An Efficient Numerical Algorithm for the L^2 Optimal Transport Problem with Applications to Image Processing

by

Louis-Philippe Saumier Demers
B.Sc., Université de Sherbrooke, 2008

A Thesis Submitted in Partial Fulfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

An Efficient Numerical Algorithm for the
L^2 Optimal Transport Problem with
Applications to Image Processing

by

Louis-Philippe Saumier Demers
B.Sc., Université de Sherbrooke, 2008

Supervisory Committee

Dr. Martial Agueh, Co-Supervisor
(Department of Mathematics and Statistics, University of Victoria)

Dr. Boualem Khouider, Co-Supervisor
(Department of Mathematics and Statistics, University of Victoria)

Supervisory Committee

Dr. Martial Agueh, Co-Supervisor
(Department of Mathematics and Statistics, University of Victoria)

Dr. Boualem Khouider, Co-Supervisor
(Department of Mathematics and Statistics, University of Victoria)

ABSTRACT

We present a numerical method to solve the optimal transport problem with a quadratic cost when the source and target measures are periodic probability densities. This method relies on a numerical resolution of the corresponding Monge-Ampère equation. We use an existing Newton-like algorithm that we generalize to the case of a non uniform final density. The main idea consists of designing an iterative scheme where the fully nonlinear equation is approximated by a non-constant coefficient linear elliptic PDE that we discretize and solve at each iteration, in two different ways: a second order finite difference scheme and a Fourier transform (FT) method. The FT method, made possible thanks to a preconditioning step based on the coefficient-averaged equation, results in an overall $\mathcal{O}(P\text{Log}P)$-operations algorithm, where P is the number of discretization points. We prove that the generalized algorithm converges to the solution of the optimal transport problem, under suitable conditions on the initial and final densities. Numerical experiments demonstrating the robustness and efficiency of the method on several examples of image processing, including an application to multiple sclerosis disease detection, are shown. We also demonstrate by numerical tests that the method is competitive against some other methods available.

Contents

Supervisory Committee **ii**

Abstract **iii**

Table of Contents **iv**

List of Tables **vi**

List of Figures **vii**

Acknowledgements **ix**

Introduction **1**

1 The Optimal Transport Problem **4**
1.1 Two versions 4
1.2 The Quadratic Case 6
1.3 The Monge-Ampère Equation 8
1.4 The Periodic Setting 11

2 Resolution of the Monge-Ampère equation **13**
2.1 The Algorithm 13
2.2 Linear Second Order Elliptic PDE in a Periodic Setting 15
2.3 Preliminaries to the Proof of Convergence 23
2.4 Proof of Convergence of the Algorithm 29
2.5 Remarks on the Proof 38

3 Numerical Discretization **41**
3.1 A Finite Differences Implementation 41
3.2 A Fourier Transform Implementation 49

3.3 Numerical Tests 54
3.3.1 Experiment 1 55
3.3.2 Experiment 2 62
3.4 A Posteriori Stability Analysis 68

4 Application to Medical Imaging **71**

5 Overview of Some Existing Numerical Methods **78**
5.1 A Fluid Dynamics Reformulation 78
5.2 A Gradient Descent on the Dual Problem 82
5.3 A Projection on the Mass Preservation Constraint 85

Conclusions **91**

A Function Spaces and Norms **94**

B Convex analysis **98**

Bibliography **101**

List of Tables

Table 2.1 Quantities involved in the bounds on ϵ for $c = \Lambda/\lambda$ 28

Table 3.1 Error Comparison for experiment 1 56
Table 3.2 BICG tolerance comparison for $N = 64$ 59
Table 3.3 Average number of BICG and GMRES iterations and total computing time 66
Table 3.4 Computed $\rho(L_h^{-1})$ for the two experiments 69

Table 5.1 Different interpolation scenarios for $\|u - u_n\|_{l^\infty}$ 89
Table 5.2 Haber et al.'s results for different grid sizes of the phantom experiment 90

List of Figures

Figure 3.1 Stencil for $h(\theta_x, \theta_y)$ 42
Figure 3.2 Stencil for $h^2(\theta_{xx}, \theta_{yy}, \theta_{xy})$ 42
Figure 3.3 Stencil for $h(u_x, u_y)$ 43
Figure 3.4 Stencil for $h^2(u_{xx}, u_{yy}, u_{xy})$ 44
Figure 3.5 Contour Plots for N=16 54
Figure 3.6 Contour Plots for N=64 55
Figure 3.7 Error between f and $\tilde{f}_n$ on a semilog scale 56
Figure 3.8 Number of BICG Iterations for Different Grid Sizes 57
Figure 3.9 Behavior comparison for different derivatives when $N = 64$ . . 58
Figure 3.10 Number of BICG iterations for a 0.1 tolerance 59
Figure 3.11 Error between f and $\tilde{f}_n$ for the FT case 60
Figure 3.12 Running time for the two methods on a semilog scale 61
Figure 3.13 Error for the FD implementation with tol=10^{-4} 62
Figure 3.14 Error for the FT implementation with tol=10^{-4} 63
Figure 3.15 3d plots for $N = 32$ in the FD case 64
(a) u_n and u 64
(b) $u - u_n$ 64
Figure 3.16 Error for the FD implementation with tol=10^{-1} 65
Figure 3.17 Error for the FT implementation with tol=10^{-1} 66
Figure 3.18 Error obtained with different values of τ for the FT case and $N = 128$ 67

Figure 4.1 Two slices of the same brain depicting the presence of MS . . 72
(a) Initial density f: MRI scan of a normal brain 72
(b) Final density g: MRI scan of the same brain with Multiple Sclerosis lesions 72
Figure 4.2 Surface plot of $\mathrm{div}(u_3)$ 73
Figure 4.3 Filtered contour plot of $\mathrm{div}(u_3)$ 74

Figure 4.4 Different slices from the same brain showing scars 75
(a) Initial density f: MRI scan of a normal brain 75
(b) Final density g: MRI scan of the same brain with multiple sclerosis lesions 75
Figure 4.5 Surface plot of $\mathrm{div}(u_4)$ for the second set of images 76
Figure 4.6 Scan of the healthy brain on which was superposed the colored filtered contour plot of $\mathrm{div}(u_4)$ for the second set of images . . 77

Figure 5.1 Contour plots of the initial and target Gaussian densities . . . 81
(a) f 81
(b) g 81
(c) f_{10} 81
Figure 5.2 Iterations of the Lena to Tiffany warp 84
(a) g 84
(b) f_2 84
(c) f_5 84
(d) f_{10} 84
(e) f_{20} 84
(f) f 84
Figure 5.3 The 256×256 pixels phantom image experiment with a nearest neighbor interpolation. 88
(a) f 88
(b) g 88
(c) f_{10} 88
Figure 5.4 Contour plot of $\mathrm{div}(u_{10})$ 89

ACKNOWLEDGEMENTS

I would like to thank:

Dr. Martial Agueh and Dr. Boualem Khouider
for supporting me, for being so patient and available, and mostly for sharing their precious wisdom, guiding my footsteps in this journey that I undertook.

Dr. Robert Russell
for being the external examiner on my committee, providing precious advices and a fresh point of view.

The National Science and Engineering Research Council of Canada
for funding me with a Postgraduate Scholarship, making my life so much easier during the last two years.

The University of Victoria and its Math. and Stat. Department
for providing me with such an incredible learning environment and also for helping me financially with two President Research Scholarships.

My Wonderful Family and Friends
for all the fruitful and inspiring discussions, for helping me in times of need, and simply for being there.

One should conceive of mathematics not as a self-contained linguistic enterprise, but as effective conceptualization of the moving geometrical spectacle of objective reality.
Gaspard Monge

Introduction

The optimal transport theory is the modern mathematical way of treating efficient reallocation problems. Originating from a simple engineering problem which can be traced back to more than 200 years ago, it is now a vibrant area of research attracting scientists and mathematicians with many different fields of expertise. One of the main reasons for this widespread interest is the broad range of applications of this theory. The first one that comes immediately to mind is in economics, where the redistribution of material in a profitable way is a common concern [11]. However, the applications are much more diverse. Indeed, it appears for example in the semi-geostrophic shallow water system used by meteorologists to model how fronts arise in large scale weather patterns [29]; in material science, where it provides information on the behavior of superconductor materials through the p-Laplacian equation [17, 28]; in computer vision where it provides a geometric framework to detect changes occurring within a scene evolving through time [36, 37]; or even in cosmology where physicists use it to model the conditions of the early universe [38].

Nowadays, the optimal transport problem is starting to be well understood in theory, thanks to the joint effort of several mathematicians. Cédric Villani, who was recently awarded the Fields medal, wrote two treatises to present these discoveries in a way that it provides a useful reference for the experts while remaining accessible to the neophyte [6, 7]. However, these results don't directly provide explicit solutions for the problem, and despite all that work, to put the theory into practice, efficient numerical methods are still in development or under investigation. One popular scenario is the so-called L^2 optimal transport problem, where we are dealing with a reallocation problem where the goal is to minimize the sum of the movements of matter measured by the square of the Euclidean distance. The reason for this popularity resides mainly in the simplicity of the results and the wide range of applications. Therefore, in this specific case, some numerical methods have been developed (see for example [2, 3, 5, 10, 35] and the references therein), but all of them have issues that call for

improvement.

We can show that under certain assumptions, solving the L^2 optimal transport problem is equivalent to solving a fully nonlinear elliptic second order partial differential equation (PDE) called the Monge-Ampère equation. What we propose to do in this work is to solve numerically this equation in an efficient way and then recover the solution of the transport problem. Our technique is inspired by the work of Loeper and Rapetti [13] who designed a damped Newton algorithm for a simpler form of the equation corresponding to the situation where one wants to uniformly redistribute the matter at hand. We thus modify their algorithm to encompass non-uniform redistributions. The hard part in this more general case is to control the possible degeneracy of the Monge-Ampère equation. Indeed, the solution of this equation yields the solution of the transport problem if and only if it is convex. Within the context of an iterative method, it is imperative to make sure that as we iterate, the solution does not loose its convexity.

The strategy we are going to employ to address this issue is to restrict ourselves to a setting involving periodic boundary conditions. By doing this, we will be able to use interior a priori regularity estimates for the solutions of the Monge-Ampère equation [9, 39] to show that our algorithm will always produce a valid answer under certain additional assumptions on the initial and final distributions of matter. Moreover, we will be able to use the classical transport results mentioned earlier since they have been adapted to the periodic case by Cordero-Erausquin [8]. Therefore, this specific framework will allow us to provide an in-depth mathematical analysis of the algorithm, which in turn will give us insight on its behavior.

The Newton-like algorithm approximates the solution of the fully nonlinear elliptic PDE by a sequence of solutions of linear elliptic PDEs. Hence, the method we choose to discretize these linear PDEs will greatly influence its performances. We will compare through theoretical arguments and numerical experiments two possible techniques for doing this; one employed in [13] and using finite differences; one created by Strain [21] and using Fourier transforms. For the rest of the implementation, we select usual fourth-order accurate finite differences schemes. As we will see later, the Newton algorithm coupled with the Fourier transform method turns out to be a very effective way of solving the optimal transport problem which requires only $\mathcal{O}(P\text{Log}P)$ operations, where P is the number of gridpoints.

After having thoroughly analyzed the algorithm and its corresponding discretizations, we will show its applicability on a practical example in medical imagery. Re-

cently, researchers realized that optimal transport could provide a powerful tool in image processing, if one could reduce its high computational cost (see for example [37]). The method we are presenting here precisely achieves this. To demonstrate it, we consider the detection of multiple sclerosis (MS) via magnetic resonance imaging (MRI) of the brain. By representing grayscale images as two-dimensional matter distributions in the optimal transport problem, taking mass to be proportional to brightness, we can apply the algorithm to two brain scans and obtain the transport map sending one to the other. Since this map gives the transformation which minimizes the total sum of the movements of brightness elements, analyzing its divergence indicates the location of the changes between the images, easily detecting the presence of scars lefts by multiple sclerosis.

This thesis is organized as follows. In Chapter 1, we present the optimal transport problem and the Monge-Ampère equation. In Chapter 2, we derive the adjusted Newton algorithm, prove that a linear second-order elliptic PDE with periodic boundary conditions has a unique (up to a constant) solution and prove that the algorithm converges. Then, in Chapter 3, we introduce the finite differences and the Fourier transform implementations and test them on two numerical experiments. Next, we present the medical imaging experiment in Chapter 4. Finally, in Chapter 5, we introduce three other existing methods to solve the optimal transport problem and draw comparisons with ours.

Chapter 1

The Optimal Transport Problem

1.1 Two versions

In 1781, the French mathematician Gaspard Monge investigated in his memoir "Sur la théorie des déblais et des remblais" [14] (On the theory of excavation and emplacements) the problem we today refer to as the mass transport problem, or optimal transport problem. It can be summarized as follows: assume that you have matter distributed with respect to a certain known initial mass density and that you need to rearrange it in a way that it respects another known final density. Moreover, you know how much it will cost you to move one particle of matter from one place to the other. Then, how do you realize this redistribution at a minimal cost?

Monge, originally motivated by very practical applications, considered his initial distribution of matter to be a pile of soil that he wanted to move in a hole in a specific way, while spending a minimal amount of energy doing it. One can think of the soil being full of rocks so when we move it to the hole, we might want all the rocks to be at the bottom for example. In this situation, the cost can be interpreted as an energetic cost (obviously, moving heavy rocks will be harder than moving light soil). Due to this scenario, the problem is also sometimes referred to as the earth-moving problem.

Nowadays, using a measure theory framework, we can state it in a more general way. Let Ω_1 and Ω_2 be two measure spaces equipped respectively with measures μ and ν to represent the initial and final densities. Having in mind that we are dealing with a redistribution problem, the total initial and final mass of matter has to be the same, and therefore, if we normalize this amount to 1, we can consider μ and ν to be

probability densities. In addition, let our cost be a measurable non-negative function $c : \Omega_1 \times \Omega_2 \to \mathbb{R}_+ \cup \{+\infty\}$.

Next, we say $T : \Omega_1 \to \Omega_2$ is an admissible transport map, or transference plan, if it respects the following mass conservation condition:

$$\nu(B) = \mu(T^{-1}(B)) \ \forall \, \text{measurable } B \subset \Omega_2. \tag{1.1}$$

When (1.1) is satisfied, we write $\nu = T_{\#}\mu$ and we say that ν is the push-forward of μ by T. Then, we can formulate Monge's problem of rearranging matter at a minimal cost:

$$\begin{cases} \text{Minimize } \displaystyle\int_{\Omega_1} c(x, T(x)) \, \mathrm{d}\mu(x) \\ \\ \text{over the set of all measurable maps } T \text{ such that } T_{\#}\mu = \nu. \end{cases} \tag{1.2}$$

The mass conservation constraint imposed on the map T being nonlinear, Monge's problem turns out to be very hard to solve, except for a few simple cases. In addition, Monge's version is not always well-posed. For example, when $X = Y = \mathbb{R}$, if μ is δ_0 the Dirac mass at 0 and $\nu = (\delta_{-1} + \delta_1)/2$, there is no T satisfying $\nu = T_{\#}\mu$.

To overcome these difficulties, the Russian mathematician Leonid Kantorovich came up with a relaxed version of it in the 1940's [27]. He considered the following:

$$\begin{cases} \text{Minimize } \displaystyle\int_{\Omega_1 \times \Omega_2} c(x, y) \, \mathrm{d}p(x, y) \\ \\ \text{for } p \in P(\Omega_1 \times \Omega_2). \end{cases} \tag{1.3}$$

where $P(\Omega_1 \times \Omega_2)$ is the space consisting of all probability measures p defined on the product space $\Omega_1 \times \Omega_2$ having marginals μ in Ω_1 and ν in Ω_2, i.e.

$$\begin{gathered} p(A \times \Omega_2) = \mu(A) \text{ and } p(\Omega_1 \times B) = \nu(B) \\ \forall \text{ measurable } A \subset \Omega_1 \text{ and } \text{ measurable } B \subset \Omega_2. \end{gathered}$$

To see why (1.3) is indeed a generalization of Monge's version (1.2), let T be an admissible transference plan and let p_T be defined by

$$\begin{aligned} p_T(A \times B) &= [(id \times T)_{\#}\mu](A \times B) \\ &:= \mu(A \cap T^{-1}(B)) \quad \text{for all } A \subset \Omega_1 \text{ and } B \subset \Omega_2. \end{aligned}$$

We can verify that $p_T \in P(\Omega_1 \times \Omega_2)$:

$$p_T(\Omega_1 \times B) = \mu(\Omega_1 \cap T^{-1}(B)) = \mu(T^{-1}(B)) = \nu(B),$$
$$p_T(A \times \Omega_2) = \mu(A \cap T^{-1}(\Omega_2)) = \mu(A \cap \Omega_1) = \mu(A).$$

From this new point of view, he discovered a very important property of the optimal transport problem, namely that it possesses a convenient dual formulation.

Theorem 1.1.1 (Kantorovich)**.** *Let $c : \Omega_1 \times \Omega_2 \to \mathbb{R}_+ \cup \{+\infty\}$ be a continuous cost function. Define Θ_c to be the set of all measurable functions $(\psi, \phi) \in L^1(d\mu) \times L^1(d\nu)$ satisfying*

$$\psi(x) + \phi(y) \leq c(x, y)$$

for $d\mu$-almost all $x \in \Omega_1$, $d\nu$-almost all $y \in \Omega_2$. Then, for $p \in P(\Omega_1, \Omega_2)$,

$$\inf_{P(\Omega_1,\Omega_2)} \left(\int_{\Omega_1 \times \Omega_2} c(x, y)\, dp(x, y) \right) = \sup_{\Theta_c} \left(\int_{\Omega_1} \psi(x) d\mu + \int_{\Omega_2} \phi(y) d\nu \right).$$

Note that the result holds with even weaker assumptions, as we can see in [6]. This dual formulation of the problem was the idea that first enabled to prove a crucial result concerning the case where the cost function is given by $|x - y|^2$, which we will present in the next section. Let's just mention before moving on that often in the literature, to pay tribute to the work of these two forefathers on the subject, the mass transport problem is referred to as the Monge-Kantorovich problem.

1.2 The Quadratic Case

Throughout this thesis, we will dedicate our efforts to one specific cost function; the square of the usual Euclidean distance $c(x, y) = \frac{|x-y|^2}{2}$ (the factor $1/2$ is introduced to simplify the subsequent calculations and notations). The optimal transport problem with a quadratic cost function has been extensively studied in the past years, mainly because of the simplicity of the results and because of its wide range of applications. Probably the most important result for the quadratic cost problem is the one due to Brenier [41], who found out that the optimal map realizing the transfer has to be the gradient of a convex function:

Theorem 1.2.1 (Brenier)**.** *Consider the optimal transport problem associated with the quadratic cost function* $c(x,y) = \frac{|x-y|^2}{2}$ *and with* μ *and* ν *two probability measures on* $\mathbb{R}^d$*. Assume* μ *and* ν *have finite second order moments:*

$$\int_{\mathbb{R}^d} \frac{|x|^2}{2} d\mu(x) + \int_{\mathbb{R}^d} \frac{|y|^2}{2} d\nu(y) < \infty.$$

If μ *and* ν *do not give mass to Lebesgue-negligible sets, then there is a unique optimal* $\check{p}$ *solving Kantorovich's optimization problem (1.3). It takes the form*

$$\check{p} = (id \times \check{T})\#\mu,$$

where $\check{T}$ *is such that:*

1) *It is uniquely determined* $d\mu$*-almost everywhere*

2) $\check{T} = \nabla\Psi$ *where* Ψ *is a convex function*

3) $\check{T}\#\mu = \nu$

4) *It is the unique solution of Monge's optimization problem (1.2)*

5) *It is invertible (in the almost everywhere sense) and its inverse* $\nabla\Phi$ *is the solution of the reverse Monge problem sending* ν *to* μ

For a more general version and more details, we strongly recommend reading [6], Cédric Villani's introductory book on optimal transport. It is also worth mentioning here that Ψ and Φ do not solve the dual problem presented in the previous section. The solution is

$$\psi = \frac{|x|^2}{2} - \Psi \quad \text{and} \quad \phi = \frac{|y|^2}{2} - \Phi. \tag{1.4}$$

Just to give the reader an idea of where this comes from, one of the main ideas for the proof of Theorem 1.2.1 resides in the condition that $\psi(x)$, $\phi(y)$ belong to Θ_c (see Theorem 1.1.1):

$$\begin{aligned} \psi(x) + \phi(y) &\leq \frac{|x-y|^2}{2} \\ \Rightarrow \qquad x \cdot y &\leq \left[\frac{|x|^2}{2} - \psi(x)\right] + \left[\frac{|y|^2}{2} - \phi(y)\right] \end{aligned}$$

and hence we can rewrite the duality condition as follows:

$$\sup_{P(\Omega_1,\Omega_2)} \left(\int_{\Omega_1 \times \Omega_2} (x \cdot y)\, dp(x,y) \right) = \inf_{\widetilde{\Theta}_c} \left(\int_{\Omega_1} \Psi(x) d\mu + \int_{\Omega_2} \Phi(y) d\nu \right).$$

where $\widetilde{\Theta}_c$ is the set containing the pairs $(\Psi, \Phi) \in L^1(d\mu) \times L^1(d\nu)$ such that $x \cdot y \leq \Psi(x) + \Phi(y)$ almost everywhere. This give us a hint to why it is possible to replace $c(x,y) = \frac{|x-y|^2}{2}$ by $c(x,y) = -x \cdot y$ in (1.2) or (1.3) without changing the value of the optimum. The ansatz (1.4) will show up quite often in future results.

We can do many useful things with the quadratic Monge-Kantorovich problem, like creating a distance function between two probability densities by taking

$$\mathcal{W}(\mu,\nu)^2 = \int_{\Omega_1} |x - \check{T}(x)|^2 \, \mathrm{d}\mu(x)$$

where $\check{T}$ is the optimal map given in Theorem 1.2.1. Indeed, is it shown in [6] that $\mathcal{W}$ defines a metric on the set of probability measures with finite moments of order 2. Actually, this is also true for more general cost functions, as long as they can be expressed as $c(x,y) = d(x,y)^p$ where d is a distance on Ω_1 and $p \in [1,\infty)$. In the literature, $\mathcal{W}$ is referred to as the quadratic Wasserstein distance (after the Russian mathematician Leonid Nasonovich Wasserstein), or quadratic Monge-Kantorovich distance.

1.3 The Monge-Ampère Equation

One of the central points of this work is the fact that under certain conditions, the optimal transport map in Monge's problem also solves a fully nonlinear second order elliptic partial differential equation called the Monge-Ampère equation. Let's see how.

From now on, we will assume that the probability densities μ and ν are absolutely continuous with respect to Lebesgue's measure; having their Radon-Nikodym derivatives given respectively by f and g (i.e. $d\mu(x) = f(x)dx$ and $d\nu(x) = g(x)dx$). Denote

$$m_f = \inf_{x \in \Omega_1} f(x), \qquad M_f = \sup_{x \in \Omega_1} f(x),$$
$$m_g = \inf_{x \in \Omega_2} g(x) \quad \text{and } M_g = \sup_{x \in \Omega_2} g(x).$$

Most of the time, we will deal with continuous f and g so the infima and suprema should be understood as minima and maxima. For convenience, we shall also refer

to f and g as the initial and final densities. From Theorem 1.2.1, we know that the optimal $\check{T}$ in the transport problem with a quadratic cost is a ($d\mu$-almost everywhere) gradient of a convex function $\nabla\Psi$ such that

$$\nu(B) = \mu\Big((\nabla\Psi)^{-1}(B)\Big) \qquad \forall \text{ measurable } B \subset \Omega_2.$$

We can rewrite this condition using the indicator function:

$$\begin{aligned}\int_{\Omega_2} \mathcal{X}_B(y)g(y)dy &= \int_{\Omega_1} \mathcal{X}_{(\nabla\Psi)^{-1}(B)}(x)f(x)dx \\ &= \int_{\Omega_1} \mathcal{X}_B(\nabla\Psi(x))f(x)dx\end{aligned}$$

Next, appealing to a classical result from measure theory (found for example in [1]) stating that every positive, bounded and measurable function can be expressed as a uniform limit of a sequence of linear combinations of indicator functions on pairwise disjoint sets, we get

$$\int_{\Omega_2} \xi(y)g(y)dy = \int_{\Omega_1} \xi(\nabla\Psi(x))f(x)dx \quad \forall \xi \in \mathcal{C}_b(\mathbb{R}^d).$$

If we assume that $\nabla\Psi$ is $\mathcal{C}^1(\Omega_1)$ and one-to-one, we can carry out the change of variable $y = \nabla\Psi$:

$$\begin{aligned}\int_{\nabla\Psi(\Omega_1)} \xi(y)g(y)dy &= \int_{\Omega_2} \xi(y)g(y)dy \\ &= \int_{\Omega_1} \xi(\nabla\Psi(x))g(\nabla\Psi(x))\det(D^2\Psi)dx,\end{aligned}$$

which implies

$$\int_{\Omega_1} \xi(\nabla\Psi(x))f(x)dx = \int_{\Omega_1} \xi(\nabla\Psi(x))g(\nabla\Psi(x))\det(D^2\Psi)dx \quad \forall \xi \in \mathcal{C}_b(\mathbb{R}^d).$$

Note that we will see in the next section under which conditions these assumptions are valid. Finally, using the fundamental lemma of calculus of variations, we reach:

$$g(\nabla\Psi(x))\det\left(D^2\Psi(x)\right) = f(x). \tag{1.5}$$

This partial differential equation is a special case of a more general one called the

Monge-Ampère equation:

$$\det\left(D^2\Psi(x)\right) = h(x, \Psi(x), \nabla\Psi(x)).$$

As we will see later, it is sometimes necessary to consider an even more general equation (by introducing other terms inside the determinant when dealing with different cost functions for example). In the literature, all these different versions are referred to as the Monge-Ampère equation, and even though we mostly deal with (1.5), we shall do the same in this work. However, we will make sure we specify which one we are using to avoid confusion.

We previously said that (1.5) was a fully nonlinear elliptic partial differential equation. Here's what we meant:

Definition 1.3.1. *Let G be a continuous function defined on $\mathbb{R}^d \times \mathbb{R} \times \mathbb{R}^d \times \mathcal{S}_d(\mathbb{R})$ where $\mathcal{S}_d(\mathbb{R})$ is the space of all d-dimensional symmetric matrices with coefficients in $\mathbb{R}$. The partial differential equation*

$$G(x, \Psi, \nabla\Psi, D^2\Psi) = 0$$

is said to be elliptic if for all choices of x, r, p, X, Y,

$$Y \geq X \Rightarrow G(x, r, p, Y) \geq G(x, r, p, X).$$

In addition, it is uniformly elliptic if there exist $Q_1, Q_2 > 0$ such that for all choices of x, r, p, m, X, Y,

$$Y \geq X \Rightarrow Q_1\, Tr(Y - X) \geq G(x, r, p, Y) - G(x, r, p, X) \geq Q_2\, Tr(Y - X).$$

In this definition, the ordering of the matrices is given by $Y \geq X$ if and only if $Y - X$ is positive semidefinite. It would be nice if our equation would be elliptic, since this type of equation usually satisfies useful properties, like maximum principles, or like the fact that smooth data produces smooth solutions.

Unfortunately, the Monge-Ampère equation (1.5) is not in general elliptic. Nonetheless, if we restrict it to the set of convex functions Ψ, we get that X and Y are

positive semidefinite. Then, we have from [34] that

$$Y \geq X \geq 0 \Rightarrow \det(Y) \geq \det(X)$$
$$\text{and } G(x,r,p,Y) - G(x,r,p,X) = g(p)\left(\det(Y) - \det(X)\right) \geq 0.$$

Observe that even with this restriction, it is not uniformly elliptic.

1.4 The Periodic Setting

We are interested here in obtaining a numerical solution of the optimal transport problem through (1.5). Our technique strongly relies on a priori regularity estimates of classical solutions of the Monge-Ampère equation. However, to the best of our knowledge, global estimates are not yet available for the second derivatives. This is (at least partially) due to the intricate nature of this equation; remember that it is not uniformly elliptic and highly nonlinear. On the other hand, as we will see later, we have access to interior a priori estimates, and in a periodic setting, we will be able to use these results on the whole domain. This being said, let us unveil the framework for the rest of this thesis.

Definition 1.4.1. *Consider $\zeta : \mathbb{R}^d \to \mathbb{R}$ and let e_i be a vector in the canonical basis of $\mathbb{R}^d$, i.e. e_i has its i-th component equal to 1 and all its other components equal to 0. We say that ζ is 1-periodic if for every i, $\zeta(x + e_i) = \zeta(x)$ for every $x \in \mathbb{R}^d$.*

Note that we only need to know a 1-periodic function on the hypercube $[0,1]^d$ to know it on the whole $\mathbb{R}^d$. We shall use both point of views in subsequent discussions. With this in mind, let $\Omega_1 = \Omega_2 = \Omega = [0,1]^d$ and assume that the initial and final densities f and g are 1-periodic. For $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, we say that y is the reflection of x with respect to x_i if $y = (x_1, \ldots, x_{i-1}, -x_i, x_{i+1}, \ldots, x_d)$. Then, from [23], we get:

Theorem 1.4.2 (Caffarelli)**.** *Let f, g be in the unit cube $\Omega = [0,1]^d$ and write the optimal transport map*

$$\check{T} = \nabla\Psi(x) = x + \nabla u(x).$$

If we extend f and g to $\widehat{f}$ and $\widehat{g}$ on a larger cube $\widehat{\Omega}$ by even reflections, then u also extends periodically to $\widehat{u}$, to the same cube $\widehat{\Omega}$ by even reflections and $\check{T}$ to the optimal transport map

$$\widehat{\check{T}} = x + \nabla\widehat{u}(x)$$

from $\widehat{\Omega}$ to $\widehat{\Omega}$. In addition, if f and g are positive and belong to $\mathcal{C}^\alpha(\Omega)$, then $\check{T}$ maps each face of the cube to itself and it has a $\mathcal{C}^{1,\alpha}$ extension across $\partial\Omega$.

This result justifies the fact that we can use interior estimates on the whole domain, since each face of Ω becomes interior after a reflection. We also find in [8] an equivalent for our periodic setting of the optimal transport theorem for the quadratic cost 1.2.1:

Theorem 1.4.3 (Cordero-Erausquin)**.** *Assume μ and ν are 1-periodic probability densities in $\mathbb{R}^d$. Then, there exist a function Ψ convex on $\mathbb{R}^d$ such that $\nabla\Psi$ transports μ on ν and $\nabla\Psi$ is additive: $\nabla\Psi(x+p) = \nabla\Psi(x)+p$ for almost all $x \in \mathbb{R}^d$ and for all $p \in \mathbb{Z}^d$. Moreover, $\nabla\Psi$ is invertible (in the almost everywhere sense) and its inverse $\nabla\Phi$ transports ν on μ. Suppose in addition that f and g are $\mathcal{C}^\alpha(\mathbb{R}^d)$ with $\alpha > 0$ and such that $m_g, M_f > 0$ (f, g are 1-periodic on $\mathbb{R}^d$ with integral 1 on Ω). In this case, Ψ is in $\mathcal{C}^{2,\beta}(\mathbb{R}^d)$ for $0 < \beta < \alpha$ and it is a convex solution of the Monge-Ampère equation (1.5).*

Note that having $\nabla\Psi$ additive is equivalent to taking $\nabla\Psi = x + \nabla u(x)$ for u 1-periodic. This theorem gives us the existence, but also indirectly the uniqueness up to a constant, of the solutions of (1.5) for $f, g \in \mathcal{C}^\alpha(\Omega)$. Indeed, by repeating the steps presented at the beginning of Section 1.3 backwards, i.e. starting with the equation, multiplying by a test function, integrating and then using a change of variable (possible since $\nabla\Psi \in \mathcal{C}^{1,\beta}(\Omega)$ is invertible) to get back to the mass conservation constraint (1.1), we realize that the gradient of a convex solution of (1.5) is the optimal map in the transport problem. Therefore, by Theorem 1.2.1, $\nabla\Psi$ must be unique, which implies that Ψ must be unique up to a constant.

Chapter 2

Resolution of the Monge-Ampère equation

2.1 The Algorithm

Loeper and Rapetti presented in [13] a numerical method based on Newton's algorithm to solve the equation

$$\det(D^2\Psi) = f(x)$$

in a periodic setting. This equation can be associated with the transport problem in the case where the density ν is uniform, which gives $g = 1$. What we propose to do is to use their arguments to cover the case of a general (smooth enough) periodic final density. In order to do this, motivated by what we saw in Chapter 1, we write $\Psi(x) = |x|^2/2 + u(x)$ and obtain the equivalent Monge-Ampère equation

$$g(x + \nabla u(x)) \det(\mathcal{I} + D^2u(x)) = f(x). \tag{2.1}$$

Let's denote this Monge-Ampère operator applied to a function u by $M(u)$. Therefore, we are looking for a periodic solution u of the equation $M(u) = f$ such that $|x|^2/2 + u(x)$ is convex on $\Omega = [0, 1]^d$.

