Università degli Studi di Padova

DIPARTIMENTO DI INGEGNERIA CIVILE, EDILE E AMBIENTALE
Corso di Laurea Magistrale in Mathematical Engineering - Ingegneria Matematica

Numerical solution of the three dimensional Optimal Transport Problem

Laureando:
Riccardo Tosi
Matricola 1129951

Relatore:
Ch.mo Prof. Mario Putti

Correlatore:
Dott. Enrico Facca

Anno Accademico 2017-2018

There's no such thing as talent, cap'n!
Only inspiration and ambition!
And mine burns with hot!
Scrooge McDuck, The Life and Times of Scrooge McDuck

Acknowledgements

First of all, I want to express my gratitude to my supervisor, prof. Mario Putti, and my co-supervisor, Enrico Facca. Their ideas, suggestions and unwavering support transmitted me enthusiasm and passion for doing research and they were an unmatched help both in the accomplishment of this thesis and in my academic growth.

None of this could have happened without my family. I am deeply grateful to my sisters for always being there for me, as siblings and friends. I will never thank enough my parents, who constantly encouraged me along these years and helped me through personal and academic choices. They taught me I have always to explore new directions in life and fill everything I do with love.

I dedicate this work to my parents, who instilled in me the desire to learn.

Abstract

In this thesis we analyze a model introduced in [14, 17] and its extensions [14, 15]. These works conjectured a new formulation of Optimal Transport, an expanding area of mathematics whose aim is the identification of the most efficient strategy to reallocate resources from one place to another. The numerical approximation of the equations of the model represents a simple yet effective numerical approach to solve Optimal Transport problems. However, the numerical scheme was analyzed only in the two dimensional case.

The aim of this thesis is to exploit the model to solve three dimensional Optimal Transport problems, where few examples of numerical solution are known from the literature. We present all the non trivial challenges required by the three dimensional extension, together with an ample series of numerical experiments, that confirms the conjectured equivalence with the Optimal Transport problem. The results show that the numerical scheme is robust and efficient, with ample space for improvement from the computational point of view.

Contents

Introduction

Optimal Transport (OT) is an expanding area of mathematics that studies how to find the least-cost strategy, with respect to a transport cost, to reallocate resources from an initial to a final configuration.

Gaspard Monge was the first to approach this problem in 1781 in "*Mémoire sur la théorie des déblais et des remblais*" [24], where he discussed the problem of finding the most efficient way to move soil from an excavation site to an embankment of equal volume. Mathematically speaking, the excavation site and the embankment are described, respectively, by two non-negative measures f^+ and f^- with equal mass. Originally, Monge considered the Euclidean distance as cost function, thus assuming the cost of transportation to be proportional to the travel distance. Nowadays, different formulations of the OT problem exist, this is due to the increasing attention to the problem after the relaxed formulation proposed by Leonid Kantorovich in 1942 [20], and nowadays this is called the Monge-Kantorovich (MK) problem. A divergence constrained formulation is another formulation of the OT problem that aims to find the optimal transportation path in case the mass accumulation along the process is either favoured or discouraged. The formulation is the following: given $\Omega \subset \mathbb{R}^d$ and $q \in (0,2)$, find among the vector-valued functions $v : \Omega \mapsto \mathbb{R}^d$, orthogonal to the boundary of Ω, the optimal v^* that solves

$$\inf_v \left\{ \int_\Omega |v|^q \ : \ \nabla\cdot v = f^+ - f^- \right\} .$$

In case $q \in (1,2)$ the problem is called Congested Transport (CT) problem, and the convex exponent in the minimization problem discourages mass accumulation along the transport. In case $q = 1$ the problem is called Beckmann problem, and it is equivalent to the Monge-Kantorovich OT problem, with cost equal to the Euclidean distance. When $q \in (0, 1)$ we have the Branched Transport (BT) problem, which encourages mass accumulation. This concentration effect leads to the formation of very singular solutions, and the integral process requires a proper characterization [35]. In the above formulation f^+ and f^- no longer represent the initial and final configuration of the transported resources, but they can be reinterpreted as a continuous injection and absorption of mass. The problem now searches for the minimizing flux that ensures mass balance and minimizes the overall "traffic", penalizing or not mass concentration along the transport depending on the value of q.

The numerical solution of the Divergence Constrained formulation finds applications in several fields. For example, the CT problem can be used to model traffic flows [8]. The Monge-Kantorovich OT problem with the Euclidean distance (equivalent to the Beckmann problem) is exploited in Machine Learning [1, 22] and Image Processing [29]. The BT problem is probably the most interesting from the physical point of view, since the solutions of this problem are spatial trees with many bifurcations, that resemble many natural and artificial complex transport systems. Tree roots represent a perfect example of this ramified structure. In fact, they explore the entire soil water horizon, typically highly heterogeneous, saving as much water and nutrients as possible [21], and as result they develop this peculiar network. On the other hand, tree branches spread out in the air to maximize the amount of light they receive from the sun for photosynthesis, minimizing their surface as protection from external factors, such as parasites or temperature changes. Another example is the circulatory system in the human body, where veins and arteries transport back and forth the blood from the heart to the whole body and exhibit the ramified structure. Therefore, this branched framework is very common in natural systems, and OT theory can be effectively used to describe these patterns as solution of minimal energy problems. However, the numerical solution of OT and especially BT problems poses very difficult issues, and only few examples in the two dimensional case exist [26].

The above examples give an idea of the importance of solving the OT problem. Unfortunately there are few examples in literature that solve the CT problem and the Beckmann problem in the three dimensional case [2, 10, 36], while there are not three dimensional solutions for the BT problem.

The main purpose of this thesis is to exploit the numerical scheme, introduced in [14, 17], to the three dimensional case. The model consists in solving the Dynamic Monge-Kantorovich (DMK) equations, a system composed by an elliptic diffusion equation for the transport potential and an ordinary differential equation for the transport density. The problem reads as follows: find the pair of functions $(\mu(t,\cdot), u(t,\cdot)) : [0,+\infty) \times \Omega \mapsto \mathbb{R}^+ \times \mathbb{R}^d$ that satisfies:

$$\begin{aligned}
&-\nabla\cdot\Big(\mu(t,x)\,\nabla\, u(t,x)\Big) = f^+(x) - f^-(x)\\
&\partial_t\mu(t,x) = [\mu(t,x)|\,\nabla\, u(t,x)|]^\beta - \mu(t,x)\\
&\mu(0,x) = \mu_0(x) > 0
\end{aligned}$$

where μ is an isotropic conductivity coefficient and u is a potential function. The dynamics of the process is modulated by the coefficient $\beta \in (0,2)$, the case $\beta \in (0,1)$ penalizes mass accumulation along the transport, while $\beta \in (1,2)$ favours the aggregation. The DMK solution is conjectured to converge toward an equilibrium configuration (μ^*, u^*) at large times, and the asymptotic vector field $v^* = -\mu^*\,\nabla\, u^*$ is conjectured to be solution of the OT problem. Unfortunately, the proof of the mathematical existence and uniqueness of the solution pair $(\mu(t), u(t))$ is still an open problem. However, in [14, 17] the authors identify a Lyapunov-candidate functional, i. e. a functional that decreases in time along $(\mu(t), u(t))$ and that reads as follows:

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

Moreover, the authors proved that in case $\beta \in (0,1]$, the minimization of $\mathcal{L}_\beta$ is equivalent to the Divergence Constrained formulation with $q = 2-\beta$, and the asymptotic configuration (μ^*, u^*) of the model is related to the solution of the p-Poisson equation. On the other hand, when $\beta \in (1, 2)$ the previous equivalence between the minimization of $\mathcal{L}_\beta$ and the Divergence Constrained formulation holds in the form of conjecture, supported by numerous numerical experiments.

The main advantage of the DMK formulation is that its numerical discretization is rather simple and efficient. Firstly, the ordinary differential equation for the transport density μ is projected onto a piecewise constant FEM space defined on a triangulation of the domain. Then, the elliptic equation is discretized using a linear Galerkin FEM method defined on the uniformly refined grid. Finally, the resulting differential-algebraic system of equations is solved exploiting the forward or the backward Euler method. The procedure is iterated in time until the relative differences on the spatial norm of the transport density are smaller than a predefined tolerance.

In this thesis we extend the two dimensional algorithm implemented in [14, 17] to verify the conjectures and the effectiveness of the solver in the three dimensional case. We consider the forward Euler scheme for the time discretization and we exploit the spatial discretization that requires to work simultaneously with a grid and the relative refined subgrid. We follow the approach described in [27] and obtain the subgrid by uniformly refining each tetrahedron into 8 tetrahedra.

All the experiments support the conjecture that the solution $(\mu(t), u(t))$ possesses a time-asymptotic equilibrium point. Moreover, we compare the asymptotic state of the DMK equations with an explicit solution of the OT problem, when available, and we prove the optimal convergence of the scheme, consistently with the two dimensional case. In addition, we are able to compute accurately the 1-Wasserstein distance, even with very coarse meshes. We also consider spherical domains, where an analytical solution for the p-Poisson equation is derived. The numerical experiments show the asymptotic configuration is related to that explicit solution, giving solidity to the scheme in the case $\beta \in (0, 1]$. Moreover, for $\beta > 1$, the optimal path, which is described by the asymptotic value μ^*, presents the ramified structures typical of the BT problem.

Chapter 1

Introduction to Optimal Mass transport theory

In this chapter we present a general introduction of the theory of Optimal Transportation (OT). We start from the original formulation made by Monge and the relaxation introduced by Kantorovich. Subsequently we focus on the Divergence Constrained formulation of the problem and we define the Monge-Kantorovich equations. Then we move to more general formulations, where mass concentration along the transport is either penalized or favoured.

1.1 Monge formulation

Gaspard Monge was the first to approach the OT problem in 1781 in "*Mémoire sur la théorie des déblais et des remblais*" [24], introducing it as a problem of military fortification construction. The Monge OT problem aims to find the least effort map to move the soil from an excavation area (*déblais*) to an embankment (*remblais*), preserving the volume and considering as transport cost the product between the distance and the mass.

Mathematically speaking, the déblais and the remblais can be considered as two non negative measures f^+ and f^- with equal mass, living in two complete and separable spaces X and Y. We denote with $\mathcal{M}_+(\Omega)$ the set of non negative measures defined on a measure space Ω. All the possible

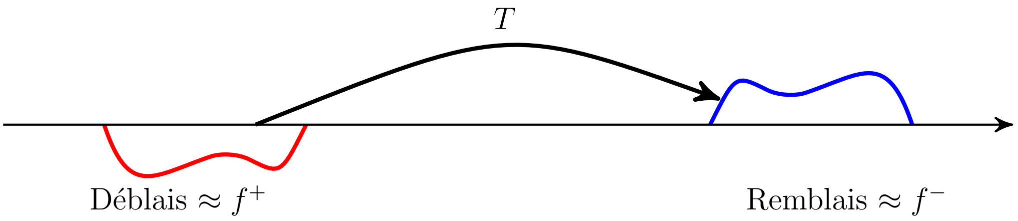


Figure 1.1: Graphic representation of the Monge problem of finding the least effort map to move the soil from an excavation area (*déblais*) to an embankment (*remblais*) of equal volume.

mass movements between X and Y with the mass-conservation assumption constitute the set of *transport maps*:

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

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

$$T_\#(f^+)(A) := f^+(T^{-1}A) \quad \text{for all measurable sets } A \in Y \ .$$

The transport cost $c : X \times Y \mapsto \mathbb{R}^+$ is a function that describes the cost of moving a unit mass from X to Y, and in its original formulation Monge considered the standard Euclidean distance $c(x,y) = |x-y|$. We can state the Monge problem as follows.

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

$$\min_{T\in\mathcal{T}(f^+,f^-)} I(T) := \int_X c(x,T(x))df^+(x) \ . \tag{1.1.1}$$

1.2 Kantorovich relaxation

Problem 1, as formulated by Monge, poses some serious issues from the mathematical point of view. In fact, standard tools of the calculus of variations can not be applied due to the high non linearity in the constraints and even the existence of a transport map is difficult to prove. A relevant improvement came around the 1940s, when Leonid Kantorovich introduced in [20] a relaxed version of the Monge problem, which eventually led to the development of linear programming and brought Kantorovich the Nobel Prize in Economics in 1975.

To simplify the exposition, we start from the finite dimensional case. Given the densities f^+ and f^- introduced in the Monge problem, consider the points $(x_i)_{i=1,n} \in \mathbb{R}^d$ and $(y_j)_{j=1,m} \in \mathbb{R}^d$. Then, we associate to each discrete point the mass $(f_i^+)_{i=1,n}$ or $(f_j^-)_{j=1,m}$, and we add the requirement $\sum_{i=1}^n f_i^+ = \sum_{j=1}^m f_j^-$. To each couple of points (x_i, y_j) we assign the real number c_{ij}, representing the cost of moving one unit of mass from x_i to y_j, i. e. the Euclidean distance between x_i and y_j. Thus we define the problem as follows.

Problem 2 (Discrete Primal Problem)
Given $\boldsymbol{c} \in \mathbb{R}^{n\times m}$, $\boldsymbol{f}^+ \in \mathbb{R}^n$ and $\boldsymbol{f}^- \in \mathbb{R}^m$, find $\boldsymbol{\gamma}^*$ that solves the minimization problem:

$$\min_{\gamma_{ij}} \sum_{i=1}^{n} \sum_{j=1}^{m} c_{ij}\gamma_{ij} \tag{1.2.1a}$$

$$\sum_{j=1}^{m} \gamma_{ij} = f_i^+ \quad \sum_{i=1}^{n} \gamma_{ij} = f_j^- \quad \gamma_{ij} \geq 0 \ . \tag{1.2.1b}$$

This problem tells that we have to identify the variables γ_{ij}, which stand for the quantity of resources moved from the initial place i to the final place j. This formulation allows the splitting of the mass in the movement from the initial configuration to the final and, due to its linearity, it can be seen as a discrete relaxed version of problem 1. Problem 2 is called "primal".

In fact, as typically in operation research, to each primal linear programming problem there is an equivalent "dual" problem. The discrete dual of problem 2 is given by:

Problem 3 (Discrete Dual Problem)
Given $\boldsymbol{c} \in \mathbb{R}^{n\times m}$, $\boldsymbol{f}^+ \in \mathbb{R}^n$ and $\boldsymbol{f}^- \in \mathbb{R}^m$, find $\boldsymbol{u}^*$ and $\boldsymbol{v}^*$ which solve the maximization problem:

$$\max_{(u_i,v_j)} \sum_{i=1}^{n} u_i f_i^+ + \sum_{j=1}^{m} v_j f_j^- \tag{1.2.2a}$$

$$u_i + v_j \leq c_{i,j} \ . \tag{1.2.2b}$$

For the sake of simplicity, to better understand the above discrete formulations, we report here a basic result of operation research (see [7]), which explains how a discrete primal minimization problem is related to its dual, that is a maximization problem.

$$\min_{\boldsymbol{x}} \left\{ \boldsymbol{c}\cdot\boldsymbol{x} \ : \ \begin{matrix} \boldsymbol{A}\boldsymbol{x} = \boldsymbol{b} \\ \boldsymbol{x} \geq 0 \end{matrix} \right\} = \max_{\boldsymbol{y}} \left\{ \boldsymbol{b}\cdot\boldsymbol{y} \ : \ \boldsymbol{A}^T\boldsymbol{y} \leq \boldsymbol{c} \right\} \tag{1.2.3}$$

$$\boldsymbol{A} \in \mathbb{R}^{m,n} \quad \boldsymbol{x},\boldsymbol{c} \in \mathbb{R}^n \quad \boldsymbol{y},\boldsymbol{b} \in \mathbb{R}^m \ .$$

Moving from the discrete to the continuous setting, the masses of the initial and of the final configurations are no more divided into different positions, but they are arranged with initial and final “densities”. The solution of the relaxed formulation will now be searched not in the set of the transport maps $\mathcal{T}(f^+, f^-)$, but among the *transport plans*:

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

where π_x and π_y are the projection maps $(x, y) \mapsto x$ and $(x, y) \mapsto y$. Thus γ and the two constraints are the analogous in the continuous case of γ_{ij} and of the constraints of equation (1.2.1). We can now formulate the Kantorovich Primal problem.

Problem 4 (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 $\gamma^* \in \Gamma(f^+, f^-)$ that solves

$$\min_{\gamma \in \Gamma(f^+,f^-)} \mathcal{K}_c(\gamma) := \int_{X \times Y} c(x,y) d\gamma(x,y) \ . \tag{1.2.4}$$

The relaxation of Monge's original problem provided by Kantorovich is such that for any map $T \in \mathcal{T}(X,Y)$ we can always define a plan γ that belongs to $\Gamma(f^+, f^-)$. A first important advantage of problem 4 is that, under very mild assumptions on the cost function, it admits a solution, as stated in the following theorem.

Theorem 5
For any $c : X \times Y \mapsto \mathbb{R}$ lower semi-continuous, problem 4 admits a solution $\gamma^* \in \Gamma(f^+, f^-)$.

The proof of the above theorem can be found in [30], and it is based on the classical direct method of the calculus of variations. A second advantage is that it admits a dual, similarly to the discrete case. Thus, defining with $\mathcal{C}_b$ the space of continuous and bounded functions and with $\mathcal{L}_c$ the set

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

we are ready to define the Kantorovich Dual problem in the continuum.

Problem 6 (Kantorovich 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 $c : X \times Y \mapsto \mathbb{R}^+$, find the pair $(u^*, v^*) \in \mathcal{L}_c$ which solves the maximization problem:

$$\sup_{(u,v) \in \mathcal{L}_c} \mathcal{I}_{(f^+,f^-)}[u,v] := \int_X u(x) df^+(x) + \int_Y v(y) df^-(y) \ . \tag{1.2.5}$$

We will see later how the dual formulation plays a fundamental role in the analysis of OT. The following is the Kantorovich duality theorem, that is mentioned here for completeness and whose proof can be found in [33].

Theorem 7 (Kantorovich Duality)
Given two non-negative finite measures f^+ and f^- on X and Y satisfying $f^+(X) = f^-(Y)$, and given a cost function $c : X \times Y \mapsto \mathbb{R}$ lower semi-continuous, the following equality holds:

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

To conclude, we remark that the Kantorovich formulation is nowadays known as the *Monge-Kantorovich Transport* problem.

