 Numerical results

In this section we illustrate the behavior of the spectral preconditioner on a sequence of linear systems arising in the discretization of (Equation (3.1a)). The code is written in Fortran 90. All the experiments were run on a 2 x Intel Xeon CPU E5645 at 2.40GHz (six core) and with 4GB RAM for each core. Times are expressed in seconds. The stopping criterion for the linear solver is independent of the preconditioner used and it is based on the relative residual:

$$\frac{\|\boldsymbol{A}_k \boldsymbol{x}_k - \boldsymbol{b}\|}{\|\boldsymbol{b}\|} < \varepsilon = 10^{-11}.$$

We solve the piecewise-constant sources test case for $\beta = 5$ because it is the case that presents the strongest time-variability of the transport density among the considered test cases. All the simulations employ a mesh $\mathcal{T}_h$ of 412417 nodes and $n_{nz} = 1647617$ nonzero elements. Efficient simulations that ensure stability of explicit Euler are obtained using an initial time step size $\Delta t^{(0)} = 10^{-3}$ and then increasing $\Delta t^{(k)}$ by a factor 1.05 at each time step up to a maximum value of $\Delta t^{(k)} = 10^{-1}$. This leads to a sequence of almost 4000 linear systems as in Equation (4.3) to reach equilibrium at the chosen tolerance τ.

For this type of problems, homogeneous Neumann boundary conditions are natural but lead to a singular system matrix, with a non trivial kernel containing the constant vectors $\boldsymbol{c}$. However, the use of unnatural Dirichlet conditions often yielded linear solver failures, due most probably to extreme matrix ill-conditioning. Hence, we employ homogeneous Neumann condition and guarantee the well-posedness of the resulting linear systems (and of the PCG process) by projecting the right hand side onto the range of $\boldsymbol{A}_k$, $R(\boldsymbol{A}_k)$, as follows:

$$\tilde{\boldsymbol{b}} = \boldsymbol{b} - \frac{\boldsymbol{c}^T \boldsymbol{b}}{\|\boldsymbol{c}\|^2} \boldsymbol{c}$$

(see [42]). Note that such projection simply corrects quadrature errors in the construction of $\boldsymbol{b}$, since f is assumed to have zero-means.

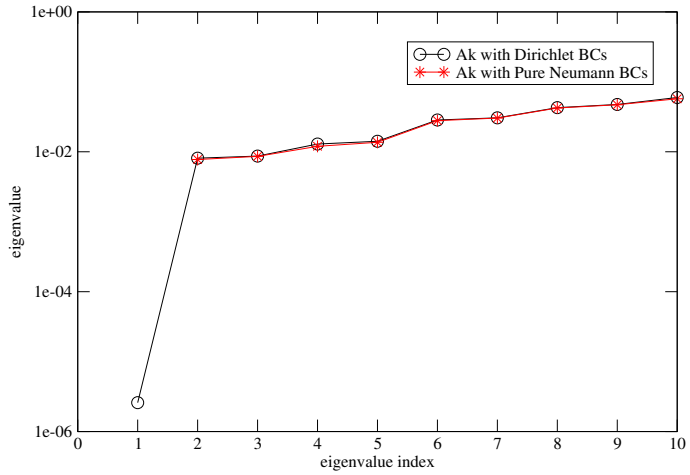

Figure 4.1: Effect of imposing Dirichlet boundary conditions on the smallest eigenvalues of $\boldsymbol{P}_k\boldsymbol{A}_k$, for system $k = 200$ in the sequence (no spectral acceleration).

Denoting with $\lambda_1 = 0 < \lambda_2 < \ldots \lambda_n$ the eigenvalues of $\boldsymbol{A}_k$ in this case, the effective spectral condition number of matrix $\boldsymbol{A}_k$ is $\kappa(\boldsymbol{A}_k) = \dfrac{\lambda_n}{\lambda_2}$ since the zero eigenvalue does not affect convergence of the PCG iteration after the projection of $\boldsymbol{b}$ on $R(\boldsymbol{A}_k)$. On the other hand, in the case of Dirichlet boundary conditions, all the eigenvalues of $\boldsymbol{A}_k^D$ change and the zero eigenvalue occurring in the Neumann case is moved by to a positive value close to zero (near 10^{-6} for the example shown in Figure 4.1), yielding a spectral condition number much greater than in the Neumann case: $\kappa(\boldsymbol{A}_k^D) = \dfrac{\lambda_n^D}{\lambda_1^D} \gg \kappa(\boldsymbol{A}_k)$.

All the linear systems have been symmetrically scaled with the diagonal of $\boldsymbol{A}_k$ in order to reduce their initial condition number, namely, defining $D = \mathrm{diag}(a_{11}, \ldots, a_{nn})$:

$$\begin{aligned} \text{solve} \quad & D^{-1/2}\boldsymbol{A}_k D^{-1/2}\boldsymbol{y}_k = D^{-1/2}\tilde{\boldsymbol{b}}, \\ \text{compute} \quad & \boldsymbol{x}_k = D^{1/2}\boldsymbol{y}_k. \end{aligned}$$

The efficacy of the proposed algorithms is verified by looking at the overall iteration count of the PCG solver and the CPU times for solving the entire linear systems sequence. We test different numbers of eigenvectors p used to build the low rank correction to the initial preconditioner for both algorithms based on the Lanczos (LAN(p)) and DACG (DACG(p)) eigensolution. We report CPU timings accounting for the computation of the preconditioner (T_{prec}), of the approximated

eigenvectors (T_{eig}), the PCG solver (T_{PCG}), and the total CPU time (T_{tot}).

4.4.1 Influence of eigenvector accuracy in DACG preprocessing

We first perform a preliminary study on the influence of accuracy of eigenvectors computations in the PCG acceleration. To this end we considered the first 200 linear systems and use three different tolerances for the relative eigenresidual test (4.5): $\tau_{DACG} \in \{0.1, 0.3, 0.5\}$. Other parameters were: $it_{eig} = 60, it_{chol} = 60, p = 20$. As a benchmark, we also solved the first 200 systems by the PCG method preconditioned by an IC factorization computed selectively (with $it_{chol} = 100$). The results are shown in Table 4.1. We find that high accuracy in eigenpairs computation is not needed to reduce the number of PCG iterations. With a very low accuracy ($\tau_{\text{DACG}} = 0.5$) the number of iterations is halved and the CPU time reduced of a factor 1.5 with respect to the fixed IC preconditioner (Table 4.1, first row).

τ_{DACG}	ITER	T_{eig}	T_{prec}	T_{PCG}	T_{tot}
–	20646	0.00	187.9	1687.8	1875.7
0.1	9907	326.9	117.9	1002.1	1446.2
0.3	10006	198.5	117.2	1011.5	1327.8
0.5	10055	150.0	117.2	1017.7	1284.9

Table 4.1: Influence of the DACG tolerance on the performance of the PCG with spectral preconditioner.

4.4.2 Smallest eigenvalues of $\boldsymbol{P}_0\boldsymbol{A}_k$

In Figure 4.2 we plot the computed eigenvalues of $\boldsymbol{P}_0\boldsymbol{A}_k$, where $\boldsymbol{P}_0$ is the IC preconditioner of $\boldsymbol{A}_k$, by the spectral Lanczos-PCG procedure. In particular we plot the 10 smallest eigenpairs of preconditioned systems $\#1, 21, 40, \ldots, 181$. From the figure we notice that the "stars" a re vertically clustered, this showing that the smallest eigenvalues of the preconditioned matrices only slightly change among systems at close simulation times.

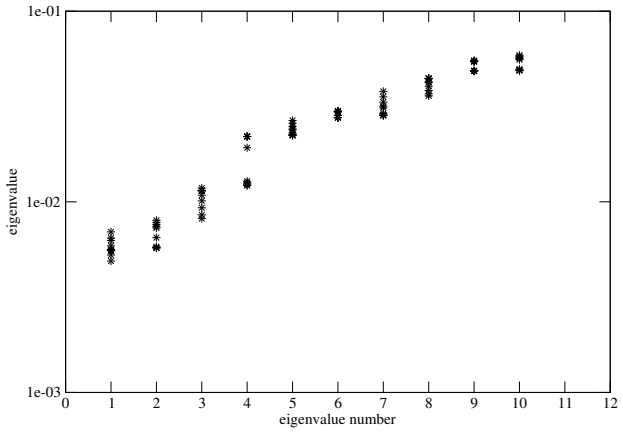

Figure 4.2: 10 smallest eigenvalues of $\boldsymbol{P}_k\boldsymbol{A}_k$ for $k = 20j + 1, j = 0, \ldots, 9$.

4.4.3 Results of the simulations

We report in Table 4.2 the results of the complete simulation corresponding to different values of the parameter β i.e. the cumulative number of PCG iterations and CPU times in solving the sequence of almost 4000 linear systems needed to reach the steady-state. In addition to the previously described parameters we used as the maximum size of the Lanczos subspace $m_{\max} = 80$, which experimentally revealed the optimal value.

Inspection of Table 4.2 reveals that both DACG and Lanczos acceleration provide an improvement in the number of iterations and total CPU time. For the easier $\beta = 1.5$ case we tried various values for parameters it_{chol} and it_{eig}. For the challenging case $\beta = 5$, the optimal spectral preconditioner turns out to be the one based on the Lanczos approach which provides a gain of more than 40% CPU time with respect to using the Cholesky preconditioner, computed at each time-step.

4.4.4 Further analysis on a portion of the simulation

To better inspect the optimal choice of the parameters we analyzed the first 800 time-steps, after which the solution is near its steady-state. From Table 4.3 we notice that the proposed low-rank update of preconditioners is effective in both variants, providing an important reduction of the number of iterations as well of the CPU time. On the average, our spectral preconditioners provide a halving of the total CPU time and a $30\% - 40\%$ reduction in the number of iterations.

Prec. (p)	it_{chol}	it_{eig}	ITER	T_{prec}	T_{eig}	T_{PCG}	T_{tot}
			$\beta = 1.5$				
DACG(10)	25	25	170111	6262.3	1869.0	14276.7	22758.7
LAN(10)	25	25	219515	11753.4	2077.5	17604.6	31800.1
IC	25	–	261964	16909.1	0.0	19348.9	36644.2
DACG(10)	30	30	190075	3453.1	1228.5	15949.9	20977.9
LAN(10)	30	30	230242	8080.1	1437.3	18699.8	28572.5
IC(10)	30	–	272779	13043.8	0.0	19981.0	33392.3
DACG(10)	40	40	232004	1660.2	811.1	19329.4	22146.1
LAN(10)	40	40	259569	3591.5	688.1	21420.5	26049.4
IC(10)	40	–	299100	9040.8	0.0	21858.1	31254.6
			$\beta = 5$				
DACG(10)	60	60	220110	1305.8	446.3	18853.5	20785.3
LAN(10)	60	60	208481	910.2	166.2	17834.9	19092.0
IC	60	–	263477	9532.6	0.0	19900.5	29632.0
IC	–	–	257219	13041.1	0.0	19415.9	32666.6

Table 4.2: Timings and iterations related to the whole sequence of linear systems corresponding to two different values of β using PCG with different preconditioners and parameters.

Using $p = 10$ or $p = 20$ eigenvectors produces only slight variations in the number of iterations/CPU time. Hence, the choice $p = 10$ seems to be preferred in terms of memory storage.

Surprisingly, the DACG variant, although affected by a CPU-intensive offline (outside the PCG algorithm) phase for the eigenvector approximation, reveals as effective as the Lanczos variant. This is mainly due to the fact that after the initial assessment of the leftmost eigenpairs, the subsequent computations are very cheap since the previously computed eigenvectors are very good initial guesses for the next systems. However, we may expect a different behavior of the two techniques in cases of higher variations of the matrices involved. In this case, the DACG preprocessing time will increase as opposite to the Lanczos technique. Moreover, the Lanczos approach can be accelerated by employing a method similar to that described in [67]. This is a topic for a future work.