As we want to develop an algorithm based on Newton's method, we need to find a linearization of (2.1). From [22], we get the formula for the derivative of the determinant and hence the expansion:

$$\det(\mathcal{I} + D^2(u + s\theta)) = \det(\mathcal{I} + D^2u) + s\,\mathrm{Tr}\,(\mathrm{Adj}(\mathcal{I} + D^2u)D^2\theta) + \mathcal{O}(s^2) \tag{2.2}$$

where $\mathrm{Adj}(A) = \det(A) \cdot A^{-1}$. Next, from the usual Taylor series, we get the expansion for g:

$$g(x + \nabla(u + s\theta)) = g(x + \nabla u) + s\, \nabla g(x + \nabla u) \cdot \nabla\theta + \mathcal{O}(s^2). \tag{2.3}$$

By multiplying the latter two we get

$$\begin{aligned} M(u + s\theta) = &\det(\mathcal{I} + D^2 u)\, g(x + \nabla u) \\ &+ s \Big[g(x + \nabla u)\, \mathrm{Tr}\left(\mathrm{Adj}(\mathcal{I} + D^2 u) D^2\theta\right) \\ &\qquad + \det(\mathcal{I} + D^2 u)\, \nabla g(x + \nabla u) \cdot \nabla\theta \Big] + \mathcal{O}(s^2) \end{aligned} \tag{2.4}$$

out of which we obtain the following formula for the derivative of the Monge-Ampère operator at u in direction θ:

$$\begin{aligned} D_u M \cdot \theta = &\, g(x + \nabla u)\, \mathrm{Tr}\left(\mathrm{Adj}(\mathcal{I} + D^2 u) D^2\theta\right) \\ &+ \det(\mathcal{I} + D^2 u)\, \nabla g(x + \nabla u) \cdot \nabla\theta. \end{aligned} \tag{2.5}$$

Observe that we can write

$$\begin{aligned} \mathrm{Tr}\left(\mathrm{Adj}(\mathcal{I} + D^2 u) D^2\theta\right) &= \sum_{i,j=1}^{d} \mathrm{Adj}(\mathcal{I} + D^2 u)_{ij} D_{ij}\theta \\ &= \det(\mathcal{I} + D^2 u) \sum_{i,j=1}^{d} (\mathcal{I} + D^2 u)^{-1}_{ij} D_{ij}\theta. \end{aligned}$$

Having this linearization formula in hand, we can at last state the Newton-like algorithm we shall use to solve our problem.

Damped Newton algorithm

$$\begin{cases} \text{With } u_0 \text{ given, loop over } n \in \mathbb{N} \\ \text{Compute } f_n = g(x + \nabla u_n)\, \det(\mathcal{I} + D^2 u_n) \\ \text{Solve the linearized Monge-Ampère equation} \\ \qquad\qquad D_{u_n} M \cdot \theta_n = \dfrac{1}{\tau}(f - f_n) \\ \text{Update the solution: } u_{n+1} = u_n + \theta_n \end{cases} \tag{2.6}$$

Since we will have to refer to this algorithm quite often, even if it's a variant of the

classical method, we shall still call it the Newton algorithm. The factor $1/\tau$ ($\tau \geq 1$) is used as a stepsize parameter to help preventing the method from diverging by taking a step that goes "too far". In Section 2.2 we will prove that (2.6) has a unique solution up to a constant. To fix this constant, we can assign the value of u at a certain point in Ω (in practice, we use $u(0) = 0$) or we can force it to satisfy $\int_\Omega u\,dx = 0$. While running the algorithm, we enforce this by asking that θ satisfies the same condition. From now on, we will denote the linearized Monge-Ampère operator at step n by L_n.

The classical Newton method strongly depends on the starting point u_0. We will show in Section 2.4 that if we start with u_0 constant, then there exists a τ such that the algorithm converges. However, if for example one has an idea a priori of how the solution should look, it is possible to start with a more general u_0 and get a faster convergence, as long as it respects the following condition (in the context of Theorem 2.4.1): for $g(x+\nabla u_0)\det(\mathcal{I}+D^2u_0) = f_0$ and $u_0 \in \mathcal{C}^{4,\alpha}(\Omega)$ where $\alpha > 0$, $\mathcal{I}+D^2u_0$ and $\mathrm{Adj}(\mathcal{I} + D^2u_0)$ are $\mathcal{C}^{2,\alpha}(\Omega)$ smooth, uniformly positive definite matrices. Note that the two other conditions required in the proof of Theorem 2.4.1 are automatically implied here. Moreover, we say that a matrix A is $\mathcal{C}^{k,\alpha}(\Omega)$ smooth if all of its entries are in $\mathcal{C}^{k,\alpha}(\Omega)$, $k \geq 0$.

2.2 Linear Second Order Elliptic PDE in a Periodic Setting

Our goal in this section is to prove that the problem

$$\begin{cases} L\theta = h \\ \\ \theta \text{ is 1-periodic,} \end{cases} \tag{2.7}$$

where L is a linear second order strictly elliptic operator with 1-periodic $\mathcal{C}^\alpha(\Omega)$ coefficients:

$$L = \sum_{i,j=1}^{d} a_{ij}(x)\partial_i\partial_j + \sum_{i,j=1}^{d} b_i(x)\partial_i + c(x), \quad a_{ij} = a_{ji} \ \forall\, i,j,$$

and h is a 1-periodic $\mathcal{C}^\alpha(\Omega)$ function with mean 0, has a (unique up to a constant) solution $\theta \in \mathcal{C}^{2,\alpha}(\Omega)$ when $c \leq 0$. Note that from this, using classical bootstrapping arguments, we get that if the coefficients are $\mathcal{C}^{k,\alpha}(\Omega)$, then the solution is in $\mathcal{C}^{k+2,\alpha}(\Omega)$. Even if this is a well-known result, we could not find a detailed proof for it in the

context of periodic boundary conditions. This is the motivating reason for this section. Note also that we base our arguments on the work of Gilbarg and Trudinger presented in [9], who proved the same result for Dirichlet boundary conditions. Let's first define what we mean by elliptic in the case of a linear second order operator:

Definition 2.2.1. *Let $A(x) = (a_{ij}(x))$ be the second order coefficients matrix of L. If*

$$\lambda(x)|\xi|^2 \leq \xi^T A(x)\,\xi \leq \Lambda(x)|\xi|^2$$

for all non-zero $\xi \in \mathbb{R}^d$, then the operator L is said to be elliptic in Ω if $\lambda > 0$ in Ω and strictly elliptic if $\lambda \geq m_\lambda > 0$ for some constant m_λ. In addition, we say it is uniformly elliptic in Ω if the ratio Λ/λ is bounded in Ω.

Observe that if the coefficients a_{ij} are constants, then ellipticity implies strict and uniform ellipticity. To show that (2.7) has a unique (up to a constant) solution, we use the classical method of continuity and relate this latter problem to the Poisson periodic boundary problem

$$\begin{cases} \Delta\theta = h \\ \\ \theta \text{ is 1-periodic.} \end{cases} \tag{2.8}$$

In order to simplify the next developments, we will fix the value of the constant by restricting the solution space $\mathcal{C}^{2,\alpha}(\Omega)$, considering only the functions θ such that $\int_\Omega \theta\, dx = 0$. The existence and uniqueness of weak solutions of (2.8) relies on the famous Lax-Milgram theorem:

Theorem 2.2.2 (Lax-Milgram)**.** *Let H be a Hilbert space, $H^\star$ its dual space, $h : H \to \mathbb{R}$ a bounded linear functional on H and ζ a coercive and continuous bilinear form, i.e.,*

1) *$\exists K_1 > 0$ constant such that $\zeta[\theta,\theta] \geq K_1\|\theta\|_H^2\ \forall\theta \in H$,*

2) *$\exists K_2 > 0$ constant such that $|\zeta[\theta, v]| \leq K_2\|\theta\|_H\|v\|_H\ \forall\theta, v \in H$.*

Then there exists a unique element $\theta \in H$ such that

$$\zeta[\theta, v] = < h, v >_{H^\star,H}\ \ \forall\, v \in H,$$

where $< h, v >_{H^\star,H}$ represents the action of $h \in H^\star$ on $v \in H$.

In our case, if we denote by $\mathcal{C}^{\infty}_{per}(\Omega)$ the restriction to $\Omega = [0,1]^d$ of 1-periodic smooth functions on $\mathbb{R}^d$, then we can define the space $L^2_{per}(\Omega)$ as the completion of $\mathcal{C}^{\infty}_{per}(\Omega)$ with respect to the L^2 norm and in a similar way $H^1_{per}(\Omega)$ where $H^1(\Omega)$ is the Sobolev space of square-integrable functions whose first-order derivatives are also square-integrable. Using this, we create

$$H = \left\{ \theta \in H^1_{per}(\Omega) \middle| \int_{\Omega} \theta \, dy = 0 \right\}$$

and its dual

$$H^{\star} = \left\{ \theta^{\star} \in (H^1_{per}(\Omega))^{\star} \middle| \ < \theta^{\star}, 1 >_{H^{\star}_{per}, H_{per}} = 0 \right\}.$$

Now, recall that θ is said to be a weak solution of Poisson's equation (2.8) if

$$\zeta[\theta, v] \equiv \int_{\Omega} \nabla\theta(x) \cdot \nabla v(x) \, dx =< h, v >_{H^{\star}, H} \ \ \forall \, v \in H.$$

The bilinear form $\zeta[\theta, v]$ satisfies the conditions of the Lax-Milgram Theorem, and therefore we conclude the existence and uniqueness of the weak solution $\theta \in H$, as long as $h \in H^{\star}$. Notice that this latter condition makes sense through the Riesz Representation Theorem. Moreover, it forces the right-hand side h to satisfy $\int_{\Omega} h \, dx = 0$. We will have to pay attention to this condition in later developments. For more details on the definitions and arguments, we refer the reader to the books [12] or [24].

Next, we appeal to another type of fundamental technique applicable to elliptic partial differential equations: use of Schauder estimates. The one we require is an interior a-priori estimate that will enable us to determine whether we can turn our weak solution into a classical one or not. It can be found in [9] or [18]. Let $\Omega_A \subset \Omega$ be open and bounded, and $\Omega_B \subset\subset \Omega_A$ (Ω_B compactly embedded in Ω_A). If θ is a weak solution of $\Delta\theta = h$ in Ω_A and if $h \in \mathcal{C}^{\alpha}(\Omega_A)$, then there is a constant $k > 0$ such that

$$\|\theta\|_{\mathcal{C}^{2,\alpha}(\Omega_B)} \leq k \left(\|h\|_{\mathcal{C}^{\alpha}(\Omega_A)} + \|\theta\|_{L^2(\Omega_A)} \right).$$

This bound implies $\mathcal{C}^{2,\alpha}$ regularity of the solution θ in a domain Ω_B inside Ω when $h \in \mathcal{C}^{\alpha}(\Omega_B)$. Remember that since we are dealing with a periodic setting, we can extend this result on the whole domain and get $\theta \in \mathcal{C}^{2,\alpha}(\Omega)$.

We are now ready to state the main theorem of this section, which links (2.8) and (2.7). Note that this result is a modified version of Theorem 6.8 of [9] for the periodic case.

Theorem 2.2.3. *Let Ω be the hypercube $[0,1]^d$. Consider the operator L to be strictly elliptic in Ω with coefficients in $\mathcal{C}^\alpha(\Omega)$ and with $c \leq 0$. If the periodic boundary problem for Poisson's equation (2.8) has a unique $\mathcal{C}^{2,\alpha}(\Omega)$ solution with mean 0 for all functions $h \in \mathcal{C}^\alpha(\Omega)$ such that $\int_\Omega h\,dx = 0$, then the periodic boundary problem (2.7) also has a unique $\mathcal{C}^{2,\alpha}(\Omega)$ solution with mean 0 for all such h.*

In order to prove this, we need a few more classical results from the theory of elliptic operators. First, we will require yet another Schauder type estimate taken from [9] to control the Hölder norm of the second derivatives of θ. Let Ω_A and Ω_B be as previously defined. If $\operatorname{dist}(\Omega_B, \partial\Omega_A) \geq r$, then there is a constant $k > 0$ such that

$$\begin{aligned} r \max_{\|i\|_1=1} \sup_{x\in\Omega_B} |D^i\theta| + r^2 \max_{\|i\|_1=2} \sup_{x\in\Omega_B} |D^i\theta| + r^{2+\alpha} \max_{\|i\|_1=2} [D^i\theta]_{\alpha,\,\Omega_B} \\ \leq k(\|\theta\|_{\mathcal{C}^0(\Omega_A)} + \|h\|_{\mathcal{C}^\alpha(\Omega_A)}) \end{aligned} \tag{2.9}$$

where $i \in \mathbb{I}_2$ is a multi-index (see Appendix A) and k depends only on the ellipticity constant λ, the $\mathcal{C}^\alpha(\Omega_A)$ norms of the coefficients of L, d, α and the diameter of Ω_A. Again, here this bound applies everywhere in the bigger domain Ω. In addition, we shall use the maximum principle, which states that the extremal values of the solution to an elliptic equation are attained on the boundary.

Theorem 2.2.4 (Maximum principle)**.** *Let L be elliptic in a bounded domain Ω and $\theta \in \mathcal{C}^2(\Omega)$. If $L\theta = 0$ and $c \leq 0$ in Ω, then*

$$\sup_\Omega |\theta| = \sup_{\partial\Omega} |\theta| \quad \textit{and} \quad \inf_\Omega |\theta| = \inf_{\partial\Omega} |\theta|.$$

We will also rely on another very useful idea for the proof of Theorem 2.2.3, namely the method of continuity, which links our two problems together so that we can draw results from the first one and apply them to the other.

Theorem 2.2.5 (Method of continuity)**.** *Let $\mathcal{B}$ be a Banach space, $\mathcal{V}$ a normed vector space and let L_0, L_1 be bounded linear operators from $\mathcal{B}$ to $\mathcal{V}$. For each $t \in [0,1]$, set*

$$L_t = (1-t)L_0 + tL_1 \tag{2.10}$$

and suppose that there is a constant C such that

$$\|\theta\|_{\mathcal{B}} \leq C\|L_t\theta\|_{\mathcal{V}} \tag{2.11}$$

for $t \in [0,1]$. Then L_1 maps $\mathcal{B}$ onto $\mathcal{V}$ if and only if L_0 maps $\mathcal{B}$ onto $\mathcal{V}$.

Finally, we need two other more technical results before starting the proof of our main theorem.

Lemma 2.2.6. *If $\theta \in \mathcal{C}^2(\Omega)$ is a 1-periodic solution of $L\theta = 0$ such that $\int_\Omega \theta \, dx = 0$, then $\theta \equiv 0$.*

Proof. We prove by induction on the dimension d of the space $\Omega = [0,1]^d$ that a $\mathcal{C}^2(\Omega)$ 1-periodic solution of $L\theta = 0$ has to be constant. For $d = 1$, by the maximum principle,

$$\sup_\Omega |\theta| = \sup_{\partial\Omega} |\theta| = |\theta(0)| = |\theta(1)| = \inf_{\partial\Omega} |\theta| = \inf_\Omega |\theta| \quad \text{since } \theta \text{ is 1-periodic,}$$

out of which we deduce that $\theta \equiv C$, a constant.

Assume now that this statement hold for a certain dimension $d \geq 1$ and consider the case where $\Omega = [0,1]^{d+1}$ is a hypercube of dimension $d+1$, or $(d+1)$-cube. Note that this cube has $2(d+1)$ hypersides, forming the boundary of Ω. Using again the maximum principle, we get that both the sup and the inf of θ are attained on this boundary.

Since θ is 1-periodic, it takes the same value on every opposite hyperside; hence, we can consider only $d+1$ of these hypersides when looking for the sup (or inf) and we can make this selection in a such a way that they will have a point in common. Denote them by S_1, ..., S_{d+1} and for $1 \leq k \leq d+1$, denote by v_k the restriction of θ to S_k, which is a 1-periodic function on S_k. By the induction hypothesis, a solution to the projected problem on S_k,

$$\begin{cases} Lv_k = 0 \\ \\ v_k \text{ is 1-periodic on } S_k, \end{cases}$$

must be $v_k \equiv c_k$ where c_k is a constant. Therefore, the value of θ on S_k is c_k for every k. In addition, the fact that all the S_k share the same point implies that the value of θ at that point is equal to all the c_k, which in turn implies that these constants are all equal to both the sup and the inf of $|\theta|$ on Ω:

$$\sup_\Omega |\theta| = \inf_\Omega |\theta| = c_k \quad \forall\ 1 \leq k \leq d+1.$$

We then conclude that θ is constant, and from $\int_\Omega \theta \, dx = 0$, it is fixed to 0. □

Lemma 2.2.7. *Every 1-periodic solution $\theta \in \mathcal{C}^{2,\alpha}(\Omega)$ of the problem $L\theta = h$ with $\int_\Omega \theta \, dx = 0$ satisfies $\|\theta\|_{\mathcal{C}^0(\Omega)} \leq C\|h\|_{\mathcal{C}^\alpha(\Omega)}$, where C is a constant.*

Proof. Assume that this condition is not satisfied. In such a case, for every constant $C_n > 0$, there exists a solution $\theta_n \in \mathcal{C}^{2,\alpha}(\Omega)$ with $\int_\Omega \theta_n \, dx = 0$ such that $\|\theta_n\|_{\mathcal{C}^0(\Omega)} > C_n\|L\theta_n\|_{\mathcal{C}^\alpha(\Omega)}$. Taking $C_n = n$ and normalizing θ_n, we get that there exists a sequence (θ_m) of functions in $\mathcal{C}^{2,\alpha}(\Omega)$ for which the next three conditions hold:

$$\int_\Omega \theta_m \, dx = 0 \ \forall m, \tag{2.12}$$

$$\left\|\theta_m\right\|_{\mathcal{C}^0(\Omega)} = 1 \ \forall m, \tag{2.13}$$

$$\text{and } \left\|L\theta_m\right\|_{\mathcal{C}^\alpha(\Omega)} \to 0. \tag{2.14}$$

Combining bound (2.9) and the fact that dealing with periodic boundary conditions enables us to use interior estimates on the whole domain, we get

$$\begin{aligned} r \max_{\|i\|_1=1} \sup_{x\in\Omega} |D^i\theta_m| + r^2 \max_{\|i\|_1=2} \sup_{x\in\Omega} |D^i\theta_m| + r^{2+\alpha} \max_{\|i\|_1=2} [D^i\theta_m]_{\alpha,\,\Omega} \\ \leq K_1(\|\theta_m\|_{\mathcal{C}^0(\Omega)} + \|L\theta_m\|_{\mathcal{C}^\alpha(\Omega)}) \end{aligned} \tag{2.15}$$

for constants $K_1, r > 0$. From the mean value theorem, we know that there exists $t \in (0,1)$ such that

$$\begin{aligned} \frac{|\theta_m(x) - \theta_m(y)|}{|x-y|^\alpha} &= \frac{|\nabla\theta_m(tx + (1-t)y)|\,|x-y|}{|x-y|^\alpha} \\ &= |\nabla\theta_m(tx + (1-t)y)|\,|x-y|^{1-\alpha}. \end{aligned}$$

Then, using this latter result,

$$\begin{aligned} [\theta_m]_{\alpha,\,\Omega} &= \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|\theta_m(x) - \theta_m(y)|}{|x-y|^\alpha} \\ &\leq \operatorname{diam}(\Omega)^{1-\alpha} \sup_{x\in\Omega} |\nabla\theta_m| \\ &= \operatorname{diam}(\Omega)^{1-\alpha} \sup_{x\in\Omega} \sqrt{\sum_{\|i\|_1=1} (D^i\theta_m)^2} \\ &\leq \operatorname{diam}(\Omega)^{1-\alpha} \sqrt{d} \sup_{x\in\Omega} \sum_{\|i\|_1=1} |D^i\theta_m| \end{aligned}$$

$$\leq \operatorname{diam}(\Omega)^{1-\alpha}\sqrt{d}\sum_{\|i\|_1=1}\sup_{x\in\Omega}|D^i\theta_m|$$
$$\leq \operatorname{diam}(\Omega)^{1-\alpha}\, d^{\frac{3}{2}}\max_{\|i\|_1=1}\sup_{x\in\Omega}|D^i\theta_m|.$$

Applying similar arguments, we also get the same result in the case of the first derivatives of θ_m:

$$\max_{\|i\|_1=1}[D^i\theta_m]_{\alpha,\Omega} \leq \operatorname{diam}(\Omega)^{1-\alpha}\, d^{\frac{3}{2}}\max_{\|i\|_1=2}\sup_{x\in\Omega}|D^i\theta_m|.$$

With this, we can now bound the $\mathcal{C}^{2,\alpha}(\Omega)$ norm of θ_m by doing the following:

$$\begin{aligned}
\big\|\theta_m\big\|_{\mathcal{C}^{2,\alpha}(\Omega)} &= \|\theta_m\|_{\mathcal{C}^2(\Omega)} + \max_{i\in\mathbb{I}_2}[D^i\theta_m]_{\alpha,\Omega} \\
&\leq \|\theta_m\|_{\mathcal{C}^0(\Omega)} + \max_{\|i\|_1=1}\sup_{x\in\Omega}|D^i\theta_m| + \max_{\|i\|_1=2}\sup_{x\in\Omega}|D^i\theta_m| \\
&\qquad + [\theta_m]_{\alpha,\Omega} + \max_{\|i\|_1=1}[D^i\theta_m]_{\alpha,\Omega} + \max_{\|i\|_1=2}[D^i\theta_m]_{\alpha,\Omega} \\
&\leq \big\|\theta_m\big\|_{\mathcal{C}^0(\Omega)} + K_2\left(\big\|\theta_m\big\|_{\mathcal{C}^0(\Omega)} + \big\|L\theta_m\big\|_{\mathcal{C}^\alpha(\Omega)}\right) \qquad \text{by (2.15)} \\
&\leq K_3 \qquad \text{by (2.14) and (2.13).}
\end{aligned}$$

where K_3 does not depend on m. Therefore (θ_m) is a uniformly bounded sequence in $\mathcal{C}^{2,\alpha}(\Omega)$; which gives us that it is equicontinuous. Appealing to the Ascoli-Arzelà theorem (see for example [24]), we know that we can extract a subsequence out of (θ_m) that converges uniformly in $\mathcal{C}^2(\Omega)$:

$$\theta_{m_j} \to \theta \in \mathcal{C}^2(\Omega). \tag{2.16}$$

Since this convergence in $\mathcal{C}^2(\Omega)$ is uniform, the limit function θ satisfies $\int_\Omega \theta\, dx = 0$. Moreover, the norm being a continuous function on its vector space, $\|L\theta_{m_j}\|_{\mathcal{C}^0(\Omega)} \to 0$ and (2.16) tells us that $L\theta = 0$, and then we invoke Lemma 2.2.6 to conclude that $\theta = 0$. However, this is a contradiction because

$$\big\|\theta\big\|_{\mathcal{C}^0(\Omega)} = \big\|\lim_{j\to\infty}\theta_{m_j}\big\|_{\mathcal{C}^0(\Omega)} = \lim_{j\to\infty}\big\|\theta_{m_j}\big\|_{\mathcal{C}^0(\Omega)} = 1 \qquad \text{by (2.12),}$$

and hence the result follows. □

We now have enough tools in our hands to prove the main theorem for this section:

Proof of Theorem 2.2.3. Consider the family of operators

$$L_t = tL + (1-t)\Delta, \quad 0 \le t \le 1.$$

and the corresponding equation $L_t\theta_t = h$. Since the coefficients of L satisfy

$$\sum_{i,j=1}^{d} a_{ij}(x)\xi_i\xi_j \ge \lambda(x)|\xi|^2 \quad \forall\, x \in \Omega,\, \xi \in \mathbb{R}^d$$
$$\left\|a_{ij}\right\|_{\mathcal{C}^\alpha(\Omega)}, \left\|b_i\right\|_{\mathcal{C}^\alpha(\Omega)}, \left\|c\right\|_{\mathcal{C}^\alpha(\Omega)} \le \sigma \quad \forall\, i,j$$

by hypothesis, we can show that the coefficients of L_t satisfy the same inequalities, replacing λ and σ by $\lambda_t = \min(1,\lambda)$ and $\sigma_t = \max(1,\sigma)$. First, if $a_{t_{ij}}$ represents the coefficients of the second order terms of L_t, we get:

$$\begin{aligned}\sum_{i,j=1}^{d} a_{t_{ij}}\xi_i\xi_j &= t\sum_{i,j=1}^{d} a_{ij}\xi_i\xi_j + (1-t)\sum_{i=1}^{d}\xi_i^2 \\ &\ge t\lambda|\xi|^2 + (1-t)|\xi|^2 \\ &= \left(t\lambda + (1-t)\right)|\xi|^2 \\ &\ge \lambda_t|\xi|^2.\end{aligned}$$

Secondly, since the coefficients b_i and c are the same for L and L_t, taking δ_{ij} to be 0 if $i \ne j$ and 1 if $i = j$, the other set of inequalities follows from

$$\begin{aligned}\|a_{t_{ij}}\|_{\mathcal{C}^\alpha(\Omega)} &= \|ta_{ij} + (1-t)\delta_{ij}\|_{\mathcal{C}^\alpha(\Omega)} \\ &\le t\|a_{ij}\|_{\mathcal{C}^\alpha(\Omega)} + (1-t)\|\delta_{ij}\|_{\mathcal{C}^\alpha(\Omega)} \\ &\le t\sigma + (1-t) \\ &\le \sigma_t.\end{aligned}$$

Then, by considering L_t as a bounded linear operator from the Banach space

$$\mathcal{B}_1 = \left\{\theta_t \in \mathcal{C}^{2,\alpha}(\Omega) : \int_\Omega \theta_t\, dx = 0\right\}$$

into the Banach space

$$\mathcal{B}_2 = \left\{h \in \mathcal{C}^\alpha(\Omega) : \int_\Omega h\, dx = 0\right\},$$

the solvability of the periodic boundary problem $L_t\theta_t = h$ becomes equivalent to the invertibility of the mapping L_t for $\theta_t \in \mathcal{B}_1$ and $h \in \mathcal{B}_2$. From Lemma 2.2.7, we have $\|\theta_t\|_{\mathcal{C}^0(\Omega)} \leq K_1\|L\theta_t\|_{\mathcal{C}^\alpha(\Omega)}$. Appealing once again to bound (2.9), we deduce that

$$\begin{aligned}\big\|\theta_t\big\|_{\mathcal{B}_1} = \big\|\theta_t\big\|_{\mathcal{C}^{2,\alpha}(\Omega)} &\leq K_2\left(\big\|\theta_t\big\|_{\mathcal{C}^0(\Omega)} + \big\|L\theta_t\big\|_{\mathcal{C}^\alpha(\Omega)}\right)\\ &\leq K_3\big\|L\theta_t\big\|_{\mathcal{C}^\alpha(\Omega)}\\ &= K_3\big\|L\theta_t\big\|_{\mathcal{B}_2}.\end{aligned}$$

Note here that K_3 is a constant that does not depend on t. By hypothesis, we assume that (2.8) has a (unique) solution in $\mathcal{B}_1$ and therefore Δ maps $\mathcal{B}_1$ onto $\mathcal{B}_2$. Using the method of continuity (Theorem 2.2.5), we obtain that L also maps $\mathcal{B}_1$ onto $\mathcal{B}_2$. Finally, from Lemma 2.2.6 (or from the bound included in Theorem 2.2.5) we conclude that L is actually an isomorphism, and thus invertible. □

2.3 Preliminaries to the Proof of Convergence

To be able to prove the convergence of our method, we mainly rely on a very important a priori estimate on the second derivative of the solution of the Monge-Ampère equation derived in its general form in [19] by Liu, Trudinger and Wang. To give the reader an idea of the conditions under which this estimate is valid, we will state it in its full generality. Consider the Monge-Ampère equation

$$\det\big(D^2\Psi(x) - A(x,\nabla\Psi)\big) = h(x,\nabla\Psi) \tag{2.17}$$

related to the optimal transport problem with a cost function $c \in \mathcal{C}^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ with

$$\begin{aligned}A(x,\nabla\Psi) &= D^2_{xx}(c(x,\check{T}(\nabla\Psi)))\\ \text{and } h(x,\nabla\Psi) &= |\det\big(D^2_{xy}c(x,y)\big)|\frac{f(x)}{g\circ\check{T}(\nabla\Psi)}\end{aligned}$$

where $\check{T}$ is the optimal transport map uniquely determined by

$$\nabla\Psi(x) = D_x(c(x,\check{T}(x))),$$

f and g being the usual initial and final densities defined respectively on Ω_1 and Ω_2. A solution u of (2.17) is said to be elliptic if $D^2\Psi(x) - A(x,\nabla\Psi)$ is a positive definite

matrix. Consider in addition the following three conditions on the cost function:

(C1) For any $x, p \in \mathbb{R}^d$, there is a unique $y \in \mathbb{R}^d$ such that $\nabla_x c(x,y) = p$, and vice-versa, for any $y, p \in \mathbb{R}^d$, there is a unique $x \in \mathbb{R}^d$ such that $\nabla_y c(x,y) = p$,

(C2) For any $(x,y) \in \mathbb{R}^d \times \mathbb{R}^d$, $\det\left(D^2_{xy} c(x,y)\right) \neq 0$,

(C3) For any $x, p \in \mathbb{R}^d$ and any $\xi, \eta \in \mathbb{R}^d$, $D^2_{p_k p_l} A_{ij}(x,p)\xi_i \xi_j \eta_k \eta_l \geq 0$.

Finally, before stating Liu, Trudinger and Wang's result, we need one extra definition:

Definition 2.3.1. *Take $\omega_h(p) = \sup\{|h(x) - h(y)| : |x - y| < p\}$, the oscillation of a function h. We say that h is Dini continuous if*

$$\int_0^1 \frac{\omega_h(p)}{p}\, dp < \infty.$$

Theorem 2.3.2 (Liu,Trudinger,Wang)**.** *Assume the cost function c satisfies **(C1)-(C3)**, f, g are Dini continuous, uniformly bounded and positive in Ω_1, Ω_2 respectively, such that $0 < \lambda \leq h \leq \Lambda$. Then if the target domain Ω_2 is c^*-convex with respect to Ω_1, any potential function $\Psi \in \mathcal{C}^2(\Omega_1)$ which is an elliptic solution of (2.17) satisfies the following interior a priori estimate:*

If $f \in \mathcal{C}^\alpha(\Omega_1)$ for some $\alpha \in (0,1)$, then

$$\|\Psi\|_{\mathcal{C}^{2,\alpha}(\Omega_{1,r})} \leq C\left[1 + \frac{\|h(x, \nabla\Psi)\|_{\mathcal{C}^\alpha(\Omega_1)}}{\alpha(1-\alpha)}\right] \tag{2.18}$$

where $\Omega_{1,r} = \{x \in \Omega_1 : dist(x, \partial\Omega_1) > r\}$ and C depends only on d, r, λ, Λ, Ω_1, Ω_2 and c. Consequently, if f, g are Hölder continuous and Ω_1, Ω_2 are c, c^*-convex with respect to each other, then the optimal mapping $\check{T}$ is a $\mathcal{C}^{1,\alpha}$ diffeomorphism from Ω_1 to Ω_2 for some $\alpha > 0$.

In our case, i.e., when we take $c(x,y) = x \cdot y$ (we already saw it is equivalent to taking $c(x,y) = |x-y|^2/2$ modulo a negative sign), we get $A = 0$, $\check{T} = \nabla\Psi = x + \nabla u(x)$, and we easily see that all the assumptions to the theorem are satisfied. We will only present the justification for why a Hölder continuous function is also Dini continuous, since it is the less obvious one. Thus, select $\gamma \in \mathcal{C}^\alpha(\Omega)$ for a certain $\alpha \in (0,1)$, which by definition means that

$$\sup_{x \neq y} \frac{|\gamma(x) - \gamma(y)|}{|x-y|^\alpha} = M < \infty.$$

Using this constant M, we get

$$\begin{aligned}\int_0^1 \frac{|\gamma(x)-\gamma(y)|}{p}\,dp &\leq \int_0^1 M\frac{|x-y|^\alpha}{p}\,dp \quad \forall\, x,y\\ &< \int_0^1 M\frac{p^\alpha}{p}\,dp \qquad\qquad \text{for } |x-y|<p\\ &= \frac{M}{\alpha} < \infty,\end{aligned}$$

and therefore γ is Dini continuous.

Before tackling the main proof for the convergence of the algorithm, we still need a few more tools in our toolbox.