1.3 L^p-OT problem: $c(x, y) = |x - y|^p$

A typical cost function used in OT is $c(x, y) = |x - y|^p$, which gives origin to the so called L^p-OT problem, with $p > 1$. Referring to [30], we report a fundamental result for this class of problems.

Proposition 8
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 $\gamma^* \in \Gamma(f^+, f^-)$ of the form $\gamma^* = (Id, T^*)_\# f^+$, with $T^* \in \mathcal{T}(f^+, f^-)$. Moreover, there exists a Kantorovich potential u such that T^* satisfies the following relation:

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

where J is the Jacobian matrix.

This proposition ensures uniqueness of an optimal transport plan that is solution of the Kantorovich Primal problem, and the existence of an optimal map. Unfortunately, we are not able to apply the above proposition when $p = 1$, i. e. when we consider the Euclidean distance as cost function, that is what happens in the problem formulated by Monge.

1.3.1 Wasserstein distance

We give now an intuitive definition of the *Kantorovich-Rubinstein-Wasserstein distance*; this is known in literature just as Wasserstein distance, so we will denote it in this way. The p-Wasserstein distance, defined in [34], reads as follows.

Definition 9 (p-Wasserstein distance)
Given Ω an open, bounded, convex, and connected domain in $\mathbb{R}^d$ with smooth boundary and two non-negative finite measures f^+ and f^- on Ω satisfying $f^+(\Omega) = f^-(\Omega)$, and $p \geq 1$. The p-Wasserstein distance between f^+ and f^- is given by:

$$W_p(f^+, f^-) := \min_{\gamma} \left\{ \int_{\Omega\times\Omega} |x-y|^p \, d\gamma(x,y) \ : \ \gamma \in \Gamma(f^+, f^-) \right\}^{\frac{1}{p}} . \quad (1.3.1)$$

We would like to highlight the relation between this distance and problem 4, when the cost function is given by $c(x,y) = |x-y|^p$, showing that OT finds the plan that minimizes the p-Wasserstein distance.

The importance of the Wasserstein measure can be easily understood considering the following simple example, taken from [1]. Given two non-overlapping Dirac densities, we want to measure their distance. The 1-Wasserstein distance provides a more precise and smoother result than other measures, which may return a non-continuous distance or infinity. It is now commonly accepted that the p-Wasserstein distance can be exploited when we need to measure distance between densities, e. g. between two probability distributions. We will se later that the model we propose computes the 1-Wasserstein distance very efficiently.

1.4 L^1-OT problem: $c(x, y) = |x - y|$

We move now from the L^p-OT problem to the L^1-OT problem, where $p = 1$ and the cost function is the Euclidean distance $c(x, y) = |x - y|$, which is the cost considered in problem 1. We highlight that, due to this choice of $c(x, y)$, the total transport cost does not depend on the intermediate phases between the starting and the final configuration of the mass transported. From now on, we will consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected and convex domain with smooth boundary. Moreover, f^+ and f^- admit L^1-density with respect to the Lebesgue measure, meaning they are integrable functions. With a small abuse of notation, we will denote these densities with f^+ and f^-. The following considerations and results can be extended in case the above hypotheses do not hold [33].

The L^1-OT problem presents more pathological behaviour than those described in proposition 8, in fact the uniqueness of an optimal plan is not ensured. In addition, it is still a matter of research the minimal assumptions needed on f^+, f^- and Ω to ensure the existence of the optimal transport map solution of the Monge problem. Despite these difficulties, L^1-OT problem has a rich mathematical theory, and it presents different analogous formulations.

The first important result we want to highlight for the L^1-OT problem is given by the following theorem, which presents a different way to describe the Kantorovich Dual problem.

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 balanced densities f^+ and f^- on Ω. Problem 6, with cost function $c(x, y) = |x - y|$, can be rewritten as follows. Find $u^* \in Lip_1(\Omega)$ that solves

$$\sup_{u \in Lip_1(\Omega)} \int_\Omega u f \, dx \tag{1.4.1}$$

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 of the theorem can be found in [33], where it is proved that the equality $(u^*, v^*) = (u^*, -u^*)$ and the constraint $(u^*, v^*) \in \mathcal{L}_c$, when $c(x,y) = |x-y|$, lead to $u^* \in \mathrm{Lip}_1(\Omega)$.

Before going on with the discussion of the L^1-OT problem, we introduce a new formulation that is crucial for the purposes of this thesis.

Problem 11 (Beckmann Problem)
Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative balanced densities f^+ and f^- on Ω. Find $v^* \in [L^1(\Omega)]^d$ solving

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

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

$$\int_\Omega \nabla \varphi \cdot v \, dx = - \int_\Omega \varphi f \, dx \quad \forall \varphi \in \mathcal{C}^1(\bar{\Omega})$$

where $\bar{\Omega}$ stands for the closure of Ω.

We are now ready to state the following equivalence.

Proposition 12
Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative balanced densities f^+ and f^- on Ω , then the following equivalence holds:

$$\sup_{u \in Lip_1(\Omega)} \int_\Omega u f \, dx = \min_{v \in [L^1(\Omega)]^d} \left\{ \int_\Omega |v| \, dx \ : \ \nabla \cdot v = f \right\}$$

where $f = f^+ - f^-$.

The proof of this equivalence result can be found in [12]. The left hand side is equivalent to the Kantorovich Dual problem when the cost function is $c(x,y) = |x-y|$, as we have seen in theorem 10, while the right hand side is the Beckmann problem. Problem 11 gives a new point of view of the

L^1-OT problem. What is really interesting about this formulation is that we have no more a static problem: now the process is continuous since we have a constant injection and absorption of mass, and the only constraint of the problem is that the vector field v satisfies $\nabla\cdot v = f = f^+ - f^-$. As already noticed, in this formulation f^+ and f^- represent mass fluxes that are continuously injected and extracted, and the equation imposes the mass balance.

1.4.1 Monge-Kantorovich equations

Proposition 12 shows the direct connection between u^*, solution of problem 6, and v^*, solution of problem 11. A finer characterization of this relation is given by the following proposition, where we introduce the *Monge-Kantorovich equations* (MK equations), that represent one of the main result of the L^1-OT problem, see e. g. [13].

Proposition 13 (Monge-Kantorovich Equations)
Given $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative balanced densities f^+ and f^- on Ω. Consider u^* and v^*, solutions of problems 6 and 11, respectively, then the following equality holds:

$$v^* = -\mu^* \nabla u^* \tag{1.4.2}$$

where $\mu^*(f^+, f^-)$ is a L^1 positive density on Ω, called OT density.
The OT density μ^* and the Kantorovich potential u^* satisfy the following system:

$$-\nabla\cdot(\mu^* \nabla u^*) = f \qquad \textit{in } \Omega \tag{1.4.3a}$$

$$|\nabla u^*| \leq 1 \qquad \textit{in } \Omega \tag{1.4.3b}$$

$$|\nabla u^*| = 1 \qquad \textit{a.e. in } \mu^* > 0 \tag{1.4.3c}$$

where $f = f^+ - f^-$.

The fundamental result of this proposition is the existence of a transport potential, so that the flux v is a sort of "function" law, widespread in many applications. We highlight that ∇u^* tells us the direction we take while moving from the source f^+ to the sink f^-, while μ^* is representative of the amount of mass we have in each point, and can be interpreted as the flux intensity.

Remark 14

According to propositions 12 and 13, it is worth to note the equivalence between the 1-*Wasserstein distance and the Beckmann problem, that leads to the following:*

$$W_1(f^+, f^-) = \int_\Omega \mu^* \, dx \ . \tag{1.4.4}$$

Historically speaking, the MK equations were introduced with two different approaches in [5] and [13]. We summarize in figure 1.2 all the different formulations of the Optimal Transport Problem we introduced up to this point. Initially we move through relaxation from the Monge problem to the Kantorovich Primal problem, and here we connect the primal with its dual, both in the discrete and in the continuous setting. After the introduction of the OT density and of equation (1.4.2), we can link the Kantorovich formulations, the MK equations and the Beckmann problem.

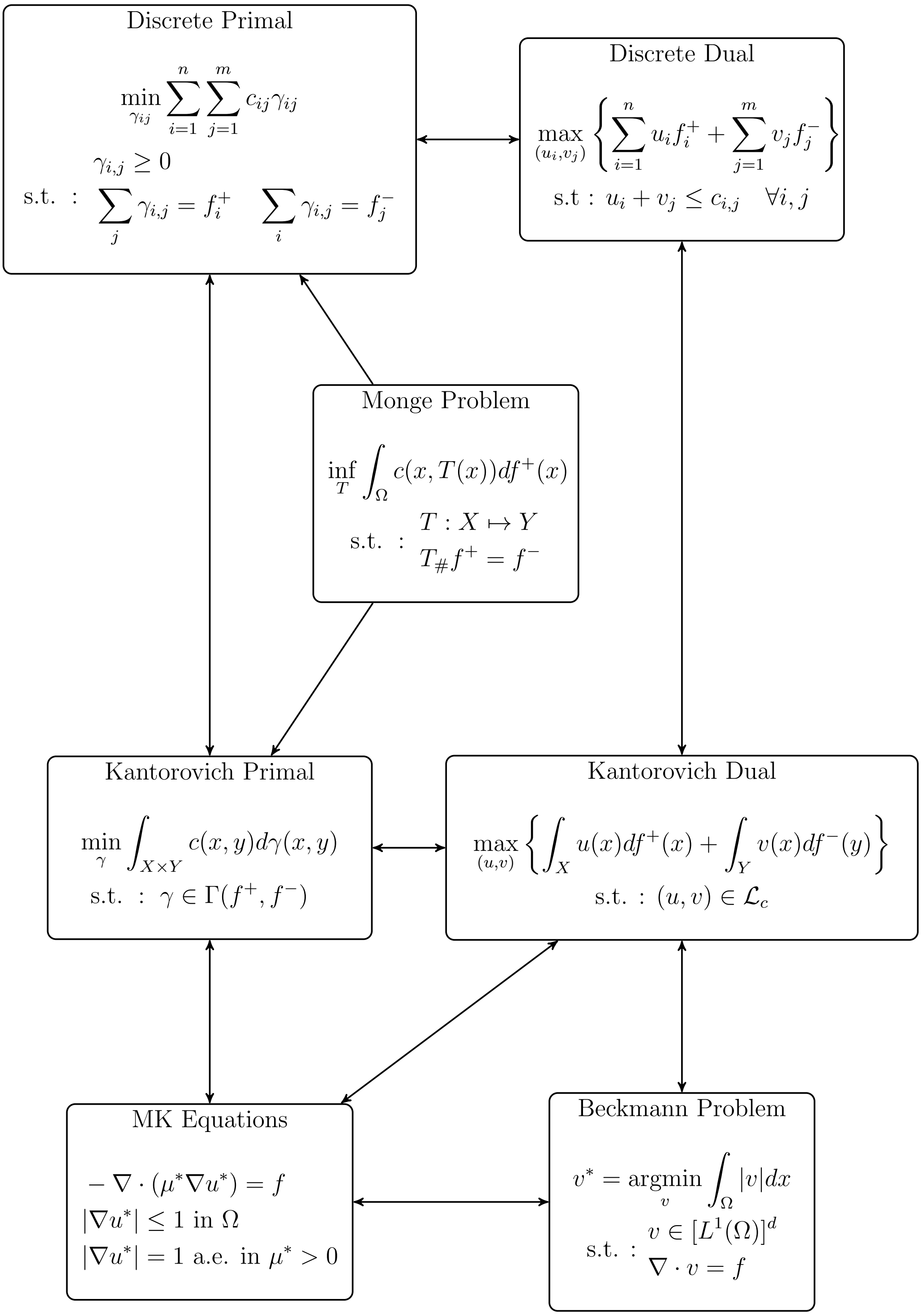

Discrete Primal
$\min_{\gamma_{ij}} \sum_{i=1}^{n}\sum_{j=1}^{m} c_{ij}\gamma_{ij}$
$\gamma_{i,j} \geq 0$
s.t. : $\sum_j \gamma_{i,j} = f_i^+ \quad \sum_i \gamma_{i,j} = f_j^-$
Discrete Dual
$\max_{(u_i,v_j)} \left\{ \sum_{i=1}^{n} u_i f_i^+ + \sum_{j=1}^{m} v_j f_j^- \right\}$
s.t : $u_i + v_j \leq c_{i,j} \quad \forall i, j$
Monge Problem
$\inf_T \int_\Omega c(x, T(x)) df^+(x)$
s.t. : $T : X \mapsto Y$
$T_\# f^+ = f^-$
Kantorovich Primal
$\min_\gamma \int_{X\times Y} c(x,y) d\gamma(x,y)$
s.t. : $\gamma \in \Gamma(f^+, f^-)$
Kantorovich Dual
$\max_{(u,v)} \left\{ \int_X u(x) df^+(x) + \int_Y v(x) df^-(y) \right\}$
s.t. : $(u,v) \in \mathcal{L}_c$
MK Equations
$-\nabla\cdot(\mu^*\nabla u^*) = f$
$|\nabla u^*| \leq 1$ in Ω
$|\nabla u^*| = 1$ a.e. in $\mu^* > 0$
Beckmann Problem
$v^* = \operatorname{argmin}_v \int_\Omega |v| dx$
s.t. : $v \in [L^1(\Omega)]^d$
$\nabla \cdot v = f$

Figure 1.2: Map of the connections among different formulations of the OT problem, with a particular focus on the L^1-OT formulation.

1.5 Divergence Constrained problem

In the Monge-Kantorovich formulation, the total transport cost depends only on the initial and final points, and not on the intermediate stages. This is not sufficient to describe all the environmental and industrial phenomena, as we can see in the following example.

Consider a courier that has to deliver two boxes from a delivery center to two different destinations. The problem is discrete, the final configuration is represented by two Dirac masses, i. e. $(f_j^- = \delta_{x_j})_{j=1,2}$, where x_j stands for the destination, while f^+ is a single Dirac source, with mass 2, located at the delivery center. L^1-OT problem answers sending each box to the corresponding destination along straight lines, as if every box goes straight from its starting to its final point, thus the path is a "V". However, this kind of transport is economically unrealistic and it is usually more convenient to aggregate the two boxes for the first part of the journey, and to split them when they are closer to the destination. Therefore, following this reasoning, the optimal transport path would have a "Y" shape. There may be also a third situation, in which the privileged routes of the boxes are independent, but more widespread than the ones of the L^1-OT problem. For this case, the optimal path has a "U" shape. Observe that in the L^1-OT problem the "V" shape is optimal since the transport cost per unit length is proportional to the amount of goods transported, while the encouragement of mass concentration and the need of a widespread path lead to the "Y" and "U" shapes, respectively.

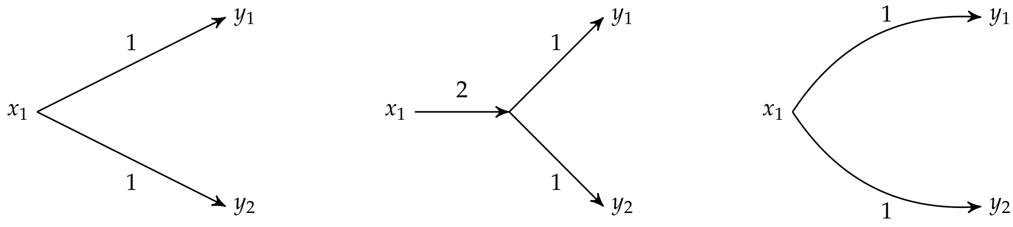


Figure 1.3: Schematic representation of the problem of delivering two boxes from a starting point x_1 to two destinations y_1 and y_2. The left pattern is the solution of the L^1-OT problem, presenting a "V" shape. The "Y" and the "U" shapes are reported in the middle and right patterns, respectively.

Therefore, the L^1-OT problem does not posses the capability to describe all the possible scenarios, and favouring or penalizing mass concentration are important factors we need to take into account. To mathematically model these two problems we introduce a new OT formulation, which comes from the generalization of the Beckmann problem and that reads as follows.

Problem 15 (Divergence Constrained Problem)
Given $0 < q \leq 2$, find among all vector-valued functions $v : \Omega \mapsto \mathbb{R}^d$, orthogonal to the boundary of Ω, the optimal v^* that solves:

$$\inf_v \left\{ \int_\Omega |v|^q \ : \ \nabla\cdot v = f^+ - f^- \right\} . \tag{1.5.1}$$

The kind of the transport depends on the value of the coefficient q. Of course we observe that when $q = 1$ we have the Beckmann problem. In the case $q \in (0,1)$, the concave exponent encourages mass accumulation, and this problem is called *Branched Transport* (BT) problem, while when $q \in (1,2]$, the convex power penalizes mass concentration along the path, and we have the *Congested Transport* (CT) problem.

1.5.1 Branched Transport problem

The BT problem encourages mass accumulation, leading to the formation of ramified structures, typical of many natural systems. These branched paths produce singular measures, thus the integration process requires a proper characterization due to regularity problems. This formulation was firstly introduced by Gilbert in [18], where he discussed the problem of finding the minimal cost network connecting cities.

The Gilbert-Steiner problem

Gilbert discussed the problem of finding the minimal cost communication network. He modelled the grid as a graph such that each edge was associated with a flow (or capacity). To this aim, he generalized the Steiner problem, which is the problem of finding the minimal length network connecting a set of points $x_1, \ldots, x_n$, using a sub-additive cost function. Denoting with $\varphi(q)$ the transport cost per unit length of an edge with flux q, this function encourages mass accumulation if it satisfies the following two properties:

$$\begin{gathered}\varphi(\max(q_1, q_2)) \leq \varphi(q_1 + q_2) \leq \varphi(q_1) + \varphi(q_2) \\ \varphi(q_1 + c) - \varphi(q_1) \leq \varphi(q_2 + c) - \varphi(q_2) \quad \forall c > 0,\ q_1 > q_2\ .\end{gathered}$$

The first property means that the transport cost increases as the transported mass grows, but it is sub-additive: it is more convenient to transport two parcels together rather than separately. In essence, it includes in the problem a sort of "economy of scale". The second property states that the marginal cost generated by adding a positive mass to a given background quantity is smaller for bigger backgrounds. A function satisfying both properties is the concave function $\varphi(x) = q^\alpha$ with $\alpha \in (0, 1)$.

To properly describe the Gilbert-Steiner problem we need to define the Transport Path.