Prec. (p)	it_{chol}	it_{eig}	ITER	T_{eig}	T_{prec}	T_{PCG}	T_{tot}
IC	–	–	64248	0.0	2841.2	5432.5	8273.7
IC	100	–	74511	0.0	447.4	6062.9	6510.3
LAN(10)	–	60	41148	29.6	2814.2	3372.1	6215.9
LAN(10)	50	70	44765	30.9	1767.2	3746.4	5544.5
LAN(10)	60	60	44041	196.0	572.7	3606.9	4375.6
LAN(20)	60	60	41738	190.4	459.2	3775.8	4425.4
DACG(10)	60	60	45502	185.4	516.0	3811.6	4512.9
DACG(20)	60	60	42050	263.5	272.2	3922.0	4457.7

Table 4.3: Timings and iterations related to the first 800 systems in the sequence corresponding to $\beta = 5$. PCG with different preconditioners and parameters.

4.4.5 Handling high density variations

As clear from Figure 4.3 there is a portion of the simulation in which the largest value of the density vector abruptly increases and rapidly reaches its maximum value. Since, as anticipated in Section 4.1, $\lambda_{\min}(\boldsymbol{A}) \leq C_1 h^2 \mu_{\min}$ and $\lambda_{\max}(\boldsymbol{A}) \geq C_2 \mu_{\max}$, the sudden increase of $\mu_{\max}$ produces a high variation in the condition number of the matrices in the sequence. Hence, in this time interval, the spectral properties of the system matrices change significantly and PCG is not able to take advantage of the spectral information provided by the previous systems.

To this aim we forced the code to recompute the Cholesky preconditioner whenever the following test on μ is satisfied:

$$\frac{\|\mu^{(k+1)} - \mu^{(k)}\|_{L^2(\Omega}}{\Delta t_k \|\mu^{(k+1)}\|_{L^2(\Omega}} > \delta \tag{4.10}$$

with $\delta = 100$ in the experiments. Also the time-step is dynamically reduced in this portion of the simulation to correctly capture the dynamics.
To appreciate the benefit of this modification we report in Figure 4.4 the number of iterations (averaged over the last 5 linear systems) needed by the PCG solver for systems #40 to #80 (corresponding to time interval: $[4, 5.43]$ where the μ variation is more pronounced). We display in the Figure results with the Cholesky preconditioner only and Lanczos(10) and DACG(10) spectral acceleration with and without test (4.10).

The new switching strategy is clearly effective and particularly so in combi-

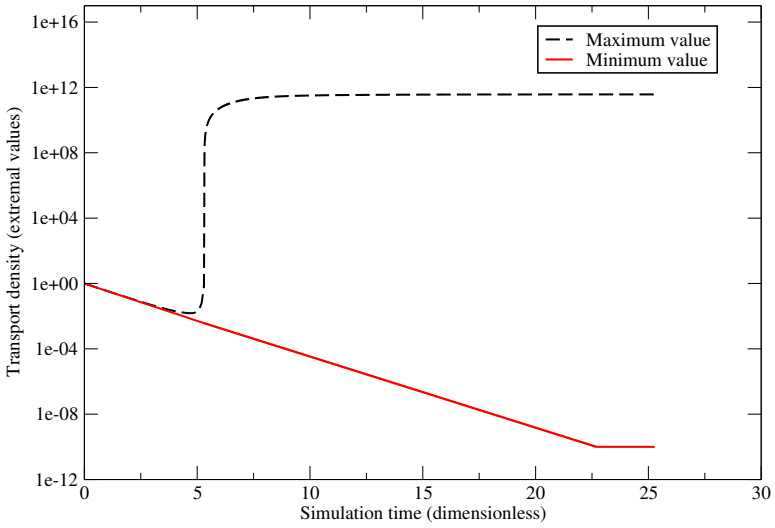

Figure 4.3: Maximum and minimum value of μ during the simulation.

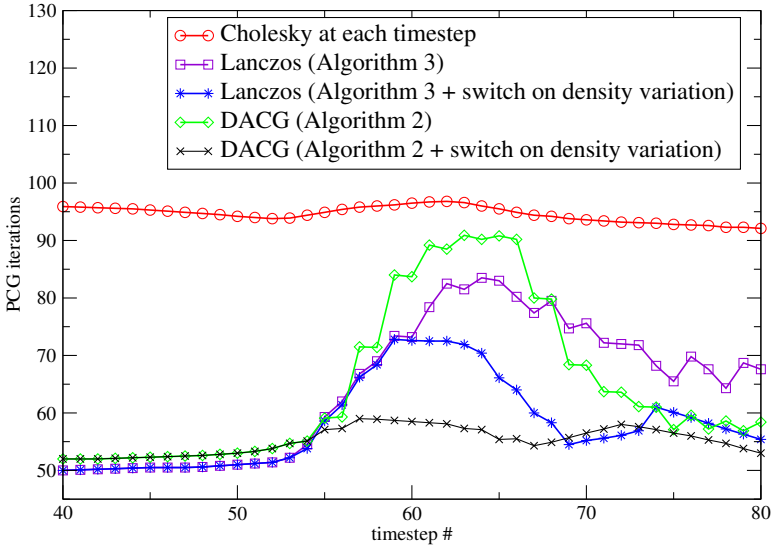

Figure 4.4: PCG iterations in solving systems #40 to #80 with various spectral and switching strategies.

nation with the DACG spectral acceleration.

Chapter 5

The Gradient Flow Approach

In this chapter we present a modification of the dynamical system Equation (3.1) which reads as

$$-\operatorname{div}(\mu(t,x)\nabla u(t,x)) = f(x) = f^+(x) - f^-(x) \tag{5.1a}$$

$$\partial_t \mu(t,x) = \mu^\beta(t,x)|\nabla u(t,x)|^2 - \mu(t,x) \tag{5.1b}$$

$$\mu(0,x) = \mu_0(x) \quad \mu(t,x)\nabla u(t,x)\cdot n_{\partial\Omega} = 0 \tag{5.1c}$$

where, unlike System 3.1, only the term μ is raised to the power β, while the $|\nabla u|^2$ is fixed. Equation (5.1) admits the Lyapunov-candidate functional given by:

$$\Phi_\beta(\mu) = \mathcal{E}_f(\mu) + \mathcal{M}_\beta(\mu) \tag{5.2}$$

$$\mathcal{M}_\beta(\mu) := \begin{cases} \frac{1}{2}\int_\Omega \ln(\mu) & \text{if } \beta = 2 \\ \frac{1}{2}\int_\Omega \frac{\mu^{2-\beta}}{2-\beta} & \text{otherwise} \end{cases}$$

If we introduced an appropriate change of variable, i.e., an appropriate metric for μ, then the formal only calculations show that Equation (5.1b) is the *Gradient Flow* (GF) of the functional Φ_β, for $0 < \beta < 2$. We are not able to present a complete proof of this claim, but many considerations presented below towards the correctness of this conjecture.

5.1 Brief Introduction to Gradient Flow

We now present a very brief introduction to the main ideas of the Gradient Flow in metric spaces. We refer the reader to [3, 65] for a complete overview on this

topic and its applications.

The GF in metric spaces may be thought of as the infinite-dimension extension of the following, finite dimensional, Cauchy Problem. Take $\Phi : \mathbb{R}^d \mapsto \mathbb{R}$ sufficiently smooth and consider the ODE

$$\partial_t x(t) = -\nabla \Phi(x(t)) \quad x(0) = x_0 \in \mathbb{R}^d$$

the equations describes a curve that follows the steepest descent trajectory of functional Φ. Fixing a time step $\tau > 0$, its backward Euler discretization consists in build the sequence $(x_k)_{k=0,\dots,}$ defined by the following recursion:

$$\frac{x_{k+1} - x_k}{\tau} = -\nabla \Phi(x_{k+1}) \quad k = 0, \dots$$

For each k this equation is the Euler-Lagrange equation of the functional

$$J_{x_k,\tau}(x) = \Phi(x) + \frac{|x - x_k|^2}{2\tau}$$

This variational structure of the implicit Euler Time-Stepping can be used also for functionals $\Phi : X \mapsto \mathbb{R}$ defined on a general metric space X. Under the proper assumptions, the linear interpolation of the sequence generated by the variational problems given by:

$$x_{k+1} = \operatorname*{argmin}_{x \in X} \mathcal{J}_{x_k,\tau}(x) = \Phi(x) + \frac{\|x - x_k\|_X^2}{2\tau}$$

converges, for $\tau \to 0$, to a curve $x(t) \in X$ for $t \in [0, +\infty[$ that solves the ODE in metric space

$$\partial_t x(t) = -\nabla_x \Phi(x(t)) \quad x(0) = x_0 \in X \tag{5.3}$$

with a proper notion of gradient in the space X. A typical requirement to ensure that the discrete sequence x_k converges to the "curve" $x(t)$ as τ tends to zero, is that the functional Φ be *geodesically λ-convex* (see [3]), i.e., there exists $\lambda \in \mathbb{R}$ such that

$$\Phi((1-s)x_0 + sx_1) \leq (1-s)\Phi(x_0) + s\Phi(x_1) - \frac{1}{2}\lambda s(1-s)\|x_0 - x_1\|_X$$
$$\forall x_0, x_1 \in X, \ \forall s \in [0,1]$$

5.2 Case $\beta = 1$ - The Hellinger/Fisher-Rao Metric

We first consider the case $\beta = 1$ and we see how to reinterpret Equation (5.1b) as a GF of the form

$$\mu'(t) = -\nabla_\mu \Phi_1(\mu(t))$$

where Φ_1 is defined in Equation (5.2). The basic idea is the introduction of a metric for μ induced by a change of variable $\mu = \Psi(\sigma)$. The same approach can used to show analogous results not only for general $\beta > 0$, but also for more complicated versions of Equation (5.1b), but their interpretation becomes at least questionable, the most basic result presented here needs to be completely unveiled before addressing more complicated dynamics.

We assume that for $\beta = 1$ the domain of Equation (5.1b) is $L^1_+(\Omega)$ and the change of variable we consider is

$$\mu = \Psi(\sigma) := \frac{\sigma^2}{2}$$

We find immediately that $\sigma \in L^2_+(\Omega)$. The transformation Ψ induces a distance in $L^1(\Omega)$ given by the following metric:

$$\|\mu_1 - \mu_2\|_{FR} := \| \underbrace{\sigma_1}_{\Psi^{-1}(\mu_1)} - \underbrace{\sigma_2}_{\Psi^{-1}(\mu_2)} \|_{L^2(\Omega)} = \sqrt{2} \left(\int_\Omega (\sqrt{\mu_1} - \sqrt{\mu_2})^2 \, dx \right)^{\frac{1}{2}} \tag{5.4}$$

called Fisher-Rao (or Hellinger) distance. Rewriting Equation (5.1b) in terms of the new variable σ we first obtain

$$\sigma(t)\sigma'(t) = \frac{1}{2} \left(\sigma^2(t) |\nabla u(t)|^2 - \sigma^2(t) \right)$$

and then, by dividing by $\sigma(t)$,

$$\begin{cases} \sigma'(t) = \frac{1}{2} \left(\sigma(t) |\nabla u(t)|^2 - \sigma(t) \right) \\ \sigma(0) = \sigma_0 = \sqrt{2\mu_0} \in L^2_+(\Omega) \end{cases} \tag{5.5}$$

with $u(t) = u(\mu(t)) = u(\Psi(\sigma(t)))$. In the following, given a functional F in μ we will use the symbol $\tilde{F}$ to indicate its composition with Ψ. Thus $\tilde{\mathcal{E}_f}(\sigma) = \mathcal{E}_f(\Psi(\sigma))$, $\tilde{\mathcal{M}} = \mathcal{M}(\Psi(\sigma))$, and $\tilde{\Phi} = \Phi(\Psi(\sigma))$. We can now formulate the following result, which claims that Equation (5.5) can be interpreted as a GF in $L^2_+(\Omega)$.