Lemma 2.3.3. *Let A, H, K be matrices in $M_d(\mathbb{R})$, the space of d-dimensional square matrices with real coefficients. Then, the second derivative (in the Fréchet sense) of the determinant operator is given by:*

$$D^2\det(A)(H,K) = \det(A)\Big[Tr(A^{-1}K)\,Tr(A^{-1}H) - Tr(A^{-1}KA^{-1}H)\Big]$$

Proof. From [22], we know that

$$\begin{aligned} D\det(A)(H) &= \mathrm{Tr}(\mathrm{Adj}(A)H) = \det(A)\mathrm{Tr}\,(A^{-1}H)\\ \text{and}\qquad DA^{-1}(H) &= -A^{-1}HA^{-1}.\end{aligned}$$

Using this with the basic differentiation rules, we get

$$\begin{aligned} D^2\det(A)(H,K) &= D(D\det(A)(H))(K)\\ &= D(\det(A)\mathrm{Tr}\,(A^{-1}H))(K)\\ &= D(\det(A))(K)\mathrm{Tr}\,(A^{-1}H) + \det(A)D(\mathrm{Tr}\,(A^{-1}H))(K)\\ &= \det(A)\mathrm{Tr}\,(A^{-1}K)\mathrm{Tr}\,(A^{-1}H) - \det(A)\mathrm{Tr}\,(A^{-1}KA^{-1}H),\end{aligned}$$

and hence the result is proven. □

Lemma 2.3.4. *Take $f(x) = g(x+\nabla u(x))\det(\mathcal{I}+D^2u(x))$, where $f\in\mathcal{C}^\alpha(\Omega)$, $g\in\mathcal{C}^{1,\alpha}(\Omega)$, $u\in\mathcal{C}^{2,\alpha}(\Omega)$ and all f, g and u are 1-periodic. If $\int_\Omega g\,dx = 1$ and if $|x|^2/2 + u(x)$ is uniformly convex, then $\int_\Omega f\,dx = 1$.*

Proof. Let $T = x + \nabla u$. Then, by the change of variable formula, we get

$$\int_{T(\Omega)} g(x)\,dx = \int_{\Omega} g(T(x))\det(DT(x))\,dx.$$

Let's analyze $T(\Omega) = T([0,1]^d)$. For e_i a vector of the canonical basis in $\mathbb{R}^d$ having its ith component equal to 1,

$$\begin{aligned} u \text{ 1-periodic} \Rightarrow \quad & u(x+e_i) = u(x) \quad \forall x \in \Omega,\, \forall i \in \{1,\dots,d\} \\ & \Rightarrow \nabla u(x+e_i) = \nabla u(x). \end{aligned}$$

Then, $T(x+e_i) = x + e_i + \nabla u(x+e_i) = x + e_i + \nabla u(x) = T(x) + e_i$. Taking $x = 0$, we get

$$\begin{aligned} T(0+e_i) &= T(0) + e_i \\ &= \nabla u(0) + e_i \quad \forall i, \end{aligned}$$

which when applied repeatedly gives $T(0 + \sum_{j=0}^{m} e_{i_j}) = \nabla u(0) + \sum_{j=0}^{m} e_{i_j}$ for all subsets of indices $\{i_0, i_1, \dots, i_m\}$ of $\{1, 2, \dots, d\}$. Therefore, the solid corresponding to $T(\Omega) = T([0,1]^d)$ has its vertices given by the points $\nabla u(0) + \sum_{j=0}^{m} e_{i_j}$. Consider now the $(d-1)$-hypersurface

$$T((x_1, x_2, \dots, x_{i-1}, 0, x_{i+1}, \dots, x_d))$$

where $x_j \in [0,1]$ for $i \neq j$. Since $T \in \mathcal{C}^{1,\alpha}(\Omega)$ and ∇u is 1-periodic, this surface will be the same as $T((x_1, x_2, \dots, x_{i-1}, 1, x_{i+1}, \dots, x_d))$, but translated one unit in the e_i direction. Since this argument holds for all i, we can conclude that the hypercube $\nabla u(0) + [0,1]^d$ contains the same points as $T(\Omega)$. Next, observe that for γ a 1D continuous 1-periodic function, integrating γ on an interval of length one is equivalent to integrating it on $[0,1]$:

$$\begin{aligned}\int_a^{a+1}\gamma(x)\,dx &= \int_a^1\gamma(x)\,dx+\int_1^{a+1}\gamma(x)\,dx \text{ assuming w.l.o.g. that } a\in[0,1]\\ &= \int_a^1\gamma(x)\,dx+\int_0^a\gamma(x-1)\,dx\\ &= \int_a^1\gamma(x)\,dx+\int_0^a\gamma(x)\,dx \qquad \text{since } \gamma \text{ is periodic}\\ &= \int_0^1\gamma(x)\,dx.\end{aligned}$$

Using this idea in a more general way for g, we get

$$\begin{aligned}\int_{T(\Omega)}g(x)\,dx &= \int_{\nabla u(0)+[0,1]^d}g(x)\,dx\\ &= \int_{u_{x_d}(0)}^{u_{x_d}(0)+1}\int_{u_{x_{d-1}}(0)}^{u_{x_{d-1}}(0)+1}\dots\int_{u_{x_1}(0)}^{u_{x_1}(0)+1}g(x)\,dx_1\,\dots\,dx_{d-1}dx_d\\ &= \int_0^1\int_0^1\dots\int_0^1 g(x)\,dx_1\,\dots\,dx_{d-1}dx_d\\ &= 1,\end{aligned}$$

hence the result. □

Next, we will require an estimate on the $C^{1,\alpha}(\Omega)$ norm of the solution of the Monge-Ampère equation. It is given to us in the next theorem, taken from [7].

Theorem 2.3.5 (Caffarelli)**.** *Let $\Omega\subset\mathbb{R}^d$ be open and let $\Psi:\Omega\to\mathbb{R}$ be a smooth convex function satisfying the Monge-Ampère equation*

$$\det\left(D^2\Psi(x)\right)=F(x,\nabla\Psi(x)) \qquad \text{in } \Omega.$$

Let $\mathcal{K}_\Omega(\Psi)$ stand for the modulus of convexity of Ψ in Ω. Then there is an $\epsilon\in(0,1)$ for which in any open subdomain Ω_A such that $\overline{\Omega}_A\subset\Omega$, one has the a priori estimate

$$\|\Psi\|_{\mathcal{C}^{1,\epsilon}(\Omega_A)}\le C\left(\Omega,\Omega_A,\|F\|_{L^\infty(\Omega)},\|\nabla\Psi\|_{L^\infty(\Omega)},\mathcal{K}_\Omega(\Psi)\right).$$

Since this estimate does not hold for all $\epsilon\in(0,1)$, one might wonder for which values it is actually valid. Forzani and Maldonado provided an answer to that question in [26]. They found out that if $0<\lambda<\det D^2\Psi<\Lambda$ (in our case, $\Lambda=\dfrac{M_f}{m_g}$ and

d	w_d	Rounded K
1	2	$64c$
2	π	$4550c$
3	$4\pi/3$	$140039c$
4	$\pi^2/2$	$2709370c$

Table 2.1: Quantities involved in the bounds on ϵ for $c = \Lambda/\lambda$

$\lambda = \dfrac{m_f}{M_g}$), then the bound is valid for

$$\epsilon = \frac{1}{1 + 2(K+1)(K+2)} \quad \text{for } K = \frac{2^{3d+2}\, w_d\, w_{d-1}\, \Lambda}{d^{-3/2}\, \lambda}$$

where

$$w_r = \frac{\pi^{r/2}}{\Gamma(r/2+1)}$$

is the volume of the r-dimensional unit ball. To give the reader an idea of these values, a few of them are presented in Table 2.1. Observe that K increases quite rapidly as we vary the dimension d. In consequence, the value presented for ϵ gets very close to 0, even if the ratio $\Lambda/\lambda = (M_f M_g)/(m_f m_g)$ is small (which would happen if one of the densities is very close to being uniform) since this ratio is always greater or equal than 1. Forzani and Maldonado also made in [26] the bound in Theorem 2.3.5 explicit. Under the same assumptions as in Theorem 2.3.2, for $k = 2(K+1)(K+2)$ (i.e. $\epsilon = 1/(1+k)$), we have

$$\frac{|\nabla\Psi(z) - \nabla\Psi(y)|}{|z-y|^{\frac{1}{1+k}}} \leq R_y \left(\frac{K}{m(\psi_y^*, 0, R_y)}\right)^{\frac{1}{1+k}} \left(\frac{M(\Psi, y, r)}{r}\right)^{\frac{1}{1+k}} \tag{2.19}$$

for $|z-y| \leq r$ where $\psi_y(x) = \Psi(x+y) - \Psi(y) - \nabla\Psi(y)\cdot x$,

$$R_y = \max\left\{1, \left(\frac{K M(\Psi, y, r)}{m(\psi_y^*, 0, 1)}\right)^{\frac{k}{1+k}}\right\},$$

and $M(\Psi, y, r)$, $m(\Psi, y, r)$ denote respectively the max and min of $\psi_y(z-y)$ taken over the points z such that $|z-y| = r$.

Lastly, we shall use a maximum principle for the second derivatives of the Monge-Ampère equation which can be found in [9, 39] or for the more general version including different costs in [31].

Theorem 2.3.6. *Let Ψ be an elliptic solution of (2.17) in Ω. Suppose that the cost function satisfies **(C1),(C2),(C3)** and h is a positive function in $\mathcal{C}^2(\overline{\Omega} \times \mathbb{R} \times \mathbb{R}^d)$. Then*

$$\sup_{\Omega} |D^2\Psi| \leq C \left(1 + \sup_{\partial\Omega} |D^2\Psi|\right)$$

where C depends only on $A, h, \Omega, \check{T}(\Omega)$ and on the sup of Ψ and $\nabla\Psi$ on Ω.

2.4 Proof of Convergence of the Algorithm

We are now ready to state and prove a theorem about the convergence of algorithm (2.6), under some suitable conditions. Note that the arguments for the proof of this theorem are similar to the ones presented by Loeper and Rapetti in [13], but the more general case of a non-uniform g presents some new difficulties that are worthwhile exposing here. In addition, the proof provides important information that can be used to gain some intuition on the performance of the algorithm in practice.

Theorem 2.4.1. *Consider $\Omega = [0,1]^d$ and let f, g be two positive 1-periodic probability densities bounded away from 0. Assume that the initial guess u_0 for the Newton algorithm (2.6) is constant. If $f \in \mathcal{C}^{2,\alpha}(\Omega)$ and $g \in \mathcal{C}^{3,\alpha}(\Omega)$ for any $0 < \alpha < 1$, then there exists $\overline{\tau} \geq 1$ such that (u_n) converges in $\mathcal{C}^{4,\beta}(\Omega)$, for any $0 < \beta < \alpha$ and any $\tau \geq \overline{\tau}$, to the unique – up to a constant – solution u of the Monge-Ampère equation (2.1). Moreover, τ depends only on α, d, $\|f\|_{\mathcal{C}^{2,\alpha}(\Omega)}, \|g\|_{\mathcal{C}^{3,\alpha}(\Omega)}$ and M_f, M_g, m_f, m_g.*

Proof. Unless otherwise stated, we only need to consider $f \in \mathcal{C}^{\alpha}(\Omega)$ and $g \in \mathcal{C}^{2,\alpha}(\Omega)$. The main step of the proof consists in showing by induction that the following holds for all $n \in \mathbb{N}$:

(1) $\mathcal{I} + D^2u_n$ and $\mathrm{Adj}(\mathcal{I} + D^2u_n)$ are $\mathcal{C}^{\alpha}(\Omega)$ smooth, uniformly positive definite matrices.

(2) $\dfrac{f}{C_1} \leq f_n \leq C_1 f$ where C_1 does not vary with n.

(3) $\|f - f_n\|_{\mathcal{C}^{\alpha}(\Omega)} \leq C_2$ where C_2 does not vary with n.

Note that **(1)** implies that L_n is a strictly elliptic linear operator and that $|x|^2/2+u_n$ is uniformly convex. Furthermore, **(2)** yields $0 < \lambda \leq \det(\Psi_n) \leq \Lambda$ which is required in order to apply the regularity estimates presented earlier, while **(3)** ensures that the right-hand side in the linearized Monge-Ampère equation has the correct regularity assumptions to employ the Schauder estimates used in Section 2.2. Now, starting with $u_0 = 0$, we get $f_0 = g$. Observe that

$$\mathcal{I} + D^2 u_0 = \operatorname{Adj}\left(\mathcal{I} + D^2 u_0\right) = \mathcal{I},$$

and therefore both matrices are smooth and uniformly positive definite. Using this,

$$\begin{aligned}\sum_{i,j=1}^{d} g\left(x + \nabla u_0\right) \operatorname{Adj}\left(\mathcal{I} + D^2 u_0\right)_{ij} \xi_i \xi_j &= g\left(x + \nabla u_0\right) |\xi|^2 \\ &\geq m_g |\xi|^2, \qquad\qquad m_g > 0.\end{aligned}$$

Hence L_0 is an elliptic operator. Next, since $D^2(u_0 + |x|^2/2) = \mathcal{I}$ is uniformly positive definite, we know that $u_0 + |x|^2/2$ is a uniformly convex function. Now, by taking

$$C_1 = \max\left(\frac{M_f}{m_g}, \frac{M_g}{m_f}\right),$$

we see that $f/C_1 \leq m_g \leq g$ and $g \leq M_g \leq C_1 f$ so $f/C_1 \leq f_0 \leq C_1 f$. Finally,

$$\|f - f_0\|_{\mathcal{C}^\alpha(\Omega)} \leq \|f\|_{\mathcal{C}^\alpha(\Omega)} + \|g\|_{\mathcal{C}^\alpha(\Omega)} = C_2 < \infty$$

since both f and g are in $\mathcal{C}^\alpha(\Omega)$. Therefore, all the statements are true for $n = 0$ with our choice of u_0.

Let's assume they hold for a certain $n \in \mathbb{N}$ and prove them for $n+1$. One should consider for now the stepsize parameter to possibly vary with n. We shall prove later that we can actually take it to be constant without affecting any result. Since $f \in \mathcal{C}^\alpha(\Omega)$, $g \in \mathcal{C}^{2,\alpha}(\Omega)$ and since **(1)** holds by the induction hypothesis, we get that the coefficients of L_n are in $\mathcal{C}^\alpha(\Omega)$. In addition, **(3)** implies that it is also the case for the right-hand side $(f - f_n)/\tau$. Recalling that θ_n solves $L_n \theta_n = (f - f_n)/\tau$, we can

use (2.9) again; where

$$r \max_{\|i\|_1=1} \sup_{x\in\Omega} |D^i\theta_n| + r^2 \max_{\|i\|_1=2} \sup_{x\in\Omega} |D^i\theta_n| + r^{2+\alpha} \max_{\|i\|_1=2} [D^i\theta_n]_{\alpha,\Omega} \\ \leq K_1 \left(\|\theta_n\|_{\mathcal{C}^0(\Omega)} + \frac{1}{\tau}\|f - f_n\|_{\mathcal{C}^\alpha(\Omega)} \right)$$

for a constant $K_1 > 0$. Moreover, we can select θ_n such that it satisfies $\int_\Omega \theta_n \, dx = 0$ ($\theta_n \in \mathcal{C}^{2,\alpha}(\Omega)$ by Section 2.2) and appeal to Lemma 2.2.7 to get

$$\|\theta_n\|_{\mathcal{C}^0(\Omega)} \leq \frac{K_2}{\tau}\|f - f_n\|_{\mathcal{C}^\alpha(\Omega)}.$$

With this and bound **(3)** in mind, we can use the mean value theorem in the same way we did in the proof of Lemma 2.2.7 to find that there exists a constant $k_{\theta_n} > 0$ such that

$$\left\|\theta_n\right\|_{\mathcal{C}^{i,\alpha}(\Omega)} \leq \frac{k_{\theta_n}}{\tau}\left\|f - f_n\right\|_{\mathcal{C}^\alpha(\Omega)} \leq \frac{k_{\theta_n} C_2}{\tau}, \quad i = 1, 2. \tag{2.20}$$

Remembering that $u_{n+1} = u_n + \theta_n$, we deduce that $\mathcal{I} + D^2 u_{n+1}$ is $\mathcal{C}^\alpha(\Omega)$ smooth. In addition, the terms of $\mathrm{Adj}(\mathcal{I} + D^2 u_{n+1})$ are products of $\mathcal{C}^\alpha(\Omega)$ functions and thus, by Lemma A.0.1, we know that they also have the same regularity properties. This implies that $\mathrm{Adj}(\mathcal{I} + D^2 u_{n+1})$ is also $\mathcal{C}^\alpha(\Omega)$ smooth. Now, since $\mathcal{I} + D^2 u_n$ is uniformly positive definite by assumption, $\exists\, K_3 > 0$ constant for which

$$\xi^T(\mathcal{I} + D^2 u_n)\xi \geq K_3|\xi|^2 \ \forall \xi \in \mathbb{R}^d.$$

We have the same result for u_{n+1} since

$$\begin{aligned}
\xi^T(\mathcal{I}+D^2u_{n+1})\xi &= \xi^T(\mathcal{I}+D^2u_n)\xi + \xi^T(D^2\theta_n)\xi \\
&\geq K_3|\xi|^2 + \sum_{i,j=1}^{d} D_{ij}\theta_n\xi_i\xi_j \\
&\geq K_3|\xi|^2 - \frac{k_{\theta_n}}{\tau}\|f-f_n\|_{\mathcal{C}^\alpha(\Omega)}\sum_{i,j=1}^{d}\xi_i\xi_j \\
&\geq K_3|\xi|^2 - \frac{k_{\theta_n}}{2\tau}\|f-f_n\|_{\mathcal{C}^\alpha(\Omega)}\sum_{i,j=1}^{d}\left(\xi_i^2+\xi_j^2\right) \\
&= K_3|\xi|^2 - \frac{k_{\theta_n}2d}{2\tau}\|f-f_n\|_{\mathcal{C}^\alpha(\Omega)}|\xi|^2 \\
&= \left[K_3 - \frac{k_{\theta_n}d}{\tau}\|f-f_n\|_{\mathcal{C}^\alpha(\Omega)}\right]|\xi|^2 \\
&\geq K_4|\xi|^2,\ K_4>0 \text{ for } \tau \text{ large enough.}
\end{aligned}$$

Hence $\mathcal{I}+D^2u_{n+1}$ is uniformly positive definite which in turn tells us that $\frac{|x|^2}{2}+u_{n+1}(x)$ is uniformly convex (see Theorem B.0.5). Let's carry on by proving that the bounds stated in **(2)** and **(3)** also hold for $n+1$. First, recall the formula for the linearized Monge-Ampère equation $L_n\theta_n = (f-f_n)/\tau$ at step n (see (2.5)):

$$\begin{aligned}
&g(x+\nabla u_n)\,\mathrm{Tr}\left(\mathrm{Adj}(\mathcal{I}+D^2u_n)D^2\theta_n\right) \\
&\qquad + \det\left(\mathcal{I}+D^2u_n\right)\nabla g(x+\nabla u_n)\cdot\nabla\theta_n = \frac{f-f_n}{\tau}.
\end{aligned} \tag{2.21}$$

We can write f_{n+1} in terms of f_n in the following way (c.f. (2.6)):

$$\begin{aligned}
f_{n+1} &= g(x+\nabla u_{n+1})\det\left(\mathcal{I}+D^2u_{n+1}\right) \\
&= g(x+\nabla u_n)\det\left(\mathcal{I}+D^2u_n\right) + L_n\theta_n + r_n \\
&= f_n + \frac{f-f_n}{\tau} + r_n.
\end{aligned} \tag{2.22}$$

We can now bound the residual r_n. For a function $\zeta : X \to Y$ where X and Y are two Banach spaces, we get from the Taylor expansion with integral remainder in a general Banach space setting (see for example [4]) that

$$\zeta(a+h) = \zeta(a) + \mathrm{D}\zeta(a)\cdot h + \int_0^1 \mathrm{D}^2\zeta(a+th)\cdot(h,h)\,dt\,.$$

Applying this to our case, using Lemma 2.3.3, we write the Taylor expansions of g and of the determinant about θ_n, multiply them, gather the second order terms and obtain the following formula for the residual:

$$r_n = \int_0^1 \Big[\mathrm{A}(u_n, \theta_n, t) + \mathrm{B}(u_n, \theta_n, t) + \mathrm{C}(u_n, \theta_n, t) \Big] \, dt, \tag{2.23}$$

where

$$\begin{aligned}
\mathrm{A}(u_n, \theta_n, t) = {} & g(x + \nabla u_n + t\nabla\theta_n) \det(\mathcal{I} + D^2 u_n + tD^2\theta_n) \\
& \times \left[\mathrm{Tr}\left((\mathcal{I} + D^2 u_n + tD^2\theta_n)^{-1} D^2\theta_n \right)^2 \right. \\
& \qquad \left. - \mathrm{Tr}\left(((\mathcal{I} + D^2 u_n + tD^2\theta_n)^{-1} D^2\theta_n)^2 \right) \right], \\
\mathrm{B}(u_n, \theta_n, t) = {} & \nabla g(x + \nabla u_n + t\nabla\theta_n) \cdot \nabla\theta_n \det(\mathcal{I} + D^2 u_n + tD^2\theta_n) \\
& \times \mathrm{Tr}\left((\mathcal{I} + D^2 u_n + tD^2\theta_n)^{-1} D^2\theta_n \right),
\end{aligned}$$

$$\mathrm{C}(u_n, \theta_n, t) = \frac{1}{2} \nabla\theta_n{}^T \cdot D^2 g(x + \nabla u_n + t\nabla\theta_n) \cdot \nabla\theta_n \det(\mathcal{I} + D^2 u_n + tD^2\theta_n).$$

Notice that A, B and C all consist of a sum of terms involving products of at least two second derivatives of θ_n with g or its derivatives, with second derivatives of $|x|^2/2 + u_n$ and with some powers of t. We know from (2.20) that we can bound the $\mathcal{C}^\alpha(\Omega)$ norm of the second derivatives of θ_n by a constant times $\|f - f_n\|_{\mathcal{C}^\alpha(\Omega)}/\tau$. In addition, since $g \in \mathcal{C}^{2,\alpha}(\Omega)$, the Hölder norm of g and its first and second derivatives are all uniformly bounded (bounded by a constant that does not depend on n). From this, we go back to the remainder r_n, apply all these bounds and estimates and get the bound we were looking for:

$$\begin{aligned}
\|r_n\|_{\mathcal{C}^\alpha(\Omega)} &\leq \frac{k_A}{\tau^2} \|f - f_n\|^2_{\mathcal{C}^\alpha(\Omega)} + \frac{k_B}{\tau^2} \|f - f_n\|^2_{\mathcal{C}^\alpha(\Omega)} + \frac{k_C}{\tau^2} \|f - f_n\|^2_{\mathcal{C}^\alpha(\Omega)} \\
&\leq \frac{k_{r_n}}{\tau^2} \|f - f_n\|^2_{\mathcal{C}^\alpha(\Omega)}
\end{aligned} \tag{2.24}$$

where k_{r_n} depends on $d, M_f, m_f, M_g, m_g, k_{\theta_n}, \alpha, \|f\|_{\mathcal{C}^\alpha(\Omega)}, \|g\|_{\mathcal{C}^{2,\alpha}(\Omega)}$ and also potentially on the Hölder norms of the first and second derivatives of Ψ_n. Next, using

(2.22) and (2.24);

$$\begin{aligned}\|f - f_{n+1}\|_{\mathcal{C}^\alpha(\Omega)} &= \|f - f_n - \frac{1}{\tau}(f - f_n) - r_n\|_{\mathcal{C}^\alpha(\Omega)} \\ &\leq \left(1 - \frac{1}{\tau}\right)\|f - f_n\|_{\mathcal{C}^\alpha(\Omega)} + \|r_n\|_{\mathcal{C}^\alpha(\Omega)} \\ &\leq \left(1 - \frac{1}{\tau}\right)\|f - f_n\|_{\mathcal{C}^\alpha(\Omega)} + \frac{k_{r_n}}{\tau^2}\|f - f_n\|^2_{\mathcal{C}^\alpha(\Omega)} \\ &\leq \|f - f_n\|_{\mathcal{C}^\alpha(\Omega)}\left(1 - \frac{1}{\tau} + \frac{k_{r_n}C_2}{\tau^2}\right).\end{aligned}$$

If $\tau \geq k_{r_n}C_2$, then $\dfrac{k_{r_n}C_2}{\tau} \leq 1$ and therefore

$$\begin{aligned}\|f - f_{n+1}\|_{\mathcal{C}^\alpha(\Omega)} &\leq \|f - f_n\|_{\mathcal{C}^\alpha(\Omega)}\left(1 - \frac{1}{\tau} + \frac{1}{\tau}\right) \\ &= \|f - f_n\|_{\mathcal{C}^\alpha(\Omega)} \\ &\leq C_2,\end{aligned}$$

which shows that bound **(3)** is preserved (for τ large enough). Actually, if we take τ a little bit bigger ($\tau \geq 2k_{r_n}C_2$), we get

$$\|f - f_{n+1}\|_{\mathcal{C}^\alpha(\Omega)} \leq \left(1 - \frac{1}{2\tau}\right)\|f - f_n\|_{\mathcal{C}^\alpha(\Omega)}. \tag{2.25}$$

This shows that we can take the stepsize τ such that the sequence of bounds K_4 created recursively will converge to a constant strictly greater than 0. Let's verify bound **(2)**. From the induction hypothesis, we know that $f/C_1 \leq f_n$ and it implies that

$$f - f_n \leq f - \frac{f}{C_1} = f\left(1 - \frac{1}{C_1}\right).$$

Using (2.24) and bound **(3)** at step n;

$$\begin{aligned}f - f_{n+1} &= \left(1 - \frac{1}{\tau}\right)(f - f_n) - r_n \\ &\leq \frac{\tau - 1}{\tau}(f - f_n) + \frac{k_{r_n}}{\tau^2}\|f - f_n\|^2_{\mathcal{C}^\alpha(\Omega)} \\ &\leq \frac{\tau - 1}{\tau}f\left(1 - \frac{1}{C_1}\right) + \frac{k_{r_n}C_2^2}{\tau^2}.\end{aligned}$$

By taking $\tau \geq \dfrac{k_{r_n} C_2^2}{m_f \left(1 - \frac{1}{C_1}\right)}$, or $\dfrac{1}{\tau} \leq \dfrac{m_f \left(1 - \frac{1}{C_1}\right)}{k_{r_n} C_2^2}$, we get:

$$\begin{aligned} f - f_{n+1} &\leq \frac{\tau - 1}{\tau} f \left(1 - \frac{1}{C_1}\right) + f \left(1 - \frac{1}{C_1}\right) \frac{1}{\tau} \\ &= f \left(1 - \frac{1}{C_1}\right) \end{aligned}$$

from which we deduce that $f/C_1 \leq f_{n+1}$. We follow a similar approach to obtain the other part of **(2)**. The induction hypothesis yields $f_n \leq C_1 f$ and then $f_n - f \leq (C_1 - 1)f$. Assuming now that $\tau \geq \dfrac{k_{r_n} C_2^2}{(C_1 - 1) m_f}$,

$$\begin{aligned} f_{n+1} - f &\leq \left(1 - \frac{1}{\tau}\right)(f_n - f) + r_n \\ &\leq \left(1 - \frac{1}{\tau}\right)(C_1 - 1)f + \frac{k_{r_n} C_2^2}{\tau} \\ &\leq \left(1 - \frac{1}{\tau}\right)(C_1 - 1)f + \frac{(C_1 - 1)f}{\tau} \\ &\leq (C_1 - 1)f. \end{aligned}$$

We can conclude that $f_{n+1} \leq C_1 f$ and hence **(2)** holds for $n + 1$. We now finish the proof of the first statement. From Appendix B, $\det\left(\mathcal{I} + D^2 u_{n+1}\right) > 0$ and therefore $\mathcal{I} + D^2 u_{n+1}$ is invertible. Let's prove that this inverse is uniformly positive definite:

$$\begin{aligned} \xi^T \left(\mathcal{I} + D^2 u_{n+1}\right)^{-1} \xi &= \left((\mathcal{I} + D^2 u_{n+1}) y\right)^T \left(\mathcal{I} + D^2 u_{n+1}\right)^{-1} \left((\mathcal{I} + D^2 u_{n+1}) y\right) \\ &\qquad \text{by taking } \xi = (\mathcal{I} + D^2 u_{n+1}) y \\ &= y^T \left(\mathcal{I} + D^2 u_{n+1}\right) y \\ &\geq K_5 |y|^2 \\ &= K_5 |(\mathcal{I} + D^2 u_{n+1})^{-1} \xi|^2 \end{aligned}$$

Select the induced matrix norm defined by

$$|A| = \max_{y \neq 0} \frac{|Ay|}{|y|}. \tag{2.26}$$

Using the norm inequality $|AB| \leq |A|\,|B|$ with the matrices A and B given respec-

tively by $\mathcal{I} + D^2 u_{n+1}$ and $(\mathcal{I} + D^2 u_{n+1})^{-1}\xi$, we obtain

$$|\xi| \leq |\mathcal{I} + D^2 u_{n+1}|\,|(\mathcal{I} + D^2 u_{n+1})^{-1}\xi|.$$

Next, motivated by the equivalence of norms, we play with the bounds we derived so far to get

$$\begin{aligned}
|\mathcal{I} + D^2 u_{n+1}| &\leq d \max_{i,j} \left\{ |(\mathcal{I} + D^2 u_{n+1})_{ij}| \right\} \\
&\leq d\Big(1 + \|u_n\|_{\mathcal{C}^{2,\alpha}(\Omega)} + \|\theta_n\|_{\mathcal{C}^{2,\alpha}(\Omega)}\Big) \\
&\leq K_6.
\end{aligned}$$

Then,

$$\begin{aligned}
\xi^T \left(\mathcal{I} + D^2 u_{n+1}\right)^{-1} \xi &\geq \frac{K_5}{K_6^2}|\xi|^2 \\
&= K_7|\xi|^2, \;\; K_7 > 0.
\end{aligned}$$

We can now use all this to show that L_{n+1} is a strictly elliptic operator:

$$\begin{aligned}
&\sum_{i,j=1}^{d} g(x + \nabla u_{n+1}) \mathrm{Adj}\left(\mathcal{I} + D^2 u_{n+1}\right)_{ij} \xi_i \xi_j \\
&\qquad = g(x + \nabla u_{n+1}) \det(\mathcal{I} + D^2 u_{n+1}) \sum_{i,j=1}^{d} \left(\mathcal{I} + D^2 u_{n+1}\right)^{-1}_{ij} \xi_i \xi_j \\
&\qquad \geq f_{n+1} K_7 \,|\xi|^2 \\
&\qquad \geq \frac{f}{C_1} K_7 \,|\xi|^2 \qquad\qquad \text{by bound } \mathbf{(2)} \\
&\qquad \geq K_8 \,|\xi|^2, \qquad\qquad K_8 > 0.
\end{aligned}$$

Note that by removing g from the previous development we get that $\mathrm{Adj}(\mathcal{I} + D^2 u_{n+1})$ is a uniformly positive definite matrix, which finalizes the proof of **(1)** for $n+1$. We conclude that by induction, **(1)** to **(3)** holds for all $n \in \mathbb{N}$.

We now prove the claim that the stepsize τ can be taken constant. Indeed, **(1)**

gives $\Psi_n \in \mathcal{C}^{2,\alpha}(\Omega)$ by construction while **(2), (3)** yield $f_n \in \mathcal{C}^{\alpha}(\Omega)$ and

$$0 < \frac{m_f}{C_1 M_g} \leq \frac{f_n(x)}{g(x + \nabla u_n)} \leq \frac{C_1 M_f}{m_g}.$$

Therefore, all the conditions for the estimate (2.3.2) are satisfied at every step. Recalling that $\Omega_r = \{x \in \Omega : \text{dist}(x, \partial\Omega) > r\}$, we get the following interior bound:

$$\|\Psi_n\|_{\mathcal{C}^{2,\alpha}(\Omega_r)} \leq K_9 \left[1 + \frac{\|h_n(x, \nabla\Psi_n)\|_{\mathcal{C}^{\alpha}(\Omega)}}{\alpha(1-\alpha)}\right]. \tag{2.27}$$

Since f_n and g are Hölder continuous with $m_g > 0$ on $[0,1]^d$, we can use the results from Lemma A.0.1 and get:

$$\begin{aligned}\|h_n\|_{\mathcal{C}^{\alpha}(\Omega)} &= \left\| \frac{f_n}{g(\nabla\Psi_n)} \right\|_{\mathcal{C}^{\alpha}(\Omega)} \\ &\leq \|f_n\|_{\mathcal{C}^{\alpha}(\Omega)} \left\| \frac{1}{g} \right\|_{\mathcal{C}^{\sqrt{\alpha}}(\Omega)} \left(1 + \|\nabla\Psi_n\|^{\sqrt{\alpha}}_{\mathcal{C}^{\sqrt{\alpha}}(\Omega)}\right)\end{aligned}$$

From Lemma 2.3.4, we observe that $\nabla\Psi_n$ is the transport map moving f_n to g. Thus, we can refer to Theorem 1.4.3 to deduce that $\nabla\Psi_n$ is invertible and $\nabla\Psi_n \in \Omega$, which in turn yields $\Psi_n \in [0, \sqrt{d}]$ when $\Omega = [0,1]^d$. At this point, the only remaining challenge is to bound the $\sqrt{\alpha}$ - Hölder coefficient of $\nabla\Psi_n$. It can be achieved through the second interior estimate (2.19). We see that the maximum terms $M(\Psi_n, y, a)$ are going to be uniformly bounded and that the only problem could come from the minimum terms $m(\psi^*_{n_y}, 0, a)$, $a = 1$ or $R_y \geq 1$.