Definition 16 (Transport Path)
Consider two atomic measures $f^+ = \sum_{i=1}^n f_i^+ \delta_{x_i}$ and $f^- = \sum_{j=1}^m f_j^- \delta_{x_j}$, where $\sum_{i=1}^n f_i^+ = \sum_{j=1}^m f_j^-$, (x_i, y_j) are points in $\Omega \subset \mathbb{R}^d$ and δ_p is the Dirac measure centered in p. An admissible Transport Path (G, q) 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 mass balance (Kirchhoff law):

$$\sum_{e \in \sigma(v)} q_e = \begin{cases} f_i^+ & \textit{if } v = x_i \textit{ for some } i \\ -f_j^- & \textit{if } v = y_j \textit{ for some } j \\ 0 & \textit{otherwise} \end{cases} \tag{1.5.2}$$

where $\sigma(v)$ is the star of v, i. e. the set of edges having vertex v in common.

Thus the Gilbert-Steiner problem reads as follows.

Problem 17 (Gilbert-Steiner Problem)

Given two balanced atomic masses f^+ and f^- and $\alpha \in [0,1]$, find the Transport Path (G,q) minimizing the Gilbert-Steiner energy

$$E_\alpha(G,q) = \sum_{e \in E(G)} (q_e)^\alpha L_e \tag{1.5.3}$$

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

The parameter α may vary in the interval $[0,1]$, and the external situations $\alpha = 0$ and $\alpha = 1$ correspond, respectively, to the Steiner problem and to a discrete version of the L^1-OTP. When $\alpha \in (0,1)$ we get the branched structures, typical of the BT formulation.

The main problem of the BT problem is that it is really hard to identify a minimizer, because we need to consider all the possible configurations, whose number grows with the number of source and sink points. The BT problem attempts to describe the branching structures typical of many natural systems as solution of minimal energy problems. For example, the formulation of problem 17 finds application in the study of river networks, where the principle of minimum energy is widespread in the study area of optimal channel networks. This is well described in [23] and [28]. River networks are solution of the following minimization problem:

$$\min_Q \sum_{e \in E(G)} Q_e^{\frac{1}{2}} L_e$$

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. The analogy with problem 17 is evident, the minimization principle corresponds to equation (1.5.3), while the flux Q_e satisfies Kirchhoff law.

Extension to the continuum

Problem 17 can also be extended to two positive mass densities f^+ and f^-, defined on the domain $\Omega \subset \mathbb{R}^d$. The intuitive idea is to take two atomic approximations f_n^+ and f_n^- of f^+ and f^-, i. e.

$$\begin{aligned} f_n^+ &\rightharpoonup f^+ \\ f_n^- &\rightharpoonup f^- \end{aligned}$$

where

$$\begin{aligned} f_n^+ &= \sum_{i=1}^{n} f_i^+ \delta_{x_i} \\ f_n^- &= \sum_{j=1}^{n} f_j^- \delta_{y_j} \ . \end{aligned}$$

Consider now the Transport Path (G_n, q_n) from f_n^+ to f_n^-, solution of problem 17. Then we can define the Transport Path from f^+ to f^- as the limit, with $n \to \infty$, of (G_n, q_n). For this purpose, we need to introduce the concepts of 1-rectifiable set in $\mathbb{R}^d$ and of 1-dimensional Hausdorff measure $\mathcal{H}^1$. The first can be seen as a countable union of Lipschitz curves, while the $\mathcal{H}^1$ measure of a simple curve is the length of the curve.

We can give now the definition of the BT problem for two positive mass densities f^+ and f^- defined on Ω, as it is given in [35].

Problem 18 (Branched Transport Problem)
Find $v^* \in [\mathcal{M}(\Omega)]^d$ solving

$$\min_{v \in [\mathcal{M}(\Omega)]^d} \left\{ \int_{E \subset \Omega} |v|^q \, d\mathcal{H}^1 \ : \ \nabla \cdot v = f^+ - f^- \right\}$$

where $E \subset \Omega$ is the 1-rectifiable union of the graph edges, $\mathcal{H}^1$ is the 1-dimensional Hausdorff measure, $q \in (0,1)$ and $\mathcal{M}(\Omega)$ is the set of measures defined on a measure space Ω. The divergence constraint on v is in the sense of distributions.

1.5.2 Congested Transport problem

In this section we analyze the CT problem, which penalizes mass concentration along the path, and it is defined as follows.

Problem 19 (Congested Transport Problem)
Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative balanced densities f^+ and f^- on Ω. Given $q \in (1,2]$, find the optimal vector field $v^* : \Omega \mapsto \mathbb{R}^d$ that solves

$$\min_{v \in [\mathcal{L}^q(\Omega)]^d} \left\{ \int_\Omega |v|^q \, dx \ : \ \nabla\cdot v = f^+ - f^- \right\} . \tag{1.5.4}$$

It can be shown that problem 19 is equivalent to the non-linear elliptic p-Poisson equation, which reads as follows.

Problem 20 (p-Poisson Equation)
Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative balanced densities f^+ and f^- on Ω. 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 define the p-Poisson equation as follows:

$$-\nabla\cdot(|\nabla u_p|^{p-2} \nabla u_p) = f^+ - f^- \tag{1.5.5}$$

complemented with zero Neumann boundary condition.

We remark that the exponent p of definition 9 is different from the exponent p of the p-Poisson equation, even if they are denoted in the same way. The following proposition affirms the equivalence between the Divergence Constrained formulation and the p-Poisson equation as a duality result.

Proposition 21
Consider $\Omega \subset \mathbb{R}^d$ an open, bounded, connected, and convex domain with smooth boundary. Take two non-negative balanced densities f^+ and f^- on Ω. 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. Then the following equivalence holds:

$$\max_{v\in[L^q(\Omega)]^d}\left\{-\int_\Omega \frac{|v|^q}{q}\,dx \ :\ \nabla\cdot v = f^+ - f^-\right\} = \min_{u\in W^{1,p}(\Omega)}\left\{\int_\Omega\left(\frac{1}{p}\left|\nabla\, u\right|^p - fu\right)dx\right\} \tag{1.5.6}$$

where $f = f^+ - f^-$ and $W^{1,p}(\Omega)$ is the Sobolev space. The solution u_p of the right-hand side problem and the solution $\bar{v}$ of the left-hand side problem satisfy the following relation:

$$\bar{v} = -\left|\nabla\, u_p\right|^{p-2}\nabla\, u_p\ .$$

The proof of the above proposition can be easily derived from the results present in [12]. Note that the right-hand side of equation (1.5.6) is the weak form of the p-Poisson equation.

The following schematic representation summarizes the different ways we can move the mass f^+ into f^-, and it highlights the relationship between the Divergence Constrained problems and the corresponding PDE formulation. As stated above, the coefficients p and q satisfy $\frac{1}{p}+\frac{1}{q}=1$, and $f = f^+ - f^-$. We want to highlight that both the Beckmann problem and the Congested Transport problem have a corresponding PDE formulation, while this does not occur for the Branched Transport problem.

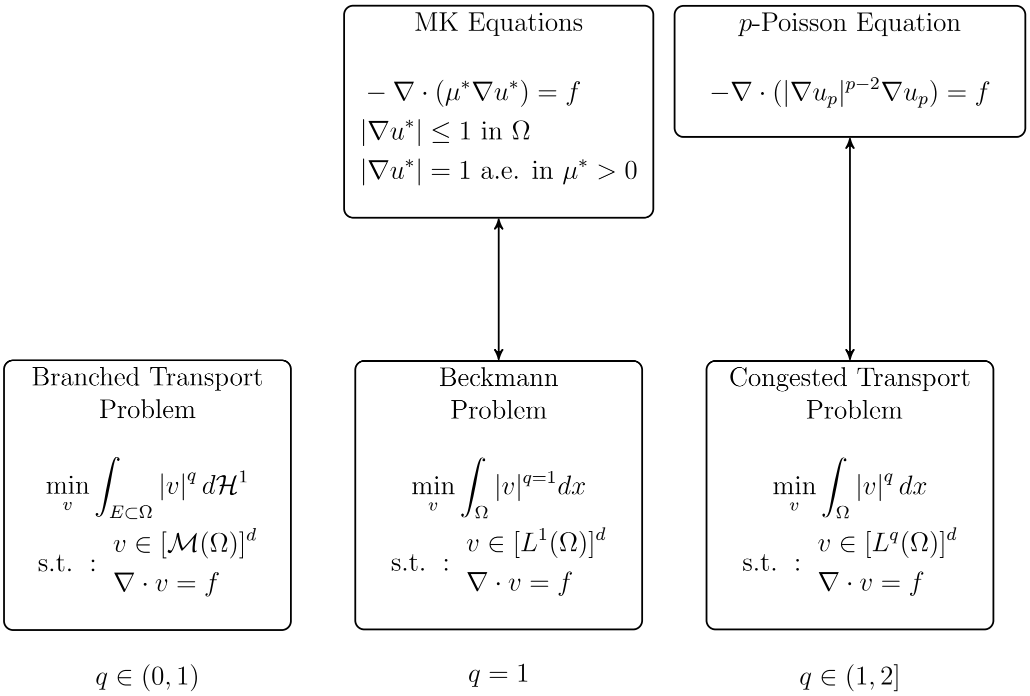

MK Equations
$-\nabla \cdot (\mu^* \nabla u^*) = f$
$|\nabla u^*| \leq 1$ in Ω
$|\nabla u^*| = 1$ a.e. in $\mu^* > 0$
p-Poisson Equation
$-\nabla \cdot (|\nabla u_p|^{p-2} \nabla u_p) = f$
Branched Transport Problem
$\min_v \int_{E \subset \Omega} |v|^q \, d\mathcal{H}^1$
s.t. : $v \in [\mathcal{M}(\Omega)]^d$, $\nabla \cdot v = f$
Beckmann Problem
$\min_v \int_\Omega |v|^{q=1} dx$
s.t. : $v \in [L^1(\Omega)]^d$, $\nabla \cdot v = f$
Congested Transport Problem
$\min_v \int_\Omega |v|^q \, dx$
s.t. : $v \in [L^q(\Omega)]^d$, $\nabla \cdot v = f$
$q \in (0,1)$
$q = 1$
$q \in (1,2]$

Figure 1.4: Schematic representation of the different ways of moving a mass f^+ into f^-, highlighting the connections between the Divergence Constrained formulations and the corresponding PDE. The exponent $q \in [0,2]$ modulates the way we move the mass from the initial to the final configurations. In case $q \in (0,1)$ mass accumulation is encouraged, while if $q \in (1,2)$ mass aggregation is discouraged; for $q = 1$ the problem is equivalent to the L^1-OT problem. The boundary values $q = 0$ and $q = 2$ correspond to the Steiner problem and to the Poisson equation, respectively.

Chapter 2

Dynamic Monge-Kantorovich formulation

In this chapter we introduce the *Dynamic Monge-Kantorovich* (DMK) equations developed in [14, 17] and further analyzed in [14, 15] and discuss how these "relaxed" dynamics can be effectively exploited to derive efficient numerical solutions for the OT problem.

2.1 Dynamics of Physarum Polycephalum

We start from the mathematical discrete model proposed in [31], which describes the behaviour of *Physarum Polycephalum* (PP): a slime mold possessing a remarkable path-finding capability in mazes. In fact, on the basis of experimental evidence (see [25]), PP is able to find the shortest route connecting food sources. A remarkable example of application of this optimization ability is the analysis of transportation networks in cities: e. g., as shown in [32], PP is able to reproduce the railroads of Tokyo.

In the experiment proposed in [25], initially a maze is filled by PP, then two food sources are added and PP starts modifying its shape, concentrating only on the shortest path connecting the two food sources. PP in the channels of the maze is schematized as an indirect planar graph $G = (V, E)$, where V is the set of vertices and E is the set of edges; furthermore we denote the positive edge length with $\{L_e\}_{e\in E}$ and the two nodal indices where the unitary food sources are located with $v = 1, n$. A conductivity function D_e is associated to each edge $e \in E$ and a potential (or pressure) function p_v to each node $v \in V$. The mathematical discrete formulation describing PP reads as follow. Find the optimal distribution of the pair (D_e, p_v) that satisfies

$$\sum_{e\in\sigma(v)} Q_e(t) = f_v = \begin{cases} +1 & v = 1 \\ -1 & v = n \\ 0 & v \neq 1, n \end{cases} \qquad \forall v \in V \qquad (2.1.1a)$$

$$Q_e(t) = D_e(t)\frac{(p_u(t) - p_v(t))}{L_e} \qquad \forall e \in E \qquad (2.1.1b)$$

$$D'_e(t) = |Q_e(t)|^\beta - D_e(t) \qquad \forall e \in E \qquad (2.1.1c)$$

$$D_e(0) = \hat{D}_e(0) > 0 \qquad \forall e \in E \qquad (2.1.1d)$$

where $e = (u, v)$ denotes the edge of G connecting vertices u and v, $\sigma(v)$ is the star of v, i. e. the set of edges having vertex v in common, and β is a non-negative coefficient. The model can be explained using a hydraulic analogy: we interpret the graph G as the pipes where a fluid flows driven by the vertex source functions. In this view, equation (2.1.1a) is the fluid mass balance, while equation (2.1.1b) is the momentum balance, affirming that the flux in each edge $e = (u, v)$ is directly proportional to the product between the discrete gradient of the potential function $(p_i)_{i=u,v}$ and the conductance coefficient D_e. We know the conductance coefficient is the inverse of the hydraulic resistance, and flow resistance is proportional to the reciprocal of the pipe diameter. Thus, the evolutive equation (2.1.1c) affirms that to allow the flow of larger fluxes with minimal energy loss the pipe diameter must increase. The decay term $-D_e(t)$ keeps the diameter bounded, compensating the growth of the hydraulic conductivity. Equation (2.1.1d) is the initial data. Using this mathematical model, in [31] the authors showed by numerical experiments that the conductivity D_e tends to localize on the edges of the shortest path between the two food sources.

In [4] we find a very important result for the above model: when time

$t \to \infty$, the distribution of the conductivity function D_e converges to the shortest path for a general planar graph G. Moreover, the authors proved the above mathematical discrete model is equivalent to the OT problem, applied on the graph G and having forcing terms satisfying:

$$\sum_{v \in V} f_v = 0 \ . \tag{2.1.2}$$

The PP problem can be reformulated as finding $Q^* = \{Q_e^*\}_{e \in E}$ such that:

$$\begin{aligned} &\min_{Q \in \{Q_e\}_{e \in E}} \sum_{e \in E} Q_e L_e \quad \text{s.t. :} \qquad (2.1.3) \\ &\sum_{e \in \sigma(v)} Q_e = f_v \qquad \text{for all } v \in V \ . \end{aligned}$$

To summarize, the solution of the discrete model describing PP converges to a stationary solution Q^*, which is solution of the above Optimal Transport problem.

2.2 Dynamic Monge-Kantorovich formulation

Moving from the discrete to the continuum, we do not have anymore a graph structure as for the PP dynamics, but an open and bounded domain $\Omega \subset \mathbb{R}^d$. This is the setting where the authors in [14, 16, 17] have worked and defined the following problem.

Problem 22 (Dynamic Monge-Kantorovich Problem)
Consider a balanced forcing function $f : \Omega \to \mathbb{R}$, i. e. $\int_\Omega f^+ = \int_\Omega f^-$, where $f = f^+ - f^-$ represents the difference between mass injected and absorbed, and thus it has to be balanced. Find the pair of functions $(\mu(t,\cdot), u(t,\cdot)) : [0, +\infty) \times \Omega \mapsto \mathbb{R}^+ \times \mathbb{R}^d$ that satisfies:

$$-\nabla\cdot \Big(\mu(t,x)\,\nabla\, u(t,x)\Big) = f(x) \tag{2.2.1a}$$

$$\partial_t \mu(t,x) = \mu(t,x)|\nabla\, u(t,x)| - \mu(t,x) \tag{2.2.1b}$$

$$\mu(0,x) = \mu_0(x) > 0 \tag{2.2.1c}$$

where μ is an isotropic conductivity coefficient and u is a potential function; the system is complemented by zero Neumann boundary conditions. Here, $\partial_t \mu$ denotes partial differentiation with respect to time, and $\nabla = \nabla_x$. We denote the system of equation (2.2.1) *as DMK equations.*

There is a close analogy between equations (2.1.1) and (2.2.1). Equation (2.2.1a) is the analogous in the continuum of the momentum balance equation (2.1.1b), where the flow is $q = -\mu \nabla u$, and equation (2.2.1b) is the evolutive equation, which is in analogy with equation (2.1.1c). The dynamics of the OT density described in equation (2.2.1b) has two components, the first term is the positive contribute, given by the flux magnitude, while the second is the decay term.

Analogously to the discrete model, in [14, 17] the authors make the following conjecture, relating equation (2.2.1) with the L^1-OT problem.

Conjecture 1
The solution pair $(\mu(t), u(t))$ of problem 22, 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, solution of the L^1-OT problem.

2.2.1 Existence and uniqueness

Conjecture 1 has not been mathematically proved yet, in fact the problem of showing existence and uniqueness of the solution pair (μ, u) of equation (2.2.1) is still open. Anyway, a proof of the local in time existence and uniqueness of the solution can be found in [14], under the assumptions of $f^+, f^- \in L^\infty(\Omega)$ and $\mu_0 \in \mathcal{C}^\delta(\Omega)$, where $\delta \in (0,1)$ and $\mathcal{C}^\delta(\Omega)$ is the set of the Hölder continuous functions in Ω:

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

2.2.2 Lyapunov-candidate functional

To add consistency to the previous conjecture, we introduce the Lyapunov-candidate functional, a function decreasing in time along $(\mu(t), u(t))$ that was firstly identified in [17]. Specifically for the L^1-OT problem, the Lyapunov-candidate functional is defined for $\mu \in L^1(\Omega)$, and it is given by

$$\mathcal{L}(\mu) := \frac{1}{2}\int_\Omega \mu |\nabla u(\mu)|^2 \, dx + \frac{1}{2}\int_\Omega \mu \, dx \ . \tag{2.2.2}$$

The Lyapunov-candidate functional is the sum of two terms: the first may be seen as the energy dissipated during the transport, while the second represents the cost of building the optimal transport infrastructure. Thus, we should look for the transport infrastructure μ^* which gives the optimal trade-off between the two components. We highlight that the Lyapunov-candidate functional is a decreasing function along the $\mu(t)$-trajectories, in fact its time derivative is given by

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

and it is easy to see that it is always non-positive. In [14, 17], the authors investigate if the Lyapunov-candidate functional $\mathcal{L}$ admits a minimum and if this is related to the p-Poisson equation, introduced in problem 20. The minimization of $\mathcal{L}$ is equivalent to problem 11, as we see in the following proposition, whoose proof can be found in [14].