Consequently the same holds in the μ variable with the metric induced by Ψ. There are several critical points in the following proof, essentially because the L^2-norm is too weak to ensure continuity of $\tilde{\mathcal{E}}_f(\sigma)$. In any case it can be a heuristic explanation to introduce the functional Φ and ensuing GF. During the derivation we will detail the critical points that prevents a full proof of this statement. We are looking for alternative avenues, but at this point we are able to show only formal calculations. We thus want to derive the fact that the ODE Equation (5.5) is a GF in $L^2_+(\Omega)$ of the form

$$\sigma'(t) = -\nabla_\sigma \tilde{\Phi}(\sigma(t)) \tag{5.6}$$

where

$$\tilde{\Phi}(\sigma) = \tilde{\mathcal{E}}_f(\sigma) + \tilde{\mathcal{M}}(\sigma)$$

The formal derivation is based on the observation that $\tilde{\Phi}(\sigma)$ is the sum of the energy functional $\tilde{\mathcal{E}}_f(\sigma)$ and the mass functional $\tilde{\mathcal{M}}(\sigma) = 1/2 \int_\Omega \sigma^2/2$.

The gradient with respect to σ of $\tilde{\mathcal{M}}(\sigma)$ is clearly $\sigma/2$. The gradient of the energy $\tilde{\mathcal{E}}_f(\sigma)$ can be calculated as follows: consider $\sigma_\varepsilon = \sigma^* + \varepsilon\zeta$ with $\zeta \in L^2(\Omega)$ and $\varepsilon > 0$ small enough. Note that unfortunately this definition is such that μ_ε does not belong to $L^2_+(\Omega)$ for all $\zeta \in L^2(\Omega)$). Denote with $u_\varepsilon = u(\mu_\varepsilon)$, $(\mu_\varepsilon = (\sigma_\varepsilon)^2/2)$ and $u^* = u(\mu^*)$ $(\mu^* = (\sigma^*)^2/2)$. The Euler-Lagrange equations of the variational problem in Equation (1.11) are:

$$\int_\Omega \left(\frac{\sigma_\varepsilon^2}{2} \nabla u_\varepsilon \cdot \nabla \varphi - f\varphi \right) dx = 0 \quad \forall \varphi \in \mathcal{C}^1(\bar{\Omega})$$

Differentiating with respect to ε and applying the chain rule (assumed to be valid in this context) we obtain

$$\int_\Omega \frac{\sigma_\varepsilon^2}{2} (\partial_\varepsilon \nabla u_\varepsilon) \cdot \nabla \varphi \, dx = -\int_\Omega \partial_\varepsilon \left(\frac{\sigma_\varepsilon^2}{2} \right) \nabla u_\varepsilon \cdot \nabla \varphi \, dx \quad \forall \varphi \in \mathcal{C}^1(\bar{\Omega}) \tag{5.7}$$

The first variation of the energy is given by

$$\begin{aligned} \frac{\partial \tilde{\mathcal{E}}_f(\sigma_\varepsilon)}{\partial \varepsilon}|_{\varepsilon=0} &= \left(\frac{1}{2} \int_\Omega \partial_\varepsilon \left(\frac{\sigma_\varepsilon^2}{2} |\nabla u_\varepsilon|^2 \right) dx \right)_{\varepsilon=0} \\ &= \left(\frac{1}{2} \int_\Omega \partial_\varepsilon \left(\frac{\sigma_\varepsilon^2}{2} \right) |\nabla u_\varepsilon|^2 + 2\frac{\sigma_\varepsilon^2}{2} \partial_\varepsilon (\nabla u_\varepsilon) \cdot \nabla u_\varepsilon \, dx \right)_{\varepsilon=0} \end{aligned}$$

Using in Equation (5.7) $\varphi = u^*$ we get (again here we are assuming this possible)

$$\begin{aligned}\frac{\partial \tilde{\mathcal{E}}_f(\sigma_\varepsilon)}{\partial \varepsilon}|_{\varepsilon=0} &= \left(\frac{1}{2}\int_\Omega \partial_\varepsilon\left(\frac{\sigma_\varepsilon^2}{2}\right)|\nabla u_\varepsilon|^2 - 2\partial_\varepsilon\left(\frac{\sigma_\varepsilon^2}{2}\right)\nabla u_\varepsilon \cdot \nabla u^* \, dx\right)_{\varepsilon=0} \\ &= \left(\frac{1}{2}\int_\Omega \sigma_\varepsilon \partial_\varepsilon(\sigma_\varepsilon)\nabla u_\varepsilon \cdot (\nabla u_\varepsilon - 2\nabla u^*)\, dx\right)_{\varepsilon=0} \\ &= \left(\frac{1}{2}\int_\Omega \sigma_\varepsilon \zeta \nabla u_\varepsilon \cdot (\nabla u_\varepsilon - 2\nabla u^*)\, dx\right)_{\varepsilon=0}\end{aligned}$$

and evaluating in $\varepsilon = 0$ we obtain (assuming $\nabla u_\varepsilon|_{\varepsilon=0} = \nabla u^*$)

$$\frac{\partial \tilde{\mathcal{E}}_f(\sigma_\varepsilon)}{\partial \varepsilon}|_{\varepsilon=0} = \int_\Omega -\zeta \frac{\sigma^*|\nabla u^*|^2}{2}\, dx$$

Since this holds for any $\zeta \in L^2(\Omega)$, this means that

$$\nabla_\sigma \tilde{\mathcal{E}}_f(\sigma) = -\frac{\sigma}{2}|\nabla u(\Psi(\sigma))|^2$$

In conclusion Equation (5.5) can be rewritten as

$$\partial_t \sigma(t) = -\nabla_\sigma \tilde{\Phi}(\sigma(t)) = -\nabla_\sigma\left(\tilde{\mathcal{E}}_f(\sigma(t)) + \tilde{\mathcal{M}}(\sigma(t))\right)$$

Several inconsistent assumptions have been made to arrive at this point, but these observations strengthen our speculation that the model we are proposing is a sound OTP model with several interesting properties.

5.3 More Gradient Flows

We now consider a further generalization of Equation (5.1b) given by

$$\mu'(t) = g(\mu(t))|\nabla u(t)|^2 - h(\mu(t)) \tag{5.8}$$

where $g, h :]0, +\infty[\mapsto]0, +\infty[$ are functions of $\mathbb{R}^+$ into itself. We assume that both functions are regular enough to allow the following computations. By using the same procedure described in the previous derivation, it is possible to rewrite Equation (5.8) in the form of a GF. To this aim we need to find a transformation $\mu = \Psi(\sigma)$ (with Ψ a diffeomorphism of $\mathbb{R}^+$ into itself) that induces a new metric for μ

$$\|\mu_1 - \mu_2\|_\Psi := \|\Psi^{-1}(\mu_1) - \Psi^{-1}(\mu_2)\| = \|\sigma_1 - \sigma_2\|_{L^2(\Omega)} \tag{5.9}$$

given by the L^2-distance. The results are given in the following proposition:

Proposition 44. *Given g, h functions from $\mathbb{R}^+$ into itself, if there exists a diffeomorphism $\Psi : \mathbb{R}^+ \mapsto \mathbb{R}^+$ with $\Psi' > 0$ (or $\Psi' < 0$) such that*

$$\frac{(\Psi'(\sigma))^2}{2} = g(\Psi(\sigma)) \tag{5.10}$$

then Equation (5.8) can be rewritten in the form of a GF

$$\mu'(t) = -\nabla_\mu \Phi_{g,h}(\mu(t)) \tag{5.11}$$

in the space $\Psi(L^2_+(\Omega))$ (the set of μ such that $\Psi^{-1}(\mu) \in L^2_+(\Omega)$) endowed with the metric defined in Equation (5.9). The functional $\Phi_{g,h}$ is defined as

$$\Phi_{g,h}(\mu) := \mathcal{E}_f(\mu) + \mathcal{M}_{g,h}(\mu) \tag{5.12}$$

with

$$\mathcal{M}_{g,h}(\mu) := \frac{1}{2}\int_\Omega I(\mu)\,dx \qquad I(s) = \int \frac{h(s)}{g(s)} ds \tag{5.13}$$

In the variable $\sigma = \Psi^{-1}(\mu)$ Equation (5.11) can be rewritten as a GF of the form

$$\sigma'(t) = -\nabla_\sigma \tilde{\Phi}_{g,h}(\sigma(t))$$

in $L^2_+(\Omega)$ where $\tilde{\Phi}_{g,h} = \Phi_{g,h} \circ \Psi$

We recall again that these are all formal calculations where all the necessary regularity hypothesis are implicitly assumed.

Proof. Assuming that a diffeomorphism Ψ exists, satisfies Equation (5.10), and is such that $\Psi' > 0$, we can substitute $\mu(t) = \Psi(\sigma(t))$ into

$$\mu'(t) = g\left(\mu(t)\right) |\nabla u(t)|^2 - h\left(\mu(t)\right)$$

Then, expanding all the derivatives and dividing by $\Psi'(\sigma(t))$, we obtain

$$\sigma'(t) = \frac{g\left(\Psi\left(\sigma(t)\right)\right)}{\Psi'(\sigma(t))} |\nabla u(t)|^2 - \frac{h\left(\Psi(\sigma(t))\right)}{\Psi'(\sigma(t))} \tag{5.14}$$

where $u(t) = u(\Psi(\sigma(t)))$. The proof follows by showing that the first and the second terms on the right hand side of Equation (5.14) are respectively the gradient of the functional $\tilde{\mathcal{E}}_f$ and $\tilde{\mathcal{M}}$, evaluated along $\sigma(t)$. Thus we first compute the gradient with respect to σ of $\tilde{\mathcal{E}}_f(\sigma)$, with the same procedure used in Section 5.2, where we obtained

$$\nabla_\sigma \tilde{\mathcal{E}}_f(\sigma) = -\frac{\sigma}{2} |\nabla u((\sigma)^2/2)|^2$$

that in this case reads as follows:

$$\nabla_\sigma \tilde{\mathcal{E}}_f(\sigma) = -\frac{\Psi'(\sigma)}{2} |\nabla u(\Psi(\sigma))|^2$$

Thanks to Equation (5.10) we obtain

$$\nabla_\sigma \tilde{\mathcal{E}}_f(\sigma) = -\frac{g(\Psi(\sigma))}{\Psi'(\sigma)} |\nabla u(\Psi(\sigma))|^2$$

Finally the gradient of $\tilde{\mathcal{M}}_{g,h}$ can be calculated formally as

$$\nabla_\sigma \tilde{\mathcal{M}}_{g,h}(\sigma) = \nabla_\mu \mathcal{M}_{g,h}(\mu)|_{\mu=\Psi(\sigma)} \Psi'(\sigma) = \frac{h(\Psi(\sigma))}{2g(\Psi(\sigma))} \Psi'(\sigma) = \frac{h(\Psi(\sigma))}{\Psi'(\sigma)}$$

where in the last step we used again Equation (5.10). □

Remark 6. *Note that the metric for μ is uniquely defined by the transformation Ψ, that depends only on function g by Equation* (5.10)*, while the function h does not affect the metric.*

5.3.1 Case $g = \mu^\beta$ $\beta > 0$, $h = \mu$

We are now able to reinterpret Equation (5.1b) as GF simply by taking $g(\mu) = \mu^\beta$ and $h(\mu) = \mu$ in Equation (5.8). The transformation Ψ defined by Equation (5.10) takes on the form

$$\mu = \Psi(\sigma) = \begin{cases} C_2(\beta)\sigma^{C_1(\beta)} & \text{if } \beta \neq 2 \\ \exp\left(\pm\sqrt{2}\sigma\right) & \text{if } \beta = 2 \end{cases} \tag{5.15}$$

with

$$C_1(\beta) = \frac{2}{2-\beta} \qquad C_2(\beta) = \begin{cases} \left(\frac{2-\beta}{\sqrt{2}}\right)^{C_1(\beta)} & \text{if } \beta < 2 \\ \left(\frac{\beta-2}{\sqrt{2}}\right)^{C_1(\beta)} & \text{if } \beta > 2 \end{cases}$$

We use the change of variable Ψ defined in Equation (5.15) and assume that $\sigma(t) \in L^2_+(\Omega)$. We can introduce the distance

$$\|\mu_1 - \mu_2\|_\beta := \|\sigma_1 - \sigma_2\|_{L^2(\Omega)} = C_\beta \left(\int_\Omega \left(\mu_1^{\frac{2-\beta}{2}} - \mu_2^{\frac{2-\beta}{2}} \right)^2 dx \right)^{1/2} \tag{5.16}$$

where we indicate with C_β a constant depending only on β. Moreover, computing $\mathcal{M}_{g,h}$ in Equation (5.13) we recover the functional $\mathcal{M}_\beta$ given in Equation (5.2).