Let's analyze $\psi^*_{n_y}$. We have $\psi_{n_y}(0) = 0$ and $\nabla\psi_{n_y}(0) = 0$. Using Fenchel's inequality (presented for example in [16]), we get $0 \leq \psi^*_{n_y}(z)$ for all $z \in \Omega$. Moreover, since Ψ_n is uniformly convex, $\psi_{n_y}(x) \geq K|x+y|^2$ and $\psi^*_{n_y}(z) \leq [|z|^2 - z \cdot y]/4K$. We also know from the theory of convex conjugates [16] that the gradient of ψ_{n_y} and $\psi^*_{n_y}$ are inverses of each other. This yields $\psi^*_{n_y}(0) = 0$, $\nabla\psi^*_{n_y}(0) = 0$ and $m(\psi^*_{n_y}, 0, a) = \min \psi^*_{n_y}(z)$ where the minimum is taken on the sphere $|z| = a$, $a \geq 1$. Note that with the periodicity, we can increase the size of Ω to include this sphere inside it and still have a uniform bound on Ψ_n and $\nabla\Psi_n$. Furthermore, due to the uniform convexity of ψ_{n_y}, $\nabla\psi_{n_y}$ is strictly monotone increasing, i.e. $(\nabla\psi_{n_y}(z_1) - \nabla\psi_{n_y}(z_2)) \cdot (z_1 - z_2) > 0$ for all $z_1, z_2 \in \Omega$. As $\nabla\psi^*_{n_y} = \nabla\psi^{-1}_{n_y}$, this property also applies to $\nabla\psi^*_{n_y}$.

We see that the only possible breakdown happens when $\nabla\psi^*_{n_y}$ converges to a function which is zero up to $|z| = a$. This happens when $|\nabla\psi_{n_y}| = |\nabla\Psi_n(x+y) -$

$\nabla\Psi_n(y)| \to \infty$ as $|x| \to 0$ and $n \to \infty$, for any y. Observe now that if we increase the regularity assumptions imposed on the densities to $f \in \mathcal{C}^{2,\alpha}(\Omega)$, $g \in \mathcal{C}^{3,\alpha}(\Omega)$, we get $f_n \in \mathcal{C}^{2,\alpha}(\Omega)$ at every step. This tells us that $\theta_n \in \mathcal{C}^{4,\alpha}(\Omega)$ (see [9]) and thus $\Psi_n \in \mathcal{C}^{4,\alpha}(\Omega)$. Therefore, we can apply the estimate presented in Theorem (2.3.6) and rule out this potential breakdown case. Indeed, by selecting the domain in a way that its boundary avoids this spike in the second derivative, we get that this situation cannot occur within our context.

We conclude that the $\mathcal{C}^{2,\alpha}(\Omega_r)$ norm of Ψ_n is uniformly bounded. From this, we have

$$\begin{aligned}\|u_n\|_{\mathcal{C}^{2,\alpha}(\Omega_r)} &= \left\|\Psi_n - \frac{|x|^2}{2}\right\|_{\mathcal{C}^{2,\alpha}(\Omega_r)} \\ &\leq \|\Psi_n\|_{\mathcal{C}^{2,\alpha}(\Omega_r)} + \left\|\frac{|x|^2}{2}\right\|_{\mathcal{C}^{2,\alpha}(\Omega_r)} \leq K_{10},\end{aligned}$$

so the same conclusion holds for the $\mathcal{C}^{2,\alpha}(\Omega_r)$ norm of u_n. By Theorem 1.4.2, we know that we can extend this result to all of Ω for u_n and then Ψ_n. Hence, we deduce that k_{r_n} and k_{θ_n} are also uniformly bounded, which tells us that we can select a $\tau \geq 1$ constant such that the three statements hold for all $n \in \mathbb{N}$. Moreover, the sequence $(u_n)_{n\in\mathbb{N}}$ is uniformly bounded in $\mathcal{C}^{2,\alpha}(\Omega)$, thus equicontinuous. By the Ascoli-Arzela theorem, it converges uniformly in $\mathcal{C}^{2,\beta}(\Omega)$ for $0 < \beta < \alpha$ to the solution u of (2.1), which is unique since we impose $\int_\Omega u\,dx = 0$. Finally, due to the fact that the initial and final densities are actually $\mathcal{C}^{2,\alpha}(\Omega)$, we know that this solution will be in $\mathcal{C}^{4,\beta}(\Omega)$. □

2.5 Remarks on the Proof

This proof by induction provides a lot of precious information concerning the properties of the iterates created by our method. First, we saw that at every step, the computed map is the solution to the transport problem sending f_n to g since $\int_\Omega f_n\,dx = 1\ \forall n$. Indeed, the algorithm does not only create a sequence of maps, it actually builds a sequence of transport problems which get closer and closer to the target problem as n grows. Next, remember that the Monge-Ampère equation we are solving is elliptic only when we restrict it to the space of convex functions. Knowing this, we realize that our algorithm is being extra careful by approximating the convex solution of the Monge-Ampère equation by a sequence of uniformly convex functions.

In addition, this guarantees the linearized equation to be strictly elliptic and thus have a unique solution (once we fix the constant). This way, we make sure that there is always only one descent direction.

When it comes to the stepsize parameter τ, it would be very useful to know a priori which value to select in order to make the algorithm converge. Such an estimate is unfortunately hard to acquire since some of the constants used through interior bounds are obtained via rather indirect arguments. However, we observe from lower bounds on τ used in the proof like

$$\tau \geq \frac{k_{r_n} C_2^2}{m_f \left(1 - \frac{1}{C_1}\right)}, \ \tau \geq \frac{k_{r_n} C_2^2}{(C_1 - 1) m_f} \text{ or } \tau \geq k_{r_n} C_2$$

where

$$C_1 = \max\left(\frac{M_f}{m_g}, \frac{M_g}{m_f}\right) \text{ and } C_2 = \|f\|_{\mathcal{C}^\alpha(\Omega)} + \|g\|_{\mathcal{C}^\alpha(\Omega)}$$

that the minimum value required on the stepsize parameter to achieve convergence could potentially be large when m_f is close to 0 or when either $\|f\|_{\mathcal{C}^\alpha(\Omega)}$ or $\|g\|_{\mathcal{C}^\alpha(\Omega)}$ is big. Through the numerous numerical experiments we conducted, we realized that τ seems to behave according to both conditions. Therefore, knowing a priori that f get close to 0 or that one density varies a lot, we can react accordingly by either increasing the value of the stepsize or by modifying the representation of the densities (which is possible in some applications). More details can be found in subsequent sections. We also obtained estimates on the speed of convergence of the method in (2.25). Indeed, if τ is larger than $2k_{r_n}C_2$, (f_n) converges to f following a geometric convergence with a rate of at least $1 - 1/2\tau$.

On top of that, we could potentially improve the performance of our algorithm by playing a little bit more with τ. Even though we take τ to be constant, it is actually possible to change its value at every iteration, creating a sequence of τ_n. By doing so, we could start with smaller stepsize values and gradually take bigger steps as we get closer to the solution. Finally, even if our proof only guarantees convergence when the update θ_n is the solution of (2.6), in practice we could hope to sometimes get good results by replacing it by the solution of one of the following:

1) $\Delta\theta_n = \dfrac{f - f_n}{\tau}$

2) $\operatorname{Tr}\left(\operatorname{Adj}(\mathcal{I} + D^2 u_n) D^2 \theta_n\right) = \dfrac{f - f_n}{\tau}$

3) $g(x + \nabla u_n)\,\mathrm{Tr}\,(\mathrm{Adj}(\mathcal{I} + D^2 u_n) D^2 \theta_n) = \dfrac{f - f_n}{\tau}$

We saw in our numerical experiments that the third one gives results that are very often as good as the full formula, which tells us that the second derivatives have a greater influence on the direction taken by the Newton algorithm than the first derivatives. However, the algorithm never converged when we used the first one and converged in some cases only for the second one. Again, for more details, one should consult Chapter 3.

Chapter 3

Numerical Discretization

3.1 A Finite Differences Implementation

We begin by presenting a two-dimensional implementation of the Newton algorithm (2.6) through finite differences. As a compromise between accuracy and complexity of the code itself, we use finite differences schemes of fourth order for the gradient and the hessian of u_n and of second order for the gradient and hessian of θ_n. These second order accurate schemes are easier to code for the case of θ_n and since this update comes from the linearization of the Monge-Ampère equation which is already an approximation, it does not need to be as precisely determined in the algorithm. We consider a uniform $N \times N$ grid with a spacestep $h = 1/N$ where we identify $x_i = 0$ with $x_i = 1$ $(i = 1, 2)$ for the periodicity. For $0 \leq i, j \leq N$ and for a function $\zeta : \Omega \to \mathbb{R}$, let's denote $\zeta(ih, jh)$ by $\zeta_{i,j}$. Then, we have the following formulas for all these derivatives, replacing u_n by u and θ_n by θ everywhere for clarity purposes (we also present the associated stencils for a better visual understanding):

$$\begin{aligned}
\theta_x(ih, jh) &= \frac{\theta_{i+1,j} - \theta_{i-1,j}}{2h} + \mathcal{O}(h^2), \\
\theta_y(ih, jh) &= \frac{\theta_{i,j+1} - \theta_{i,j-1}}{2h} + \mathcal{O}(h^2), \\
\theta_{xx}(ih, jh) &= \frac{\theta_{i+1,j} - 2\theta_{i,j} + \theta_{i-1,j}}{h^2} + \mathcal{O}(h^2), \\
\theta_{yy}(ih, jh) &= \frac{\theta_{i,j+1} - 2\theta_{i,j} + \theta_{i,j-1}}{h^2} + \mathcal{O}(h^2), \\
\theta_{xy}(ih, jh) &= \frac{\theta_{i+1,j+1} - \theta_{i-1,j+1} - \theta_{i+1,j-1} + \theta_{i-1,j-1}}{4h^2} + \mathcal{O}(h^2),
\end{aligned}$$

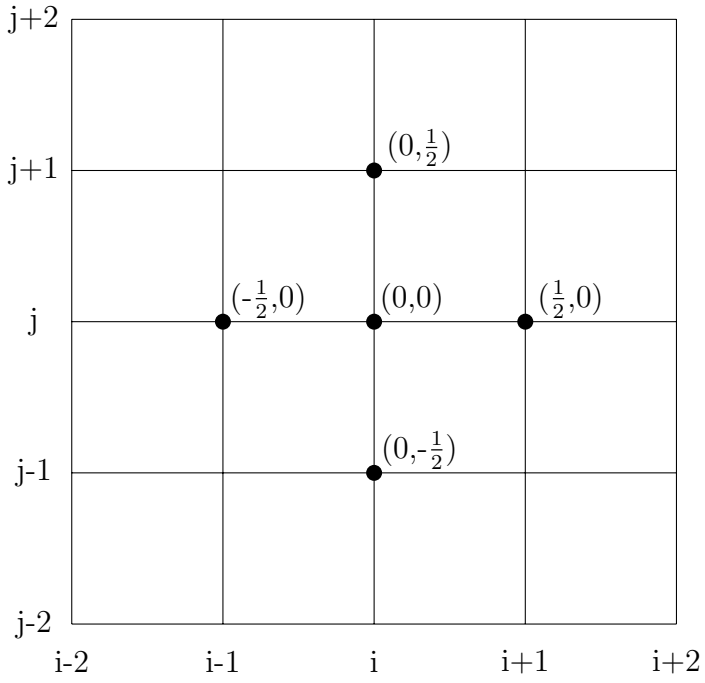


Figure 3.1: Stencil for $h(\theta_x, \theta_y)$

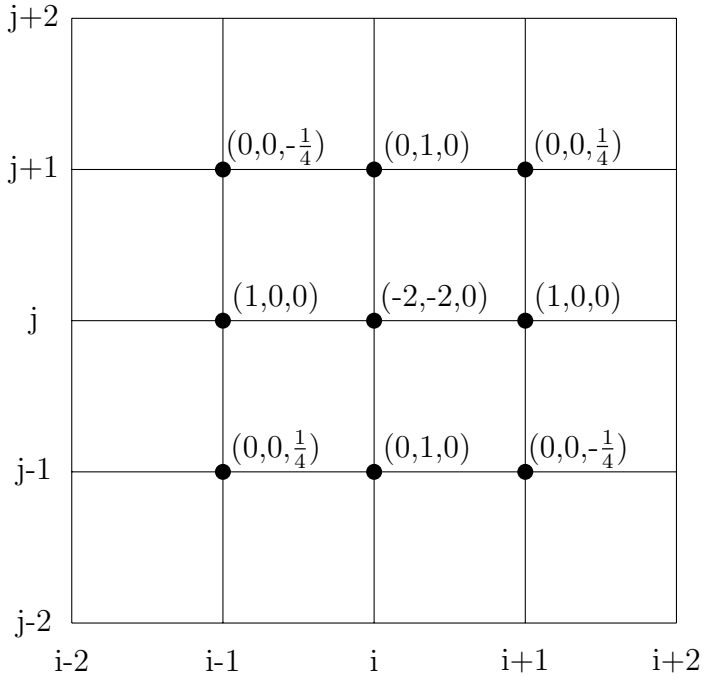


Figure 3.2: Stencil for $h^2(\theta_{xx}, \theta_{yy}, \theta_{xy})$

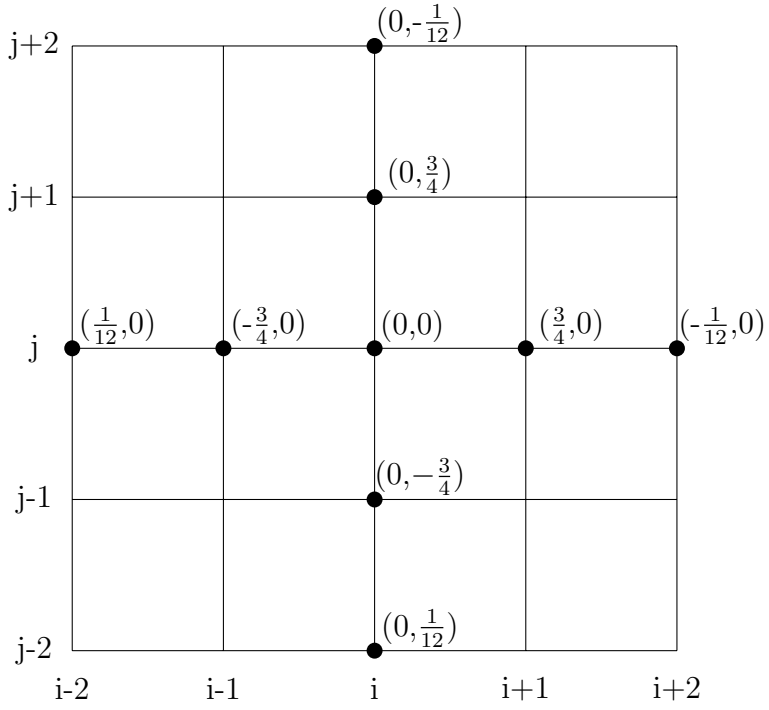


Figure 3.3: Stencil for $h(u_x, u_y)$

$$
\begin{aligned}
u_x(ih, jh) &= \frac{-u_{i+2,j} + 8u_{i+1,j} - 8u_{i-1,j} + u_{i-2,j}}{12h} + \mathcal{O}(h^4),\\
u_y(ih, jh) &= \frac{-u_{i,j+2} + 8u_{i,j+1} - 8u_{i,j-1} + u_{i,j-2}}{12h} + \mathcal{O}(h^4),\\
u_{xx}(ih, jh) &= \frac{-u_{i+2,j} + 16u_{i+1,j} - 30u_{i,j} + 16u_{i-1,j} - u_{i-2,j}}{12h^2} + \mathcal{O}(h^4),\\
u_{yy}(ih, jh) &= \frac{-u_{i,j+2} + 16u_{i,j+1} - 30u_{i,j} + 16u_{i,j-1} - u_{i,j-2}}{12h^2} + \mathcal{O}(h^4),\\
u_{xy}(ih, jh) &= \frac{u_{i+2,j+2} - 8u_{i+1,j+2} + 8u_{i-1,j+2} - u_{i-2,j+2}}{144h^2}\\
&\quad + \frac{-8u_{i+2,j+1} + 64u_{i+1,j+1} - 64u_{i-1,j+1} + 8u_{i-2,j+1}}{144h^2}\\
&\quad + \frac{8u_{i+2,j-1} - 64u_{i+1,j-1} + 64u_{i-1,j-1} - 8u_{i-2,j-1}}{144h^2}\\
&\quad + \frac{-u_{i+2,j-2} + 8u_{i+1,j-2} - 8u_{i-1,j-2} + u_{i-2,j-2}}{144h^2} + \mathcal{O}(h^4).
\end{aligned}
$$

Since the linearized Monge-Ampère equation has a unique solution only up to a constant, the linear system corresponding to its discretization has one free parameter.

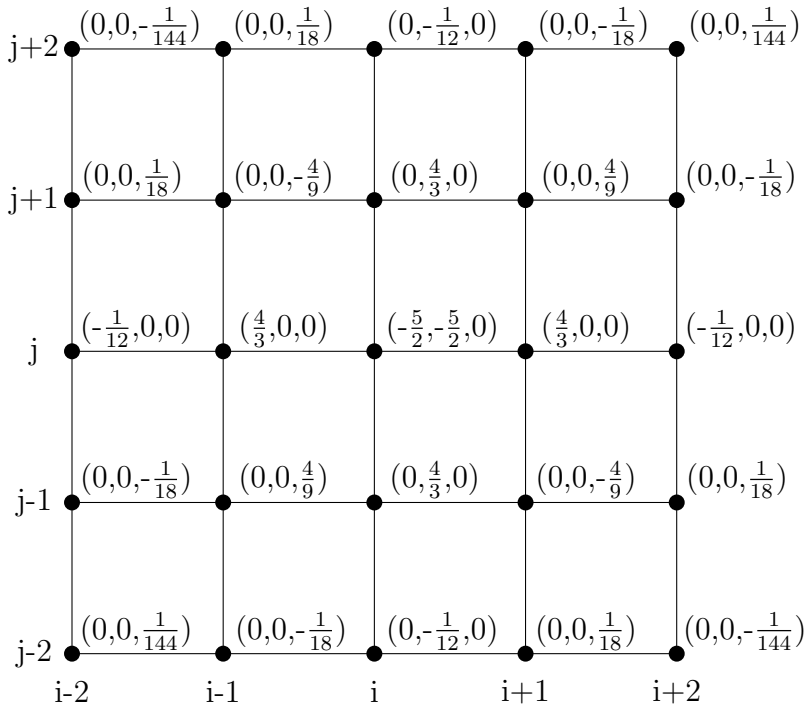


Figure 3.4: Stencil for $h^2(u_{xx}, u_{yy}, u_{xy})$

In order to fix the value of one variable to a known value through this parameter, we need to reduce the matrix to an invertible one. However, in an attempt to avoid having to row-reduce it (which would not be efficient) we use the strategy stated below.

Let $P = N^d$ be the total number of points in our grid; d being equal to 2 in this case. Starting with $Ax = b$ (where $\text{rank}(A) = P - 1$), we obtain the reduced system $\tilde{A}\tilde{x} = \tilde{b}$ by adding the first line of A and b to all their other lines and then by removing the first line of A and b and the first column of A (which corresponds to fixing the free parameter, $\theta(0)$ in this case, to 0):

$$A = \begin{pmatrix} a_{1,1} & a_{1,2} & & a_{1,P-1} & a_{1,P} \\ a_{2,1} & a_{2,2} & \dots & a_{2,P-1} & a_{2,P} \\ a_{3,1} & a_{3,2} & & a_{3,P-1} & a_{3,P} \\ & \vdots & & \ddots & \\ a_{P-1,1} & a_{P-1,2} & & a_{P-1,P-1} & a_{P-1,P} \\ a_{P,1} & a_{P,2} & \dots & a_{P,P-1} & a_{P,P} \end{pmatrix}, \qquad b = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{P-1} \\ b_P \end{pmatrix}$$

turns into

$$\tilde{A} = \begin{pmatrix} a_{2,1} + a_{1,1} & \dots & a_{2,P-1} + a_{1,P-1} \\ a_{3,1} + a_{1,1} & & a_{3,P-1} + a_{1,P-1} \\ \vdots & \ddots & \\ a_{P-1,1} + a_{1,1} & & a_{P-1,P-1} + a_{1,P-1} \\ a_{P,1} + a_{1,1} & \dots & a_{P,P-1} + a_{1,P-1} \end{pmatrix}, \qquad \tilde{b} = \begin{pmatrix} b_2 + b_1 \\ b_3 + b_1 \\ \vdots \\ b_{P-1} + b_1 \\ b_P + b_1 \end{pmatrix}.$$

Then we can solve $\tilde{A}\tilde{x} = \tilde{b}$ and subsequently form x out of $\tilde{x}$, taking $x_1 = 0$. The next lemma shows under which conditions this strategy will produce a valid answer to the system $Ax = b$.

Lemma 3.1.1. *Take $A, x, b, \tilde{A}, \tilde{x}$ and $\tilde{b}$ as defined previously. Since A has rank $P - 1 = N^2 - 1$, we know there exists real numbers $\alpha_1, \alpha_2, \dots, \alpha_P$ not all zero such that $\alpha_1 L_1 + \alpha_2 L_2 + \dots + \alpha_P L_P = 0$ where L_i is the ith line of A. If $\alpha_2 + \dots + \alpha_P \neq \alpha_1$, then $\tilde{A}$ has rank $P - 1$.*

Proof. For simplicity in exposition, we will prove this result for $P = N^2 = 4$, the arguments being the same for greater values. Let's consider two cases:
1) $\alpha_1 \neq 0$
Here we have $L_1 = -\left(\alpha_2 L_2 + \alpha_3 L_3 + \alpha_4 L_4\right)/\alpha_1$. Assume that

$$\gamma_1(L_2 + L_1) + \gamma_2(L_3 + L_1) + \gamma_3(L_4 + L_1) = 0.$$

From this, we get

$$\gamma_1 L_2 + \gamma_2 L_3 + \gamma_3 L_4 + (\gamma_1 + \gamma_2 + \gamma_3) L_1 = 0$$
$$\Rightarrow \quad \gamma_1 L_2 + \gamma_2 L_3 + \gamma_3 L_4 + (\gamma_1 + \gamma_2 + \gamma_3) \left(\frac{-1}{\alpha_1}\right) (\alpha_2 L_2 + \alpha_3 L_3 + \alpha_4 L_4) = 0$$

and then

$$\begin{cases} \gamma_1 - \dfrac{\alpha_2}{\alpha_1}(\gamma_1 + \gamma_2 + \gamma_3) = 0 \\ \gamma_2 - \dfrac{\alpha_3}{\alpha_1}(\gamma_1 + \gamma_2 + \gamma_3) = 0 \\ \gamma_3 - \dfrac{\alpha_4}{\alpha_1}(\gamma_1 + \gamma_2 + \gamma_3) = 0 \end{cases}$$

because $\text{rank}(A) = P - 1$. Multiplying the previous three equations by α_1, we form

$$\begin{pmatrix} \alpha_2 - \alpha_1 & \alpha_2 & \alpha_2 \\ \alpha_3 & \alpha_3 - \alpha_1 & \alpha_3 \\ \alpha_4 & \alpha_4 & \alpha_4 - \alpha_1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

which is equivalent to (by summing the three lines together and dividing):

$$\begin{pmatrix} \alpha_2 - \alpha_1 & \alpha_2 & \alpha_2 \\ \alpha_3 & \alpha_3 - \alpha_1 & \alpha_3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

since $\alpha_2 + \alpha_3 + \alpha_4 \neq \alpha_1$. From this point, no matter if α_2, α_3 and α_4 are equal to 0 or not, we can always reduce this matrix to the identity. This tells us that $\gamma_1 = \gamma_2 = \gamma_3 = 0$; hence the lines $L_2 + L_1$, $L_3 + L_1$ and $L_4 + L_1$ are linearly independent which implies that $\text{rank}(\tilde{A}) = P - 1$.

2) $\alpha_1 = 0$

In that case, $\alpha_2 L_2 + \alpha_3 L_3 + \alpha_4 L_4 = 0$ and there is at least one alpha different from 0. Therefore, we can assume without loss of generality that $\alpha_4 \neq 0$ and write $L_4 = -(\alpha_2 L_2 + \alpha_3 L_3)/\alpha_4$. Taking again

$$\gamma_1(L_2 + L_1) + \gamma_2(L_3 + L_1) + \gamma_3(L_4 + L_1) = 0,$$

and following the same procedure as in **1**, we obtain the system

$$\begin{pmatrix} \alpha_4 & 0 & -\alpha_2 \\ 0 & \alpha_4 & -\alpha_3 \\ 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \gamma_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

Using the fact that both $\alpha_4 \neq 0$ and $\alpha_2 + \alpha_3 + \alpha_4 \neq 0$, we can also reduce this matrix to the identity and reach the same conclusion as in **1**. □

This lemma does not hold if condition $\alpha_2 + \cdots + \alpha_P \neq \alpha_1$ is not satisfied. Take for example a matrix A that has its second line equal to the negative of the first line and that has all its other lines linearly independent. Then $\widetilde{A}$ has rank $P-2$. Nonetheless, due to the shape of our matrix (which looks like a block diagonal matrix with extra blocks appearing because of the periodic boundary conditions) and to the fact that $\alpha_2 + \cdots + \alpha_P$ is not likely to be exactly equal to α_1 numerically, it was never violated

in our numerical experiments.

For robustness matters, we could still want to make sure this condition is satisfied by A. A possible strategy to achieve that would be to verify the rank of $\tilde{A}$. If it's not $P-1$, we could multiply the lines of A (and the corresponding element in b) one by one by a constant different than one until $\tilde{A}$ has the right rank. Again, due to the (nearly) block diagonal structure of the matrix, we would not have to do this on many lines before reaching the required rank.

Then, to actually solve the system $\tilde{A}x = b$, we employ the Biconjugate Gradient (BICG) method just like Loeper and Rapetti did in their paper [13]. This choice can be justified by the fact that we are dealing with a sparse matrix which is not symmetric nor positive definite, the BICG procedure being specifically designed to deal with these conditions. In fact, A is a very sparse matrix; it contains at most 9 non-zero elements per line and hence $\tilde{A}$ has at most 18. Having in mind that A is $N^2 \times N^2$, we see that it gets much sparser as we augment the grid size.

The BICG algorithm produces a pair of biorthogonal bases for the Krylov subspaces

$$\begin{aligned} & \mathcal{K}_m = \text{span}\{r_0, Ar_0, A^2r_0, ..., A^{m-1}r_0\} \\ \text{and} \qquad & \mathcal{L}_m = \text{span}\{r_0, A^Tr_0, (A^T)^2r_0, ..., (A^T)^{m-1}r_0\} \end{aligned}$$

where r_0 is the normalization of $b - Ax_0$. It obtains an approximate solution of both $Ax = b$ and $A^{\mathrm{T}}x = b$, respectively as a projection onto $\mathcal{K}_m$ and $\mathcal{L}_m$, in a way that the dimension of the subspaces augments at every iteration. Moreover, these approximate solutions converge to the solution of the linear system (with P variables) in a maximum of P steps in theory (i.e. in exact arithmetic), provided the method does not break down before. In practice, roundoff errors could possibly slow down the convergence in some cases. However, due to the nice properties of our matrix here, converge is reached in fewer than P steps.

BICG Algorithm

$$
\begin{cases}
\textit{Compute } r_0 := b - Ax_0. \\
\textit{Set } p_0 := r_0,\ p_0^* := r_0. \\
\textit{For } j = 0, 1, ... \textit{ until convergence, Do} \\
\quad \alpha_j := (r_j \cdot r_j^*)/(Ap_j \cdot p_j^*) \\
\quad x_{j+1} := x_j + \alpha_j p_j \\
\quad r_{j+1} := r_j - \alpha_j A p_j \\
\quad r_{j+1}^* := r_j^* - \alpha_j A^T p_j^* \\
\quad \beta_j := (r_{j+1} \cdot r_{j+1}^*)/(r_j \cdot r_j^*) \\
\quad p_{j+1} := r_{j+1} + \beta_j p_j \\
\quad p_{j+1}^* := r_{j+1}^* + \beta_j p_j^* \\
\textit{End Do}
\end{cases}
$$

Observe that this algorithm takes advantages of the sparsity by only referencing the matrix through the products of A and A^T with the respective vectors. Note also that we can add a preconditioner M to the procedure if necessary by multiplying at every iteration r_j to the left by M^{-1}, r_j^* to the right by M^{-1} and for the products $r_j \cdot r_j^*$, by inserting M^{-1} in between the two vectors. For more information on the BICG algorithm or other procedures to solve a sparse linear system of equations, one might take a look at the excellent book written by Yousef Saad [42].

We say a few words now about the theoretical computational complexity of the Newton algorithm. Apart from the step where we solve the linearized Monge-Ampère equation, every other step can be done in $\mathcal{O}(P)$ operations. For the resolution of the linear partial differential equation, taking advantage of the sparsity like previously mentioned, we need $\mathcal{O}(P)$ operations per iteration in the BICG algorithm. In addition, this algorithm converges in at most P steps and therefore the computational complexity is at worst $\mathcal{O}(nP^2)$, where n is the number of Newton iterations. However, as we shall see in the next section, in practice it can be much smaller than that. Since we know that the speed of convergence of the BICG algorithm dictates the global complexity of the Newton algorithm, measuring the number of BICG iterations at every step will give us this order of convergence.

Finally, another point we should mention is that even though f_n has a total mass of 1 at every step in theory, it is not necessarily the case in the numerical experiments, due to the discretization errors. However, we need the right-hand side

of the linearized Monge-Ampère equation to integrate to 0 on the whole domain (see section 2.2). Therefore, we introduce a normalization step right after computing f_n in the implemented algorithm, taking

$$\tilde{f_n} = f_n - \frac{1}{N^2} \sum_{i,j=0}^{N-1} f_{n_{i,j}} + 1$$

instead of f_n.