Proposition 23

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 the Beckmann problem and the minimization of $\mathcal{L}$ are equivalent:

$$\min_{v \in [L^1(\Omega)]^d} \left\{ \int_\Omega |v| \, dx \ : \ \nabla\cdot v = f \quad \right\} = \min_{\mu \in L^1_+(\Omega)} \ \mathcal{L}(\mu) \tag{2.2.3}$$

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 $\mathcal{L}$.

2.3 Extended Dynamic Monge-Kantorovich formulation

L^1-OT problem is not enough to properly describe all the possible ways of moving a mass, as described in the previous chapter. For this reason in section 1.5 we extended problem 11 to problems 18 and 19. Thus, we generalize problem 22 adding an exponent β to the first term of the dynamic equation (2.2.1b), as suggested in [14, 15].

Problem 24 (Extended Dynamic Monge-Kantorovich Problem)
Consider a balanced forcing function $f : \Omega \to \mathbb{R}$, meaning $\int_\Omega f^+ = \int_\Omega f^-$, where $f = f^+ - f^-$ represents the difference between mass injected and absorbed and a coefficient $\beta \in (0, 2)$. Find the pair of functions $(\mu(t, \cdot), u(t, \cdot)) : [0, +\infty) \times \Omega \mapsto \mathbb{R}^+ \times \mathbb{R}^d$ that satisfies:

$$-\nabla\cdot\Big(\mu(t,x)\,\nabla u(t,x)\Big) = f(x) \tag{2.3.1a}$$

$$\partial_t\mu(t,x) = [\mu(t,x)|\nabla u(t,x)|]^\beta - \mu(t,x) \tag{2.3.1b}$$

$$\mu(0,x) = \mu_0(x) > 0\ , \tag{2.3.1c}$$

where μ is an isotropic conductivity coefficient and u is a potential function. The system is complemented by zero Neumann boundary conditions. Here, $\partial_t\mu$ still denotes partial differentiation with respect to time, and $\nabla = \nabla_x$. We define the system of equations (2.3.1a) *to* (2.3.1c) *as DMK equations, with a little abuse of notation.*

The evolutive equation (2.3.1b) states the time derivative of the OT density μ grows non-linearly with the flux $\mu(t, x)|\nabla u(t, x)|$. When $\beta \in (0, 1)$, the growth is sub-linear, and it penalizes the flux intensity, i. e. the OT density. Instead when $\beta \in (1, 2)$, the growth is super-linear, and the OT density accumulation is favoured. Thus, equation (2.3.1) in case $\beta \in (0, 1)$ is related to the Congested Transport problem, instead when $\beta \in (1, 2)$ the problem is in analogy with the Branched Transport problem. We have not considered in problem 24 the case $\beta = 0$, we remark now that this would lead the system to an explicit solution. Moreover, note that the restriction $\beta \in (0, 2)$

has been enforced for theoretical reasons, although numerical experiments, reported in [14], show a similar behaviour for $\beta \geq 2$.

We can find in [14] a generalization of conjecture 1.

Conjecture 2
The solution $(\mu(t), u(t))$ of problem 24 converges at large times to an equilibrium configuration $(\mu^*_\beta(\cdot), u^*_\beta(\cdot))$, as in conjecture 1.

2.3.1 Lyapunov-candidate functional

The generalization of the Lyapunov-candidate functional to a generic β reads as follows:

$$\mathcal{L}_\beta(\mu) := \frac{1}{2}\int_\Omega \mu |\nabla u(\mu)|^2 \, dx + \begin{cases} \frac{1}{2}\int_\Omega ln(\mu)\, dx & \text{if } \beta = 2 \\ \frac{1}{2}\int_\Omega \frac{\mu^{\frac{2-\beta}{\beta}}}{\frac{2-\beta}{\beta}}\, dx & \text{if } \beta \in (0,2)\ . \end{cases} \tag{2.3.2}$$

Note that typical solutions of the OT of Xia [35] and numerical solutions reported in [14, 15] show that the optimiser of $\mathcal{L}_\beta$ has a singular structure and thus the integrals in equation (2.3.2) need to be intended not in the Lebesgue sense, but accordingly to the singular measure arising from the solution. This is a still unresolved theoretical issue that will need to be addressed in future studies. The derivative of the Lyapunov-candidate functional along the $\mu(t)$ trajectory is given by:

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

and is always non-positive. We remark once more that the Lyapunov-candidate functional is the sum of two terms, the first representing the energy dissipated along the transport, the second standing for the cost of building the optimal infrastructure. Thus we have to look for the minimum of the Lyapunov-candidate functional. Therefore we focus on the existence of a minimum of the Lyapunov-candidate functional $\mathcal{L}_\beta$, and on its relation with the p-Poisson equation, as we have done in proposition 23.

Proposition 25

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

$$\min_{v \in [L^q(\Omega)]^d} \left\{ \int_\Omega \frac{|v|^q}{q} \, dx \ : \ \nabla \cdot v = f \right\} = \min_{\mu \in L_+^{P(\beta)}(\Omega)} \mathcal{L}_\beta(\mu) \tag{2.3.3}$$

where $L_+^{P(\beta)}(\Omega)$ denotes the space of the non-negative function in $L^{P(\beta)}(\Omega)$. Moreover, the functional $\mathcal{L}_\beta$ admits a unique minimizer $\mu_\beta^* \in L_+^{P(\beta)}(\Omega)$, given by

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

where u_p is the solution of the p-Poisson equation

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

with p conjugate exponent of q:

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

The proof can be found in [14]. Proposition 25 affirms the equivalence between problem 19 and the minimization of the Lyapunov-candidate functional, in case $\beta \in (0, 1)$. Concerning the relation between the Lyapunov-candidate functional and the p-Poisson equation, the authors in [14] propose the following conjecture.

Conjecture 3

When $\beta \in (0, 1)$, the solution (μ, u) of problem 24 converges, as $t \to \infty$, to the pair $(|\nabla u_p|^{p-2}, u_p)$, where u_p is the solution of the p-Poisson equation with

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

The conjecture holds for any initial condition μ_0. We can include also $\beta = 0$ in conjecture 2, due to equation (2.3.1) converging to the Poisson equation if $p = 2$.

In case $\beta \in (1,2)$, we still want to check if the Lyapunov-candidate functional $\mathcal{L}_\beta$ admits a minimum, and if $\mu(t)$ converges to this minimizer as $t \to \infty$. Unfortunately, there is no mathematical proof of this, so we are not able to state the analogous of propositions 23 and 25 when $\beta \in (1,2)$. On the other hand, in [14] the authors propose the following.

Conjecture 4
For $\beta > 1$ the solution $(\mu(t), u(t))$ of the DMK equations converges to the stationary point (μ^*_β, u^*_β), which is a minimum of the Lyapunov-candidate functional $\mathcal{L}_\beta$ and it depends on the initial condition μ_0.

To support this conjecture, the authors performed various numerical simulations showing that the solution (μ, u) of the DMK equations converges to the equilibrium solution (μ^*_β, u^*_β). Furthermore, the support of the numerical solution of the OT density μ^*_h approximates 1-dimensional structures typical of the BT problem.

To summarize, we have seen that solving problem 24 is analogous to solve the OT problem, and the value of β determines the behaviour of the mass while moving along the path. In addition, the Lyapunov-candidate functional decreases along the $\mu(t)$-trajectories, and looking for the minimum of this functional is analogous to solve the Divergence Constrained problem.

2.3.2 Relation with Wasserstein distance

In remark 14 we have seen the equivalence between the 1-Wasserstein distance $W_1(f^+, f^-)$ and the Beckmann problem. This equivalence holds when the coefficients p of the Wasserstein distance and q of the Divergence Constrained problem are equal to 1. The equivalence holds only in this specific case. We also know from propositions 23 and 25 the relation $p = \frac{2-\beta}{1-\beta}$, thus we say that equations (1.3.1) and (2.2.3) are equivalent only when $q = \beta = p = 1$. When the equality does not hold, we can only rely on propositions 23 and 25, but we don't know any relationship with the Wasserstein distance.

In figure 2.1 the above relations between coefficients are illustrated, and we remark once more that coefficients p, q and β are related, respectively, to the Wasserstein distance, the Divergence Constrained problem and the Lyapunov-candidate functional.

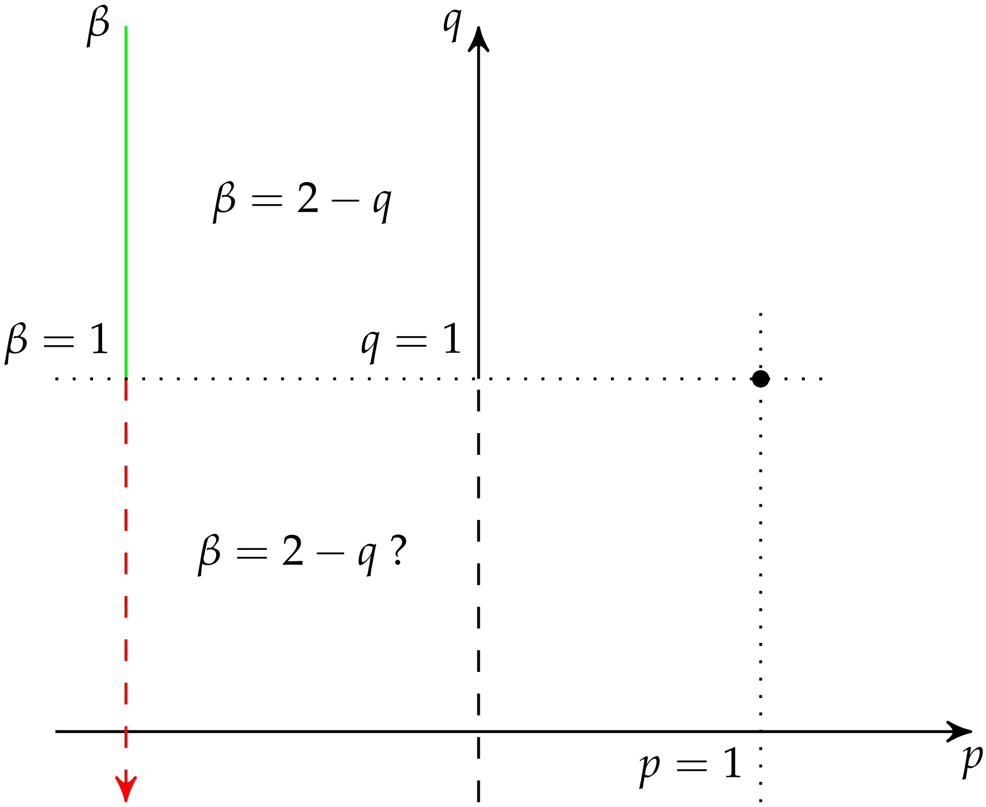


Figure 2.1: Relations among the coefficients p, q and β, respectively related to the Wasserstein distance, the Divergence Constrained problem and the Lyapunov-candidate functional. The red dashed line ($\beta > 1$) highlights the region where an analogous of propositions 23 and 25 exists only in the form of conjecture. The black dashed line ($q \in (0, 1)$) denotes the region where integration problems arise (problem 18). The circle indicates the point where the equality between equations (1.3.1) and (2.2.3) holds. This plot comes from an idea of F. Santambrogio.

Chapter 3

Three dimensional numerical solution of the DMK formulation

In this chapter we move toward the work specifically done in this thesis, describing the numerical approach we use to solve numerically the DMK equations. Firstly, we need to remark there is no mathematical proof of the convergence of the scheme because of the lack of global control on $|\nabla u|$, and what we do is to rely on numerical tests. To solve equation (2.3.1) our approach is based on the method of lines: we discretize all but one dimension, which is time that remains continuous; this leads to a system of non-linear ordinary differential-algebraic equations. Initially we focus on the projection spaces, many choices are analyzed in [14] and the most reliable and reasonable result is to look for a continuous approximation of the Kantorovich potential u and a more flexible and less regular approximation of the OT density μ. Then we focus on the uniform mesh refinement algorithm we implemented to refine the three dimensional grid. Successively, we talk about the temporal discretization we perform. The non-linear differential algebraic system is discretized by means of the first order forward Euler method. Finally, we focus on the solution of the arising system of differential algebraic equations.

3.1 Projection spaces

The spatial discretization of the DMK equations is achieved by projecting the weak formulation of the system onto a pair of finite dimensional spaces $(\mathcal{V}_h, \mathcal{W}_h)$. We denote with $\mathcal{T}_h(\Omega)$ the regular triangulation of the domain, assumed to be polygonal to avoid having to deal with the geometrical error induced by a piecewise linear approximation of the boundary. The computational mesh is characterized by N nodes, M tetrahedra and characteristic length h. We also denote with $\mathcal{P}_0(\mathcal{T}_h(\Omega)) = \text{span}\{\psi_1(x), \dots, \psi_M(x)\}$ the space of element-wise constant functions on $\mathcal{T}_h(\Omega)$, i. e. $\psi_i(x)$ is the characteristic function of the tetrahedron T_i, and with $\mathcal{P}_1(\mathcal{T}_h(\Omega)) = \text{span}\{\varphi_1(x), \dots, \varphi_N(x)\}$ the space of element-wise linear Lagrangian basis functions defined on $\mathcal{T}_h(\Omega)$. The choice of the space $\mathcal{V}_h$ for the projection of the weak formulation of the elliptic equation (2.3.1a) is $\mathcal{V}_h = \mathcal{P}_{1,h/2} = \mathcal{P}_1(\mathcal{T}_{h/2}(\Omega))$. We remark that the triangulation $\mathcal{T}_{h/2}(\Omega)$ is generated by uniformly refining each tetrahedron $T_k \in \mathcal{T}_h(\Omega)$, i. e. each element T_k is divided in 2^d sub-elements, and we will see later how this process is performed. Moving to the projection space of the dynamic equation (2.3.1b), we consider $\mathcal{W}_h = \mathcal{P}_{0,h} = \mathcal{P}_0(\mathcal{T}_h(\Omega))$. Different pairs of spaces $(\mathcal{V}_h, \mathcal{W}_h)$ were considered in [14], and the most efficient among them turned out being $(\mathcal{P}_{1,h/2}, \mathcal{P}_{0,h})$. Furthermore, some of the spaces considered in [14] present the classical lack of stability typical of a violation of an inf-sup-like constraint; unfortunately this condition is still not identified.

The discrete potential $u_h(t,x)$ and diffusion coefficient $\mu_h(t,x)$ are defined as follows:

$$
\begin{aligned}
u_h(t,x) &= \sum_{i=1}^{N} u_i(t)\varphi_i(x) \qquad && \varphi_i \in \mathcal{V}_h = \mathcal{P}_{1,h/2} = \mathcal{P}_1(\mathcal{T}_{h/2}(\Omega)) \\
\mu_h(t,x) &= \sum_{k=1}^{M} \mu_k(t)\psi_k(x) \qquad && \psi_k \in \mathcal{W}_h = \mathcal{P}_{0,h} = \mathcal{P}_0(\mathcal{T}_h(\Omega))
\end{aligned}
$$

where N and M are the dimensions of $\mathcal{V}_h$ and $\mathcal{W}_h$, respectively.

Problem 26 (Fem formulation)

For $t > 0$ find $(u_h(t,\cdot), \mu_h(t,\cdot)) \in \mathcal{V}_h \times \mathcal{W}_h$ such that

$$\int_{\Omega} \mu_h \nabla u_h \cdot \nabla \varphi_j \, dx = \int_{\Omega} f \varphi_j \, dx \qquad j = 1, \ldots, N \tag{3.1.1a}$$

$$\int_{\Omega} \partial_t \mu_h \psi_l \, dx = \int_{\Omega} \Big[(|\mu_h \ \nabla u_h|)^{\beta} - \mu_h \Big] \psi_l \, dx \qquad l = 1, \ldots, M \tag{3.1.1b}$$

$$\int_{\Omega} \mu_h(0,\cdot) \psi_l \, dx = \int_{\Omega} \mu_0 \psi_l \, dx \qquad l = 1, \ldots, M \tag{3.1.1c}$$

where $\mathcal{V}_h = \mathcal{P}_1(\mathcal{T}_{h/2}(\Omega))$, $\mathcal{W}_h = \mathcal{P}_0(\mathcal{T}_h(\Omega))$ and $\beta \in (0,2)$. In addition, we add to equation (3.1.1a) *the zero-mean constraint $\int_{\Omega} u_h \, dx = 0$ to enforce well-posedness.*

Moving toward the discrete formulation of the problem, we denote with $\boldsymbol{u}(t) = \{u_i(t)\}_{i=1,\ldots,N}$ and $\boldsymbol{\mu}(t) = \{\mu_k(t)\}_{k=1,\ldots,M}$ the vectors describing the time evolution of the projected system. The non-linear system of differential algebraic equations (DAE) is then:

$$\boldsymbol{A}[\boldsymbol{\mu}(t)]\boldsymbol{u}(t) = \boldsymbol{b} \tag{3.1.2a}$$

$$\partial_t \boldsymbol{\mu}(t) = \boldsymbol{D}[\boldsymbol{u}(t)]\boldsymbol{\mu}(t) \tag{3.1.2b}$$

$$\boldsymbol{\mu}(0) = \tilde{\boldsymbol{\mu}}_0 \ . \tag{3.1.2c}$$

The $N \times N$ stiffness matrix $\boldsymbol{A}[\boldsymbol{\mu}(t)]$ and the $M \times M$ diagonal matrix $\boldsymbol{D}[\boldsymbol{u}(t)]$ are, respectively:

$$A_{ij}[\boldsymbol{\mu}(t)] = \sum_{k=1}^{M} \mu_k(t) \int_{T_k} \psi_k \nabla \varphi_i \cdot \nabla \varphi_j \, dx \tag{3.1.3a}$$

$$D_{k,k}[\boldsymbol{u}(t)] = \frac{1}{|T_k|} \int_{T_k} \left(|\sum_{i=1}^{N} u_i(t) (\nabla \varphi_i)|_{T_k}|^{\beta} - 1 \right) dx \tag{3.1.3b}$$

where $|T_k|$ is the measure of the element T_k. The N components of the source vector $\boldsymbol{b}$ are

$$b_i = \int_{\Omega} f \ \varphi_i \, dx \ .$$

Besides, $\tilde{\boldsymbol{\mu}}_0$ is a M-dimensional vector whose components are given by

$$\tilde{\mu}_{0_k} = \frac{1}{|T_k|} \int_{T_k} \mu_0 \, dx \ ,$$

i. e. it is the L^2-projection of μ_0 on the tetrahedrons of $\mathcal{T}_h$.