We are now able to reinterpret system 5.1 as a GF in $L^{2-\beta}(\Omega)$ given by

$$\partial_t \mu(t) = \mu^\beta(t) |\nabla u(t)|^2 - \mu(t) = -\nabla_\mu \Phi_\beta(\mu(t)) \tag{5.17}$$

$$\Phi_\beta(\mu) := \mathcal{E}_f(\mu) + \frac{1}{2}\int_\Omega \frac{\mu^{2-\beta}}{2-\beta}\, dx$$

Using the variable σ, the ODE and the functional $\tilde{\Phi}_{g,h} = \tilde{\Phi}_{g,h} \circ \Psi$ become

$$\sigma'(t) = C_\beta \left(\sigma^{\frac{\beta}{2-\beta}}(t) |\nabla u(t)|^2 - \sigma(t) \right) = -\nabla_\sigma \tilde{\Phi}_\beta(\sigma(t))$$

$$\tilde{\Phi}_\beta(\sigma) := \tilde{\mathcal{E}}_f(\sigma) + \left(\frac{2-\beta}{2}\right)^2 \int_\Omega \sigma^2\, dx$$

The connection with OTPs for an asymptotic state of Equation (5.1) follows the same ideas described in Chapter 3. We reproposed below the statements, mutata mutandis, of Propositions 38 and 39, together with Conjectures 2 and 3

Proposition 45. *The derivative along the $\mu(t)$ trajectory of functional $\mathcal{L}_\beta$ is given by*

$$\frac{d}{dt}\left(\mathcal{L}_\beta(\mu(t))\right) = -\frac{1}{2}\int_\Omega \mu(t)^\beta \left(|\nabla u(\mu(t))|^2 - \mu^{1-\beta}(t) \right)^2 dx$$

and it is strictly decreasing in time. For $\sigma = \Psi^{-1}(\mu)$ the above expression rewrites as:

$$\frac{d}{dt}\left(\tilde{\mathcal{L}}_\beta(\sigma(t))\right) = -\frac{1}{2}\int_\Omega \left(\partial_t \sigma(t)\right)^2 dx$$

Proposition 46. *For $0 < \beta \leq 1$ and for $q = 2\frac{2-\beta}{3-\beta}$ the following equality holds*

$$\inf_{v \in (L^q(\Omega))^d} \left\{ \int_\Omega \frac{|v|^q}{q}\, dx \ : \ \operatorname{div}(v) = f \right\} = \inf_{\mu \in L^{2-\beta}_+(\Omega)} \Phi_\beta(\mu)$$

For $0 < \beta < 1$, $\mathcal{L}_\beta$ admits a unique minimizer $\mu^*_\beta \in L^{2-\beta}_+(\Omega)$ of given by

$$\mu^*_\beta = |\nabla u_p|^{p-2}$$

where u_p is the solution p-Laplacian

$$-\operatorname{div}(|\nabla u_p|^{p-2} \nabla u_p) = f$$

with p the conjugate exponent of q:

$$p = 2\frac{2-\beta}{1-\beta}$$

For $\beta = 1$ the OT density $\mu^*(f^+, f^-)$ is a minimum for Φ.

Conjecture 4. *For $0 < \beta \leq 1$ Equation (5.1b) represents the GF in $L^{2-\beta}(\Omega)$ of functional*

$$\Phi_\beta(\mu) = \mathcal{E}_f(\mu) + \frac{1}{2}\int_\Omega \frac{\mu^{2-\beta}}{2-\beta}\,dx$$

with metric defined in Equation (5.16). For $0 < \beta < 1$ the pair $(\mu(t), u(t))$ solution of Equation (5.1) converges toward $(|\nabla u_p|^{p-2}, u_p)$ where u_p is the solution of the p-Poisson equation with

$$p = 2\frac{(2-\beta)}{1-\beta}$$

For $\beta = 1$ the pair $(\mu(t), u(t))$ converges towards the pair (μ^*, u^*) solution associate the MK equations associated to f^+, f^-. For $0 < \beta \leq 1$ the convergence does not depend on the initial data μ_0 in Equation (5.1c).

Conjecture 5. *For $\beta > 1$, the solution $(\mu(t), u(t)$ of Equation (3.1) admits an equilibrium point (μ^*_β, u^*_β), which depends on the initial condition μ_0. This solution is a minimum of the Lyapunov-candidate functional Φ_β.*

5.4 Technical difficulties in the formal application of the GF approach

The main idea in applying the GF method is to use the Implicit Euler scheme to obtain a sequence (μ_k) whose convergence towards the minimum of Φ_β, if it exists, can be proved. To this aim, as described in [65], we first derive the variational formulation of the Implicit Euler scheme that approximates the GF in Equation (5.17). We fix $\tau > 0$ and write the k-th step of Euler scheme as:

$$\begin{aligned}\mu_{k+1} &= \underset{\mu \in L^{2-\beta}_+(\Omega)}{\operatorname{argmin}} \left\{ \Phi_\beta(\mu) + \frac{\|\mu - \mu_k\|^2_\beta}{2\tau} \right\} \\ &= \underset{\mu \in L^{2-\beta}_+(\Omega)}{\operatorname{argmin}} \left\{ \mathcal{E}_f(\mu) + \mathcal{M}_\beta(\mu) + \frac{\|\mu - \mu_k\|^2_\beta}{2\tau} \right\}\end{aligned} \tag{5.18}$$

where $\|\cdot\|_\beta$ is defined in Equation (5.16). This problem can be written in terms of the variable σ as:

$$\begin{aligned}\sigma_{k+1} &= \underset{\sigma \in L^2_+(\Omega)}{\operatorname{argmin}} \left\{ \tilde{\Phi}_\beta(\sigma) + \frac{\|\sigma - \sigma_k\|^2_{L^2(\Omega)}}{2\tau} \right\} \\ &= \underset{\sigma \in L^2_+(\Omega)}{\operatorname{argmin}} \left\{ \tilde{\mathcal{E}}_f(\sigma) + \tilde{\mathcal{M}}_\beta(\sigma) + \frac{\|\sigma - \sigma_k\|^2_{L^2(\Omega)}}{2\tau} \right\}\end{aligned} \tag{5.19}$$

Unfortunately, we are not able to prove the existence of a minimizer μ_k for the variational problems in Equations (5.18) and (5.19), already for the first time step $k = 0$.

To exemplify the difficulties encountered, consider the case $\beta = 1$, and the corresponding variational problem in Equation (5.19) written in terms of the variable $\sigma \in L^2_+(\Omega)$. The mass and distance functionals given by

$$\tilde{\mathcal{M}}_1(\sigma) = 1/2 \int_\Omega \sigma^2/2 \, dx \quad , \quad \frac{\|\sigma - \sigma_k\|^2_{L^2(\Omega)}}{2\tau}$$

are clearly lower semi-continuous and λ-geodesically convex, with λ equal to 1 and $1/2\tau$, respectively. But the energy functional $\tilde{\mathcal{E}}_f(\sigma)$, given by:

$$\tilde{\mathcal{E}}_f(\sigma) = \sup_{\varphi \in \mathcal{C}^1(\bar{\Omega})} \int_\Omega \left(f\varphi - \frac{\sigma^2}{2} \frac{|\nabla \varphi|^2}{2} \right) dx$$

is not lower semi-continuous since the term $\int_\Omega \sigma^2/2 |\nabla \varphi|^2/2 \, dx$ is not. However, if we look at the variational Equation (5.18) written in terms of the variable $\mu \in L^1(\Omega)$, the functional $\mathcal{E}_f(\mu)$ is lower semi-continuous (being the supremum of functionals that are linear with respect to μ) and convex, and the same holds for the corresponding mass and distance functionals

$$\mathcal{M}_1(\mu) \quad , \quad \frac{\|\mu - \mu_k\|^2_{FR}}{2\tau}$$

In this case, since $L^1(\Omega)$ is not reflexive, we cannot apply the direct method of the calculus of variations. On the other hand, the dichotomy described above obtained by switching variables, whereby one functional is lower semi-continuous in one variable but not the other, seems to suggest that these are technical difficulties that may be overcome by means of finer variational tools, possibly on appropriately relaxed versions of the problem. The case $\beta > 1$ is evidently more complicated because of the presence of the concave factor $\int_\Omega \mu^{2-\beta} \, dx$. We leave these issues for future work.

Chapter 6

Applications

In this chapter we discuss some applications of the optimal transport theory that were initiated but not completed during the course of this study. Some of these applications motivated the initial interest in the OT theory. In fact, our early interests concerned in particular two phenomena: i) geomorphology of river networks and ii) dynamics of plant roots in soils. These two problems, of great importance in the general field of climate research, have a long scientific history. The former started at the end of the 19th century with the first experimental observations of Horton [41, 46], arriving to more recent work of [59], summarized in the context of Optimal Transport by [61]. The history of plant root models is more recent but not less rich. The second half of the last century, starting with the work of [40], has seen a great activity on the modeling of root dynamics. The complex mechanisms leading to root branching structures has been studied in details [76], and several attempts at modeling these structures have been proposed [28, 43]. We noted that the theory of BTP seems to be ideally suited to these applications.

More broadly, we have identified a number of applications such as angiogenesis, formation of the Purkinje network, and in general in the dynamics of complex natural transport structures. After the two main application topics of geomorphology and root dynamics, we report a short discussion on general issues related to complex networks. In the overview of the BTP presented in Section 1.5 we mentioned how these types of mathematical problems have been introduced trying to give a common explanation to the recurrent emergence of branching structures in natural systems. According to the least-action principle, this attempt is done

describing these branching structures as solutions of minimal energy problems, fitting perfectly within the general framework of our BTP models.

6.1 Geomorphology of river basins

In the study of river network the idea of looking for general energy minimizing principle is pervasive in the all the work on Optimal Channel Networks well described in [48] and summarized in [60]. There it is suggested that river networks are solution of the following minimization problem:

$$\min_{Q} \sum_{e \in G} Q_e^{\frac{1}{2}} L_e$$

where G denotes a graph that schematizes the river network, while Q_e and L_e are, respectively, the flux of water passing through each edge and the edge length. The flux Q_e satisfies the water conservation principle stating that rainfall volumes must be conserved through the domain.