3.2 A Fourier Transform Implementation

The first implementation we employed to solve the linearized Monge-Ampère equation was motivated by Loeper and Rapetti's choice, but there might be better methods for doing this. Indeed, there exist much cheaper ways to solve a linear second-order strictly elliptic equation with such boundary conditions. The one we are going to explore here is due to Strain [21] and in practice requires $\mathcal{O}(P\text{Log}P)$ operations through the use of Fourier series (and hence through the Fast Fourier Transform algorithm, or FFT algorithm). Let's start by considering the equation

$$L\theta(x) = \sum_{i,\,j=1}^{d} a_{ij}(x)\partial_i\partial_j\theta(x) + \sum_{i=1}^{d} b_i(x)\partial_i\theta(x) + c(x)\theta(x) = h(x)$$

where the coefficients of the operator L are $\mathcal{C}^\alpha(\Omega)$ smooth, 1-periodic and where the a_{ij} satisfy the uniform ellipticity condition given in Definition 2.2.1. L is invertible once we restrict c to be less than or equal to 0 (provided the solution has mean zero, see Section 2.2), and Strain's technique works only for such c. As we are dealing here with an operator having $c = 0$, we will present his method for this particular case only (the general one does not differ much, see the original paper for more details). Let $\overline{L}$ be the operator

$$\overline{L} = \sum_{i,\,j=1}^{d} \overline{a}_{ij}(x)\partial_i\partial_j + \sum_{i=1}^{d} \overline{b}_i(x)\partial_i$$

with constant coefficients given by the averages $\overline{a}_{ij} = \int_\Omega a_{ij}\,dx$ and $\overline{b}_i = \int_\Omega b_i\,dx$. Then, we introduce the new unknown σ and rewrite the problem as a system:

$$\begin{cases} L\overline{L}^{-1}\sigma(x) = h(x) \\ \qquad \overline{L}\theta(x) = \sigma(x). \end{cases} \tag{3.1}$$

Next, we expand σ in Fourier Series by taking

$$\sigma(x) = \sum_{\substack{k\in\mathbb{Z}^d \\ k\neq 0}} \widehat{\sigma}(k) e^{2\pi\iota k\cdot x}$$

$$\text{and} \qquad \widehat{\sigma}(k) = \int_\Omega \sigma(x) e^{-2\pi\iota k\cdot x}\,dx;$$

ι representing $\sqrt{-1}$ and $\widehat{\sigma}(k)$ being the usual Fourier coefficients. From this expansion, we get

$$\overline{L}\sigma(x) = \sum_{\substack{k\in\mathbb{Z}^d \\ k\neq 0}} \left(\sum_{i,j=1}^{d} \overline{a}_{ij} 2\pi\iota k_i 2\pi\iota k_j + \sum_{i=1}^{d} \overline{b}_i 2\pi\iota k_i \right) \widehat{\sigma}(k) e^{2\pi\iota k\cdot x}$$

which tells us that as long as the sum inside the brackets is not 0, the coefficients of the inverse operator $\overline{L}^{-1}$ are the multiplicative inverses of this sum. However, if it is equal to 0, the corresponding coefficient of the inverse is simply set to 0 (in our case, this should not happen as the the densities are positive and the operator is strictly elliptic). This yields the following formula:

$$L\overline{L}^{-1}\sigma(x) = \sum_{\substack{k\in\mathbb{Z}^d \\ k\neq 0}} \left(\sum_{i,j=1}^{d} a_{ij}(x) 2\pi\iota k_i 2\pi\iota k_j + \sum_{i=1}^{d} b_i(x) 2\pi\iota k_i \right) \overline{\rho}(k)\widehat{\sigma}(k) e^{2\pi\iota k\cdot x}.$$

where

$$\overline{\rho}(k) = \begin{cases} 1 \div \left(\displaystyle\sum_{i,j=1}^{d} \overline{a}_{ij} 2\pi\iota k_i 2\pi\iota k_j + \sum_{i=1}^{d} \overline{b}_i 2\pi\iota k_i \right) & \text{if the sum is not } 0 \\ 0 & \text{otherwise.} \end{cases}$$

Finally, we swap the sums on i and j and the one on k to be able to express the sums on k as functions of x through Fourier transforms:

$$\begin{aligned} L\overline{L}^{-1}\sigma(x) &= \sum_{i,j=1}^{d} a_{ij}(x) \sum_{\substack{k\in\mathbb{Z}^d \\ k\neq 0}} 2\pi\iota k_i 2\pi\iota k_j \overline{\rho}(k)\widehat{\sigma}(k) e^{2\pi\iota k\cdot x} \\ &\qquad + \sum_{i=1}^{d} b_i(x) \sum_{\substack{k\in\mathbb{Z}^d \\ k\neq 0}} 2\pi\iota k_i \overline{\rho}(k)\widehat{\sigma}(k) e^{2\pi\iota k\cdot x} \\ &= \sum_{i,j=1}^{d} a_{ij}(x)\alpha_{ij}(x) + \sum_{i=1}^{d} b_i(x)\beta_i(x) \end{aligned}$$

For the discretized problem, knowing the value of σ, we can compute $\widehat{\sigma}$ with one application of the FFT algorithm and then compute α_{ij} and β_i with $d(d+1)$ applications of the inverse FFT algorithm to be able to get the value of $L\overline{L}^{-1}\sigma(x)$ in $\mathcal{O}(P\text{Log}P)$ operations. Therefore, we can use an iterative method to solve the first part of system (3.1) at a cost of $\mathcal{O}(P\text{Log}P)$ operations per iteration. The one Strain chose is the Generalized Minimal Residual method, or GMRES. Just like BICG, it is an efficient way of solving a linear system of equations where the matrix is non-symmetric and non-positive definite. It consists of two phases; first, it generates an orthornormal basis set for the Krylov subspace

$$\mathcal{K}_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2 r_0 ..., A^{m-1} r_0\}$$

where $r_0 = b - Ax_0$ and $1 \leq m \leq \kappa$. The parameter κ is chosen at the beginning and dictates the dimension of the biggest Krylov subspace we will project the solution on. Then it forms a linear combination out of these basis vectors by solving a least squares minimization problem to get the coefficients of the projection of the solution on that subspace.

GMRES(κ) Algorithm

Given x_0, *compute* $r_0 := b - Ax_0, \xi := |r_0|$ *and* $v_1 := r_0/\xi$.
For $j = 1, 2, \ldots, \kappa$ *Do*
  Compute $w_j := Av_j$
  For $i = 1, \ldots, j$ *Do*
    $h_{ij} := w_j \cdot v_i$
    $w_j := w_j - h_{ij}v_i$
  End Do
  $h_{j+1,j} = |w_j|$
  If $h_{j+1,j} = 0$
    Set $m := j$ *and exit loop.*
  Else
    $v_{j+1} = w_j/h_{j+1,j}$
  End If
End Do
Compute y_m *the minimizer of* $|\xi e_1 - H_m y|$ *and* $x_m = x_0 + V_m y_m$.
If the stopping criterion is satisfied
  Stop.
Else
  Set $x_0 := x_m$ *and start over*
End If

Here, V_m denotes the matrix formed with column vectors $v_1, \ldots, v_m$ and H_m is a $(m+1) \times m$ matrix of the Hessenberg type (almost triangular) formed by the h_{ij}, its other entries being 0. In the second phase of the algorithm, the usual resolution process is to turn H_m into an upper triangular matrix by applying m Givens rotations, making the system easily solvable. The resulting matrix will have an extra line of zeros that needs to be removed (it is an $(m+1) \times m$ matrix) and the corresponding element in the right-hand side vector will be the residual, which can be used as a stopping criterion. Again, more details and variants of the GMRES algorithm can be found in Saad's book [42].

We remark that, unlike the BICG procedure, GMRES does not use projections on the Krylov subspace generated by the transposed matrix. This makes it easier to code for the particular setting we are dealing with since we do not form A directly; we

reference it instead through the result of its product with a given vector σ. Moreover, Strain proved in his work (and confirmed in his numerical experiments) that the number of GMRES iterations required did not vary with N, which yielded a global complexity of $\mathcal{O}(P\text{Log}P)$. For these two reasons, we will also use GMRES as an iterative method for the resolution of the first part of system (3.1).

After computing $\sigma(x)$, we need to solve $\overline{L}\theta(x) = \sigma(x)$. This can be easily achieved since we already know how to compute the inverse of $\overline{L}$. Therefore, we get

$$\theta(x) = \sum_{\substack{k\in\mathbb{Z}^d \\ k\neq 0}} \overline{\rho}(k)\widehat{\sigma}(k)e^{2\pi\iota k\cdot x}$$

and we only require another application of the (inverse) FFT algorithm to obtain θ. Strain's method is generally very effective. We shall compare its actual efficiency to the efficiency of the other implementation in several numerical examples in subsequent sections. One should note that another of its advantages is its spectral accuracy; i.e. the error decreases faster than any power of the grid size as the spacestep size goes to 0.

Before we move on to the numerical experiments, we need to mention a few more things. Let's remember first that the linear system of equations has a unique solution up to a constant. In order to fix it and thus get an invertible matrix, we can employ a similar strategy than in the first implementation. Consider the discrete equivalent to forcing the value of the integral of σ to be 0:

$$\sum_{i,j=0}^{N} \sigma(ih, jh) = 0.$$

We can take this extra equation and add it to all the other equations just like we did in the previous section (Lemma 3.1.1 still provides a justification to this approach). Next, to compute the averages of the operator's coefficients $\overline{a_{ij}}$ and $\overline{b_i}$, we will use Simpson's numerical integration formula. Finally, let's observe that so far in this second implementation we only changed the way we solve the linearized Monge-Ampère equation. We used again fourth-order accurate finite differences for the discretization of the derivatives of u_n, and everything else is done in the same way.

3.3 Numerical Tests

We present in this section two detailed numerical tests. In order to observe the behavior of the algorithm and to compare the effect of its different parameters with the different implementations, we chose functions for which we knew the analytical expression (and hence could compute the exact expression of the gradient of the final density). However, we do not always possess such analytical expression, we sometimes know the value of g only on the grid points. In these situations, we need to approximate the values of $g(x + \nabla u_n)$ and $\nabla g(x + \nabla u_n)$. We will deal with this case through applications in subsequent sections. For the sake of clarity, we shall refer to the finite differences implementation as the FD implementation and to the Fourier transform implementation as the FT implementation.

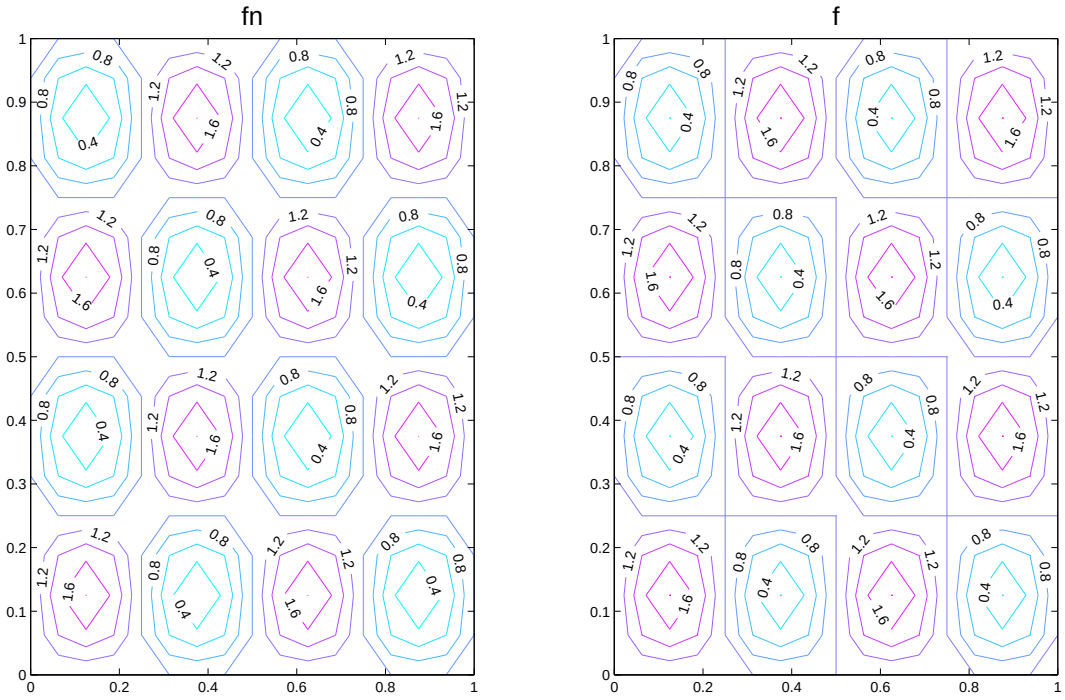


Figure 3.5: Contour Plots for N=16

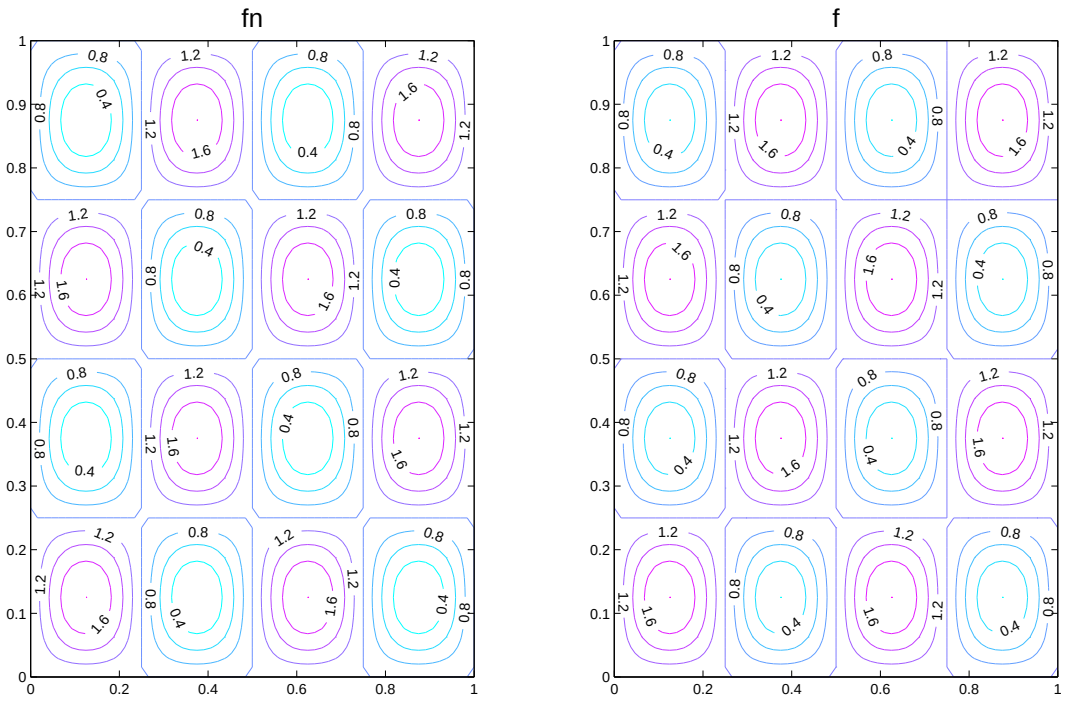


Figure 3.6: Contour Plots for N=64

3.3.1 Experiment 1

Consider initial and final densities given by

$$f(x,y) = 1 + \beta \sin(2\pi\gamma x)\sin(2\pi\gamma y)$$
$$g(x,y) = 1 + \alpha \cos(2\pi\rho x)\cos(2\pi\rho y)$$

where $\alpha = 0.4$, $\beta = 0.8$, $\gamma = 2$ and $\rho = 4$. We will first do all the analyses with the FD implementation. We run 18 iterations of the algorithm with $\tau = 1$ for grid sizes ranging from $N = 16$ to $N = 256$, with a pre-fixed maximum of 1000 BICG iterations per Newton iteration and with a tolerance of 10^{-8} for the BICG algorithm. First, we can qualitatively compare f with the f_n obtained after these iterations by glancing at the contour plots presented in Figures 3.5 and 3.6.

As we can see, the computed f_n looks really close to the target f. To quantify the quality of the approximation, we compute the l^2 norm of the difference between f and f_n and between f and the normalized f_n (which we previously called $\tilde{f}_n$) after 18 iterations for different grid sizes. The results are presented in Table 3.1. The errors we observe in this table confirm what we deduced from the contour plots. They get

N	$\|f - f_n\|_{l^2}$	$\|f - \tilde{f}_n\|_{l^2}$
16	0.0012	5.5776e-10
32	5.9876e-05	7.0387e-11
64	1.5193e-05	2.5579e-12
128	1.0726e-06	2.5773e-14
256	6.7445e-08	7.3015e-14

Table 3.1: Error Comparison for experiment 1

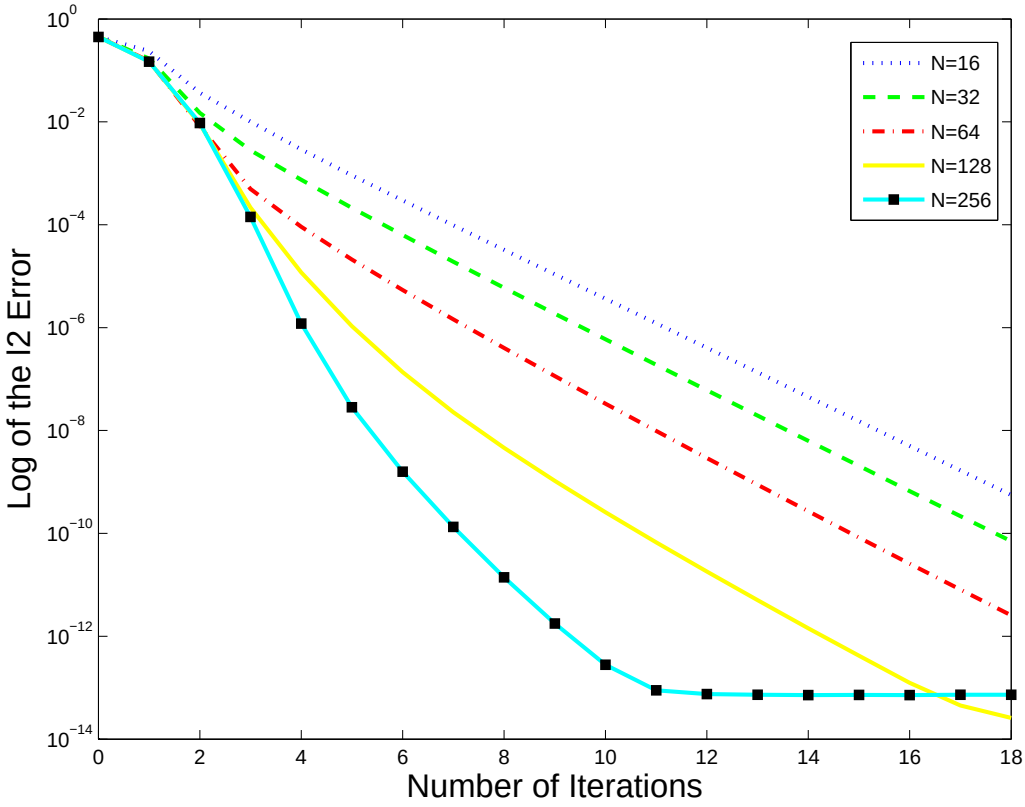


Figure 3.7: Error between f and $\tilde{f}_n$ on a semilog scale

smaller as we augment the grid size, except for the one corresponding to $\|f - \tilde{f}_n\|_{l^2}$ for $N = 256$. In this case, we see on Figure 3.7 that the error settles close to the machine epsilon ($2.2204e^{-16}$ in our case) after about 11 iterations. The small difference in the maximum precision reached (as opposed to the case $N = 128$) might be due to the fact that the BICG algorithm needed more than 1000 iterations to converge to a higher precision for such a big matrix (see Figure 3.8).

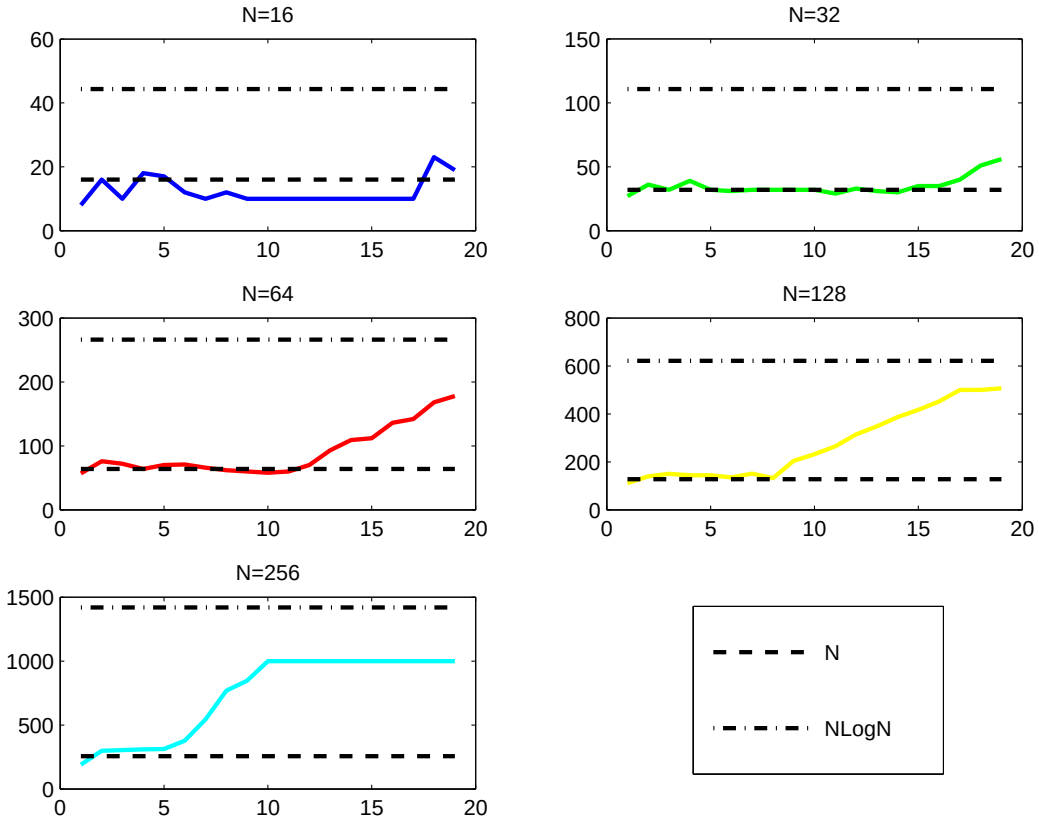


Figure 3.8: Number of BICG Iterations as a function of the Newton iterations for Different Grid Sizes, the dashed and dotted lines indicating the number of grid points per dimension as references ($P = N^2$)

To verify the order of accuracy of the discretization scheme, let e_N be the error $\|f - f_n\|_{l^2}$ for a $N \times N$ grid. The discretization error has the form

$$e_N \approx kh^\rho$$

where k is a constant, and hence we compute the order of accuracy by taking

$$\rho_{N/2,\,N} = \frac{\log\left(\frac{e_{N/2}}{e_N}\right)}{\log 2}.$$

For the values of e_N presented in Table 3.1, we get $\rho_{16,\,32} = 4.3249$, $\rho_{32,\,64} = 1.9786$, $\rho_{64,\,128} = 3.8242$ and $\rho_{128,\,256} = 3.9913$, which are getting very close to 4. This is to be expected as we implemented the discretizations of u_n with fourth-order finite differences schemes.

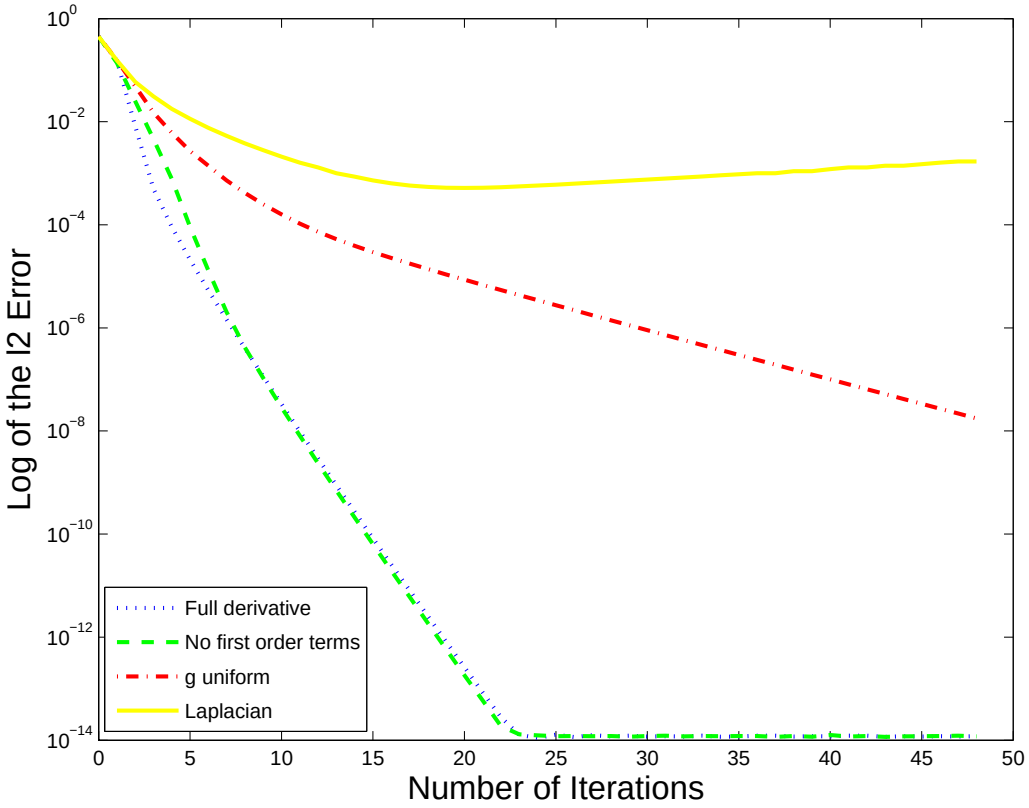


Figure 3.9: Comparison of the behavior for different derivatives when $N = 64$

We can analyze the behavior of $\|f - \tilde{f}_n\|_{l^2}$ iteration per iteration by looking again at Figure 3.7. We see that the error is always smaller for a bigger grid size and that it decreases linearly on the log-plot after a few iterations, depending on the grid size. Actually, the ratio between the error at iteration $n+1$ and the one at iteration n gets very close to $1/3$ for the first four curves when they do settle on a linear trend. This confirms a geometric convergence with a rate of about a third.

In Figure 3.8, we present the number of BICG Iterations required to solve the linearized Monge-Ampère equation for a given Newton Iteration. We realize that even though in theory, the BICG algorithm needs to perform at worst $\mathcal{O}(P^2)$ operations to converge ($\mathcal{O}(P)$ operations per step and then at worst P steps), for the current experiment this number stays in between N and $N\text{Log}N$. This yields an approximate computational complexity of $\mathcal{O}(nP\sqrt{P}\text{Log}(\sqrt{P})) = \mathcal{O}(nP^{3/2}\text{Log}(P^{1/2}))$.

We also tried for this case to replace the linearized equation with simpler ones to observe the difference. We considered four cases: the full derivative of the Monge-Ampère equation, the equation containing only the second derivatives of θ_n, the case where $g \equiv 1$ (which is the same as the derivative in the Loeper-Rapetti paper [13])

Tolerance	$\|f - f_n\|_{l^2}$	$\|f - \tilde{f}_n\|_{l^2}$	Average Number of BICG Iterations
10^{-8}	1.5193e-05	2.5579e-12	90.7368
10^{-6}	1.5193e-05	2.5579e-12	60.0526
10^{-4}	1.5193e-05	2.5581e-12	33.9474
10^{-2}	1.5193e-05	3.7237e-12	23.9474
10^{-1}	1.5193e-05	2.8880e-11	15.3684

Table 3.2: BICG tolerance comparison for $N = 64$ and 18 Newton iterations

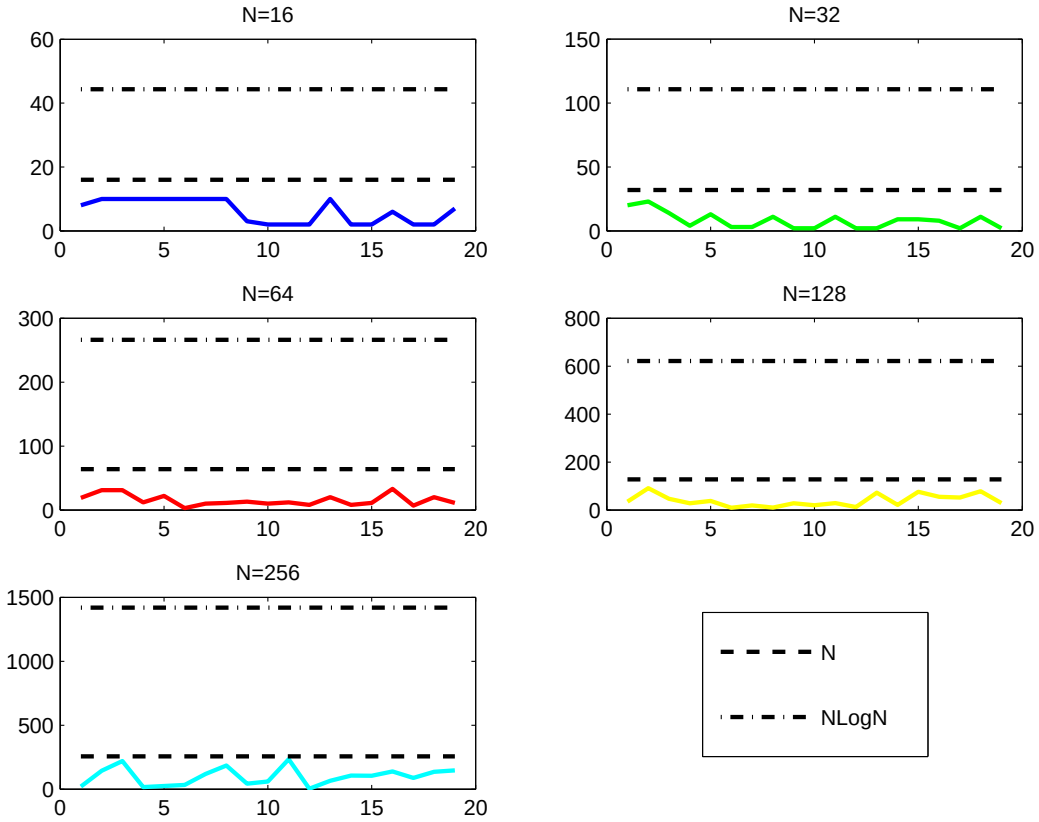


Figure 3.10: Number of BICG iterations for a tolerance of 0.1

and finally the Laplacian case (c.f. Section 2.5). The results can be found in Figure 3.9 where 50 iterations were presented to get a better idea of the convergence in all these different scenarios. As we can see, the laplacian case does not converge, but all the other ones do. Interestingly, the case where we remove the part involving the gradient of θ_n displays a similar behavior as the full case, which suggests that the

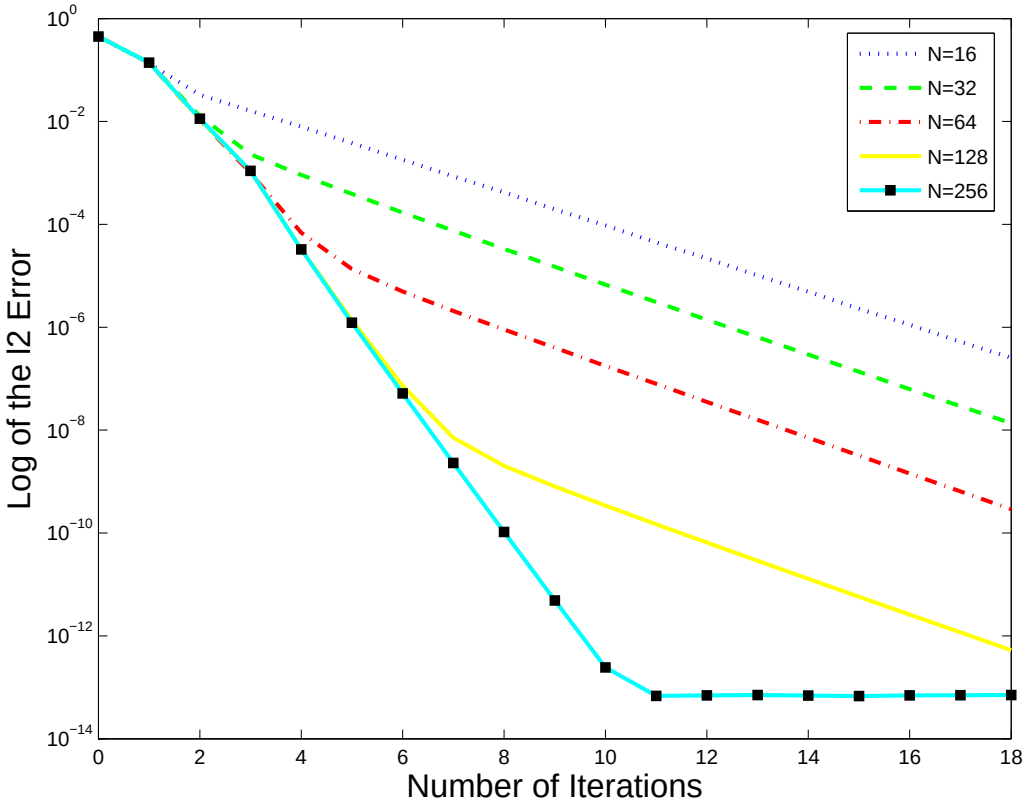


Figure 3.11: Error between f and $\tilde{f}_n$ on a semilog scale for the FT case

second derivatives mostly dictates the direction of steepest descent.

The last thing we are going to look at for the current experiment is the effect of the tolerance in the BICG algorithm. From Table 3.2, we see that as we loosen the tolerance, we don't loose any precision on $\|f - f_n\|_{l^2}$. The number of BICG iterations required also significantly drops as expected. This means that for this example, we don't need a very precise update on the solution at every step, and therefore the algorithm can be much faster than we first thought. We shall see that this conclusion also applies to all the other numerical experiments we conducted, and thus, in practice we can select a loose tolerance and expect good results.

As we observe in Figure 3.10, number of BICG iterations drops significantly for a tolerance of 0.1 and stays below N. This yields a computational complexity of about $\mathcal{O}(nP^{3/2})$ which is what Loeper and Rapetti observed in the case where the final density is uniform. Actually, if we compute the average number of iterations, we get a complexity that is very close to the optimal $\mathcal{O}(nP\text{Log}P)$.