3.2 Mesh refinement

We use a spatial discretization that requires to work simultaneously with a grid with mesh parameter h and the relative subgrid with mesh parameter $\frac{h}{2}$. In this section we explain the idea we exploited to refine the tetrahedral mesh. We followed the approach described in [27] to implement our own algorithm that uniformly refines a given grid.

We start from a tetrahedral grid for which we know the topology and the ensuing geometric properties for each cell: coordinates of the vertices, volume, baricenter, edges, faces, surfaces of the faces, neighbouring cells and nodes of each surface and edge. We denote with N_i the number of nodes relative to the grid i, with M_i the number of cells, and with E_i the number of edges. We refer to the original grid with $i = 1$ and with $i = 2$ to the subgrid. For the sake of simplicity, we refer to figure 3.1 to explain the refinement process, where we can graphically see it for one tetrahedron; the extension to the case of multiple tetrahedrons is straightforward. We subdivide the tetrahedron into 8 tetrahedra, and to do so we need to add new points: these are the middle points of each edge. Therefore, the number of nodes N_2 will be equal to $N_1 + E_1$. We remark that subdividing this way each edge we are sure that now the mesh parameter is $\frac{h}{2}$, consistently with the space discretization. We analyze now the building of the topology of the subgrid. The number of tetrahedra of the refined mesh is given by $M_2 = M_1 \times 8$. The 4 tetrahedra given by $(1, 5, 6, 7)$, $(2, 5, 8, 9)$, $(3, 6, 8, 10)$ and $(4, 7, 9, 10)$ are always present, regardless of the regularity of the cell. On the other hand, the other tetrahedra depend on the shape of the cell, and we choose the configuration that minimizes the Euclidean distance between the opposite pair of nodes $(5, 10)$, $(6, 9)$ and $(7, 8)$, thus satisfying the Delaunay property

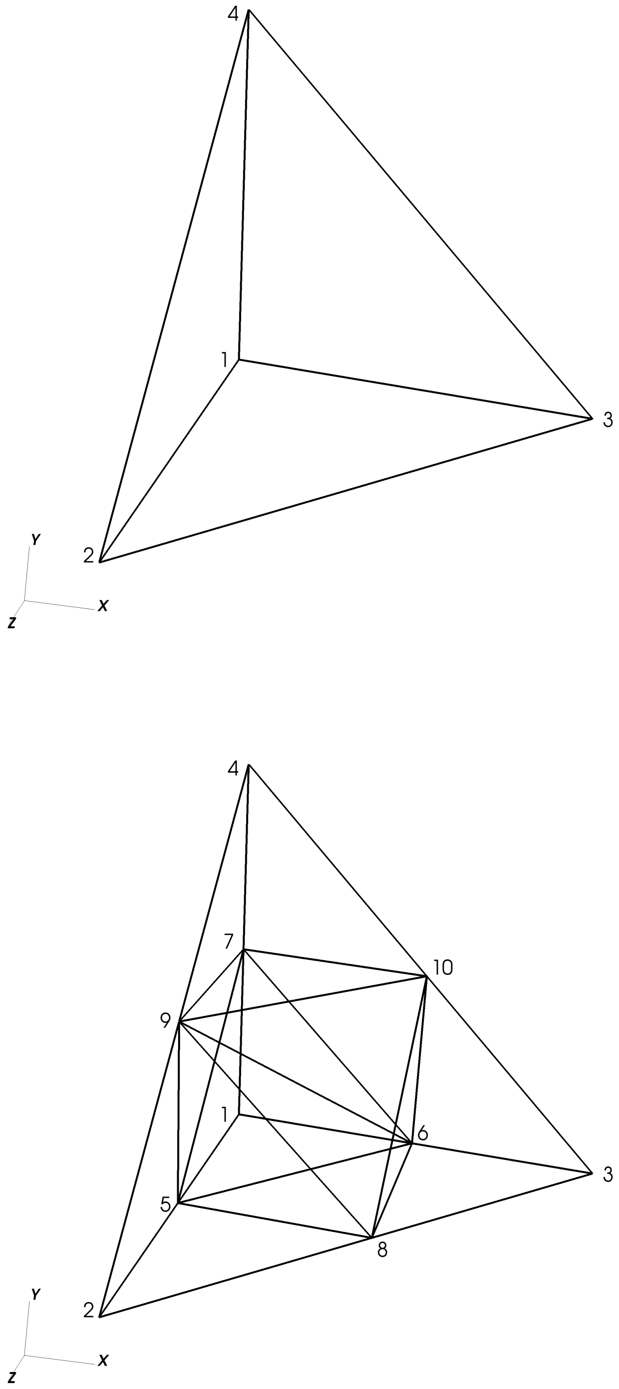


Figure 3.1: Single tetrahedron refinement example. For this refinement we have $N_1 = 4,\ N_2 = 10,\ M_1 = 1$ and $M_2 = 8$.

[11]. In figure 3.1 the pair $(6, 9)$ minimizes such distance, so we connect that vertices and we build the four tetrahedrons. The topology of the two meshes is summarized in the following table.

$(\text{topol})_1$	$(\text{topol})_2$
(1,2,3,4)	(1,5,6,7)
	(2,5,8,9)
	(3,6,8,10)
	(4,7,9,10)
	(5,6,7,9)
	(5,6,8,9)
	(6,7,9,10)
	(6,8,9,10)

In case of different choice of pair of nodes minimizing the Euclidean distance, the process is analogous.

The main problem we face exploiting such refinement is that the bandwidth we obtain is large, thus after the building of the subgrid we do a reorder of the mesh nodes following the Cuthill-McKee approach [9], to minimize the bandwidth of the stiffness matrix.

3.3 Time discretization

In order to solve the DAE equation (3.1.2) we introduce time discretization exploiting the forward Euler scheme. The approximate solution at time t_k can be written as

$$u_h^k(x) = \sum_{i=1}^{N} u_i^k \varphi_i(x)$$

$$\mu_h^k(x) = \sum_{l=1}^{M} \mu_l^k \psi_l(x)$$

where we define $(\boldsymbol{u}^k, \boldsymbol{\mu}^k) = (\boldsymbol{u}(t_k), \boldsymbol{\mu}(t_k)) = \left(\left\{u_i^k\right\}_{i=1,N}, \left\{\mu_l^k\right\}_{l=1,M}\right)$. Moreover, denoting with Δt_k the time-step size, i. e. $t_{k+1} = t_k + \Delta t_k$, and recalling

$\mu_h(t,\cdot) \in \mathcal{P}_{0,h}$, the forward Euler scheme reads as follows:

$$\begin{aligned}
&\boldsymbol{A}[\boldsymbol{\mu}^k]\ \boldsymbol{u}^k = \boldsymbol{b} \\
&\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k + \Delta t_k \big[\boldsymbol{D}[\boldsymbol{u}^{\boldsymbol{k}}](\boldsymbol{\mu}^k)^\beta - \boldsymbol{\mu}^k\big] \\
&\boldsymbol{\mu}^0 = \tilde{\boldsymbol{\mu}}_0
\end{aligned}$$

where the matrices $\boldsymbol{A}[\boldsymbol{\mu}^k]$ and $\boldsymbol{D}[\boldsymbol{u}^{\boldsymbol{k}}]$ are defined in equations (3.1.3a) and (3.1.3b). We remark here that using an explicit procedure like the forward Euler method introduces limitations on the time step size, which needs to be small enough to ensure numerical stability. In a future work we aim to exploit the backward Euler method, and to solve the implicit equation we will rely on the Newton algorithm.

3.4 Algorithm

The aim of our model is to optimally transport f^+ into f^-, and to do so we look for the pair $(\mu(t,x), u(t,x))$ solution of the DMK equations. We control the achievement of the large time equilibrium by monitoring the relative variation of μ_h between two successive time steps, defined as follows:

$$\operatorname{var}(\mu_h(t)) := \frac{\rho(\mu_h(t), \mu_h(t-\Delta t))}{\Delta t} \qquad (3.4.1)$$

where

$$\frac{\rho(\mu_h(t), \mu_h(t-\Delta t))}{\Delta t} := \frac{\|\mu_h(t) - \mu_h(t-\Delta t)\|_{L^2(\Omega)}}{\Delta t \|\mu_h(t-\Delta t)\|_{L^2(\Omega)}} .$$

and Δt is the distance between two successive time steps. Operatively, we start from the projected initial data μ_h^0 and we progress in time until $\operatorname{var}(\mu_h(t))$ is below a fixed threshold τ_{T}, or when we exceed a maximum number of time steps. When the first condition is achieved, we assume the conjectured asymptotic state is reached.

3.5 Solution of the linear system

The DAE equation (3.1.2) leads to a large sparse symmetric linear system, which is positive semi-definite because of homogeneous Neumann boundary conditions. We solve the system through the Preconditioned Conjugate Gradient (PCG) method, and we exploit the approach suggested in [19] to construct the Krylov subspace orthogonal to the null space of the system matrix. A Krylov subspace of order r, generated by a matrix $\boldsymbol{M}$ and a vector $\boldsymbol{c}$, is the linear subspace $K_r = \text{span}\left\{\boldsymbol{c}, \boldsymbol{M}\boldsymbol{c}, \boldsymbol{M}^2\boldsymbol{c}, \ldots, \boldsymbol{M}^{r-1}\boldsymbol{c}\right\}$. The idea is to maintain the generators of K_r always orthogonal to the kernel of $\boldsymbol{A}$. Since during the dynamical process some of the $\mu_l^k \to 0$, we evaluate a "near-kernel" of $\boldsymbol{A}$ by calculating the eigenvectors relative to eigenvalues that are smaller than a threshold (e. g. 10^{-10}) and use these as generators of Ker$(\boldsymbol{A})$. This is coupled with an effective spectral preconditioner, developed ad hoc in [3].

Convergence of the PCG method is considered achieved when the Euclidean norm of the residual relative to the initial residual norm is smaller than a fixed tolerance τ_{CG}. Operatively, we start from u_h^k, i. e. the solution at the previous time step, and we use an incomplete Cholesky 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 imposing $\mu_h \geq 10^{-10}$ everywhere. This is sufficient to guarantee the coercivity of the FEM bilinear form, which is a hypothesis of the Lax-Milgram theorem, and to keep bounded the system condition number, so that the PCG method converges within a limited number of iterations.

Chapter 4

Numerical experiments

In this chapter we focus on the numerical experiments we performed in order to verify the accuracy of our model. Firstly, we perform a test case for the L^1-OT problem, where we compare the numerical solution with a known analytical one, refining progressively the mesh and checking how the error behaves as the mesh parameter h decreases. Successively, the relation between the OT problem and the p-Poisson equation is verified for different values of $\beta \in (0, 1]$, matching the numerical with the analytical derived solution of the p-Poisson equation. Then, we solve the same problem of the first test case as a Branched Transport problem, and this numerical simulation allows us to observe the 1-dimensional structures typical of the BT formulation.

4.1 L^1-OT problem test case

Performing this test case we check the convergence of the numerical solution of our scheme toward the closed-form solution of the MK equations. The coefficient $\beta = 1$ and the problem we are solving is the L^1-OT problem. We consider a cubic domain in $\mathbb{R}^3$, $\Omega = [0, 1] \times [0, 1] \times [0, 1]$, a zero-mean function f, whose supports are two parallelepipeds Q^+ and Q^- contained in Ω. The forcing f assumes different signs in the two parallelepipeds, in fact one is the source and the other the sink. In addition, we identify with Q^c the support of the OT density μ^* in between the source and the sink. The three supports are the following:

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

This problem consists in optimally transporting $f^+ = f^+(Q^+)$ into $f^- = f^-(Q^-)$. We solve problem 22, recalling that , when $t \to \infty$ and $\beta = 1$, the solution of the DMK equations is conjectured to become stationary and to approximate the solution of the MK equations. We consider a steady-state configuration is achieved when the relative variation of the numerical density μ_h between two successive time iterations is smaller than a fixed tolerance τ_{T}, i. e.

$$\operatorname{var}(\mu_h^{k+1}) = \frac{\rho(\mu_h^{k+1}, \mu_h^k)}{\Delta t_k} = \frac{\|\mu_h^{k+1} - \mu_h^k\|_{L^2(\Omega)}}{\|\mu_h^k\|_{L^2(\Omega)} \Delta t_k} < \tau_{\mathrm{T}} \ .$$

We denote with t^* the time at which the numerical solution becomes stationary and with μ_h^* the corresponding μ_h^k, where k is the time step relative to time t^*. The forcing function f is constant and assumes opposite sign in Q^+ and Q^-. Specifically,

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

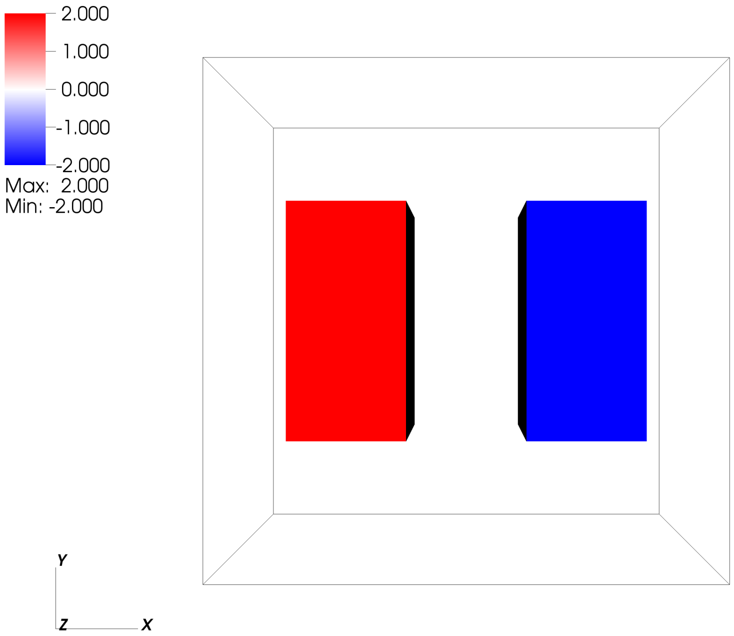


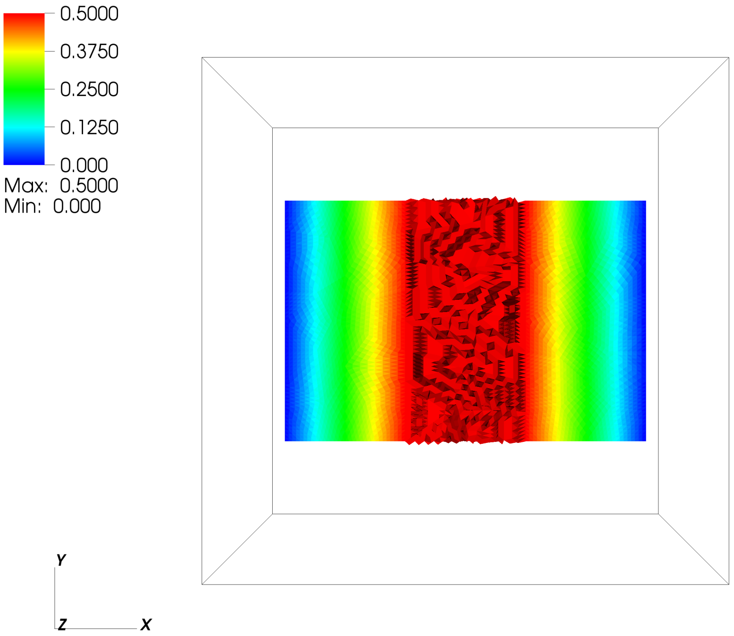
Figure 4.1: The forcing function f of the test case. We highlight that the mesh is aligned with the forcing function, and this is true for both mesh 1 and mesh 2.

and the corresponding OT density $\mu^*(f)$ is given by [6]:

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

We are solving the L^1-OT problem, whose aim is to minimize the Euclidean distance, thus the support of $\mu^*(f)$ is given by $Q^\mu = Q^+ \cup Q^c \cup Q^-$. We set as initial condition $\mu_0(x,y,z) = 1$ in the whole domain.

Two different mesh families are considered and we denote them as mesh 1 and mesh 2. Mesh 1 is aligned only with the two forcing parallelepipeds Q^+ and Q^- and we call it Q^f-aligned. On the other hand, mesh 2 is aligned



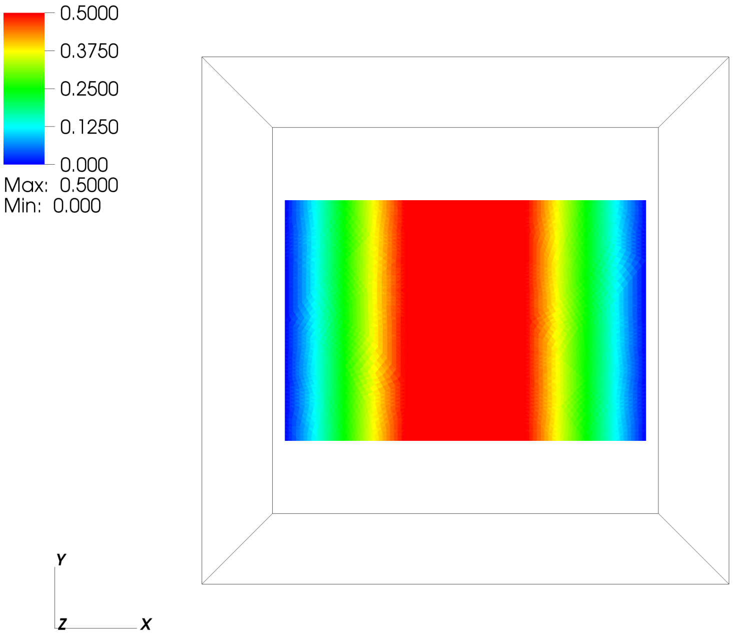


Figure 4.2: The two pictures represent the spatial distribution of the OT density of the L^1-OT problem test case. In the upper figure we consider mesh 1, while in the lower picture we have mesh 2. The two meshes present different tonalities of red in Q^c, which imply different orientations of the tetrahedrons. In fact, a light illuminates the meshes, and the non smooth surface produces the shadows we see in the top panel.

with both the forcing function in Q^+ and Q^- and the OT density in Q^c and we define it as Q^μ-aligned. Both meshes are aligned with the support of f, thus the constraint $\sum_i \int_\Omega f(x)\varphi_i \, dx = 0$ can be imposed exactly. The difference between the two meshes is shown in figure 4.2, where we show the view of the μ_h^* distribution from a vertical section locate far from Q^c. The different shades indicate different illumination of the tetrahedra faces that are not orthogonal to the lighting rays. The bottom figure shows that the Q^μ-aligned mesh has regular illumination planes, indicating alignment with Q^c. In order to asses the FEM convergence, we uniformly refine our grids a number of times. We expect that, for mesh 1, the convergence should be influenced also by the geometric convergence of the mesh boundary elements toward the support Q^μ of μ^*, and not only by the mesh parameter h.