These results clearly connect the study of river network with the Gilbert-Steiner Problem described in Problem 23. The minimal-energy properties of the river network is deduced by looking at the steady state equations of a dynamics describing the landscape evolution of river basin. Since this is the same idea behind our model (the optimal solution is an equilibrium point), our claim is that the coupled system of equations of our model (or some ad-hoc modifications) can represent the evolution of river catchments leading naturally to a model for the generation of the river network.

We found a first attempt that points to this direction in [21], where the authors encode into a system of three partial differential equations the main phenomena shaping the formation of river network that, according to literature on this topic, are erosion, sedimentation, and creep, together with the conservation laws for water and sediment.

Trying to recast everything into the framework of our BTP model, f^+ and f^- represent the rain-fall rate on the river basin and the water efflux into the estuary or the delta, respectively. The elliptic equation plays the role of a stationary balance stating conservation of sediments and water flowing together. The diffusion coefficient μ represents the spatially varying flow capacity of the basin. The dynamics for $\mu(t)$ should encase all the main evolutive equations that governs

the formation of the river networks. Then the branching structures of the river networks arise as minimizers of the Lyapunov-candidate functional $\mathcal{L}_\beta$.

To test the capability of our BTP model, we proceed to simulate the formation of the Po river network in Northern Italy. The aim here is to test the hypothesis that with minimal geometric and hydrologic information the BTP model is able to reconstruct a network structure with branching characteristics that are similar, at least in an intuitive way, to observations. To this aim we have drastically simplified the geometry, shown in Figure 6.1, assuming an unrealistically symmetric pattern of the mountain and plain regions. We have separated the Po network basin from the delta area in the Adriatic Sea to represent, always schematically, the presence of the Adige river that does not contribute to the Po river basin. Rainfall rates are assumed unitary in the mountains and are halved in the central plain, to simulate the effect of groundwater recharge with the hypothesis that half of the rainfall infiltrates. The mass balance is imposed introducing a sink term f^- supported in the right rectangle in Figure 6.1 that approximately delimits the area of the Po Delta. We used a BTP exponent $\beta = 1.5$ corresponding the the exponent suggested by [60]. Despite the crude approximations just described, the numerical results of our model (Figure 6.1, lower panel) show a promising ability to capture the main features of the Po river network. A central stream connects the source-aggregating areas to the distributive region, collecting the rainfall via a network of hierarchically distributed tributaries. Striking similarities of these results with observable features of the real network show affluents that change direction according to the geometry of the external arch, a clear pattern of the true network. However, channel meandering is not visible in the numerical results , which is believed that these features can indeed be represented with our framework, but they are phenomena acting at a different temporal scales with respect to the network (and thus valley) formation, and are not incorporated in the current simulations.

Another problem is that our model cannot reproduce is the formation of loops in the delta area. Our patterns are (as theory suggests) tree-like networks. However, we would like to draw the attention of the reader to the last section of this chapter where a possible suggestion for the formation of loops as equilibrium points is proposed.

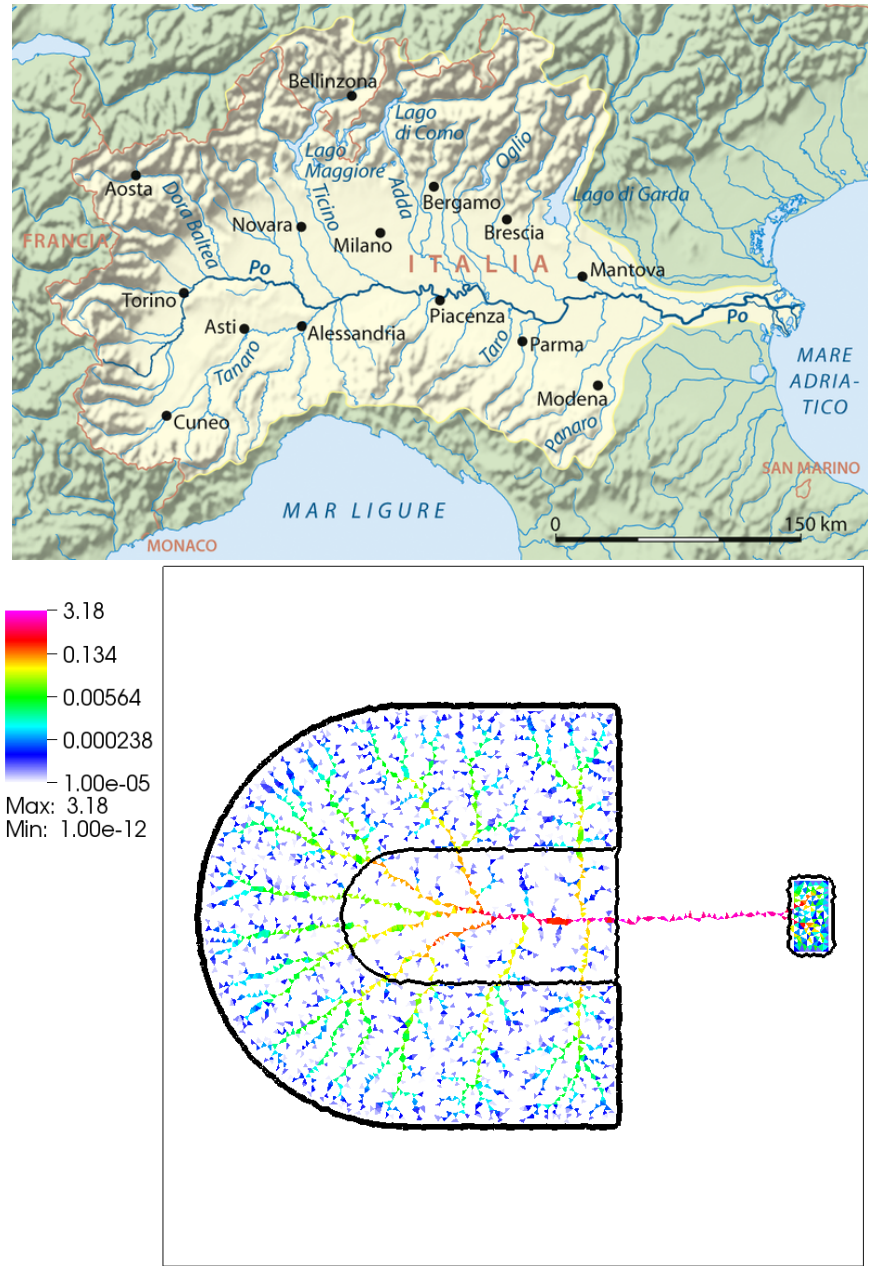

Figure 6.1: Domain of the Po River basin simulation. The upper panel reports the river basin of the Po river in Northern Italy (image taken from [23]). The lower panel shows the spatial distribution of μ_h^* using $\beta = 1.5$ and $\mu_0 \equiv 1$. The source term f^+ assumes two values. The value $1/2$ is imposed in the in the center of the horse-shoe shape representing the Po river basin, honoring the assumption that in the central plain half of the rainfall water infiltrates and does not contribute to the river discharge. In the external arch, representing the surrounding mountains, we imposed $f^+ = 1$. Mass conservation is ensured by the presence of a sink term f^- supported in the right rectangle that represents the Po delta.

6.2 Modeling Plant-Root Dynamics

Another field of study that we want to approach with our model is the simulation of plant-root dynamics. Roots plays a crucial role in the so-called Soil-Plant-Atmosphere Continuum (SPAC), an integrated approach introduced in [56] to describe the interaction between soil, biota, and atmosphere in terms of mass (water, carbon, nutrients) and energy exchanges. The SPAC concept is used for modeling both above-ground atmospheric processes and below-ground soil processes.

The roots are an important component of the biota that contributes to the control and modulation of the mass and energy exchanges between the ground surface and the atmosphere. In meteo-climatic models the rhizosphere is taken into account using simplified models and parametrizations, the so-called"land-surface models". This over-simplification is thought to be the cause of inaccuracies [51, 12] in weather predictions, in particular in terms of rainfall spatial distribution and rates, with obvious consequential effects on climate modeling.

Despite the fundamental rule played by plant roots in the SPAC concept, little is know about their real behavior. While there is a tendency to include root activity in plant modeling using an absolutely continuous density distribution [28], the highly nonlinear character of the coupling between soil water and root channels prevents an easy solution to this problem. On the other hand, the modeling of branching roots has received some attention in the literature but no satisfactory approach has been proposed until now [43].

From an evolutionary point of view, the adaptation capability of roots, that seems to point towards optimal response to the environmental stresses, is often postulated in the description of what roots do or aim at, but rarely, if not ever, truly verified. We think that our model provides a structured answer to this last fundamental question, supplying the necessary mathematical framework that connects energy and biomass allocation functionals. The idea, intrinsic to our branched transport model, of finding the optimal trade-off between the minimal cost of transportation, explicated in terms of dissipated energy, and cost of building the transport infrastructure, explicated in terms of root biomass allocation, finds an ideal setting within the optimality theories stating that plants try to maximize carbon (biomass) allocation in the aerial system while minimizing at the same time energy expenditure to secure enough water and nutrient

uptake [45, 47].

Starting from the coupled model of water flow in the soil and plant dynamics developed in [47], we recast all the relevant variables into our branched model. Thus, we reinterpret μ as the density distribution of roots in the soil space. The forcing function $f = f^+ - f^-$ is given by the difference between the "potential root-water uptake" (f^-) provided by the plant model for a given atmospheric condition. The soil water content f^+ is given by the solution of Richards' equation governing flow of water in partially saturated soils. The plant model, given the plant parameters, the atmospheric conditions, and the water content distribution, returns the root water uptake and the additional total biomass to be allocated to the root network. The DMK model takes as input the water content distribution (in our mind equal to f^+) and the root water uptake (f^-) and the constraint on the total marginal biomass, and returns a new root network. Note that the latter DMK problem is in reality a mass-optimization problem to be cast within our framework miming the shape-optimization problem of [15] as described in Section 1.4.2. This process constitutes a nonlinear feedback that can be solved by iteration.

In this test case, schematically represented in Figure 6.2 (left) we consider a two-dimensional slice of a heterogeneous soil formed by a sand in the left half and a silty clay in the other half. A plant with standard features is located in the center of the domain and is equipped with a root density distribution (shown in the figure with contour lines) that decays with depth. A fixed groundwater table depth is considered together an initial equilibrium water content vertical distribution above the water table. We first run the CATHY simulator (coupled Richards equation-plant model) described in [47] to evaluate the root water uptake fluxes. The resulting spatial distribution is shown in the right panel of Figure 6.2 in the background. Clearly, the plant prefers to take up the water from the sandy soil as the corresponding dissipated energy is smaller than in the case of the silty-clayey soil. The corresponding optimal network calculated by an off-line application of our DMK is shown in the foreground. Most of the large network structure is visible in the sandy portion of the domain. However, a few minor branches are exploring the clayey area, ready to become active in the occurrence of a drought and exploit the high residual saturation and low seepage velocities characteristics of clays.