Let's now see what happens when we redo these tests with the FT implementation. For this case, we used again a tolerance of 10^{-1} for the GMRES algorithm and a τ

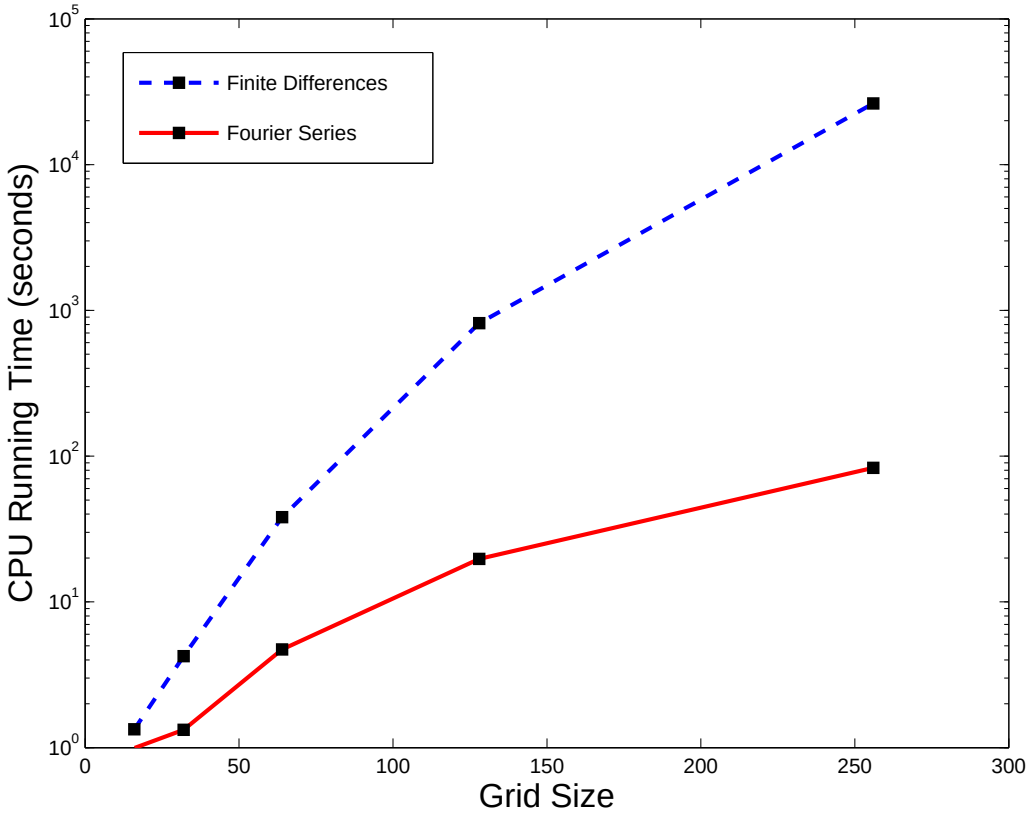


Figure 3.12: Running time for the two methods on a semilog scale

equal to 1. We also set the restarting threshold to a maximum of 10 inner iterations. In Figure 3.11, we present the value of $\|f - \tilde{f}_n\|_{l^2}$ for every iteration for all the grid sizes. We observe that the error decreases slower than the first method for smaller grid sizes, but this rate eventually catches up for bigger grid sizes. We note that for the same number of iterations, we get the exact same values of $\|f - f_n\|_{l^2}$ as in Table 3.1. Moreover, when we try to vary the tolerance for a selected grid size, the results are analogous to the ones presented in Table 3.2.

The biggest advantage remains the speed of this new implementation. Just as Strain observed in his experiments, the number of GMRES iterations remained nearly constant for every iteration and every grid size (1 outer iteration and 3 inner iterations), only sometimes varying by 1 inner iteration. By looking at Figure 3.12, we can compare the actual computing time between the first method with a tolerance of 10^{-1} for the BICG algorithm and the second method, still with a tolerance of 10^{-1} for the GMRES algorithm. As we see, the second method clearly outperforms the first one on that issue. All these computations were done using MATLAB on a single core of an Intel Zeon server running at 2.33 Ghz per CPU.

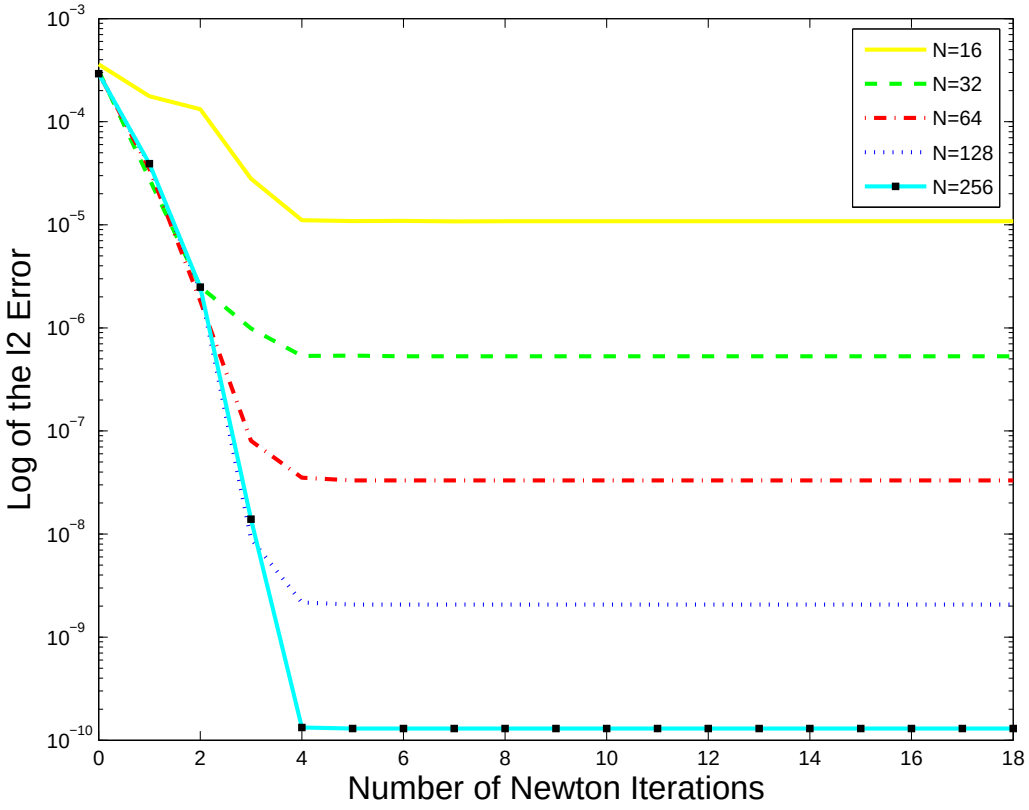


Figure 3.13: Error between u and u_n for the FD implementation with tol=10^{-4} on a semilog scale

3.3.2 Experiment 2

In this experiment, instead of only comparing the computed f_n to the target f, we compare also the solution u of our Monge-Ampère equation with the u_n reached after 18 iterations. We start with a known u and a known g, compute the corresponding right-hand side f, and then run the algorithm to obtain u_n. We select

$$u(x,y) = \frac{1}{2(\gamma\pi)^2}\cos(2\pi\gamma x)\sin(2\pi\gamma y)$$
$$g(x,y) = 1 + \alpha\cos(2\pi\rho x)\cos(2\pi\rho y)$$

where $\gamma = 4, \alpha = 0.9$ and $\rho = 4$.

For the FD implementation, we set a maximum of 1000 BICG iterations and a tolerance of 10^{-4}. For the FT implementation, we take the restarting threshold to be at a maximum of 10 inner iterations and also set the tolerance to 10^{-4}. The

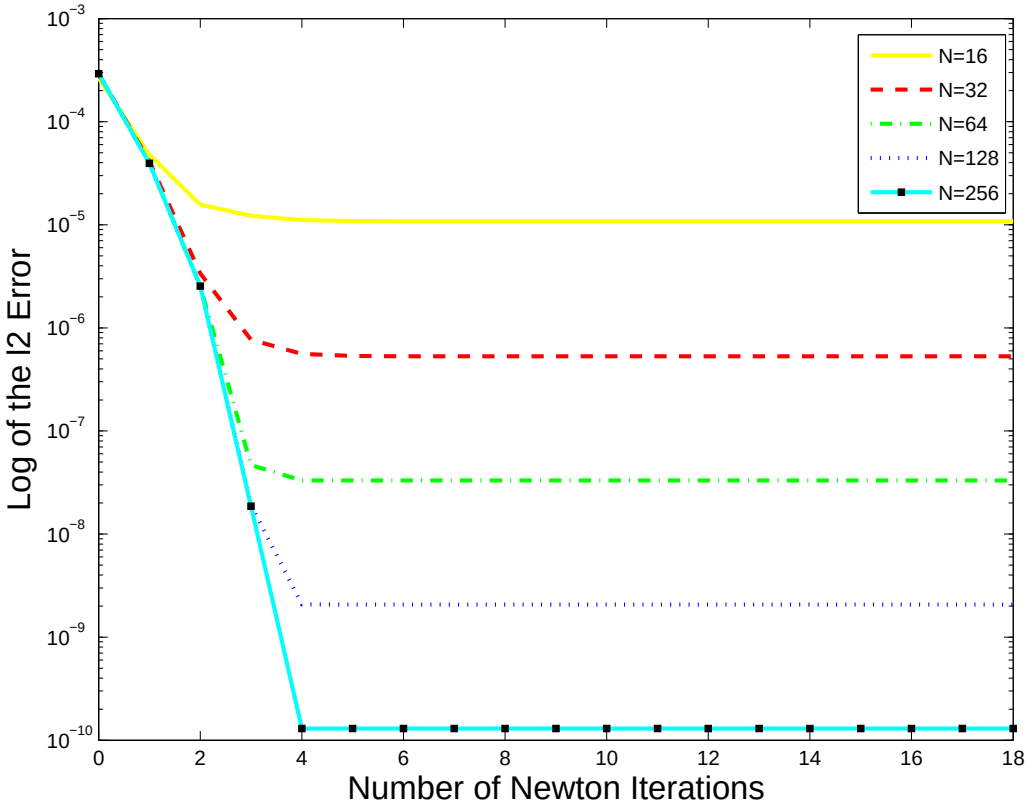
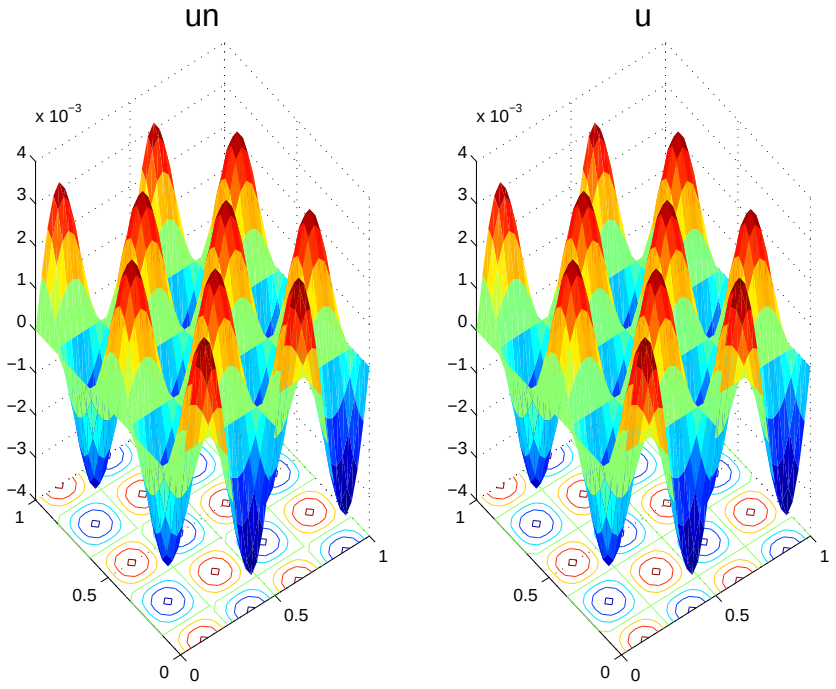


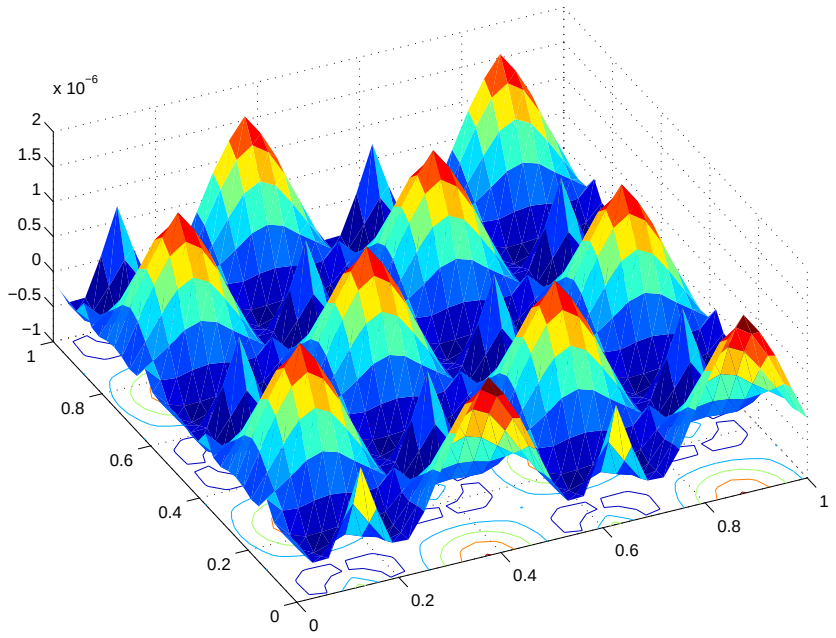
Figure 3.14: Error between u and u_n for the FT implementation with tol=10^{-4} on a semilog scale

corresponding comparison of the iteration per iteration error can be found for different grid sizes, respectively in Figure 3.13 and in Figure 3.14. This time again, these results are presented for a value of $\tau = 1$, which was enough to achieve convergence (we shall see later what happens when we vary this parameter). Observe that in both cases, for any grid size, the error settles on the discretization error after only 4 Newton iterations. This error of course gets smaller as N gets bigger. Computing the observed order of accuracy from the errors between u and u_n, we get from smaller to bigger grid sizes; $4.3521, 4.0035, 3.9965$ and 3.9990. This confirms the fourth-order consistent with the order of the finite differences scheme used to compute the right-hand side.

In addition, one can take look at the 3d plots of the analytical solution u and of the numerical solution u_n in Figure 3.15(a). The 3d plot of $u - u_n$ is also presented in Figure 3.15(b) to get an idea of the distribution of the errors. Both of these figures are results obtained for the FD implementation, the corresponding ones for the FT implementation being visually identical. As we can see, the errors seem evenly



(a) u_n and u



(b) $u - u_n$

Figure 3.15: 3d plots for $N = 32$ in the FD case

distributed on the whole domain. If we plot again the error between f and $\tilde{f}_n$ as a function of the number of iterations on a semilog plot, we observe that after the

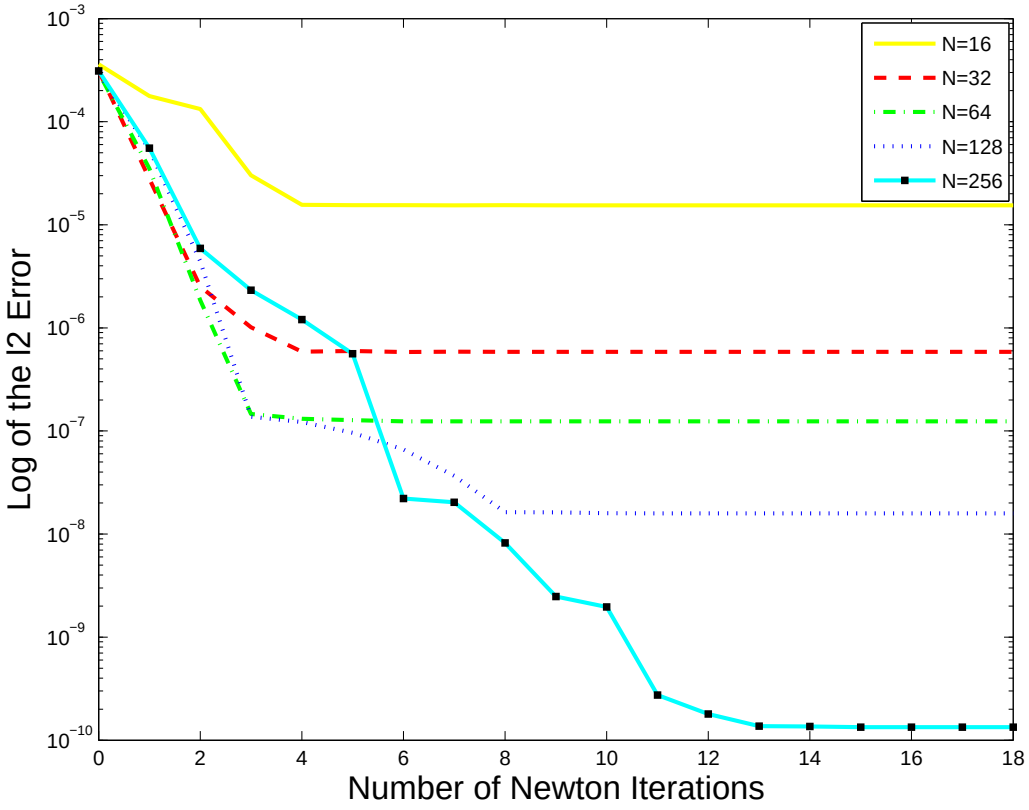


Figure 3.16: Error between u and u_n for the FD implementation with tol=10^{-1} on a semilog scale

first 4 Newton iterations, the convergence of $\|f - \tilde{f}_n\|_{l^2}$ follows a linear slope with a convergence rate slightly faster than a half. The estimated ratio is actually about 0.45 in the Fourier transforms case and is about 0.33 in the finite differences case, so the convergence is faster in this latter case for this final stage. Note that these rates are similar to the ones obtained in the first experiment.

Here again, in order to investigate whether we can decrease the computing time without losing too much precision in the results, we try to run the same experiment with a tolerance 10^{-1} (see Figure 3.16 and Figure 3.17). Due to the looser tolerance employed, the results are a bit erratic for the FD implementation, but overall still very good. Note that for this new tolerance, the computational cost of one iteration is now much less expensive, and the global computing time decreases a lot in both cases. We can quantify this by looking at Table 3.3. Observe that the BICG algorithm required less operations than the worst case scenario $\mathcal{O}(nP^2)$. Actually, it is still very close to $\mathcal{O}(nP^{3/2})$. This being said, we realize at first glance that the FT method is much faster than the FD method. The number of GMRES iterations per Newton iteration

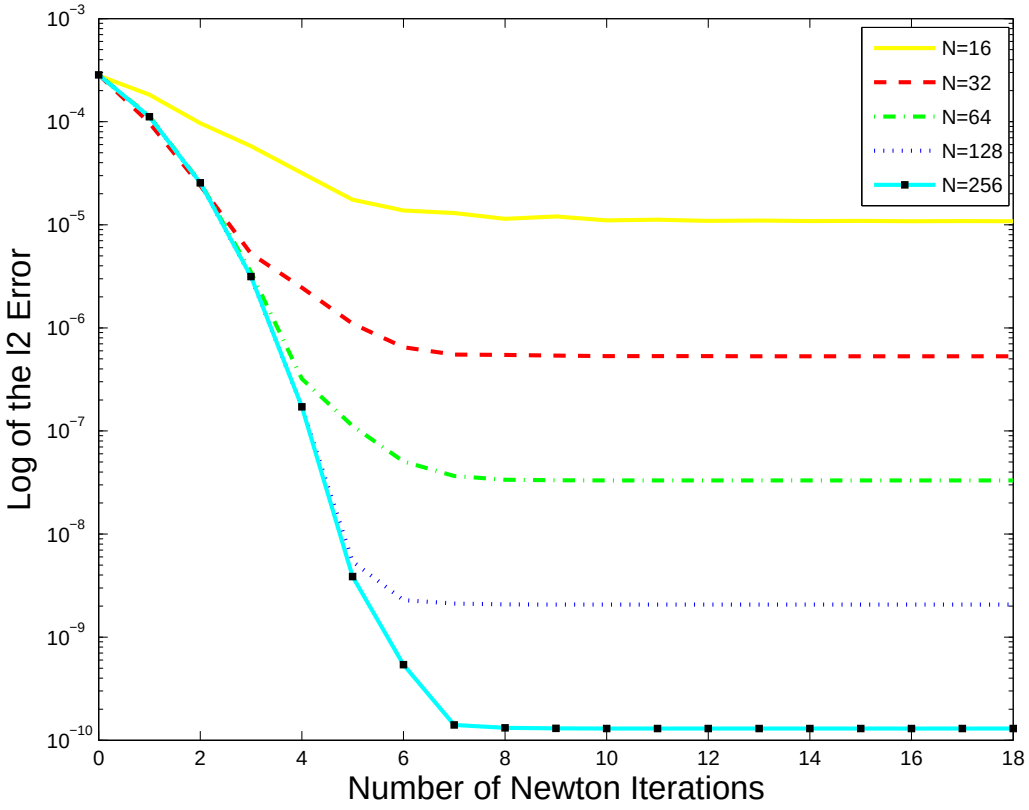


Figure 3.17: Error between u and u_n for the FT implementation with tol=10^{-1} on a semilog scale

N	Average number of GMRES iterations	Total computing time for FT	Average number of BICG iterations	Total computing time for FD
16	5.32	1.07	14.21	2.21
32	6.37	1.94	17.79	8.70
64	7.32	8.06	31.11	79.17
128	7.95	34.38	63.32	1221.10
256	8.05	145.07	134.63	34639.82

Table 3.3: Average number of BICG and GMRES iterations per Newton iteration and total computing time in seconds for the whole experiment (18 iterations) when the tolerance is set to 10^{-1}. We used a MATLAB implementation on an Intel Xeon running at 2.33 GHZ.

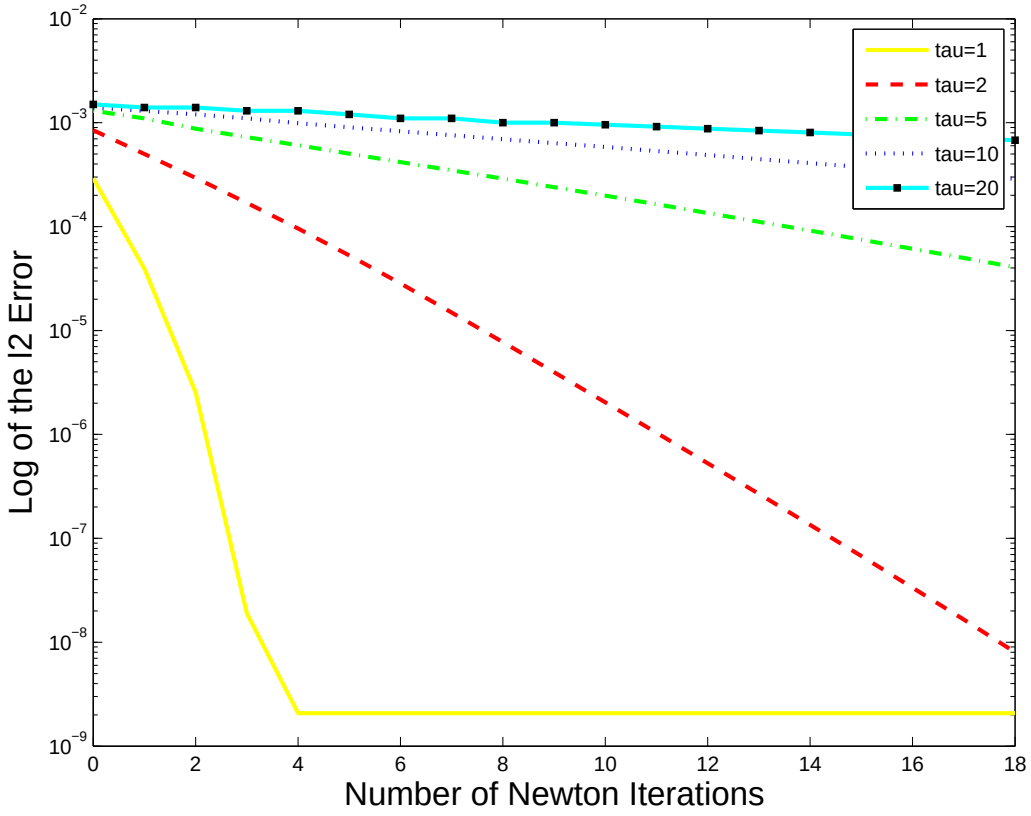


Figure 3.18: Error between u and u_n obtained with different values of τ for the FT case and $N = 128$

stayed nearly constant as we increased the grid size, which confirms the $\mathcal{O}(nP\text{Log}P)$ computational complexity. Moreover, we see in this experiment that the number of Newton iterations required to achieve convergence is independent of the grid size. This tells us that the global computational complexity is in fact $\mathcal{O}(P\text{Log}P)$.

Let's take a look now at what happens when we vary the parameter τ for a fixed grid size ($N = 128$). The results for the FT case are shown in Figure 3.18, taking again a tolerance of 10^{-1}. The results behave accordingly to our expectations, with a slower convergence for a bigger τ. We also obtain similar results for the FD case. At this point, the reader might wonder if the algorithm will always converge for $\tau = 1$. Unfortunately, we shall see later that the answer is no. However, in all the numerical experiments we conducted related to the upcoming image processing applications, we never had to pick a τ greater than 2 to achieve convergence in the FT case. For the FD implementation, the results were not always as good, and bigger values were sometimes necessary.

3.4 A Posteriori Stability Analysis

One of the main concerns when analyzing an algorithm is to know whether or not the solution u_h of the discretized problem converges to the solution of the continuous problem u. Two necessary conditions to ensure this convergence are the consistency and stability of the numerical scheme. To explain this briefly, consider the solution obtained by the Newton algorithm, written as a series of the updates, $u = \sum_{i=0}^{\infty} \theta_i$. When we discretize, we hope to obtain the solution u_h as $\sum_{i=0}^{\infty} \theta_{h_i}$, where θ_{h_i} is the solution of the discretized linear Monge-Ampère equation at step i of the Newton algorithm, $L_{h_i}\theta_{h_i} = (f_h - f_{h_i})/\tau$. However, due to discretization errors, the series $\sum_{i=0}^{\infty} \theta_{h_i}$ could diverge as $h \to 0$. We know that if it converges, then

$$\lim_{h\to 0} \lim_{i\to\infty} \|\theta_{h_i}\| = 0.$$

Note that this limit can be bounded by using the inequality

$$\|\theta_{h_i}\| \leq \|L_{h_i}^{-1}\| \cdot \|(f_h - f_{h_i})/\tau\|.$$

Then, we say that a scheme is consistent if $\|f_h - f\|$ converges to 0 as $h \to 0$. In addition, a scheme is said to be stable if $\|L_h^{-1}\|$ is uniformly bounded as $h \to 0$. These two quantities should be seen as giving the behavior of $\|f_{h_i} - f_i\|$ and $\|L_{h_i}^{-1}\|$ for i large. We observe that if one of these conditions fails, then we are not guaranteed to see the discretized solution u_h converge at all. Since we are using second and fourth order accurate finite differences, we know that our scheme is consistent. It is unfortunately hard to obtain a uniform bound on the norm of L_h^{-1}. What we propose to do here is to compute the norm of $L_{h_i}^{-1}$ numerically for i large in the case of the two previous numerical experiments and thus get at least an idea of the stability properties of our methods. One approximate way to do this is to measure the spectral radius ρ, i.e. the largest eigenvalue, of this matrix. Indeed, for any induced matrix norm, $\rho(A) \leq \|A\|$ and in fact, $\rho(A)$ is the infimum of the induced norms.

For the FD implementation, we have an explicit representation of the matrix corresponding to the discretization of the linearized Monge-Ampère operator at every Newton step n and therefore we can compute its inverse and then the corresponding spectral radius of this new matrix. However, for the FT implementation, we do not possess such representation and we have to use an indirect way of computing ρ. The one we selected is the power method (or power iteration). For a matrix A, this iterative

N	First Experiment		Second Experiment	
	FD case	FT case	FD case	FT case
8	2.5	0.04	9.7	0.13
16	6.7	0.03	11.6	0.04
32	26	0.03	33.3	0.03
64	103.6	0.03	127.9	0.03
128	414.3	0.03	500.1	0.03

Table 3.4: Computed $\rho(L_{h_i}^{-1})$ as an approximation of $\rho(L_h^{-1})$ taking $i = 18$ for the two experiments and for different grid sizes

algorithm starts with a vector b_0 and computes the iterates $b_{k+1} = Ab_k/\|Ab_k\|$. If A has a dominant eigenvalue and if b_0 has a non-zero component in the direction of the eigenvector associated with this largest eigenvalue, then the sequence (b_k) converges to this eigenvector, from which we can deduce the spectral radius of A (see [15]). We apply this technique to the inverse matrix produced at every step. More specifically, for a given n, we start with a b_0 randomly generated with components in $[0, 1]$. Then, using the method presented in Section 3.2, we compute the product $A_n^{-1}b_k$ and then the iterate with $A_n^{-1}b_k/\|A_n^{-1}b_k\|_2$.

The results of this exercise are presented in Table 3.4. We observe that for the FD implementation, the spectral radius clearly increases with the grid size. This indicates that this method is not stable and thus convergence of the numerical solution to the actual solution of the partial differential equation is not guaranteed as the grid size is increased. However, the numerical results presented here, for the analytical tests, do not seem to suffer from this instability. The present studies warants about a potential non-convergence in some case and therefore care should be taken when this method is used in practice. For the other implementation, results are much more promising. Indeed, we see that in both experiments, the spectral radius converges to a very small constant, 0.03 (more precisely, in the first case it is about 0.0251 and in the second case about 0.0301). This suggests that the FT implementation is in fact stable. Actually, in every numerical experiment presented in this work (there are still a lot to come), and even when we pushed the grid size to much bigger values like N=512, this implementation always displayed the same behavior and the spectral

radius always stayed close to 0.03. Of course we cannot directly conclude stability from these observations and even if we could, convergence of u_h to u would not be certain due to the nonlinear nature of our equation. However, it gives us a good indication to whether we can expect this convergence from the discretized solution to the continuous one. Clearly, further investigation is required here. Note also that we can possibly explain this observed stability by the fact that in the FT case, we used the averaged operator to act as a preconditioner for the linearized Monge-Ampère operator, as opposed to the FD case, where we did not use any preconditioner, before or after discretization. Finally, one can find more information on the stability and consistency of numerical schemes in [25] or [33].

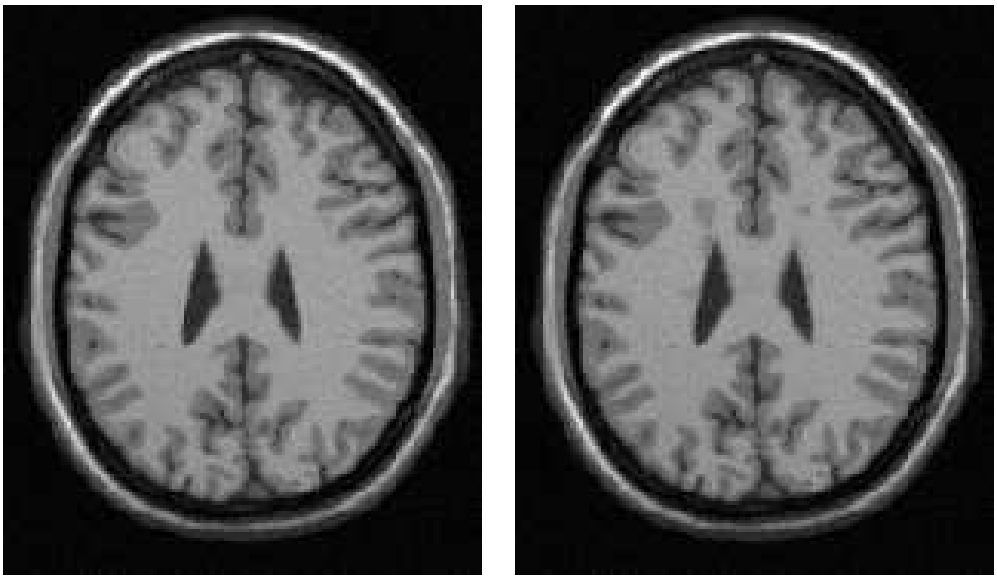
Chapter 4

Application to Medical Imaging

Out of the numerous applications of mass transport, the one that is going to be of interest to us lies in the world of image processing. There, one of the most common tasks is to determine a geometric correspondence between images taken from the same scene in order to compare them or to integrate the information they contain; hence obtaining more meaningful data. One could think of pictures acquired at different times, with different equipment or from different viewpoints. This process falls into the category of what is referred to as image registration. There are two main types of image registration methods: the rigid ones which involve translations or rotations and the nonrigid ones where some stretching of the image is required to map it on the other one.