We highlight that our aim does not lie in maximizing computational speed, but only in assessing the numerical behaviour of the system, and only once this is checked a successive step will be increasing the computational speed. Thus we do not limit the minimum time step size, the number of time iterations and the number of iterations for solving the linear system of algebraic equations through the PCG algorithm. The tolerances we impose are the followings: $\tau_{\mathrm{CG}} = 10^{-11}$ for the PCG exit, $\tau_{\mathrm{T}} = 5 \times 10^{-5}$ for stationarity. In all the simulations we adopted a varying time step Δt_k, whose size is tuned in according to the value $\mathrm{var}(\mu_h^k)$. The upper threshold for Δt_k we impose ensures the stability of the forward Euler scheme, to be more specific $\Delta t_k \in [0.005, 0.5]$ for each iteration. To asses the convergence of the FEM scheme, we look at the time behaviour of the $L^2(\Omega)$ relative error

$$\mathrm{err}(\mu_h(t), f) := \frac{\|\mu_h(t) - \mu^*(f)\|_{L^2(\Omega)}}{\|\mu^*(f)\|_{L^2(\Omega)}} \tag{4.1.1}$$

as $h \to 0$. In addition, we compute also the Wasserstein error, which evaluates the correctness of proposition 23. This error is given by:

$$\mathrm{err}_{W1}(\mu_h(t), f) := \frac{\mathcal{L}(\mu_h(t)) - \|\mu^*(f)\|_{L^1(\Omega)}}{\|\mu^*(f)\|_{L^1(\Omega)}} \tag{4.1.2}$$

where $\mathcal{L}(\mu_h(t))$ is the Lyapunov-candidate functional evaluated at time t.

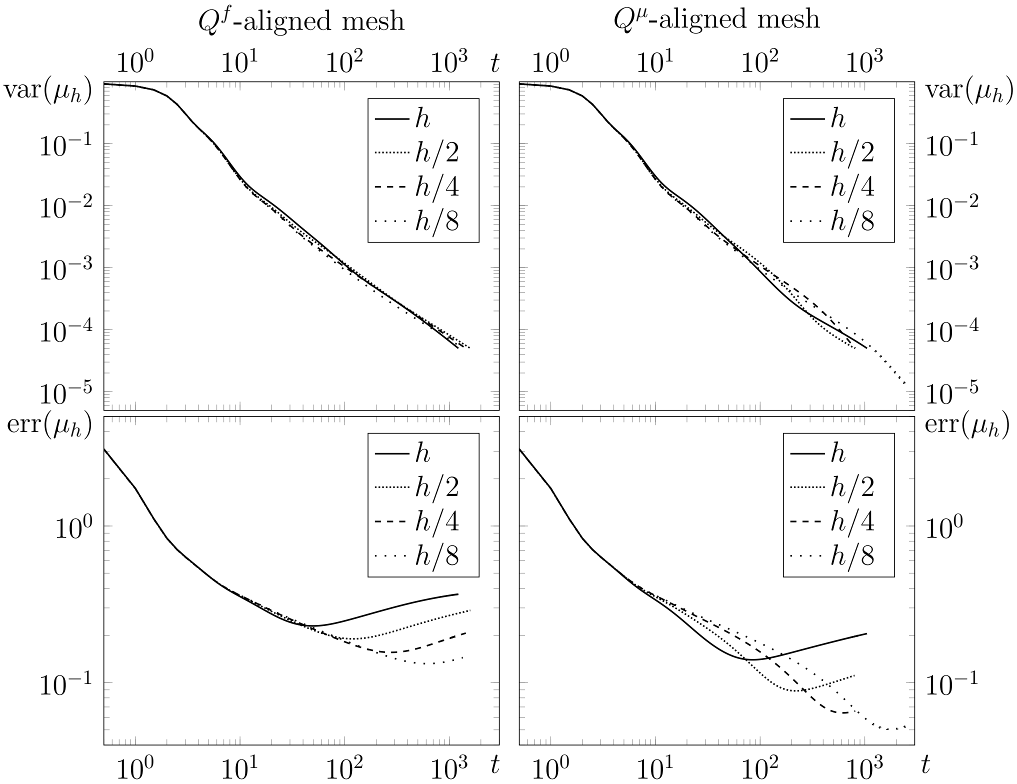


Figure 4.3: Convergence toward equilibrium for both the Q^f-aligned and the Q^μ-aligned meshes. The log-log plots of var($\mu_h(t)$) and err($\mu_h(t)$) vs. time are reported.

4.1.1 Convergence toward steady-state equilibrium

Initially we analyse the numerical convergence of the solution (μ_h, u_h) toward the equilibrium by looking at the time evolution of var($\mu_h(t)$) and err($\mu_h(t)$), which are reported in figure 4.3.

The μ_h variation displays a decreasing monotone behaviour for every refinement, with an expected convergence rate toward steady-state. For both the Q^f-aligned and Q^μ-aligned meshes, initially the var($\mu_h(t)$) plots coincide, while at larger times they deviate as a consequence of the higher spatial accuracy of the finer meshes. Moreover, we observe that also the err($\mu_h(t)$) curves initially coincide, and they start diverging when the corresponding spatial accuracy limit is attained. Accuracy saturation in the error plots (err($\mu_h(t)$)) vs. t occurs at the same time at which var($\mu_h(t)$) starts diverging. Differently from the two dimensional case, for both mesh families $\mathrm{err}(\mu_h(t))$ reaches a point of minimum and then begins a slight growth, and the reason of this trend is still unknown. Moreover, $\mathrm{err}(\mu_h(t))$ presents a more peculiar behaviour in the non-stationary interval for the Q^μ-aligned mesh, but the final value of $\mathrm{err}(\mu_h(t))$ gets consistently smaller as long as the mesh is refined. We also remark that $\mathrm{err}(\mu_h(t))$ of the finest of the Q^μ-aligned meshes required $\tau_\mathrm{T} = 1.2 \times 10^{-5}$ to reach the minimum value. Therefore, an idea to improve the computational speed of the model may be to initially use the coarsest grid, and progressively move to more refined grids only once the spatial accuracy limit is reached.

4.1.2 Convergence of the spatial discretization

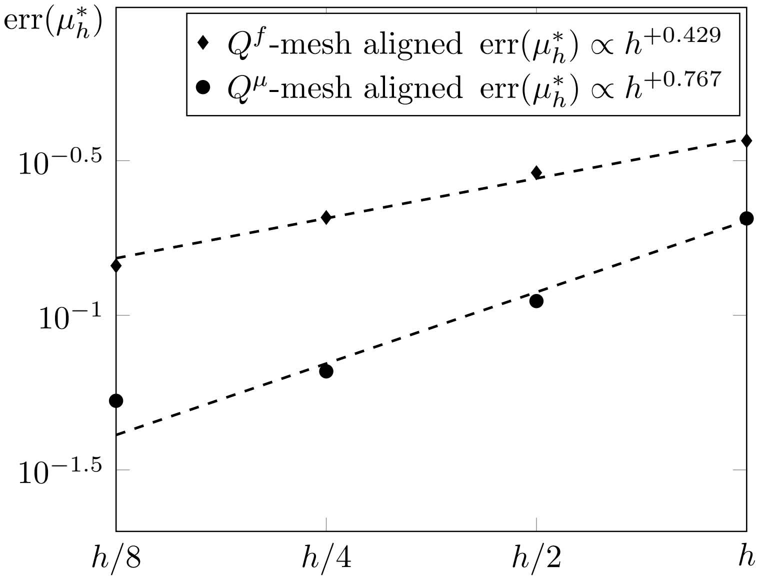


Figure 4.4: Log-log plot of $\mathrm{err}(\mu_h^*)$ vs. h for both the Q^f-aligned and the Q^μ-aligned meshes. The average experimental convergence rate are reported in the legend for both meshes.

We have just seen the L^2 relative error decreases as long as the mesh becomes finer. The experimental convergence profiles for the different meshes are reported in figure 4.4.

4.1.3 Dynamics of $\mathcal{L}(\mu(t))$ and Wasserstein error

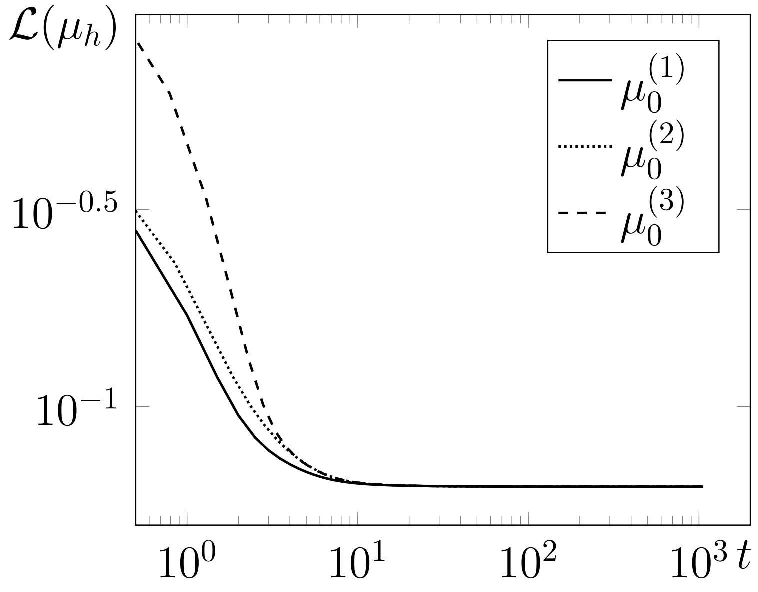


Figure 4.5: Time behaviour of the Lyapunov-candidate functional for different initial data. The log-log plot of $\mathcal{L}(\mu_h(t))$ vs. time is reported.

In figure 4.5 we look at the time behaviour of the Lyapunov-candidate functional $\mathcal{L}(\mu_h(t))$ for one of the Q^μ-aligned meshes, the result for the other meshes being practically indistinguishable. Initially $\mathcal{L}$ decreases monotically and then it becomes stationary; moreover different initial data lead to the same stationary value. The three different initial conditions for the OT density are:

$$\begin{aligned}
&\mu_0^{(1)} = 1\\
&\mu_0^{(2)} = 0.1 + 4((x-0.5)^2 + (y-0.5)^2 + (z-0.5)^2)\\
&\mu_0^{(3)} = 3 + 2\sin(8\pi x)\sin(8\pi y)\sin(8\pi z)\ .
\end{aligned}$$

We recall that the initial condition for the OT density for the other numerical simulations is $\mu_0^{(1)}$.

Figure 4.6 reports the Wasserstein error, defined in equation (4.1.2), for both the Q^f-aligned and the Q^μ-aligned meshes. We see that the Lyapunov-

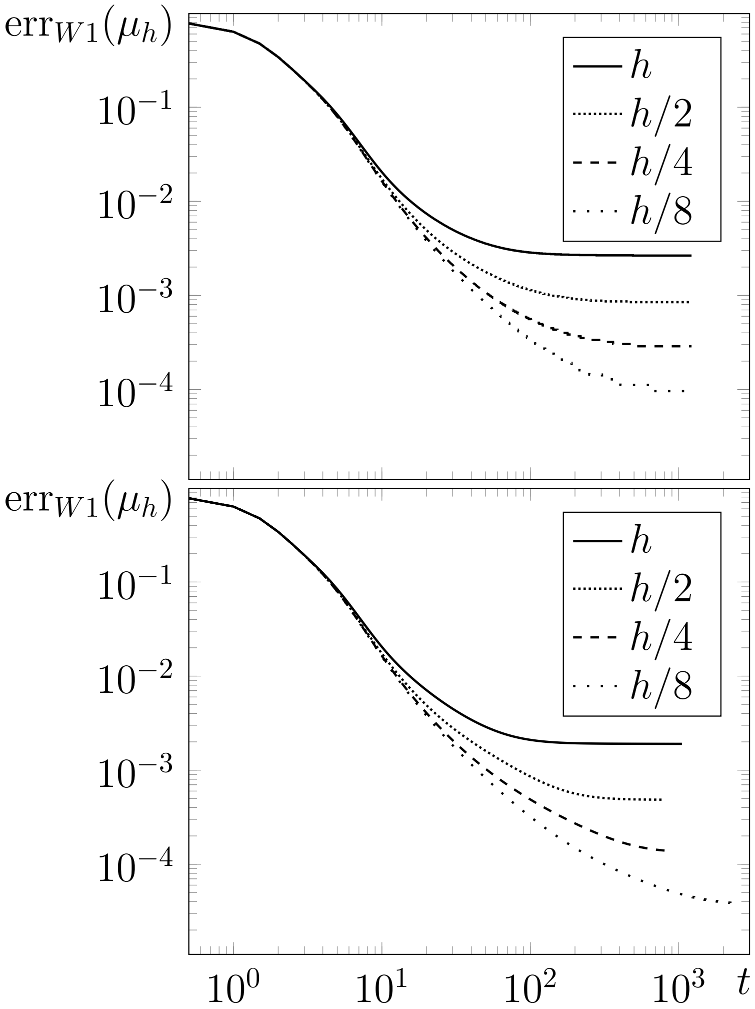


Figure 4.6: Log-log plot of the Wasserstein error $\mathrm{err}_{W1}(\mu_h(t))$ vs. time for both Q^f-aligned and Q^μ-aligned meshes. The top plot refers to the Q^f-aligned mesh, the bottom plot to the Q^μ-aligned mesh.

candidate functional is a good approximation of $W_1(f^+, f^-)$, the error decreases consistently as h gets smaller and $\mathrm{err}_{W1}(\mu_h(t))$ of the Q^μ-aligned-mesh is smaller than the one of the Q^f-aligned mesh.

4.1.4 Computational cost

In this thesis we are interested in evaluating the accuracy of our approach, which is still not optimized for computational speed. Nonetheless, we include

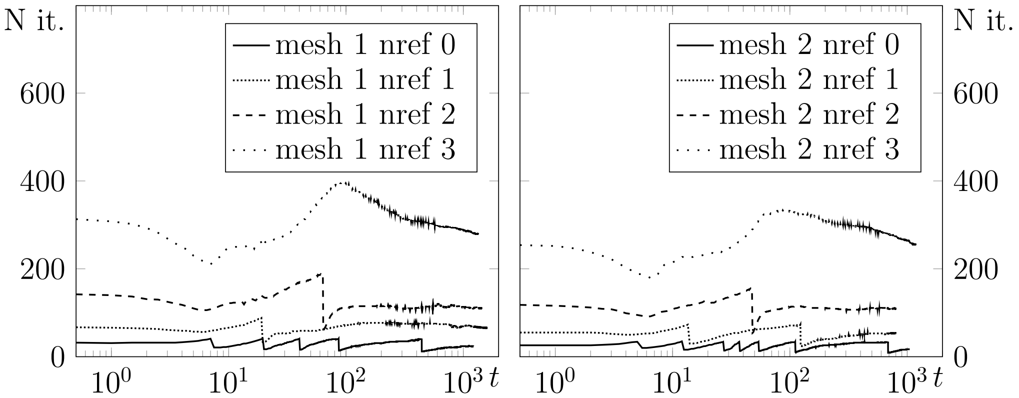


Figure 4.7: Number of iterations (N it.) to solve the linear system for each time step vs. time. The linear system is solved through the PCG scheme. The left plot is referred to the non-aligned mesh, the righ plot to the mesh aligned with the OT density in Q^c. With nref we denote the number of refinements.

a little discussion about the computational cost. We evaluate the computational cost with the total number of iterations required to solve the linear system arising from the discretization of the elliptic equation, and the time behaviour of the number of PCG iterations is reported in figure 4.7. These number of iterations grow with the size of the matrices, so with the mesh refinements, and they are more or less constant around their mean values for the whole time. We highlight that the number of iterations for the aligned mesh are slightly smaller than the ones of the non-aligned mesh.

Some ideas to improve the computational speed are the use of coarse-mesh solutions to extrapolate initial guess, or the use of an implicit scheme to improve stability and allow larger time step size.

4.2 p-Poisson test case

This test case evaluates the validity of conjecture 3, comparing $\mu_h(t, x)$ with $\mu^*_\beta := |\nabla u_p|^{p-2}$, where u_p is the solution of the p-Poisson equation, for which an explicit solution is developed. We denote with μ^*_h the long-time limit of $\mu_h(t, x)$. The domain is formed by three concentric spheres of radius

1, $\frac{2}{3}$ and $\frac{1}{3}$. The forcing term is balanced and radially symmetric, meaning that $f(x, y, z) = F(r)$, where $r = \sqrt{x^2 + y^2 + z^2}$, and $F : (0, 1) \mapsto \mathbb{R}$.