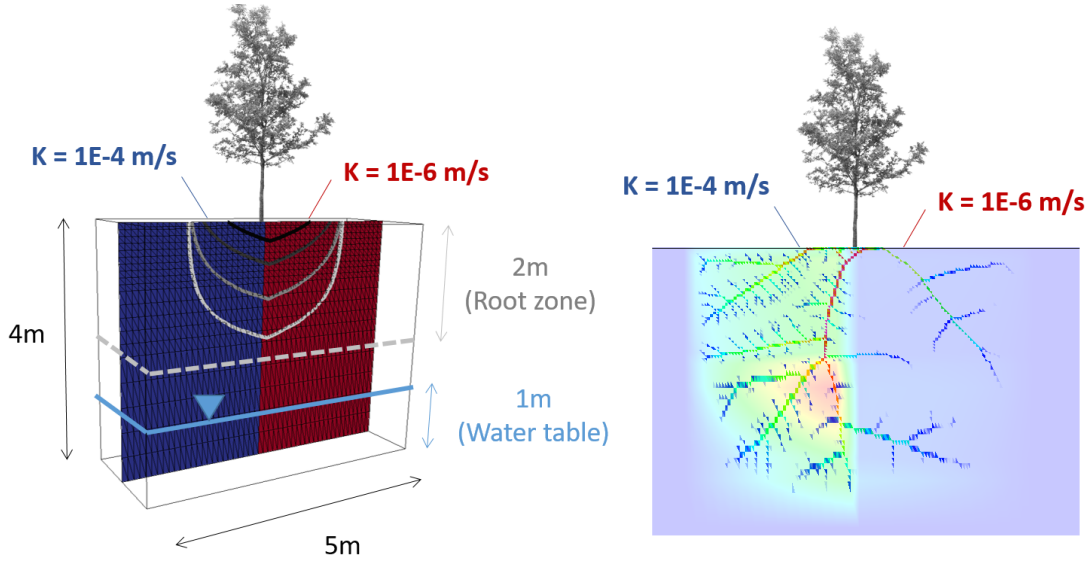

Figure 6.2: Setup of the hydrological (Richards equation) simulation of the plant dynamics (left). Spatial distribution (right) of the root water uptake estimated by the plant model coupled to Richards equation with superimposed the DMK network "optimally" transporting the soil water satisfying the atmospheric demand. No iterations of the integrated CATHY-DMK solvers are performed in this initial simulation, and only a one-way coupling with the DMK is considered.

6.3 Time varying forcings

It is well known in the field of complex networks that network trees are often optimal, but the presence of loops is fundamental to ensure the robustness of a network. In fact, in a tree any broken connection separates the network into two disjoint components. In the presence of loops, network are more robust in the sense that one single cut will never divide the network into separate components.

At the end of Section 6.1 we reported a critical incongruence in our model of river networks, namely that, contrary to observations, no loops are formed in the delta areas. This behavior prompted a closer inspection of the dynamics leading to the final equilibrium configuration. Looking at the μ_h^* distribution at intermediate times in test case 2 with Dirac forcings (Section 3.4.3), we note that loops actually form in the support of μ_h, as shown in Figure 6.3. After this intermediate phase, loops are destroyed by the dynamics and the system converges invariably to a tree-like structure.

The conjecture we can gain from the previous observation is that time-dependent

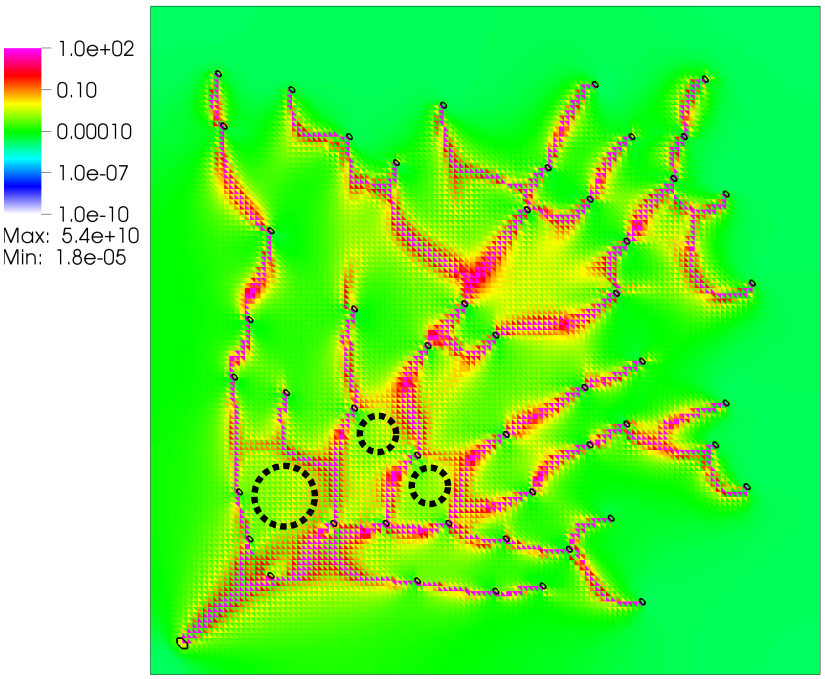

Figure 6.3: Test case TC2. Spatial distribution of μ_h at intermediate times before the equilibrium configuration μ_h^* is achieved. The black points highlight the center of the loops that are surviving in this intermediate phase.

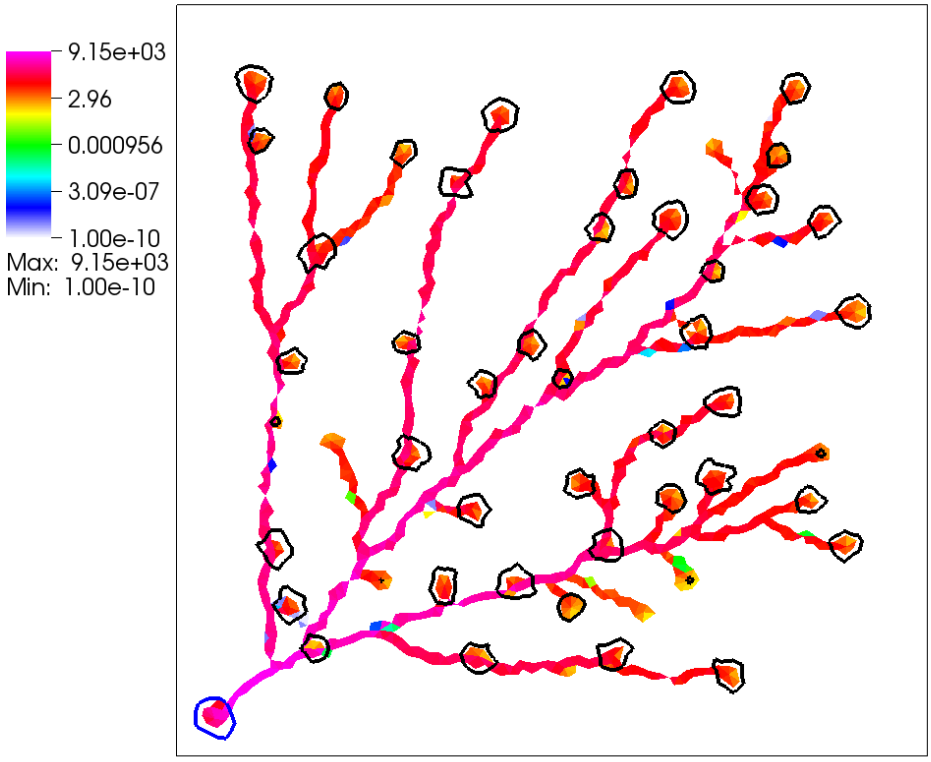

Figure 6.4: Spatial distribution of μ_h for test case E1. We report the large-time meta-equilibrium configuration μ_h^*, considered achieved when the topological structure does not change. The black circles indicate the approximate positions of the source points, while the circle size is proportional to the actual (current time) value of f^+.

forcings in our DMK model may lead to "meta-equilibrium" configurations that present loops. We refer here as "meta-equilibrium" a state where μ may vary but its support, i.e., the topological structure of the network, does not. To test this hypothesis, we introduce source and sink terms that vary in time faster then the μ dynamics. Thus we devised two experiments based on TC2 in Section 3.4.3. In the first experiment (E1) the value of the 50 source points is rapidly varying according to a Brownian motion reflected between zero and two, while in the second experiment (E2) the reflections is between minus one and plus one. Thus in E1 the 50 points in $[0.1, 0.9] \times [0.1, 0.9]$ always play the role of source terms, while in E2 sources and sinks are rapidly exchanging. In both cases, at each time t, a sink/source point located in $(0.05, 0.05)$ balances the system.

Figures 6.4 and 6.5 show the results of test cases E1 and E2, respectively. In

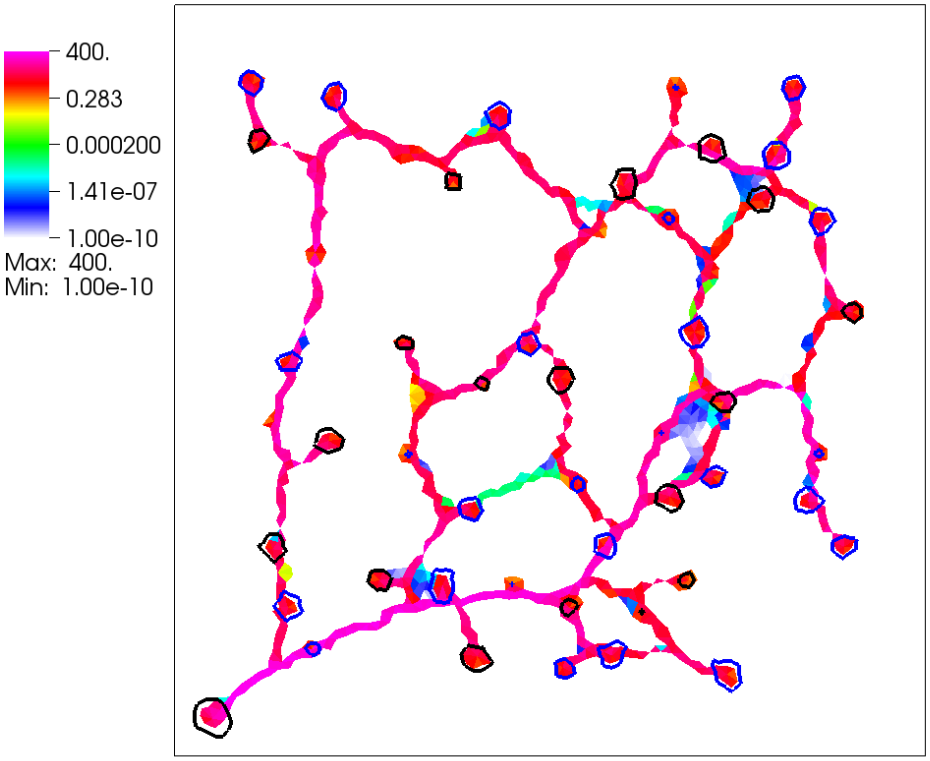

Figure 6.5: Spatial distribution of μ_h for test case E2. We report the large-time meta-equilibrium configuration μ_h^*, considered achieved when the topological structure does not change. The black (blue) circles indicate the approximate positions of the source (sink) points, while the circle size is proportional to the actual (current time) value of f^+ $(Sink)$.

test case E1, even if the value of f^+ is changing in all source points, no switch between source and sink functioning occurs. The transport density μ_h converges towards a tree-shaped equilibrium configuration μ_h^* that connects all the points where f is concentrated. On the other hand, in test case E2, the support μ_h shrinks towards a network structure connecting all points, but the continuous sink/source switching preserves the initial loops. Intuitively, we can explaining this phenomenon by observing that, if the forcings vary sufficiently rapidly, the fluxes settle to a nonzero value corresponding to a nonzero μ. If the time variation of f is too slow, then the decay term in the dynamics of the transport density dominates over the flux term, and transient loops are removed. The Lyapunov-candidate functional oscillates around a decreasing average behavior in test case E1, and converges towards an "average" equilibrium that seems to be time stationary. In the E2 test case, the behavior turns out to be similar, albeit more oscillatory.

The results allow us to conjecture that indeed loops form whenever the forcings oscillate with zero temporal mean, while they disappear with nonzero-average forcings. These observations would imply that natural systems always tend to an optimal equilibrium point given by a tree-like network. Non optimal, but robust, loops emerge as a response of the system to non-stationary forcings characterized by a zero average (i.e. always balanced) and a fast oscillatory behavior.