Recently, people working on the optimal mass transport problem realized that it could provide a good nonrigid image registration technique. Indeed, take for example two grayscale images. We could think of them as representing a mass distribution of the amount of light "piled up" at a given location. A bright pixel on that image would then represent a region with more mass whereas a darker pixel would correspond to a region with less mass. Computing the optimal map between these images and analyzing the rate of change of that map could reveal the best way (in terms of minimizing the transportation distance) of moving the mass from the first density to the second, precisely showing what is changing on the images and how it is happening.

In [37], Rehman et al. actually lists several advantages of the optimal mass transport method for image registration. However they also stress the fact that it is computationally expensive and that is one reason why it is important to find efficient numerical methods to solve this problem, which is precisely what we are trying to accomplish in this work.



(a) Initial density f: MRI scan of a normal brain

(b) Final density g: MRI scan of the same brain with Multiple Sclerosis lesions

Figure 4.1: Two slices of the same brain depicting the presence of MS

Let's now focus our attention on the specific field of medical imagery to test our algorithm. In healthcare, the process of establishing accurate diagnostics is crucial. Think of a patient developing a malignant tumor. Early detection and effective monitoring are essential to increase that person's chance of survival. In certain situations, the speed of that detection also matters a lot. A surgeon performing brain surgery could use real-time imaging with an automated change detection procedure to quickly detect the apparition of any cerebral aneurysm while undergoing surgery and react appropriately. Therefore, having access to a fast and accurate method that does such a thing could be of great help to the specialists. Optimal transport presents a potentially good way to achieve this, if one could reduce its high computational cost.

With this in mind, consider the two brain MRI (Magnetic Resonance Imaging) scans presented in Figure 4. These images were taken from the BrainWeb Simulated Brain database at McGill University [32] and represent a slice of a healthy brain and a slice of the same brain where the multiple sclerosis disease, more commonly known as MS, is spreading. This nervous system disease damages the myelin sheaths around the axons of the brain and leaves scars (or sclerosis) visible on an MRI. We chose MS as a test case since its actual detection process relies on neuroimaging by trying to identify the scars and since its presence leaves traces similar to multiple tumors.

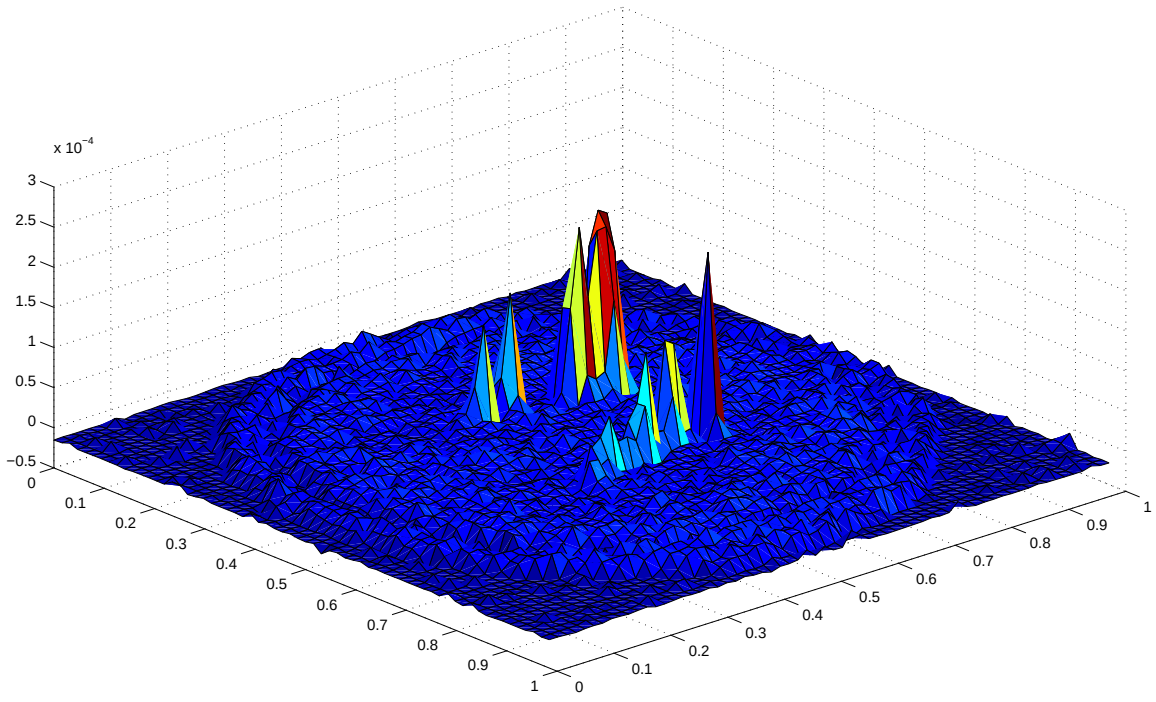


Figure 4.2: Surface plot of div(u_3)

Note that because the scans are dark, their representation in greyscale contains many values close to or equal to 0. Our method require the densities to be normalized and bounded away from 0, and thus, we applied a translation of the form of (3.1) on both of them before initiating the algorithm. In addition, we rescaled them so that they would be exactly square (256×256 pixels). This is not required, but helps simplify the code.

We can observe in Figures 4.2 and 4.3 the results reached after only 3 Newton iterations with $\tau = 1$ and a tolerance of 10^{-2} for the FT implementation. The l^2 norm of the error between f and f_n was reduced to 9.9151×10^{-04}. We filtered the contour plots a bit to rule out some of the noise, and thus to be able to have a better view of the important changes. It is easy to see the spikes corresponding to the variations in brightness where the scars are appearing. The number of GMRES iterations required per Newton iteration was very small and nearly constant (only 1 outer iteration and about 6 inner ones). Moreover, even if our code was not necessarily optimized in terms of speed, it only took about 30 seconds to compute these results on an Intel

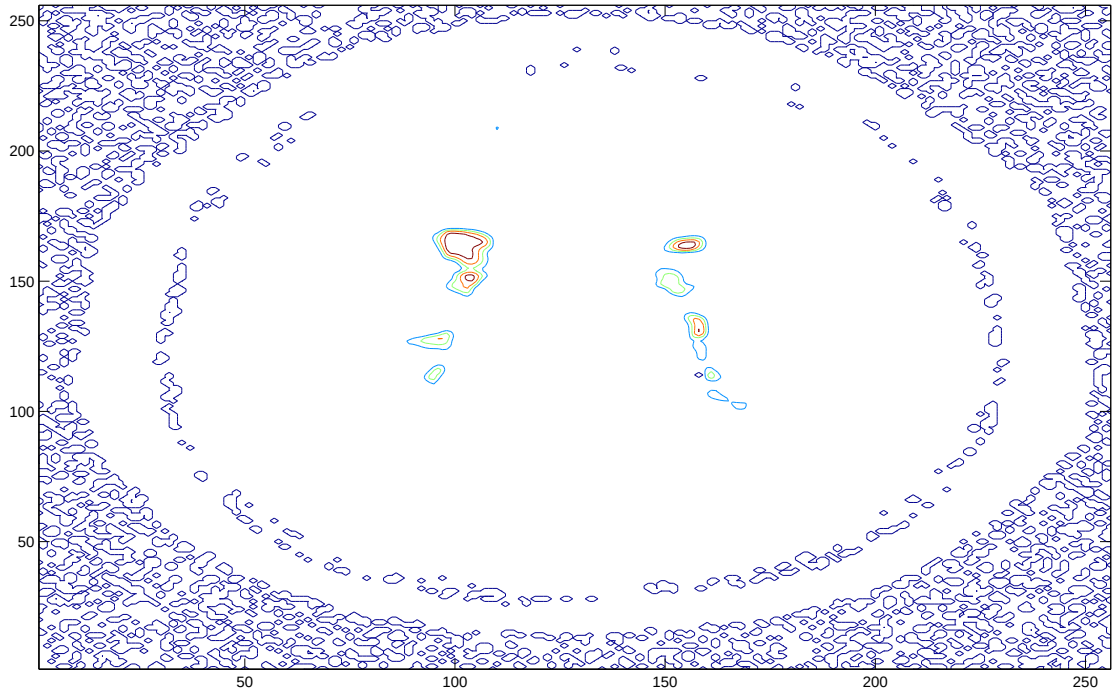
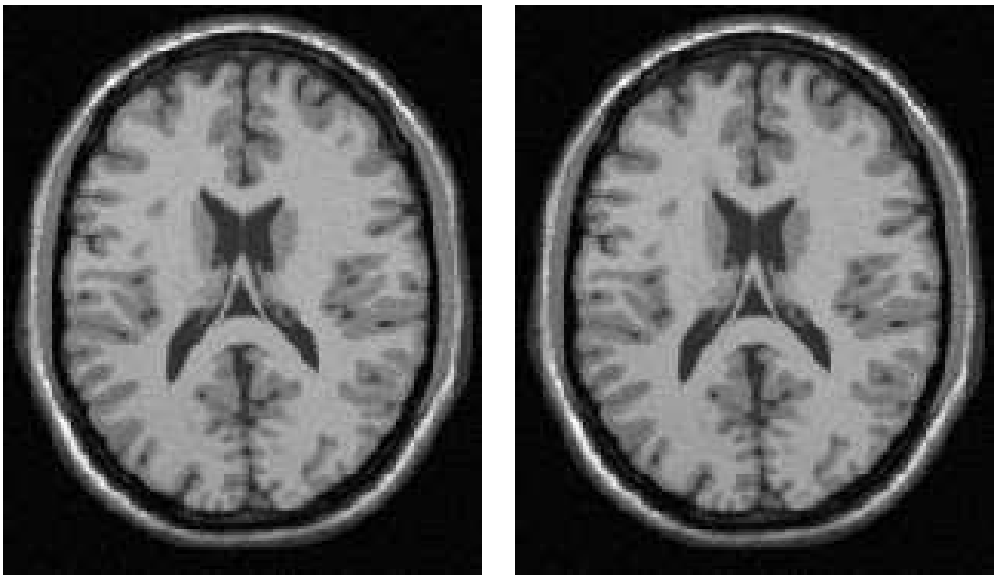


Figure 4.3: Filtered contour plot of $\mathrm{div}(u_3)$

Xeon with 2.33 GHZ of RAM.

We also need to mention that we don't have access here to an analytical expression for f or g. Therefore, an approximation is required for $g(x + \nabla u_n(x))$ and $\nabla g(x + \nabla u_n(x))$ at every $x = (ih, jh)$. When $x + \nabla u_n(x)$ does not land exactly on one of the grid points, we can interpolate the values of g and ∇g at such point by using polynomial interpolation, for example. However, we found that we could obtain good results with only a closest neighbor approximation, which is much less computationally demanding.

If we try to run the algorithm with the same conditions on a second set of images corresponding to a different slice of the same brain where the sclerosis lesions are a little bit less obvious to detect by eye, we still reach a very satisfactory outcome. Again, this can be seen from the surface plot of the divergence of u_4 in Figure 4.5. We also graphed the filtered contour plot like in the first case, but this time to get a better visual understanding of the situation, we colored the inside of the contour lines corresponding to the affected regions and we superposed this image on the MRI



(a) Initial density f: MRI scan of a normal brain

(b) Final density g: MRI scan of the same brain with multiple sclerosis lesions

Figure 4.4: Different slices from the same brain showing scars

scan of the normal brain (see Figure 4.6). Note that these results were obtained after 4 Newton iterations, where $\|f - f_n\|_{l^2}$ was reduced to 0.001. The number of GMRES iterations was again nearly constant, close to 1 outer and 6 inner iterations.

In addition to that change detection, the optimal transport map $\check{T} = x + \nabla u(x)$ actually gives us precise information on the amount of variation from one MRI scan to the other. Remember from Section 1.2 that we can define a metric between probability densities from the solution to the transport problem, the distance being

$$\int_\Omega |x - \check{T}(x)|^2 f(x)\,\mathrm{d}x = \int_\Omega |\nabla u(x)|^2 f(x)\,\mathrm{d}x.$$

This could quantify the magnitude of the change between the two images and thus help monitor the growth of the disease. In our experiment, we got for the first case a value of 4.75×10^{-10} and for the second case a value of 4.14×10^{-10}. These numbers validate our initial visual estimate which suggests that the sclerosis are more widely spread in the first slice of the afflicted brain than in the second one.

In conclusion, we saw in this section that our algorithm also performed well on more practical examples. Due to our limited knowledge of the field of medical imagery, we cannot say that optimal transport has to be the method of choice for performing

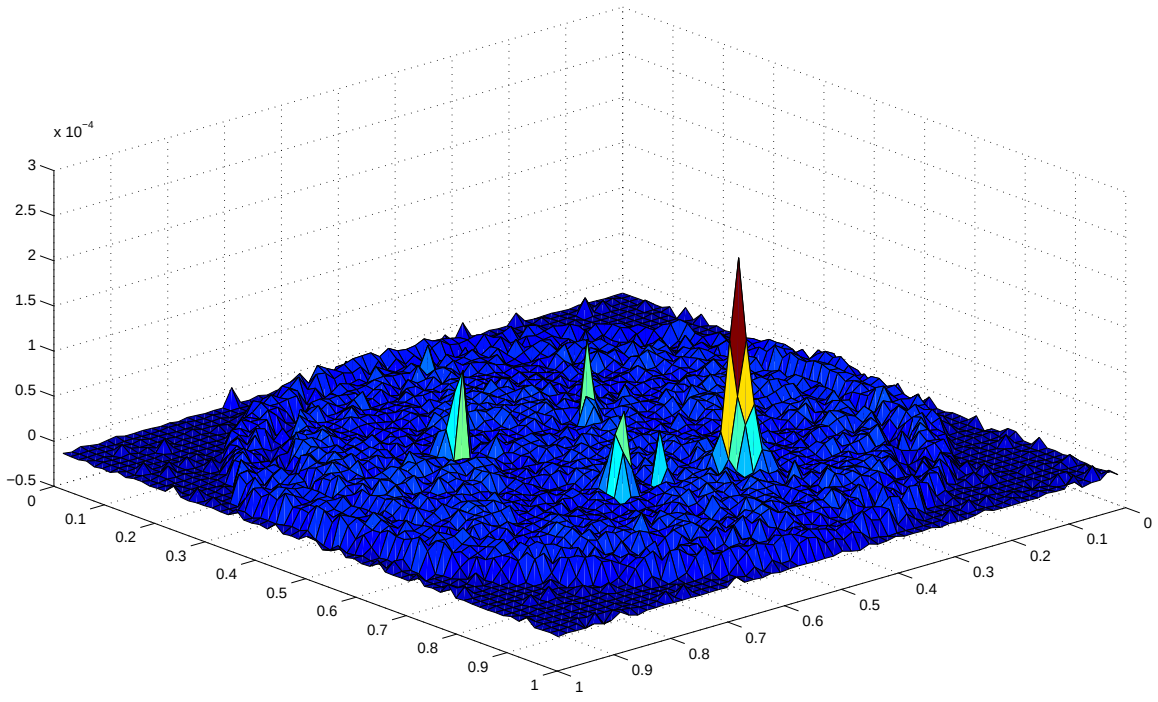


Figure 4.5: Surface plot of div(u_4) for the second set of images

the tasks we presented. However, it provides an interesting option that could be made practical with the idea behind our method. Recall that even though we implemented it only in 2D, in theory it is valid in any dimension. Therefore, it could also be applicable on 3D datasets which would be much more realistic when it comes to analyzing a biological phenomenon similar to the ones we treated here.

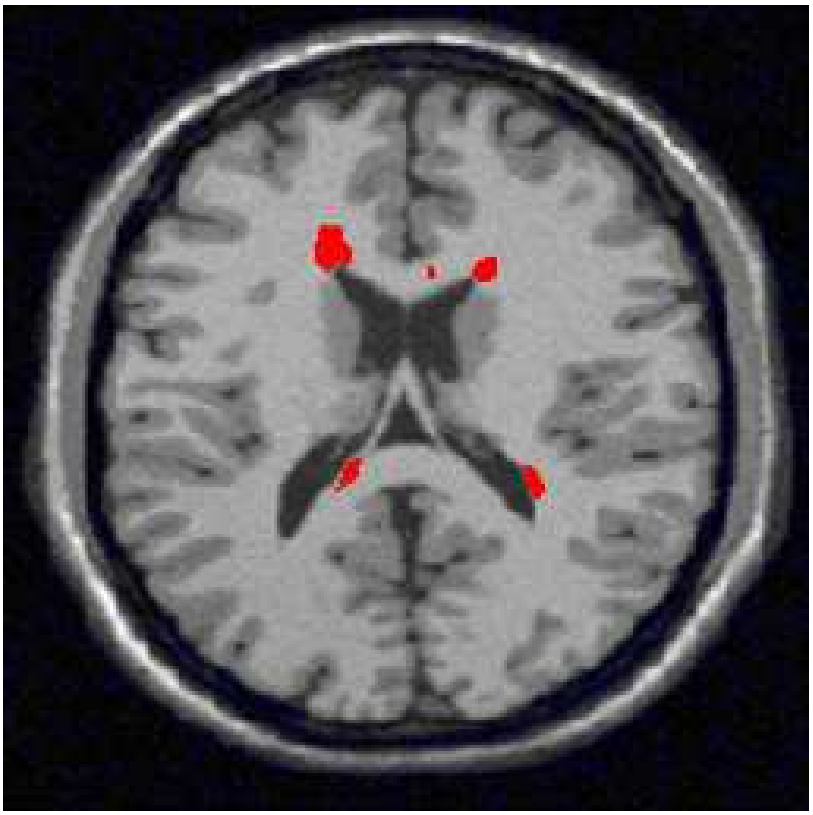


Figure 4.6: Scan of the healthy brain on which was superposed the colored filtered contour plot of div(u_4) for the second set of images

Chapter 5

Overview of Some Existing Numerical Methods

The optimal transport problem is starting to be well understood theoretically, especially for the quadratic distance case. In spite of that, it is only very recently that people started working on developing algorithms to obtain numerical solutions of this problem. In order to judge if our method performs well compared to the other ones already available, we are going here to present three of these algorithms and then try to see from their test case if our method is efficient. While we selected three methods, our list is not at all a comprehensive one. The number of new methods has been growing fast lately, making it difficult to keep track of developments in the subject. We are going to present those methods only very briefly. For more information one should consult the original papers. In addition, we did not implement any of them, so we base our observations only on what's available in the papers themselves.

5.1 A Fluid Dynamics Reformulation

We start with the method of Benamou and Brenier who presented in [5] one of the first approaches to deal with the numerical resolution of the Monge-Kantorovich problem. Their approach consisted in reformulating the problem into a fluid dynamics framework where the square of the quadratic Wasserstein distance $\mathcal{W}(\nu, \mu)$ is equal to the infimum of

$$T \int_{\mathbb{R}^d} \int_0^T \rho(t,x)|v(t,x)|^2 dxdt,$$

which is taken on all density fields $\rho(t,x) \geq 0$ and all velocity fields $v(t,x) \in \mathbb{R}^d$ satisfying the continuity equation

$$\frac{\partial \rho}{\partial t} + \operatorname{div}(\rho v) = 0$$

for $0 < t < T$, $x \in \mathbb{R}^d$ and with the Cauchy data

$$\rho(0,\cdot) = \mu, \ \ \rho(T,\cdot) = \nu.$$

Then, they write this space-time constrained minimization problem as a saddle-point problem using a Lagrange multiplier ϕ for the two previous constraints:

$$\inf_{\rho,m} \sup_{\phi} L(\phi,\rho,m) \tag{5.1}$$

where

$$\begin{aligned} L(\phi,\rho,m) = \int_0^T \int_\Omega \left(\frac{|m|^2}{2\rho} - \frac{\partial}{\partial t}\phi\rho - \nabla_x \phi \cdot m \right) dx\, dt \\ - \int_\Omega \left(\phi(0,x)\mu - \phi(T,x)\nu\right)\, dx. \end{aligned}$$

Here, m is the momentum: $m = \rho v$. Next, they show that we can rewrite (5.1) as

$$\begin{aligned} &\sup_{(\rho,m)} \inf_{\phi,(a,b)} \widehat{L}(\phi,a,b,\rho,m) \\ &\quad = \sup_{(\rho,m)} \inf_{\phi,(a,b)} \left[F(a,b) + \int_\Omega \left(\phi(0,x)\mu - \phi(T,x)\nu\right)\, dx \right. \\ &\qquad\qquad \left. + \int_0^T \int_\Omega (\rho,m)\cdot\left(\nabla_{t,x}\phi - (a,b)\right) dx\, dt \right] \end{aligned}$$

for $(a,b) : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R} \times \mathbb{R}^d$ and F defined by

$$F(a,b) = \begin{cases} 0 & \text{if } a + \dfrac{|b|^2}{2} \leq 0 \text{ pointwise} \\ +\infty & \text{otherwise.} \end{cases}$$

The resulting new Lagrangian is then modified using the classical technique of the augmented Lagrangian, which consists of adding a penalty term to combine the advantages of the Lagrange multipliers and of the penalty methods. The problem

becomes

$$\sup_{(\rho,m)} \inf_{\phi,(a,b)} L_r(\phi, a, b, \rho, m)$$
$$= \sup_{(\rho,m)} \inf_{\phi,(a,b)} \Bigg[\widehat{L}(\phi, a, b, \rho, m) + \frac{r}{2} \int_0^T \int_\Omega (\nabla_{t,x}\phi - (a,b)) \cdot (\nabla_{t,x}\phi - (a,b))\, dx\, dt \Bigg]. \quad (5.2)$$

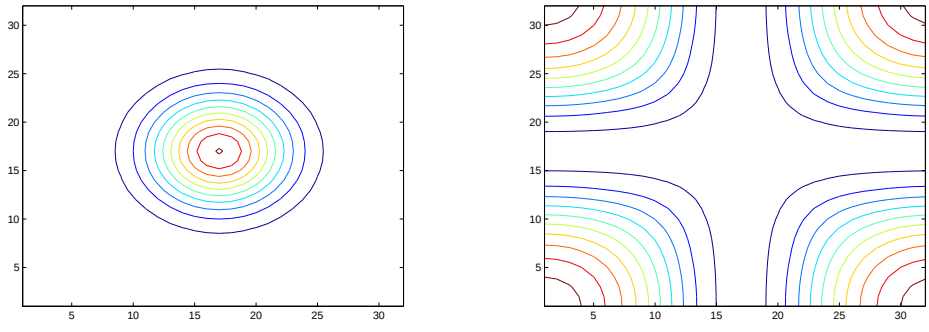
The interested reader can find more details concerning the augmented Lagrangian technique and the related resolution algorithms in [30], an introductory textbook on the subject. Benamou and Brenier used a procedure commonly referred to in that context as ALG2 to compute a numerical solution of (5.2):

ALG2 algorithm

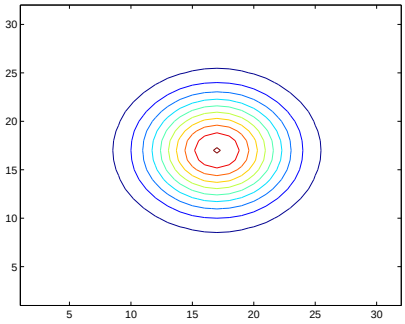
$$\begin{cases} \text{Given } \phi_0, a_0, b_0, \rho_1 \text{ and } m_1, \text{ loop over } n \in \mathbb{N} \\ \text{Step A: Find } \phi_n \text{ such that:} \\ \qquad L_r(\phi_n, a_{n-1}, b_{n-1}, \rho_n, m_n) \leq L_r(\phi, a_{n-1}, b_{n-1}, \rho_n, m_n), \quad \forall \phi. \\ \text{Step B: Find } (a_n, b_n) \text{ such that:} \\ \qquad L_r(\phi_n, a_n, b_n, \rho_n, m_n) \leq L_r(\phi, a, b, \rho_n, m_n), \quad \forall (a,b). \\ \text{Step C: Update} \\ \qquad (\rho_{n+1}, m_{n+1}) = (\rho_n, m_n) + r(\nabla_{t,x}\phi_n - (a_n, b_n)) \end{cases}$$

They show that the resolution of Step A is equivalent to the resolution of a space-time Laplace equation, which can be solved in $\mathcal{O}(M\log M)$ operations, M being the size of the space-time grid. For Step B, the minimization can be done pointwise and therefore only requires $\mathcal{O}(M)$ operations. Finally, Step C is only the update which can also be achieved in linear time with respect to M. The total computational cost of that method is thus $\mathcal{O}(nM\log M)$ where n is the number of iterations of ALG2 required to reach convergence.

We are now going to reproduce one of their test cases and make a few comments on the performance of our method. Consider the initial and final densities to be given



(a) f (b) g



(c) f_{10}

Figure 5.1: Contour plots of the initial and target Gaussian densities

by the periodic extension of Gaussian curvatures centered at different points in space:

$$f(x,y) = Ce^{-0.5\left[(x-0.5)^2+(y-0.5)^2\right]/\sigma_1^2},$$
$$g(x,y) = Ce^{-0.5\left[x^2+y^2\right]/\sigma_2^2}.$$

with parameters $C = 2$, $\sigma_1 = 0.12$ and $\sigma_1 = 0.15$. Since our method requires f and g to be bounded away from 0 and to have the same mass, we need to translate these densities so that they have an integral equal to 1. This does not affect the behavior of the solution to the transport problem for the kind of applications we saw in the previous chapter since it only affects the internal representation of the image in the computer.

Then, we compute the optimal map on a 32×32 grid with our Newton algorithm by selecting $\tau = 1$, tol$= 10^{-1}$ and the FT implementation. Convergence of the error $f - f_n$ to 7×10^{-05} in the l^2 norm was reached after 10 iterations with only about 4

GMRES iterations per Newton iterations. For our other implementation, the number of BICG iterations required to get to the same precision was also close to 10, but it obviously took longer to compute. In [5], the authors mention that they required about 30 iterations of ALG2 to get good approximate solutions. Observe that the smaller number needed in our case relies on the fact that it converged for a τ of only 1.

When we compare the two methods, we realize that even though they are both designed to compute the value of the quadratic optimal transport, in the end they provide much different information on the transport itself. Benamou and Brenier's algorithm show how the reallocation of matter is done through time, but does not directly provide the optimal map, whereas ours was specifically built for that purpose. By introducing a time variable, they also increase the computational capacity to $\mathcal{O}(M\log M)$ where $M = N^d \times N_T$, N_T being the time grid. As we saw before, ours usually only required $\mathcal{O}(N^d \log N^d)$, which potentially would yield a substantial increase in efficiency.

5.2 A Gradient Descent on the Dual Problem

The second method we choose to present is due to Chartrand, Wohlberg, Vixie and Bollt [35]. If one is familiar with the theory surrounding the Monge-Kantorovich problem, it is much simpler to derive than the previous one. Instead of solving directly the mass transport problem, they seek a solution of the dual (in the Kantorovich sense) introduced in Theorem 1.1.1. Recall from Sections 1.1 and 1.2 that this dual problem is equivalent to minimizing

$$\int_\Omega \Psi(x) f(x)\, dx + \int_\Omega \Phi(y) g(y)\, dy$$

amongst all Ψ and Φ satisfying

$$\Psi(x) + \Phi(y) \geq x \cdot y \quad \forall\, x, y \in \Omega.$$

It is known that the minimizing functions are convex conjugates of each other:

$$\Psi(x) = \Phi^*(x) := \inf_{y \in \Omega} (x \cdot y - \Phi(y))$$

and vice-versa, $\Phi(y) = \Psi^*(y)$. Their idea is then to do a steepest descent on the functional

$$F(\Psi) = \int_\Omega \Psi(x) f(x)\, dx + \int_\Omega \Psi^*(y) g(y)\, dy$$

where the derivative is given by the Monge-Ampère equation

$$F'(\Psi) = f - g(\nabla \Psi^{**}) \det(D^2 \Psi^{**}).$$

This yields an update on the solution of the form

$$\Psi_{n+1} = \Psi_n - \frac{1}{\tau_n} F'(\Psi_n).$$

Just like in our method, τ_n is a parameter that dictates the size of the step. Note that the proof that F' is indeed the correct derivative can be found in their paper. Moreover, they replaced in practice Ψ^{**} with Ψ to compute $F'(\Psi)$.

One of the numerical experiments they have performed uses of two famous pictures in the image processing community, namely, Lena and Tiffany. They computed the transport map using their algorithm, applied it to the first image to get the second one and then visually compared the result with the target image. We do the same thing with our method.

First, we do a little bit of preprocessing by translating the initial pictures and by scaling them to exactly 256×256 pixels. Then, we select $\tau = 2$, tol$= 10^{-1}$ and run 20 iterations of the Newton algorithm with the FT implementation. The output is presented in Figure 5.2. Again, the number of GMRES iterations stayed nearly constant (only about 4 inner iterations) which made the computing time very small. This time, our method did not converge for τ equals to 1. This can possibly be explained by the fact that the two images are really different and therefore the transport map has to vary a lot more than in the previous examples. Moreover, when we repeated the same experiment with the FD implementation, the τ required jumped to 8 and we had to iterate about 60 times to get good results. This is yet another argument in favor of the second implementation. Observe also that our algorithm is designed in a way that at every step we have access to the transport map that sends f_n to g, and it is also true for the other one presented here.

For the same experiment, Chartrand et al. had to run 190 iterations of their method with a $\tau_n = 1$ for all n and even then, the resulting image was not visually identical to the target one. They noted that numerical artifacts started to appear

(a) g

(b) f_2

(c) f_5

(d) f_{10}

(e) f_{20}

(f) f

Figure 5.2: Iterations of the Lena to Tiffany warp

and worsened as they kept iterating. In an attempt to improve the quality of the result, they tried to select smoothed versions of the initial images, run the algorithm

for a certain number of iterations, replace the initial map with the final one from the previous step and start over with images smoothed to a lesser degree. By employing this multiresolution until the images are (nearly) not smoothed anymore, they obtain a better quality warp, but we can still visually observe discrepancies, which is not the case for our experiment. We conclude that even if the computational complexity of their method is only $\mathcal{O}(nN)$, it required a much bigger n than in our case for the current experiment, and the results obtained are not as good.

5.3 A Projection on the Mass Preservation Constraint

The last algorithm we consider for the Monge-Kantorovich problem was introduced by Haber, Rehman and Tannenbaum [10]. Their idea is to start from an initial transformation which is not mass preserving and then by projecting on the Monge-Ampère equation, reach a map satisfying this condition. We can write this problem as follows:

$$\begin{cases} \min\limits_{T} \|T(x) - x\|^2_{L^2(d\mu)} \\ \text{s.t.} \quad g(T)\det(\nabla T) = f, \end{cases} \tag{5.3}$$

where $\|\cdot\|_{L^2(d\mu)}$ is the μ-weighted (or f-weighted) L_2 norm. There is one complication with this approach: global convergence of the projection process is not guaranteed (it could yield only a local minimum). To overcome this difficulty, they propose to use the fact that the optimal map needs to be the gradient of a convex function, i.e., $\check{T} = \nabla\Psi$. In that case, we know that the corresponding vector field has to be curl free: $\nabla \times \check{T} = 0$. Therefore, if we add this as an extra penalty term in (5.3), it will give a bias towards an irrotational solution without changing the value of the minimum. The resulting problem becomes

$$\begin{cases} \min\limits_{T} \dfrac{1}{2}\displaystyle\int_{\Omega} \Big(|T - x|^2 f(x) + \beta |\nabla \times T|^2 \Big)\, dx \\ \text{s.t.} \quad g(T)\det(\nabla T) = f, \end{cases} \tag{5.4}$$

where $\beta > 0$ is an initial penalty parameter. Then, to solve this constrained optimization problem, they write the corresponding Lagrangian with Lagrange multiplier

p:

$$L(T,p) = \frac{1}{2}\int_\Omega \Big(|T-x|^2 f(x) + \beta|\nabla \times T|^2\Big)\,dx + p\Big(g(T)\det(\nabla T) - f\Big).$$

In addition, instead of dealing directly with T, they chose to work with $U = T - x$ just like we did (we took $U = \nabla u = \nabla\Psi - x = T - x$). They argue that this choice, without affecting the curl, leads to numerically smaller perturbations.