4.2.1 Analytical solution for the ball

We show now the calculations we performed to obtain the analytical expression of the OT density, solution of the p-Poisson equation for a ball centred in $(0, 0)$. We start from the p-Poisson equation

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

where $u_p = u_p(x, y, z)$ and $f = f(x, y, z)$. We move to spherical coordinates, thus $u_p(x, y, z) = U(r)$ and $f(x, y, z) = F(r)$. The above equation becomes

$$r^{1-d} \frac{d}{dr} \left(r^{d-1} |U'(r)|^{p-2} U'(r) \right) = F(r)$$

where d is the dimension of the problem. Moving r^{1-d} to the right hand side and integrating between 0 and r, we get

$$r^{d-1} |U'(r)|^{p-2} U'(r) = \frac{\int_0^r t^{d-1} F(t) dt}{r^{d-1}} \ ,$$

then we apply the absolute value to both sides of the equation and we raise to the power $\frac{p-2}{p-1}$. We obtain

$$|U'(r)|^{p-2} = \frac{\left|\int_0^r t^{d-1} F(t) dt\right|^{\frac{p-2}{p-1}}}{r^{(d-1)\frac{p-2}{p-1}}} \ .$$

Moving back to Cartesian coordinates, and knowing from proposition 25 that, for $\beta \in (0, 1)$, $p = \frac{2-\beta}{1-\beta}$ and the pair $(\mu(t), u(t))$, solution of equation (2.3.1), converges to the pair $(|\nabla u_p|^{p-2}, u_p)$, we finally obtain the explicit formula for μ^*_β:

$$\mu^*_\beta(x,y,z) = |Z(r)|^{\frac{p-2}{p-1}} = |Z(r)|^\beta \tag{4.2.1}$$

$$Z(r) = \frac{1}{r^{d-1}}\int_0^r t^{d-1}F(t)dt\ . \tag{4.2.2}$$

4.2.2 Numerical simulation

The forcing function we consider is piecewise constant, positive in the interval $r \in (0, \frac{1}{3})$, null in $r \in (\frac{1}{3}, \frac{2}{3})$ and negative when $r \in (\frac{2}{3}, 1)$. Moreover, f verifies the mass conservation principle, i. e. $\int_0^{\frac{1}{3}} F(r)dr = \int_{\frac{2}{3}}^1 F(r)dr$. We performed the experiments for a sequence of conformally refined grids and for four different values of β: $0.25,\ 0.5,\ 0.75$ and the limit value of 1. The grid refinement considered are characterized by a ratio between successive mesh parameters of 1.2, and not of 2 as in the others numerical experiments; i. e. for this test case we have $\frac{h_i}{h_{i+1}} = 1.2$. It is trivial to say conjecture 3 holds also for the boundary case $\beta = 1$. In particular, for this value equation (4.2.1) represents the OT density solving the MK equations.

The tolerances we impose are as follows: $\tau_{\mathrm{T}} = 5 \times 10^{-5}$ for stationarity and $\tau_{\mathrm{CG}} = 10^{-11}$ for the PCG exit. The initial condition is $\mu_0 = 1$ in the whole domain. In all the simulations we adopted a varying time step Δt_k, whose size is tuned in according to the value $\mathrm{var}(\mu_h^k)$. The upper threshold for Δt_k we impose ensures the stability of the forward Euler scheme.

We evaluate the relative variation and the relative L^2 error, which we recall are given by:

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

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

where f is the balanced forcing function.

Evaluation of the Wasserstein error

In addition to the L^2 error, we also compute the time behaviour of another error, that evaluates the correctness of propositions 23 and 25. With a little abuse of notation, we call it Wasserstein error, remarking that, in case $\beta = 1$, it coincides with equation (4.1.2). We start from proposition 25, and we denote with v^* the solution of

$$\min_{v \in [L^q(\Omega)]^d} \left\{ \int_\Omega \frac{|v|^q}{q} dx \ : \ \nabla\cdot v = f \right\} .$$

From proposition 25 we know $q = 2 - \beta$, thus we have

$$\int_\Omega \frac{|v^*|^{2-\beta}}{2-\beta} .$$

Moreover, from proposition 21 we know that $v^* = -|\nabla u_p|^{p-2} \nabla u_p$, where u_p is the solution of the p-Poisson equation. In conjecture 3 we affirmed that at large times μ^*_β converges to $|\nabla u_p|^{p-2}$, and knowing from proposition 25 that the equality $p = \frac{2-\beta}{1-\beta}$ holds, we arrive at the following optimal "Wasserstein" distance:

$$\int_\Omega \frac{(\mu^*)^{\frac{2-\beta}{\beta}}}{2-\beta} = \| \frac{\mu^*(f)^{\frac{2-\beta}{\beta}}}{2-\beta} \|_{L^1(\Omega)}$$

where without any risk of misunderstanding we denote μ^*_β with μ^*.

Finally, the error reads as follows:

$$\mathrm{err}_{W1}(\mu_h(t), f) := \frac{\mathcal{L}_\beta(\mu_h(t)) - \| \frac{\mu^*(f)^{\frac{2-\beta}{\beta}}}{2-\beta} \|_{L^1(\Omega)}}{\| \frac{\mu^*(f)^{\frac{2-\beta}{\beta}}}{2-\beta} \|_{L^1(\Omega)}} \tag{4.2.3}$$

where $\mathcal{L}_\beta(\mu_h(t))$ is the Lyapunov-candidate functional evaluated at time t, and $\| \frac{\mu^*(f)^{\frac{2-\beta}{\beta}}}{2-\beta} \|_{L^1(\Omega)}$ is derived above.

Convergence toward steady-state and convergence profiles

In figure 4.8 we observe the behaviour of $\mathrm{var}(\mu_h)$, $\mathrm{err}(\mu_h)$, and $\mathrm{err}_{W1}(\mu_h)$ for this test case. The errors decrease consistently refining the mesh and for all

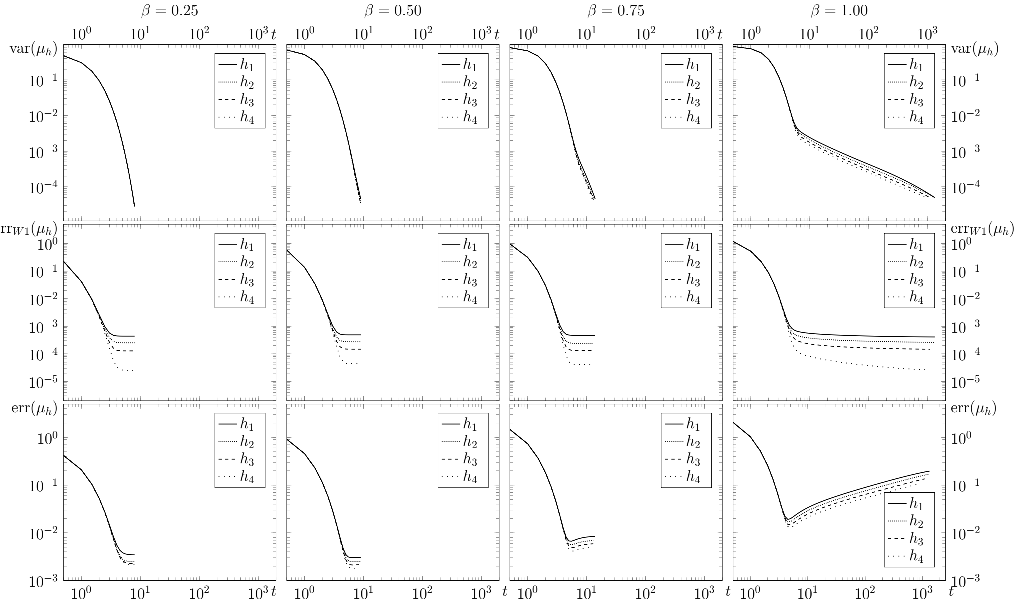


Figure 4.8: Log-log time behaviour of var($\mu_h(t)$), err($\mu_h(t)$) and err$_{W1}$($\mu_h(t)$). The ratio between successive mesh parameters h is $\frac{h_i}{h_{i+1}} = 1.2$.

the simulations the equilibrium configuration is achieved. In case $\beta = 1$, err(μ_h) starts growing after the point of minimum, and we suppose this is due to symmetry errors between the mesh-aligned and the explicit solutions. For the CT problem, i. e. for $\beta < 1$, the system reaches stationarity very quickly in time with respect to the case $\beta = 1$.

In figure 4.9 we report the values of err(μ_h^*) for successive refinements for all the values of β considered. err(μ_h^*) decreases consistently with the mesh refinement, and the experimental rate of convergence of the scheme seems to be proportional to h^m, where h is the mesh characteristic length and m a coefficient that grows with β.

4.3 BT problem test case

In this section we present a numerical experiment we performed imposing $\beta = 1.5$ in equation (2.3.1). The domain, the forcing function, and the support of the forcing function we consider are the same of the L^1-OT problem

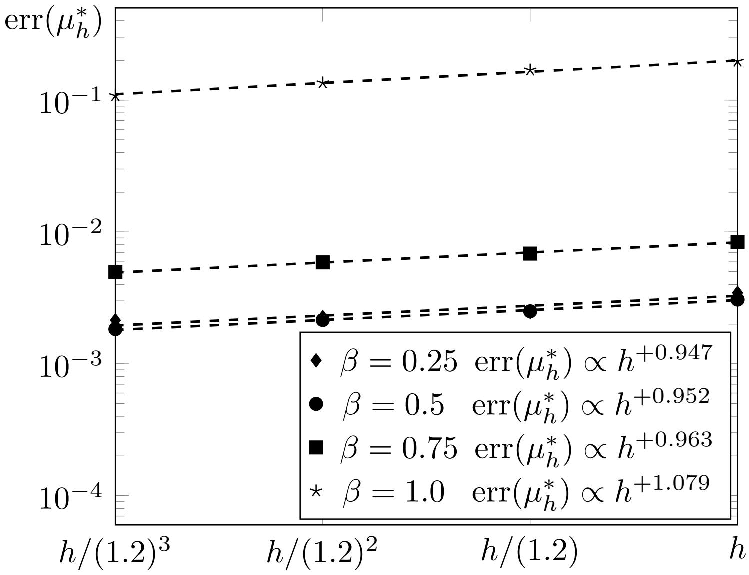


Figure 4.9: Log-log plot of $\mathrm{err}(\mu_h^*)$ vs. the mesh parameter h, for all the values of β considered. In the legend we report the average experimental rate of convergence for each β. The ratio between successive mesh parameters h is $\frac{h_i}{h_{i+1}} = 1.2$.

test case; which we recall are:

$$\begin{aligned}
&\Omega = \{(x,y,z) \in \Omega : (x,y,z) \in [0,1] \times [0,1] \times [0,1]\}\\
&f(x,y,z) = \begin{cases} 2 & \text{in } Q^+ \\ -2 & \text{in } Q^- \\ 0 & \text{elsewhere} \end{cases}\\
&Q^+ = \left\{(x,y,z) \in \Omega : (x,y,z) \in \left[\frac{1}{8},\frac{3}{8}\right] \times \left[\frac{1}{4},\frac{3}{4}\right] \times \left[\frac{1}{4},\frac{3}{4}\right]\right\}\\
&Q^- = \left\{(x,y,z) \in \Omega : (x,y,z) \in \left[\frac{5}{8},\frac{7}{8}\right] \times \left[\frac{1}{4},\frac{3}{4}\right] \times \left[\frac{1}{4},\frac{3}{4}\right]\right\} .
\end{aligned}$$

Again, the tolerances we impose are $\tau_{\mathrm{T}} = 5 \times 10^{-5}$ for stationarity and $\tau_{\mathrm{CG}} = 10^{-11}$ for the PCG exit. The initial condition for the OT density is $\mu_0 = 1$ in the whole domain. As in the previous test cases, the original mesh is conformally refined up to three times. In all the simulations we adopted a varying time step Δt_k, whose size is tuned according to the value $\mathrm{var}(\mu_h^k)$.

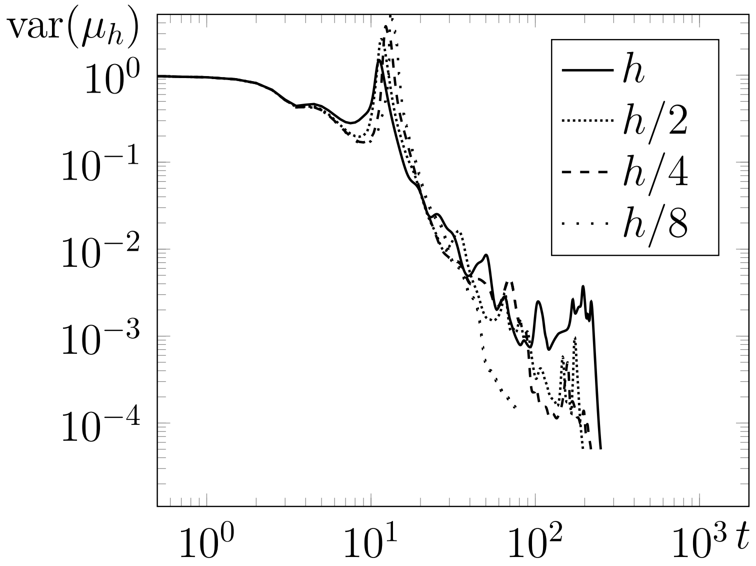


Figure 4.10: Log-log plot of var($\mu_h(t)$) vs. time for $\beta = 1.5$.

The upper threshold for Δt_k we impose ensures the stability of the forward Euler scheme. An explicit solution for the OT density is not known, so we cannot evaluate the errors as in the previous test cases, but only the variation of the OT density μ_h between successive time steps.

After the initial transient, var($\mu_h(t)$) presents an irregular behaviour, and is decreasing in time non monotonically and there are oscillations, as we see in figure 4.10. Despite the irregular behaviour, this test case reaches convergence toward the equilibrium configuration (μ_h^*, u_h^*) for every refinement.

4.3.1 Branched structures

The solution of the problem displays an irregular pattern made by narrow channels, that connects Q^+ to Q^-, and whose shape tends to approximate the 1-dimensional structures typical of the BT problem. The equilibrium configuration μ^*, which is representative of the flow capacity, is not homogeneous in the domain, but its pattern suggests the presence of a hierarchical structure. There are sub-channels with low OT density that branch into each other, and the flow capacity is maximum in the ramified channels connecting f^+ to f^-.

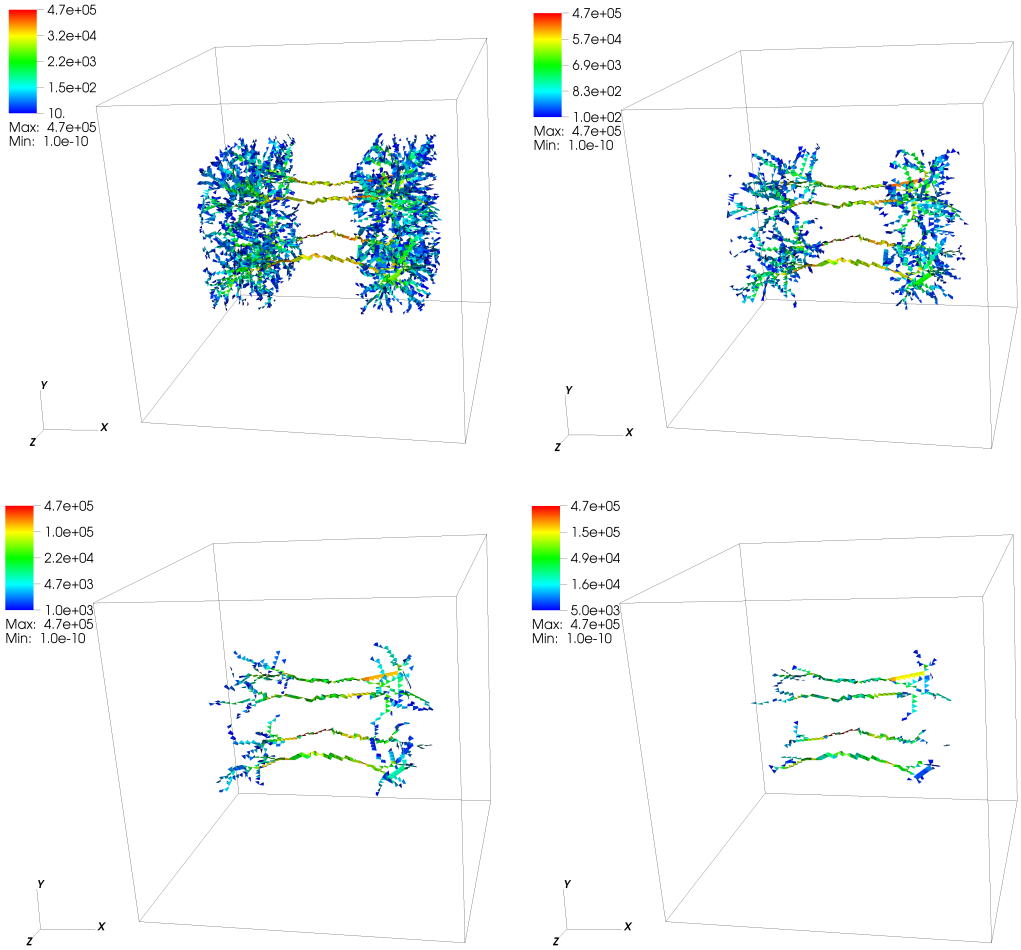


Figure 4.11: Numerical approximation of μ_h^* for $\beta = 1.5$. The four pictures are obtained for the most refined mesh, i. e. with mesh parameter $\frac{h}{8}$; the OT density is plotted only if above the minimum thresholds of 10, 10^2, 10^3 and 5×10^3, respectively.

In figure 4.11 the spatial distribution of μ_h^* is reported for the most refined mesh for different minimum thresholds of μ^*, and we can easily observe the network described above. Note that these structures appear only with a carefully selected logarithmic colour scale, showing a range in μ_h^* ranging between several orders of magnitude. Inside Q^+ and Q^- the hierarchical branching structure appears, while in Q^c μ_h^* concentrates on a series of connected tetrahedra, creating a tight channel with high flow capacity. Looking at figure 4.12, we observe that, as long as we refine the grid, the values of μ_h^* in the channels connecting f^+ to f^- grow. In fact the mass we need to transport is the same and the channels are tighter. In addition, increasing the spatial accuracy the mass does not change its trajectory, it still passes through the same points. Thus, as $h \to 0$, the trajectory approximates always better the 1-dimensional structure, typical of the BT problem.