Appendix A

Appendix

A.1 Convex analysis

In this chapter we present a brief introduction to the Convex Analysis Theory. We try to compress in a few pages the definitions and the results that will be used to develop our work. We start from the basic basic concepts followed by well known result in minimization problems. We conclude with some results of the duality theory. We do not report the proofs of the results but we address the reader to [29], from which this chapter borrows.

A.1.1 Definitions of Convex Analysis

We recall here some basic notions of convex analysis taken from Chapter 1 of [29]. We try to follow the same notation of the book, with minimal changes to adapt to the variables and the functionals in this thesis. We hereafter denote with $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ and with V a real vector space.

Definition 47 (Convex Set)**.** *A set $\mathcal{C} \subseteq V$ is said convex if*

$$(1-\lambda)u_0 + \lambda u_1 \in \mathcal{C} \quad \forall u_0, u_1 \in \mathcal{C} \quad \forall\, 0 \leq \lambda \leq 1 \tag{A.1}$$

Definition 48 (Convex and Proper Functionals)**.** *Given $\mathcal{C}$ a convex subset in V, a functional $J : \mathcal{C} \mapsto \bar{\mathbb{R}}$ is* convex *if*

$$J((1-\lambda)u_0 + \lambda u_1) \leq (1-\lambda)J(u_0) + \lambda J(u_1) \quad \forall\, 0 \leq \lambda \leq 1 \tag{A.2}$$

for all $u_0, u_1 \in \mathcal{C}$. The functional is strictly convex *if $u_0 = u_1$ when strict*

inequality holds in Equation (A.2)*. Moreover J is said to be* proper *if it nowhere takes the value $-\infty$.*

We now consider the topological properties of V that will be here after a locally convex vector space with Hausdorff topology. We recall that V is a locally convex space is the origin posses a fundamental system of convex convex neighborhood, and that V is a topological vector space if it is endowed with a topology for which the sum operator between elements of V and the multiplication with real scalars are continuous.

Definition 49 (Lower Semi-Continuous Functionals)**.** *A functional $J : V \mapsto \bar{\mathbb{R}}$ is said to* lower semi-continuous *(l.s.c.) if*

$$\forall a \in \mathbb{R} \quad \{u \in V \mid J(u) \leq a\} \quad \textit{is closed} \tag{A.3}$$

$$\forall \bar{u} \in V \qquad \liminf_{u \mapsto \bar{u}} J(u) \geq J(\bar{u}) \tag{A.4}$$

Definition 50 (Sub-differential)**.** *A functional $J : V \mapsto \bar{\mathbb{R}}$ is said to be sub-differentiable at $u \in V$ if it has a continuous affine minorant that is exact at u. The slope $u^* \in V^*$ of such minorant is called a sub-gradient of J at u, and the set of sub-gradients at u is called the sub-differential at u and it is denoted by $\partial J(u)$. If J is not sub-differentiable at u, we have $\partial J(u) = \emptyset$. Thus we can state thee following characterization of the sub-differential of J at u^*:*

$$u^* \in \partial J(u) \textit{ if and only if } J(u) \textit{ is finite} \tag{A.5}$$

$$\textit{and}$$

$$\langle v - u, u^* \rangle + J(u) \leq J(u) \quad \forall v \in V$$

Definition 51 (Legendre-Fenchel Transform)**.** *Given a functional $J : V \mapsto \bar{\mathbb{R}}$, the* Legendre-Fenchel Transform *of J is defined as*

$$J^* : V^* \mapsto \bar{\mathbb{R}}$$

$$J^*(u^*) := \sup_{u \in V} \{\langle u^*, u \rangle_{V^*,V} - J(u)\}$$

A.1.2 Minimization of Convex Function

One of the main arguments of study in the convex analysis is to ensure existence and uniqueness of solution for problems of the form

$$\inf_{u \in \mathcal{C}} J(u) \tag{A.6}$$

where $\mathcal{C}$ is a convex subset of V and $J : \mathcal{C} \mapsto \bar{\mathbb{R}}$ is a l.s.c.convex and proper functional. A standard tool used to ensure the existence of a minimum of in problem A.6 is the direct method of the calculus of variations, that when V is a reflexives space applies as follows ([29, Proposition 1.2 Chapter 2]).

Proposition 52. *Let V be a reflexive space (a Banach space with coincides with its bi-dual). Given $\mathcal{C} \subseteq V$ and l.s.c., convex and proper functional $J : \mathcal{C} \mapsto \bar{\mathbb{R}}$. Assume that is $\mathcal{C}$ is bounded or that J satisfies the following hypothesis*

$$\lim_{\substack{\|u\|_V \mapsto +\infty \\ u \in \mathcal{C}}} J(u) = +\infty \tag{A.7}$$

then problem in Equation (A.6) *admits at least one solution, which is unique if J is strictly convex.*

Remark 7. *Problem A.6 can be always reformulated on the whole space V. We just have to replace J with*

$$\tilde{J}(u) := \begin{cases} J(u) & \text{if } u \in \mathcal{C} \\ +\infty & \text{if } u \notin \mathcal{C} \end{cases}$$

For this reason, hereafter we will consider functionals J defined on the whole domain V.

A.1.3 Duality

Given a convex, l.s.c.and proper functional $J : V \mapsto \bar{\mathbb{R}}$, we denote as the *Primal Problem* the following minimization problem

$$\mathcal{P} \quad : \quad \inf_{u \in V} J(u) \tag{A.8}$$

Consider another Hausdorff topological vector space Y and a functional $\Phi : V \times Y \mapsto \bar{\mathbb{R}}$ (we call it perturbation functional), such that

$$J(u) = \Phi(u, 0) \quad \forall u \in V \tag{A.9}$$

The *Dual Problem* $\mathcal{P}^*$ of the primal problem A.8 is defined as

$$\mathcal{P}^* \quad : \quad \sup_{p \in Y} \{-\Phi^*(0, p^*)\} \tag{A.10}$$

where

$$\Phi^* : V^* \times Y^* \mapsto \bar{\mathbb{R}}$$

is the conjugate function of Φ in the duality pairing between $V \times Y$ and $V^* \times Y^*$

Proposition 53. *([29, Proposition 2.1, Chapter 3])*

$$-\infty \leq \inf_{u \in V} J(u) \leq \sup_{p \in Y} \{-\Phi^*(0, p^*)\} \leq +\infty \tag{A.11}$$

Proposition 54. *[29, Proposition. 2.4, Chapter 3] If problems in $\mathcal{P}, \mathcal{P}^*$ admit solutions and*

$$\inf_{u \in V} J(u) = \sup_{p \in Y} \{-\Phi^*(0, p^*)\} =< +\infty \tag{A.12}$$

then the following, called extremality relation*, holds for all solution $\bar{u}$ of Equation* (A.8) *and all solutions $\bar{p}^*$ of Equation* (A.10)

$$\Phi(\bar{u}, 0) + \Phi^*(0, \bar{p}^*) = 0 \tag{A.13}$$

or

$$(0, \bar{p}^*) \in \partial\Phi(\bar{u}, 0) \tag{A.14}$$

Conversely if $\bar{u} \in V$ and $\bar{p}^* \in Y^*$, then $\bar{u}$ is a solution of $\mathcal{P}$, $\bar{p}^*$ is a solution of $\mathcal{P}^*$ and Equation (A.12) *holds.*

A.1.3.1 A case of direct interest

We now focus on a particular, important, problem. Given V and Y two Banach spaces, let Λ be a linear operator from V to Y ($\Lambda \in \mathcal{L}(V, Y)$), $F : V \mapsto \bar{\mathbb{R}}$, and $G : Y \mapsto \bar{\mathbb{R}}$. Next, we consider a particular form for J

$$J(u) = F(u) + G(\Lambda(u)) \tag{A.15}$$

In this case the perturbation functional Φ can be defined as

$$\Phi(u, p) = F(u) + G(\Lambda(u) - p) \tag{A.16}$$

and the dual problem in Equation (A.10) becomes

$$\sup_{p^* \in Y^*} \{-F^*(\Lambda^*(p^*)) - G^*(-p^*)\} \tag{A.17}$$

where $\Lambda^* : \mathcal{L}(Y^*, V^*)$ indicates the adjoint of operator Λ. Under these assumptions we can state the following theorem:

Theorem 55. *[29, Section 4, Chapter 3] Assume that F, G are convex functions as above, assume that there exists $u_0 \in V$ such that $F(u_0) < +\infty$, $G(\Lambda(u_0)) < +\infty$, G being continuous at $\Lambda(p_0)$ then*

$$\sup_{p^* \in Y^*} \{-F^*(\Lambda^*(p^*)) - G^*(-p^*)\} = \inf_{u \in V} F(u) + G(\Lambda(u)) \tag{A.18}$$

and Equation (A.10) *has at least one solution* $\bar{p}^*$*. The extramality relation in Equation* (A.13) *reads as*

$$F(\bar{u}) + F^*(\Lambda^*(\bar{p}^*)) - \langle \Lambda^*(\bar{p}^*), u \rangle_{V^*,V} = 0 \tag{A.19}$$

$$G(\Lambda(\bar{u})) + G^*(-\bar{p}^*) - \langle \bar{p}^*, \Lambda\bar{u} \rangle_{Y^*,Y} = 0 \tag{A.20}$$

These last conditions amount to saying that

$$\Lambda^*(\bar{p}^*) \in \partial F(\bar{u}) \qquad \bar{p}^* \in \partial G\left(\Lambda(\bar{u})\right) \tag{A.21}$$

Bibliography

[1] A. Adamatzky, *Physarum Machines*, Computers from Slime Mould, World Scientific, 2010.

[2] L. Ambrosio, *Lecture Notes on Optimal Transport Problems*, in Lecture Notes in Mathematics, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 1–52.

[3] L. Ambrosio, N. Gigli, and G. Savare, *Gradient Flows: In Metric Spaces and in the Space of Probability Measures*, Lectures in Mathematics. ETH Zürich, Birkhäuser Basel, 2005.

[4] J. W. Barrett and W. Liu, *Finite element approximation of the p-laplacian*, Math. Comp., 61 (1993), pp. 523–537.

[5] J. W. Barrett and L. Prigozhin, *A mixed formulation of the Monge-Kantorovich equations*, ESAIM-Math. Model. Num., 41 (2007), pp. 1041–1060.

[6] L. Bergamaschi, E. Facca, A. Martínez Calomardo, and M. Putti, *Spectral preconditioners for the efficient numerical solution of a continuous branched transport model*, J. Comput. Appl. Math., Submitted (2017).

[7] L. Bergamaschi, G. Gambolati, and G. Pini, *Asymptotic convergence of conjugate gradient methods for the partial symmetric eigenproblem*, Numer. Lin. Alg. Appl., 4 (1997), pp. 69–84.

[8] L. Bergamaschi and A. Martínez, *Two-stage spectral preconditioners for iterative eigensolvers*, Numer. Lin. Alg. Appl., 24 (2017), pp. 1–14.

[9] L. Bergamaschi, A. Martínez, and G. Pini, *Parallel Rayleigh Quotient optimization with FSAI-based preconditioning*, J. Applied Mathematics, 2012, Article ID 872901, 14 pages (2012).

[10] L. Bergamaschi and M. Putti, *Numerical comparison of iterative eigensolvers for large sparse symmetric matrices*, Comp. Methods App. Mech. Engrg., 191 (2002), pp. 5233–5247.

[11] P. Bochev and R. B. Lehoucq, *On the finite element solution of the pure Neumann problem*, SIAM Rev., 47 (2005), pp. 50–66.