To solve the problem, they used a version of the inexact Sequential Quadratic Programming method. The SQP algorithm roughly applies an iteration to solve a nonlinear problem in a way that the new iterate is generated by minimizing a quadratic approximation of L. More specifically, starting with the usual KKT conditions, we get the first order optimality conditions

$$\nabla L(T,p) = \begin{pmatrix} \nabla\Theta(T) + J(T)^{\mathrm{T}} p \\ c(T) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

where Θ is the objective function, c is the constraint function and J is the Jacobian of c. If we compute the Jacobian of this vector, we get the matrix

$$\begin{pmatrix} \nabla^2_{TT} L(T,p) & J(T)^{\mathrm{T}} \\ J(T) & 0 \end{pmatrix}. \tag{5.5}$$

Hence, in the context of an iterative process of the form

$$\begin{pmatrix} T_{n+1} \\ p_{n+1} \end{pmatrix} = \begin{pmatrix} T_n \\ p_n \end{pmatrix} + \begin{pmatrix} \delta_{T_n} \\ \delta_{p_n} \end{pmatrix},$$

the (steepest) descent direction is given by the solution of

$$\begin{pmatrix} \nabla^2_{TT} L(T_n,p_n) & J(T_n)^{\mathrm{T}} \\ J(T_n) & 0 \end{pmatrix} \begin{pmatrix} \delta_{T_n} \\ \delta_{p_n} \end{pmatrix} = \begin{pmatrix} -\nabla\Theta(T_n) - J(T_n)^{\mathrm{T}} p_n \\ -\,c(T_n) \end{pmatrix}. \tag{5.6}$$

The "Quadratic" part in SQP is due to the fact that this latter system is equivalent

to the first order optimality conditions for the problem

$$\begin{cases} \min\limits_{\delta_T} \ \dfrac{1}{2}\delta_T^{\mathrm{T}} \nabla^2_{TT} L(T_n, p_n)\, \delta_T + \nabla\Theta(T_n)^T \delta_T \\ \text{s.t. } J(T_n)\delta_T + c(T_n) = 0. \end{cases} \tag{5.7}$$

Indeed, if we identify δ_{T_n} with the optimal solution $\delta_T^\star$ and δ_{p_n} with the corresponding optimal Lagrange multiplier (say $\zeta^\star$) from which we subtract p_n ($\zeta^\star = p_{n+1} = p_n + \delta_{p_n}$), we get (5.6) back. Therefore, by finding the descent direction from (5.6), we are essentially solving (5.7).

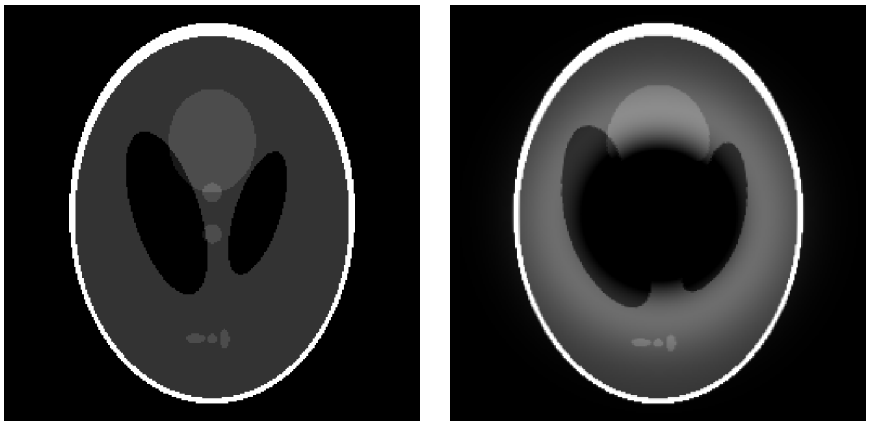
Unfortunately, some difficulties arise in practice: matrix (5.5) might not be explicitly available or system (5.6) might be expensive to solve. One way to deal with this is to use instead an approximation of (5.5) or an approximation of the system itself, depending on the problem in hand. In such a case, people refer to the SQP algorithm as the inexact SQP. If the reader wants to get more acquainted with this method and other related ones, we suggest looking at [20]. As we previously mentioned, Haber et al. employed an inexact SQP algorithm, approximating $\nabla^2_{TT} L(T_n, p_n)$. They also solved the linear system with the GMRES algorithm we presented earlier.

In their paper, the authors discretized the whole thing and proceeded to the analysis of the consistency and stability, which we will not repeat here as we only want to present an overview of the method. Let's just say that in order to make it stable, they add another penalty term to the Lagrangian of the form $-\gamma\,|\nabla p|^2/2$ where $\gamma = h^2$. This had the effect of adding an extra term in (5.5), replacing the zeroth term. On top of that, they also applied a few more numerical techniques to improve the performance of the procedure, such as the use of a block preconditionner to solve (5.6). For more details we refer to [10].

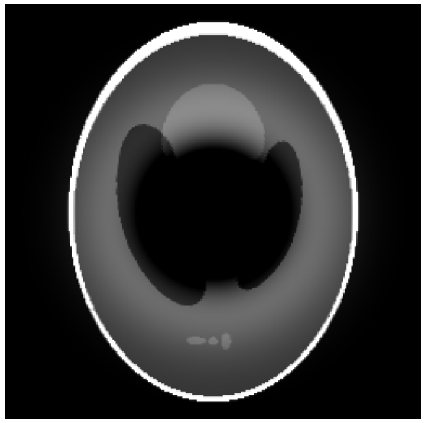
Once again, we try to reproduce one of the numerical experiments the authors presented with our own method. They selected a classic MATLAB test image depicting a phantom as an initial density. To create a target density, they considered a transport map in the form of a Gaussian distribution, which they applied to the phantom image. In our version of the experiment, we took more specifically

$$u(x, y) = Ce^{-0.5\left[(x-0.5)^2+(y-0.5)^2\right]/\sigma^2}$$

with parameters $C = 1/75$ and $\sigma = 0.12$. Taking $\tau = 2$, tol$= 10^{-1}$ and the FT implementation yields the results presented in Figure 5.3. After 10 iterations, we



(a) f (b) g



(c) f_{10}

Figure 5.3: The 256×256 pixels phantom image experiment with a nearest neighbor interpolation.

reached a point where the error in the l^∞ norm (using the same norm as the authors) would not significantly increase as we kept iterating, and we could not visually observe any difference between the target f and the computed f_n. We can also observe from Figure 5.4 the contour plot of the divergence of u_{10} which shows, if we rule out the noise, the effect of the Gaussian deformation. The number of GMRES iterations still stayed almost constant, close to only 6 inner ones per Newton iteration.

By taking a closer look at our results, we realize that the precision obtained for $\|u - u_n\|_{l^\infty}$ is not as good as theirs, probably because of the approximations we made. Indeed, if we try to compute the compositions $g(x + \nabla u_n)$ and $\nabla g(x + \nabla u_n)$ with linear interpolation instead of closest neighbor, we get better results (see Table 5.1 and 5.2). However, to reach such a precision or not does not really matter for the practical examples we presented previously since by that point we cannot distinguish

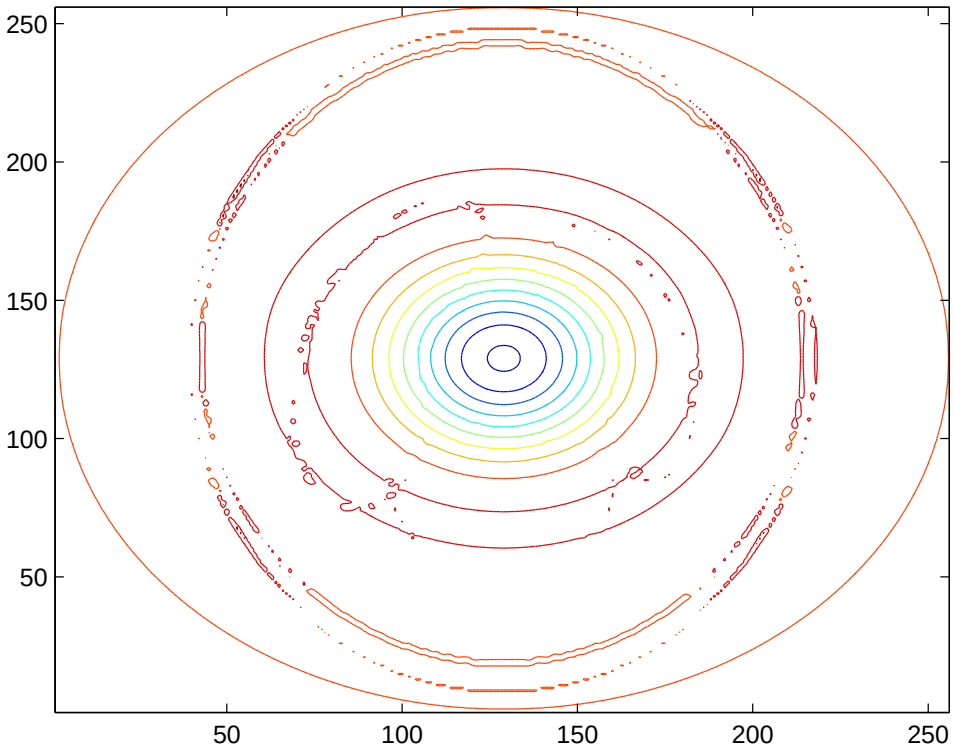


Figure 5.4: Contour plot of div(u_{10})

Grid Size	Nearest Neighbor	n_{nearest}	Linear	n_{linear}
8	5.15×10^{-04}	7	2.53×10^{-04}	8
16	4.06×10^{-04}	8	2.78×10^{-05}	11
32	1.99×10^{-04}	9	9.36×10^{-06}	15
64	7.14×10^{-04}	7	9.64×10^{-06}	16
128	4.74×10^{-04}	7	9.64×10^{-06}	15
256	1.25×10^{-04}	10	9.27×10^{-06}	15

Table 5.1: Different interpolation scenarios for $\|u-u_n\|_{l^\infty}$ with the number of Newton iterations required to reach the presented precision, n_{nearest} representing this number for the nearest neighbor interpolation case and n_{linear} being the one for the linear interpolation case.

Grid Size	$\|u - u_{n_{\text{proj}}}\|_{l^\infty}$	n_{proj}
8	1.1×10^{-04}	56
16	2.5×10^{-05}	32
32	6.0×10^{-06}	28
64	2.4×10^{-07}	27
128	5.9×10^{-08}	28
256	1.2×10^{-08}	26

Table 5.2: Haber et al.'s results for different grid sizes of the phantom experiment. Here, n_{proj} is the number of projections of their algorithm required to obtain a curl of the solution 4 times smaller than the initial one.

any difference visually. Nonetheless, this requires further investigation.

When it comes to the speed of the two methods, they both required a similar number of steps; about 15 Newton iterations for ours and about 25 for theirs. Haber et al. noticed just like we did that the number of GMRES iterations required to solve system (5.6) was almost mesh independent. Hence, the computational complexity of their method depends on the resolution of (5.6) and the way they propose to do this can also be done in $\mathcal{O}(P\text{log}P)$.

In conclusion, the main advantage of our algorithm over theirs is the fact that convergence of the method (before discretization) is guaranteed in theory, thanks to Theorem (2.4.1). In practice, they pushed the analysis further than we did and applied some stabilization techniques, which does not seem necessary here as our method appears to be stable in the FT case. Moreover, out of the many experiments we conducted, we were always able to obtain very good results only by varying τ, like the theory predicted. Recall also that the biggest τ we had to pick to obtain convergence was only 2 in the case of the FT implementation.

Conclusions

We presented in this work an efficient numerical method to solve the L^2 optimal transport problem via the Monge-Ampère equation in the case where the initial and final densities are periodic. It was created as a generalization of the one presented by Loeper and Rapetti [13] who treated the specific case of a uniform target density. Using a variant of the classical Newton algorithm, it employs a stepsize parameter τ for which we proved that it can be selected in a way that, given two densities regular enough, the method will be guaranteed to converge. This parameter also turns out to be crucial in order to control the possible degeneracy of the Monge-Ampère equation, ensuring that the sequence of approximate solutions consisted of uniformly convex functions.

We selected and compared two different two-dimensional implementations for the algorithm which differed only in the resolution process of the corresponding linearized Monge-Ampère equation required to give the direction of descent. The finite differences (FD) implementation, which was motivated by the work of Loeper and Rapetti, displayed a computational complexity close to $\mathcal{O}(P^{3/2})$ and did not seem stable. On the other hand, the Fourier transform (FT) implementation provided an $\mathcal{O}(P\mathrm{Log}P)$ method which appeared to be stable. Actually, in the numerical experiment we conducted, the latter outperformed the former on almost every point of comparison.

The two principal drawbacks of our algorithm did not turn out to be major hindrances for practical usage. First, even if we do not possess precise a priori knowledge of the value of the stepsize parameter τ required to make the algorithm converge, this value was never bigger than 2 in our examples for the FT implementation. The results were not as good in the FD case, where the maximum τ we had to select to obtain satisfactory results was 8. Recalling that the bigger the τ, the slower the algorithm, this is yet another argument in favor of the FT implementation. Next, in the context of image processing, the limitation to densities bounded away from 0 and to periodic boundary conditions did not seem to be a serious shortcoming for applying this al-

gorithm to important practical examples. Indeed, it performed well not only in the medical imaging experiment, but also in the cases of Lena to Tiffany warp and of the phantom deformation.

Concerning the medical imaging application, we point out that the algorithm was able to quickly detect the presence of the scars left by multiple sclerosis in both examples. In only 3 or 4 Newton iterations (with a τ equal to 1), the divergence of the transport map clearly displayed the location of these scars. This could provide a first step towards an efficient and fully automated disease diagnosis method based on optimal transport that could be applicable to several different kinds of illnesses like brain tumors. However, for this to happen, we would need to implement it in three dimensions. This would not be a problem in theory, since the method we developed is valid in any dimension. Moreover, both discretization techniques employed could be easily generalized to higher dimensions.

After comparing our algorithm against other ones available in the literature, we realized that it is very competitive. Indeed, Benamou and Brenier's use of the fluid dynamics reformulation of optimal transport introduces an extra time variable to the problem which is useful in some scenarios (when the transport path is required) but creates a non-necessary cost for the practical purposes presented in this work. The gradient descent on the dual problem introduced by Chartrand et al. while being easy to understand and implement, does not always produce satisfactory results (see for example the Lena to Tiffany warp). Finally, the projection algorithm on the mass preservation constraint of Haber et al. is as of now probably the best method to solve the L^2 optimal transport problem. Just like the method we developed here, it enjoys efficiency and stability properties. However our method is guaranteed to converge in theory (before discretization), thanks to Theorem 2.4.1. This is not necessarily guaranteed in their case. Note that we are not implying here that our algorithm is the best one available. What we mean is that there are practical situations where ours is a very good alternative (for example, when we are dealing with grayscale images).

To conclude, we mention several directions we would like to take as a sequel to this work. First, as we previously mentioned, we want to implement a three-dimensional version of our method to apply to more realistic examples in image processing. Furthermore, in a modern context, it would also be interesting to derive an implementation suitable for a parallel architecture in order to increase the performances even more . Finally, as we saw in Section 2, the estimates we were using also hold for more general cost functions, and therefore, it might be possible to generalize our technique

to encompass a wide range of costs. This could provide a method to effectively solve the general optimal transport problem associated with a wide range of cost functions.

Appendix A

Function Spaces and Norms

We shall briefly review here the main spaces of interest in this thesis, i.e. the spaces of differentiable functions $\mathcal{C}^m(\Omega)$ and the Hölder spaces $\mathcal{C}^{m,\alpha}(\Omega)$ with their usual norms. We start with $\mathbb{R}^d$, where we will employ the three following norms:

$$\|x\|_1 = \sum_{j=1}^{d} x_j \,,$$

$$\|x\|_2 = |x| = \sqrt{\sum_{j=1}^{d} x_j^2} \,,$$

$$\|x\|_\infty = \max(|x_1|, |x_2|, ..., |x_d|).$$

Note that we choose to associate $|x|$ with $\|x\|_2$ in d dimensions since it is the norm we are going to use the most. Next, let Ω be a closed subset of $\mathbb{R}^d$ and $m \in \mathbb{N}$. Let $\mathbb{I}_m$ be the set of multi-indices having a sum of m:

$$\mathbb{I}_m = \{i = (i_1, i_2, ..., i_d) \in \mathbb{N}^d : \|i\|_1 = m\}.$$

With this in mind, we write

$$D^i f = \frac{\partial^{\|i\|_1} f}{\partial x_1^{i_1} \partial x_2^{i_2} ... \partial x_d^{i_d}}$$

for $i \in \mathbb{I}_m$. Then, the set of functions $f : \Omega \to \mathbb{R}$ for which the partial derivatives D^i are continuous when $i \in \mathbb{I}_k$ and $0 \leq k \leq m$ is denoted by $\mathcal{C}^m(\Omega)$. It is equipped with

the norm

$$\|f\|_{\mathcal{C}^m(\Omega)} = \max_{0\leq\|i\|_1\leq m} \sup_{x\in\Omega} |D^i f(x)|.$$

In addition, we define the space of smooth functions on Ω to be

$$\mathcal{C}^\infty(\Omega) = \bigcap_{m=0}^{\infty} \mathcal{C}^m(\Omega).$$

To properly introduce Hölder spaces, we require the so-called Hölder coefficient of a function f:

$$[f]_{\alpha,\,\Omega} = \sup_{\substack{x,y\in\Omega \\ x\neq y}} \frac{|f(x)-f(y)|}{|x-y|^\alpha},$$

where $\alpha \in (0,1)$. Then, we define the Hölder space $\mathcal{C}^{0,\,\alpha}(\Omega)$ (or simply $\mathcal{C}^\alpha(\Omega)$) to be the space of functions $f \in \mathcal{C}^0(\Omega)$ for which $[f]_{\alpha,\,\Omega} < \infty$. In this case, we have the norm

$$\|f\|_{\mathcal{C}^\alpha(\Omega)} = \|f\|_{\mathcal{C}^0(\Omega)} + [f]_{\alpha,\,\Omega}\,.$$

More generally, $\mathcal{C}^{m,\,\alpha}(\Omega)$ is the set of all functions $f \in \mathcal{C}^m(\Omega)$ such that $[D^i f]_{\alpha,\,\Omega} < \infty$ for every $i \in \mathbb{I}_m$. The corresponding norm is

$$\|f\|_{\mathcal{C}^{m,\,\alpha}(\Omega)} = \|f\|_{\mathcal{C}^m(\Omega)} + \max_{i\in\mathbb{I}_m} [D^i f]_{\alpha,\,\Omega}\,.$$

Finally, we will require a few inequalities involving Hölder norms. They are presented in the following lemma:

Lemma A.0.1. *Let $\gamma \in \mathcal{C}^\alpha(\Omega)$, $\sigma \in \mathcal{C}^\alpha(\Omega)$ and $\phi \in \mathcal{C}^\beta(\Omega)$.*

1) *The product $\gamma\sigma \in \mathcal{C}^\alpha(\Omega)$ and*

$$\|\gamma\sigma\|_{\mathcal{C}^\alpha(\Omega)} \leq \|\gamma\|_{\mathcal{C}^\alpha(\Omega)} \|\sigma\|_{\mathcal{C}^\alpha(\Omega)}.$$

2) *If there exists a constant m_σ such that $0 < m_\sigma \leq \sigma$, then $1/\sigma \in \mathcal{C}^\alpha(\Omega)$.*

3) *The composition $\gamma \circ \phi \in \mathcal{C}^{\alpha\beta}(\Omega)$ and*

$$\|\gamma\circ\phi\|_{\mathcal{C}^{\alpha\beta}(\Omega)} \leq \|\gamma\|_{\mathcal{C}^\alpha(\Omega)} \left[1 + \left(\|\phi\|_{\mathcal{C}^\beta(\Omega)}\right)^\alpha\right].$$

Proof. **1)** For any x different than y;

$$\frac{|\gamma(x)\sigma(x)-\gamma(y)\sigma(y)|}{|x-y|^\alpha} = \frac{|\gamma(x)\sigma(x)-\gamma(y)\sigma(x)+\gamma(y)\sigma(x)-\gamma(y)\sigma(y)|}{|x-y|^\alpha}$$
$$\leq \frac{|\gamma(x)-\gamma(y)|}{|x-y|^\alpha}|\sigma(x)| + \frac{|\sigma(x)-\sigma(y)|}{|x-y|^\alpha}|\gamma(y)|$$

$$\Rightarrow \sup|\gamma\sigma| + \sup_{x\neq y}\frac{|\gamma(x)\sigma(x)-\gamma(y)\sigma(y)|}{|x-y|^\alpha} \leq \sup|\gamma\sigma|$$
$$+ [\gamma]_{\alpha,\Omega}\|\sigma\|_{\mathcal{C}^0(\Omega)} + [\sigma]_{\alpha,\Omega}\|\gamma\|_{\mathcal{C}^0(\Omega)}$$
$$\leq \|\gamma\|_{\mathcal{C}^\alpha(\Omega)}\|\sigma\|_{\mathcal{C}^\alpha(\Omega)}.$$

2) We know that there exists a positive constant k_σ such that

$$|\sigma(x)-\sigma(y)| \leq k_\sigma|x-y| \ \forall x \neq y \in \Omega.$$

Dividing both sides by $\sigma(x)$ and $\sigma(y)$, we get

$$\left|\frac{1}{\sigma(x)} - \frac{1}{\sigma(y)}\right| \leq \frac{k_\sigma}{m_\sigma^2}|x-y| \ \forall x \neq y \in \Omega,$$

hence the result.
3) If $\phi(x) \neq \phi(y)$, then

$$\frac{|\gamma\circ\phi(x)-\gamma\circ\phi(y)|}{|\phi(x)-\phi(y)|^\alpha} \leq [\gamma\phi]_{\alpha,\Omega}$$

Therefore, for any $x \neq y$;

$$\frac{|\gamma\circ\phi(x)-\gamma\circ\phi(y)|}{|x-y|^{\alpha\beta}} \leq \left(\sup_{x\neq y}\frac{|\gamma(x)-\gamma(y)|}{|x-y|^\alpha}\right)\left(\frac{|\phi(x)-\phi(y)|}{|x-y|^\beta}\right)^\alpha.$$

Note that the latter inequality holds for any value of $\phi(x)$ and $\phi(y)$. Using this, we

get:

$$\begin{aligned}
\|\gamma \circ \phi\|_{\mathcal{C}^{\alpha\beta}(\Omega)} &\leq \|\gamma\|_{\mathcal{C}^0(\Omega)} + [\gamma]_{\alpha,\Omega}\left([\phi]_{\beta,\Omega}\right)^{\alpha} \\
&\leq \|\gamma\|_{\mathcal{C}^{\alpha}(\Omega)} + \|\gamma\|_{\mathcal{C}^{\alpha}(\Omega)}\left([\phi]_{\beta,\Omega}\right)^{\alpha} \\
&\leq \|\gamma\|_{\mathcal{C}^{\alpha}(\Omega)}\left[1 + \left(\|\phi\|_{\mathcal{C}^{\beta}(\Omega)}\right)^{\alpha}\right]
\end{aligned}$$

which is the desired inequality. □

Appendix B

Convex analysis

The goal in this section is to refresh the reader's memory on a few basic facts concerning convex analysis (in the context of optimal transport). This way, we shall avoid any ambiguity that could arise from different definitions of the same concepts given by different authors. We first start with some definitions for convex sets.

Definition B.0.2. *A subset S of $\mathbb{R}^d$ is said to be convex if for every $x, y \in S$, the line segment linking x and y; $tx + (1-t)y$ for all $t \in [0,1]$, lies in S. Next, let $c \in \mathcal{C}^\infty(\mathbb{R}^d \times \mathbb{R}^d)$ and take S^* another subset of $\mathbb{R}^d$. In such a case, S is said to be c-convex with respect to S^* if for every $y \in S^*$, the image $\nabla_y c(\cdot, y)(S)$ is convex in $\mathbb{R}^d$. Conversely, S^* is said to be c^*-convex with respect to S if for every $x \in S$, $\nabla_x c(x, \cdot)(S^*)$ is convex in $\mathbb{R}^d$.*

Note that in the case of a quadratic cost function $c(x,y) = \frac{|x-y|^2}{2}$, or $c(x,y) = x \cdot y$, the concept of c-convexity goes back to the usual notion of convexity. Let's now state the convexity concepts we will require for functions:

Definition B.0.3. *Take S a convex subset of $\mathbb{R}^d$ and $f : S \to \mathbb{R}$. We say that*

1) *f is convex if for any $x, y \in S$ and all $t \in (0,1)$, we have*

$$f(tx + (1-t)y) \leq tf(x) + (1-t)f(y),$$

2) *f is strictly convex if for any $x \neq y \in S$ and all $t \in (0,1)$, we have*

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y),$$

3) *f is uniformly convex if there exists a constant c such that for any $x \neq y \in S$ and all $t \in (0,1)$, we have*

$$f(tx + (1-t)y) < tf(x) + (1-t)f(y) - \frac{1}{2}ct(1-t)|x-y|^2.$$

Motivated by the third case and all the constants c making f uniformly convex, we define a function $\mathcal{K}_S(f)(x,\rho) : \mathbb{R}^d \times \mathbb{R}_+ \to \mathbb{R}$, called the modulus of convexity, by taking

$$\mathcal{K}_S(f)(x,\rho) = \inf_{y \in S} \left\{ \Theta_f(x,y,t) : t \in (0,1),\ |y-x| = \rho \right\} \tag{B.1}$$

where

$$\Theta_f(x,y,t) = \frac{tf(y) + (1-t)f(x) - f(ty + (1-t)x)}{t(1-t)}.$$

Observe that $\mathcal{K}_S(f)(x,\rho)$ measures in some sense the lack of linearity of a function. It is also guaranteed to be non-negative if f is convex. When the function f is $\mathcal{C}^2(S)$, we can link the convexity concepts to the "positivity" of the Hessian matrix of f. Let's see what we mean by that, but first we need some extra definitions.

Definition B.0.4. *Let A be an $d \times d$ symmetric matrix with real coefficients. We say that A is*

1) *positive semidefinite if $\xi^T A \xi \geq 0\ \forall \xi \in \mathbb{R}^d$,*

2) *positive definite if $\xi^T A \xi > 0\ \forall \xi \in \mathbb{R}^d$,*

3) *uniformly positive definite if there exists a constant $k > 0$ such that $\xi^T A \xi \geq k|\xi|^2\ \forall \xi \in \mathbb{R}^d$.*

One convenient way of verifying if a matrix is positive definite or not is by looking at its principal minors. Indeed, it is positive definite if and only if all the principal minors are positive. As a consequence, the determinant of such a matrix has to be greater than 0, and therefore, it is invertible. Finally, here is the relationship between convexity and positivity:

Theorem B.0.5. *Let S be an open convex subset of $\mathbb{R}^d$ and f be a twice continuously differentiable function on S. Consider H_f the Hessian matrix of f. Then*

1) *f is convex if and only if H_f is positive semidefinite.*

2) *f is strictly convex if and only if H_f is positive definite.*

1) *f is uniformly convex if and only if H_f is uniformly positive definite.*

For further references on the convexity concepts, we refer to [40].

Bibliography

[1] J.B.Bruckner A.M.Bruckner and B.S.Thomson. *Real Analysis*. Prentice-Hall, 1997.

[2] A.Oberman. Wide stencil finite difference schemes for the elliptic Monge-Ampre equation and functions of the eigenvalues of the Hessian. *Discrete and Continuous Dynamical Systems series B*, 10(1):221–238, 2008.

[3] A.Oberman and B.Froese. Fast finite difference solvers for singular solutions of the elliptic Monge-Ampre equation. *J. Comput. Phys.*, 230(3):818–834, 2011.

[4] A.Wouk. *A Course of Applied Functional Analysis*. Wiley-Interscience, 1979.

[5] J-D Benamou and Y.Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84:375–393, 2000.

[6] C.Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[7] C.Villani. *Optimal Transport:Old and New*. Springer-Verlag, 2009.

[8] D.Cordero-Erausquin. Sur le transport de mesures périodiques. *C. R. Acad. Sci. Paris*, I(329):199–202, 1999.

[9] D.Gilbarg and N.S.Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, 2001.

[10] T.Rehman E.Haber and A.Tannenbaum. An efficient numerical method for the solution of the l^2 optimal mass transfer problem. *SIAM J. Sci. Comput.*, 32(1):197–211, 2010.

[11] F.Santambrogio. Models and applications of optimal transport in economics, traffic and urban planning. `http://arxiv.org/abs/1009.3857`, preprint 2010.

[12] G.A.Pavliotis and A.M.Stuart. *Multiscale Methods: Averaging and Homogenization*, volume 53 of *Texts in Applied Mathematics*. Springer, 2008.

[13] G.Loeper and F.Rapetti. Numerical solution of the Monge-Ampère equation by a Newton's algorithm. *C. R. Acad. Sci. Paris*, I(340):319–324, 2005.

[14] G.Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris, avec les Mémoires de Mathématique et de Physique pour la même année*, pages 666–704, 1781.

[15] G.H. Golub and C.F. Van Loan. *Matrix computations*. JHU Press, 1996.

[16] J. B. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms: Fundamentals, Volume 1*. Springer-Verlag, 1996.

[17] H.M.Yin. On a p-Laplacian type of evolution system and applications to the Bean model in the type-II superconductivity theory. *Quarterly of Applied Mathematics*, LIX(1):47–66, 2000.

[18] J.Jost. *Partial Differential Equations*. Graduate Texts in Mathematics. Springer, 2007.

[19] N.S.Trudinger J.Liu and X.J.Wang. Interior $\mathcal{C}^{2,\alpha}$ regularity for potential functions in optimal transportation. *Comm. Part. Diff. Eq.*, 35:165–184, 2010.

[20] J.Nocedal and S.Wright. *Numerical Optimization*. Springer Series in Operations Research. Springer-Verlag, 1999.

[21] J.Strain. Fast spectrally-accurate solution of variable-coefficients elliptic problems. *Proc. of the AMS*, 122(3):843–850, 1994.

[22] K.B.Peterson and M.S.Pederson. The matrix cookbook. `http://matrixcookbook.com`, 2008.

[23] L.A.Caffarelli. Monotonicity of optimal transportation and the fkg and related inequalities. *Commun. Math. Phys.*, 214:547–563, 2000.

[24] L.C.Evans. *Partial Differential Equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, 1998.

[25] R.J. LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*. SIAM, 2007.

[26] L.Forzani and D.Maldonado. Properties of the solutions to the Monge-Ampère equation. *Nonlinear Analysis*, 57(5–6):815–829, 2004.

[27] L.V.Kantorovich. On a problem of Monge. *Uspekhi Mat. Nauk.*, 3:225–226, 1948.

[28] M.Agueh. Rates of decay to equilibria for p-Laplacian type equations. *Nonlinear Analysis*, 68(7):1909–1927, 2008.

[29] M.Cullen and W.Gangbo. A variational approach for the 2-d semi-geostrophic shallow water equations. *Arch. Rat. Mech. and Anal.*, 156:241–273, 2001.

[30] M.Fortin and R.Glowinski. *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, volume 15 of *Studies in Mathematics and its Applications*. Elsevier, 1983.

[31] N.S.Trudinger and X.J.Wang. On the second boundary value problem for Monge-Ampère type equations and optimal transportation.

[32] McConnell Brain Imaging Centre (BIC) of the Montreal Neurological Institute. Brainweb simulated brain database. `http://www.bic.mni.mcgill.ca/brainweb/`.

[33] P.Linz. *Theoretical Numerical Analysis: An Introduction to Advanced Techniques*. Dover publications, 2001.

[34] R.A.Horn and C.R.Johnson. *Matrix Analysis*. Cambridge University Press, 1985.

[35] K.Vixie R.Chartrand, B.Wholberg and E.Bollt. A gradient descent solution to the Monge-Kantorovich problem. *Applied Mathematical Sciences*, 3(21–24):1071–1080, 2009.

[36] E.Pichon S.Angenent and A.Tannenbaum. Mathematical methods in medical image processing. *Bull. of the AMS*, 43:365–396, 2006.

[37] G.Pryor J.Melonakos T.Rehman, E.Haber and A.Tannenbaum. 3d nonrigid registration via optimal mass transport on the gpu. *Medical Image Analysis*, 13:931–940, 2009.

[38] R.Mohayaee U.Frisch, S.Matarrese and A.Sobolevski. A reconstruction of the initial conditions of the universe using optimal transportation. *Nature*, 417:260–262, 2002.

[39] J. Urbas. On the second boundary value problem for equations of Monge-Ampère type. *J. Reine Angew Math.*, 487:115–124, 1997.

[40] W.Sun and Y.X.Yuan. *Optimization Theory and Methods: Nonlinear Programming*. Optimization and its Applications. Springer, 2006.

[41] Y.Brenier. Polar factorization and monotone rearrangements of vector valued functions. *Comm. Pure Appl. Math.*, 44:375–417, 1991.

[42] Y.Saad. *Iterative methods for sparse linear systems, Second Edition*. SIAM, 2003.