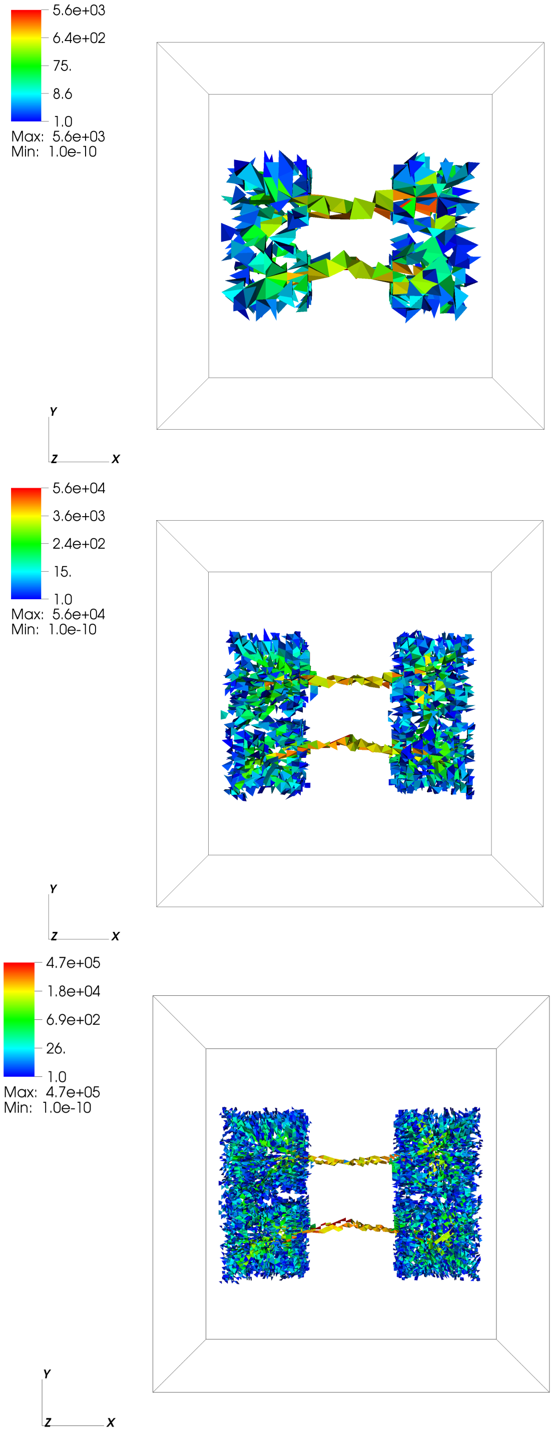


Figure 4.12: Numerical approximation of μ_h^* for $\beta = 1.5$. The three figures show the branched structures for three successive refinements.

4.3.2 Lyapunov functional

In this subsection we focus on the time behaviour of the Lyapunov-candidate functional $\mathcal{L}_\beta(\mu_h(t))$, that for $\beta = 1.5$ is given by equation (2.3.2). We recall that this functional is the sum of two terms: $\mathcal{E}(\mu(t))$, which represents the energy dissipated along the transport, and $\mathcal{M}(\mu_h(t))$, that stands for the cost of building the optimal infrastructure. Figure 4.13 reports the time evolution of the Lyapunov-candidate functional and its two contributes. The behaviour of $\mathcal{L}_\beta$ for different refinements and for different initial data μ_0 is shown in the upper plots, while the two lower images are the contributes $\mathcal{E}$ and $\mathcal{M}$ for different mesh refinements and initial data $\mu_0 = 1$. First of all, we observe that the Lyapunov-candidate functional decreases monotically

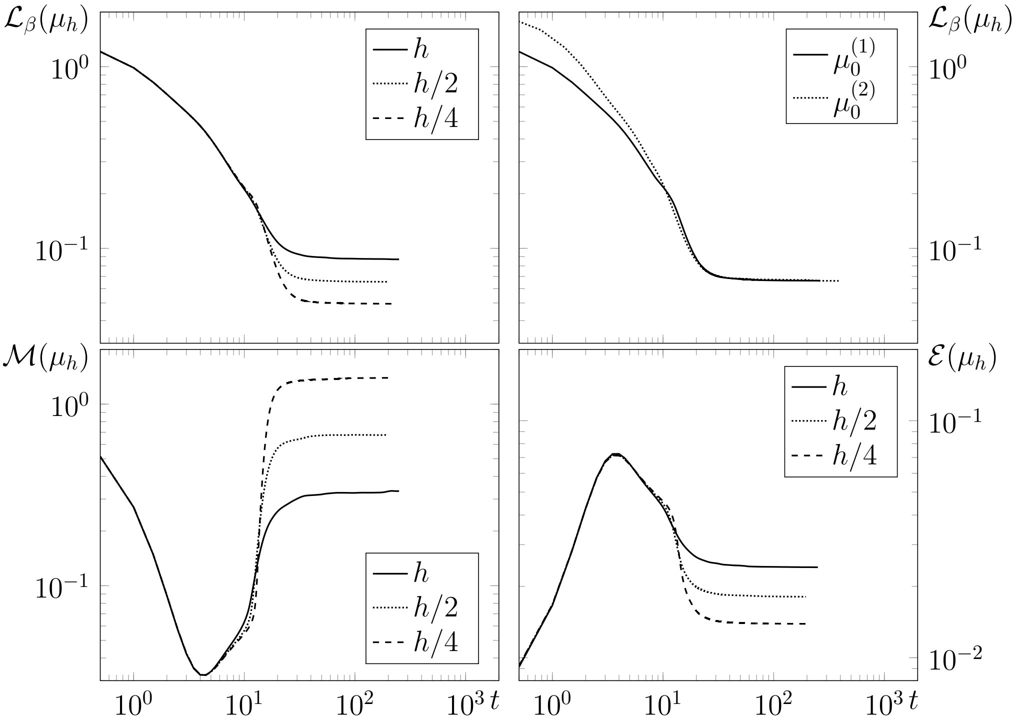


Figure 4.13: Log-log time behaviour of $\mathcal{L}_\beta(\mu_h(t))$, $\mathcal{E}(\mu(t))$ and $\mathcal{M}(\mu_h(t))$ for the BT problem test case. The top left plot reports the time evolution of the Lyapunov-candidate functional for different mesh refinements and initial condition $\mu_0^{(1)}$, the top right plot for different initial data μ_0. The two bottom images are computed for different mesh refinements, starting from the initial condition $\mu_0^{(1)}$.

	$\mathcal{L}_\beta(\mu^*)$
$\mu_0^{(1)}$	0.663×10^{-1}
$\mu_0^{(2)}$	0.661×10^{-1}

Table 4.1: Lyapunov-candidate functional equilibrium value for the considered initial conditions.

in time, consistently with what we expect. We start analysing the case with different mesh refinements and initial data $\mu_0 = 1$. $\mathcal{L}_\beta(\mu_h)$ reaches an equilibrium value $\mathcal{L}_\beta(\mu_h^*)$ for every mesh refinement, but, differently from the L^1-OT problem test case, this value scales with the mesh parameter h. Also the two functionals $\mathcal{E}(\mu(t))$ and $\mathcal{M}(\mu_h(t))$ present a time behaviour similar to $\mathcal{L}_\beta(\mu_h)$, in fact they both reach an equilibrium point which changes with h. The behaviour of $\mathcal{E}(\mu(t))$ and $\mathcal{M}(\mu_h(t))$ is opposite as long as we refine the mesh, in fact the first decreases as h gets smaller, while the latter is inversely proportional to h. On the other hand, the equilibrium values $\mathcal{L}_\beta$ are slightly different for the two different initial conditions considered, that are

$$\begin{aligned}\mu_0^{(1)} &= 1\\ \mu_0^{(2)} &= 3 + 2\sin(8\pi x)\sin(8\pi y) sin(8\pi z)\ .\end{aligned}$$

In table 4.1 we observe the value $\mathcal{L}_\beta(\mu^*)$ for each refinement. The different initial conditions influence the branching displacements even for coarse meshes, as we observe in figure 4.14. Thus this test case presents dependence on the initial data μ_0, consistently with conjecture 4.

4.3.3 Computational cost

The computational cost is given by the number of PCG iterations required to solve the linear system arising from the FEM discretization of the problem. The time behaviour of the number of iterations is reported in figure 4.15. The mean value of the number of iterations for each different refinement grow consistently with the size of the system, and the resulting behaviour is

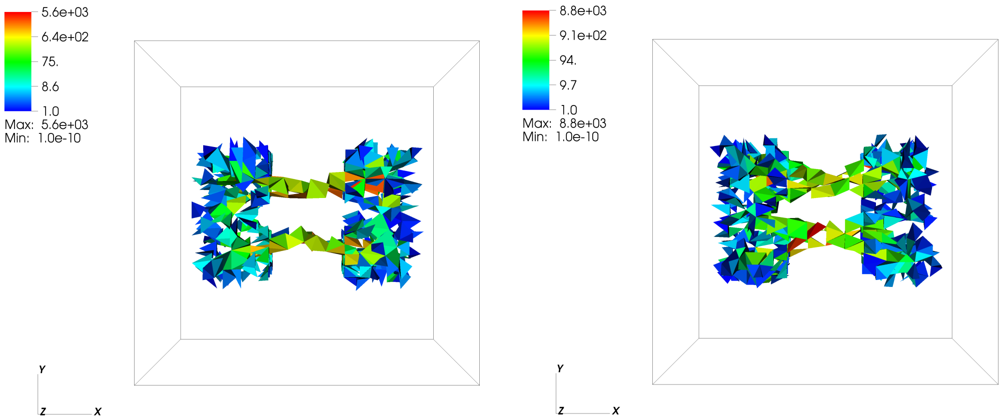


Figure 4.14: Numerical approximation of μ^* for different initial conditions μ_0 and $\beta = 1.5$. The initial data are $\mu_0^{(1)}$ and $\mu_0^{(2)}$, respectively for the left and the right pictures.

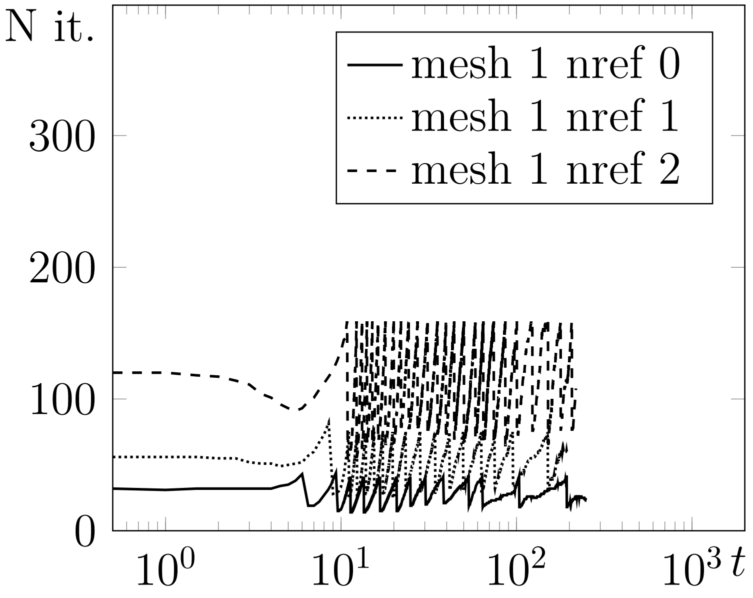


Figure 4.15: Number of iterations (N it.) to solve the linear system for each time step vs. time. The linear system is solved through the PCG scheme. With nref we denote the number of refinements.

strongly oscillatory around the mean value.

Chapter 5

Conclusions

In this thesis we derived and analyzed experimentally a three dimensional numerical extension of the DMK formulation, which is an innovative way proposed in [14, 17] to solve Optimal Transport problems. The aim of this thesis work is to verify the accuracy and robustness of this extension. Further developments can be easily identified for the improvement of the computational speed. Two possible developments may consist in moving the present work toward a parallel implementation, and involving an implicit time integration, exploiting BDF (Backward Differentiation Formulae) and the quadratic convergent Newton method.

We performed several test cases, presented in chapter 4, and the results we obtained are satisfactory. Our three dimensional numerical model is consistent with both the two dimensional model and the conjectures stated in chapter 2. Our results show that the L^1-OT problem test case verifies the idea that the DMK equations converge at large times toward an asymptotic configuration, and this equilibrium is related to the solution of the MK equations. Then we performed the p-Poisson test case, where we observe the relation between $(\mu(t), u(t))$ and u_p, solutions of the DMK equations and of the p-Poisson equation, respectively. This relation is verified for both the CT problem and the Beckmann problem. We observe that also for the CT problem the solution reaches an equilibrium configuration. Next, we addressed the BT problem test case, and for this problem the solution of the DMK equations again reaches an equilibrium, and our idea is that this stationary configuration is related to the BT problem solution. The one dimensional branching structures arising as solution and the behaviour of the Lyapunov-candidate functional give consistency to our conjectures.

Of course, the main future development is the application of the model to physical problems, and one of the first future purposes may be the study of the plant-root dynamics through an OT approach. It is a matter of fact that OT presents extremely wide fields of application. In fact, we believe that the cardiovascular system, animal brain, transportation networks and climate change processes are examples of areas where OT theories can effectively and successfully be applied toward a better understanding of their functioning mechanism. The three dimensional numerical model developed in this thesis will contribute substantially to the study of these fundamental and complex problems.

Bibliography

[1] M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein GAN*, arXiv preprint arXiv:1701.07875, (2017).

[2] J.-D. Benamou and G. Carlier, *Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations*, J. Optim. Theory Appl., 167 (2015), pp. 1–26.

[3] L. Bergamaschi, E. Facca, A. Martínez, and M. Putti, *Spectral preconditioners for the efficient numerical solution of a continuous branched transport model*, J. Comput. Appl. Math., (2018).

[4] V. Bonifaci, K. Mehlhorn, and G. Varma, *Physarum can compute shortest paths*, J. Theor. Biol., 309 (2012), pp. 121 – 133.

[5] G. Bouchitté, G. Buttazzo, and P. Seppecher, *Shape optimization solutions via Monge-Kantorovich equation*, CR MATH, 324 (1997), pp. 1185–1191.

[6] G. Buttazzo and E. Stepanov, *On regularity of transport density in the Monge-Kantorovich problem*, SIAM J. Control Optim., 42 (2003), pp. 1044–1055.

[7] M. Caramia, S. Giordani, F. Guerriero, R. Musmanno, and D. Pacciarelli, *Ricerca operativa*, De Agostini Scuola SpA, Novara, Italy, 2014.

[8] G. Carlier, C. Jimenez, and F. Santambrogio, *Optimal transportation with traffic congestion and wardrop equilibria*, SIAM J. Control Optim., 47 (2008), pp. 1330–1350.

[9] E. Cuthill and J. McKee, *Reducing the bandwidth of sparse symmetric matrices*, ACM, (1969), pp. 157–172.

[10] M. Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transportation distances*, Adv. Neural Inf. Process. Syst., 26 (2013).

[11] B. N. Delaunay, *Sur la sphère vide*, Bulletin of Academy of Sciences of the USSR, (1934), pp. 793–800.

[12] I. Ekeland and R. Téman, *Convex Analysis and Variational Problems*, Classics in Applied Mathematics, SIAM, Philadelphia, PA, USA, 1999.

[13] L. C. Evans and W. Gangbo, *Differential equations methods for the Monge-Kantorovich mass transfer problem*, American Mathematical Soc., 137 (1999).

[14] E. Facca, *Biologically inspired formulation of Optimal Transport Problems*, PhD thesis, Università degli Studi di Padova, 2018.

[15] E. Facca, F. Cardin, and M. Putti, *Extended Dynamic Monge-Kantorovich equation for Congested and Branched Optimal Transport problems*, SIAM J. Appl. Math., submitted, (2018).

[16] ——, *Towards a stationary Monge-Kantorovich dynamics: the Physarum Policephalum experience*, SIAM J. Appl. Math., 75 (2018), pp. 651 – 676.

[17] E. Facca, S. Daneri, F. Cardin, and M. Putti, *Numerical solution of Monge-Kantorovich equations via a dynamic formulation*, SIAM J. Sci. Comput., (2017).

[18] E. N. Gilbert, *Minimum cost communication networks*, Bell Labs Technical Journal, 46 (1967), pp. 2209–2227.

[19] E. F. Kaasschieter, *Preconditioned conjugate gradients for solving singular systems*, J. Comput. Appl. Math., 24 (1988), pp. 265–275.

[20] L. V. Kantorovich, *On the translocation of masses*, C. R. (Doklady) Acad. Sci. USSR, 321 (1942), pp. 199–201.

[21] 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.

[22] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde, *Optimal mass transport: Signal processing and machine-learning applications*, IEEE Signal Process. Mag., 34 (2017), pp. 43–59.

[23] A. Marani, R. Rigon, and A. Rinaldo, *A Note on Fractal Channel Networks*, Water Resour. Res., 27 (1991), pp. 3041–3049.

[24] G. Monge, *Mémoire sur la théorie des déblais et des remblais*, De l'Imprimerie Royale, 1781.

[25] T. Nakagaki, H. Yamada, and A. M. Tóth, *Path finding by tube morphogenesis in an amoeboid organism*, Biophys. Chem., 92 (2001), pp. 47 – 52.

[26] E. Oudet and F. Santambrogio, *A Modica-Mortola approximation for branched transport and applications*, Arch. Ration. Mech. An., 201 (2011), pp. 115–142.

[27] A. Plaza, M. Padrón, J. Suarez, and S. Falcon, *The 8-tetrahedra longest-edge partition of right-type tetrahedra*, Finite Elem. Anal. Des., 41 (2004), pp. 253–265.

[28] I. Rodríguez-Iturbe and A. Rinaldo, *Fractal River Basins: Chance and Self-Organization*, Cambridge University Press, 2001.

[29] Y. Rubner, C. Tomasi, and L. J. Guibas, *The earth mover's distance as a metric for image retrieval*, Int. J. Comput. Vis., 40 (2000), pp. 99–121.

[30] F. Santambrogio, *Optimal Transport for Applied Mathematicians*, Birkäuser, NY, 2015.

[31] 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.

[32] 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.

[33] C. Villani, *Topics in Optimal Transportation*, vol. 58 of Graduate studies in mathematics, AMS, Providence, R.I., 2003.

[34] ——, *Optimal Transport: Old and New*, Springer Science & Business Media, Berlin, Heidelberg, 2008.

[35] Q. Xia, *Optimal Paths related to Transport Problems*, CCM, 5 (2003), pp. 251–279.

[36] ——, *Numerical simulation of optimal transport paths*, 2010 Second International Conference on Computer Modeling and Simulation, 1 (2010), pp. 521–525.