[12] S. Bonetti, G. Manoli, J.-C. Domec, Putti, Mario, M. Marani, and G. G. Katul, *The influence of water table depth and the free atmospheric state on convective rainfall predisposition*, Water Resour. Res., 51 (2015), pp. 2283–2297.

[13] V. Bonifaci, K. Mehlhorn, and G. Varma, *Physarum can compute shortest paths*, J Theor Biol, 309 (2012), pp. 121–133.

[14] G. Bouchitté and G. Buttazzo, *Characterization of optimal shapes and masses through Monge-Kantorovich equation*, Journal of the European Mathematical Society", 3 (2001), pp. 139–168.

[15] G. Bouchitté, G. Buttazzo, and P. Seppecher, *Energies with respect to a measure and applications to low dimensional structures*, Calc. Var. Partial Differential Equations, 5 (1996), pp. 37–54.

[16] ——, *Shape optimization solutions via Monge-Kantorovich equation*, CR MATH, 324 (1997), pp. 1185–1191.

[17] L. Brasco, *Geodesics and pde methods in transport models*, 2010.

[18] G. Buttazzo and E. Stepanov, *On regularity of transport density in the Monge-Kantorovich problem*, SIAM J. Control Optim, 42 (2003), pp. 1044–1055.

[19] S. Campanato, *Equazioni ellittiche del secondo ordine e spazi* $\mathfrak{L}^{(2,\lambda)}$, Annali di Matematica, 69 (1965), pp. 321–381.

[20] B. Carpentieri, I. S. Duff, and L. Giraud, *A class of spectral two-level preconditioners*, SIAM J. Sci. Comput., 25 (2003), pp. 749–765 (electronic).

[21] A. Chen, J. Darbon, G. Buttazzo, F. Santambrogio, and J. M. Morel, *On the equations of landscape formation*, Interfaces Free Bound, 16 (2014), pp. 105–136.

[22] M. Colombo and G. Mingione, *Regularity for double phase variational problems*, Arch. Ration. Mech. Anal, 215 (2015), pp. 443–496.

[23] W. Commons, *File:po bacino idrografico.png — wikimedia commons, the free media repository*, 2015. [Online; accessed 27-October-2017].

[24] L. De Pascale, L. Evans, and A. Pratelli, *Integral estimates for transport densities*, Bull. London Math. Soc., 36 (2004), pp. 383–395.

[25] L. De Pascale and A. Pratelli, *Sharp summability for Monge transport density via interpolation*, ESAIM Control Optim. Calc. Var., 10 (2004), pp. 549–552.

[26] G. L. Delzanno, L. Chacón, J. M. Finn, Y. Chung, and G. Lapenta, *An optimal robust equidistribution method for two-dimensional grid adaptation based on Monge-Kantorovich optimization*, J. Comp. Phys., 227 (2008), pp. 9841–9864.

[27] I. S. Duff, L. Giraud, J. Langou, and E. Martin, *Using spectral low rank preconditioners for large electromagnetic calculations*, Int. J. Numer. Methods Engrg., 62 (2005), pp. 416–434.

[28] L. X. Dupuy, P. J. Gregory, and A. G. Bengough, *Root growth models: towards a new generation of continuous approaches.*, J. Exp. Botany, 61 (2010), pp. 2131–2143.

[29] I. Ekeland and R. Témam, *Convex Analysis and Variational Problems*, Classics in Applied Mathematics, SIAM, Philadelphia, PA, 1999.

[30] L. C. Evans, *Partial differential equations and Monge-Kantorovich mass transfer*, CDM, 1997 (1997), pp. 65–126.

[31] L. C. Evans, *Partial differential equations*, AMS, Providence, R.I., 2010.

[32] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge-Kantorovich mass transfer problem*, vol. 653, American Mathematical Soc., 1999.

[33] E. Facca, F. Cardin, and M. Putti, *A continuous model of slime mold dynamics*, SIAM J. Appl. Math., Accepted (2017).

[34] E. Facca, S. Daneri, F. Cardin, and M. Putti, *Numerical solution of Monge-Kantorovich equations via a dynamic formulation*, SIAM J. Sci. Comput., Submitted (2017).

[35] M. Feldman and R. J. McCann, *Uniqueness and transport density in Monge's mass transportation problem*, Calc. Var. Partial Differential Equations, 15 (2002), pp. 81–113.

[36] I. Fragalà, M. S. Gelli, and A. Pratelli, *Continuity of an optimal transport in Monge problem*, J. Math. Pures Appl. (9), 84 (2005), pp. 1261–1294.

[37] M. Giaquinta and L. Martinazzi, *An Introduction to the Regularity Theory for Elliptic Systems, Harmonic Maps and Minimal Graphs*, Springer Science & Business Media, Pisa, July 2013.

[38] E. N. Gilbert, *Minimum cost communication networks*, Bell Labs Technical Journal, 46 (1967), pp. 2209–2227.

[39] L. Giraud, S. Gratton, and E. Martin, *Incremental spectral preconditioners for sequences of linear systems*, Applied Numerical Mathematics, 57 (2007), pp. 1164 – 1180. Numerical Algorithms, Parallelism and Applications (2).

[40] C. Hackett and D. A. Rose, *A model of the extension and branching of a seminal root of barley, and its use in studying relations between root dimensions i. the model.*, Aust J Biol Sci, 25 (1972), pp. 669–679.

[41] R. E. Horton and C. Jarvis, *Drainage-basin characteristics*, EOS, Trans. Am. Geophys. Union, 13 (1932), pp. 350–361.

[42] E. F. Kaasschieter, *Preconditioned conjugate gradients for solving singular systems*, J. Comput. Appl. Math., 24 (1988), pp. 265–275.

[43] D. I. Kalogiros, M. O. Adu, P. J. White, M. R. Broadley, X. Draye, M. Ptashnyk, A. G. Bengough, and L. X. Dupuy, *Analysis of root growth from a phenotyping data set using a density-based model*, Journal of Experimental Botany, 67 (2016), pp. 1045–1058.

[44] L. V. Kantorovich, *On the translocation of masses*, C. R. (Doklady) Acad. Sci. USSR, 321 (1942), pp. 199–201.

[45] G. G. Katul, S. Manzoni, S. Palmroth, and R. Oren, *A stomatal optimization theory to describe the effects of atmospheric CO2 on leaf photosynthesis and transpiration.*, Ann. Bot., 105 (2010), pp. 431–442.

[46] W. B. Langbein, *Topographic Characteristics of Drainage Basins*, U.S. Geol. Surv. Prof. Pap., (1947), pp. 125–158.

[47] G. Manoli, S. Bonetti, J.-C. Domec, Putti, Mario, G. G. Katul, and M. Marani, *Tree root systems competing for soil moisture in a 3D soil–plant model*, Adv. Water Resources, 66 (2014), pp. 32–42.

[48] A. Marani, R. Rigon, and A. Rinaldo, *A note on fractal channel networks*, Water Resour. Res., 27 (1991), pp. 3041–3049.

[49] A. Martínez, *Tuned preconditioners for the eigensolution of large spd matrices arising in engineering problems*, Numer. Lin. Alg. Appl., 23 (2016), pp. 427–443.

[50] J. Mas, J. Cerdán, N. Malla, and J. Marín, *Application of the Jacobi-Davidson method for spectral low-rank preconditioning in computational electromagnetics problems*, SeMA. J., 67 (2015), pp. 39–50.

[51] R. M. Maxwell and S. J. Kollet, *Interdependence of groundwater dynamics and land-energy feedbacks under climate change*, Nature Geosc., 1 (2008), pp. 665–669.

[52] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, De l'Imprimerie Royale, 1781.

[53] C. B. Morrey Jr, *Second-order elliptic systems of differential equations*, in Contributions to the theory of partial differential equations, Princeton University Press, Princeton, N. J., 1954, pp. 101–159.

[54] T. Nakagaki, H. Yamada, and A. Toth, *Maze-solving by an amoeboid organism*, Nature, 407 (2000), pp. 470–470.

[55] E. Oudet and F. Santambrogio, *A Modica-Mortola approximation for branched transport and applications*, Arch. Ration. Mech. An., 201 (2011), pp. 115–142.

[56] J. R. Philip, *Plant water relations: some physical aspects*, Ann. Rev. Plant Physiol., 17 (1966), pp. 245–268.

[57] M. Putti and C. Cordes, *Finite element approximation of the diffusion operator on tetrahedra*, SIAM J. Sci. Comput., 19 (1998), pp. 1154–1168.

[58] A. Quarteroni and A. Valli, *Numerical approximation of partial differential equations*, vol. 23 of Springer Series in Computational Mathematics, Springer-Verlag, Berlin, 1994.

[59] A. Rinaldo, R. Rigon, J. R. Banavar, A. Maritan, and I. Rodriguez-Iturbe, *Evolution and selection of river networks: statics, dynamics, and complexity.*, Proc. Nat. Acad. Sci., 111 (2014), pp. 2417–2424.

[60] I. Rodríguez-Iturbe and A. Rinaldo, *Fractal river basins: chance and self-organization*, Cambridge University Press, 2001.

[61] F. Santambrogio, *Optimal channel networks, landscape function and branched transport*, Interfaces Free Bound., 9 (2007), pp. 149–169.

[62] ——, *Absolute continuity and summability of transport densities: simpler proofs and new estimates*, Calc. Var. Partial Differential Equations, 36 (2009), pp. 343–354.

[63] ——, *A Modica–Mortola approximation for branched transport*, Comptes Rendus Mathematique, 348 (2010), pp. 941–945.

[64] ——, *Optimal transport for applied mathematicians*, Birkäuser, NY, 2015.

[65] F. Santambrogio, $\{$*Euclidean, Metric, and Wasserstein*$\}$ *gradient flows: an overview*, ArXiv e-prints, (2016).

[66] A. Schrijver, *Combinatorial Optimization : Polyhedra and Efficiency (Algorithms and Combinatorics)*, Springer, July 2004.

[67] A. Stathopoulos and K. Orginos, *Computing and deflating eigenvalues while solving multiple right-hand side linear systems with an application to quantum chromodynamics*, SIAM J. Sci. Comput., 32 (2010), pp. 439–462.

[68] A. Tero, R. Kobayashi, and T. Nakagaki, *A mathematical model for adaptive transport network in path finding by true slime mold*, J. Theor. Biol., 244 (2007), pp. 553–564.

[69] A. Tero, S. Takagi, T. Saigusa, K. Ito, D. P. Bebber, M. D. Fricker, K. Yumiki, R. Kobayashi, and T. Nakagaki, *Rules for biologically inspired adaptive network design.*, Science, 327 (2010), pp. 439–442.

[70] G. Troianiello, *Elliptic Differential Equations and Obstacle Problems*, University Series in Mathematics, Springer, 1987.

[71] A. M. Vershik, *Long history of the Monge-Kantorovich transportation problem*, Math. Intelligencer, 35 (2013), pp. 1–9.

[72] C. Villani, *Topics in Optimal Transportation (Graduate Studies in Mathematics, Vol. 58)*, AMS, Providencem, R.I., 2003.

[73] C. Villani, *Optimal Transport*, vol. 338 of Old and New, Springer Science & Business Media, Berlin, Heidelberg, 2008.

[74] Q. Xia, *Optimal paths related to transport problems*, CCM, 5 (2003), pp. 251–279.

[75] Q. Xia, *On landscape functions associated with transport paths*, Discrete Contin. Dyn. Syst, 34 (2014), pp. 1683–1700.

[76] L. M. York, A. Carminati, S. J. Mooney, K. Ritz, and M. J. Bennett, *The holistic rhizosphere: integrating zones, processes, and semantics in the soil influenced by roots*, J Exp Bot., 67 (2016), pp. 3629–3643.