PSL UNIVERSITÉ PARIS

THÈSE DE DOCTORAT
DE L'UNIVERSITÉ PSL

Préparée à Université Paris Dauphine

Finite volume approximation of optimal transport and Wasserstein gradient flows

Soutenue par
Gabriele TODESCHI
Le 13/12/2021

École doctorale n°543
L'École Doctorale SDOSE

Spécialité
Mathématiques

Composition du jury :

Jean-David BENAMOU Directeur de recherche, UNIVERSITÉ PARIS DAUPHINE - INRIA	*Directeur*
Clément CANCÈS Directeur de recherche, INRIA	*Co-encadrant*
Thomas GALLOUËT Chargé de recherche, INRIA	*Co-encadrant*
Marie-Therese WOLFRAM Professor, UNIVERSITY OF WARWICK	*Rapporteur*
Giuseppe BUTTAZZO Professor, UNIVERSITÀ DI PISA	*Rapporteur*
Quentin MÉRIGOT Professeur des universités, UNIVERSITÉ PARIS-SACLAY	*Président du jury*
Virginie EHRLACHER Maître de conférences, ÉCOLE DES PONTS PARISTECH	*Membre du jury*
Daniel MATTHES Professor, TECHNISCHE UNIVERSITÄT MÜNCHEN	*Membre du jury*

Dauphine | PSL UNIVERSITÉ PARIS

Preface

Motivations

Many physical, biological, socio-economical problems can be written as Wasserstein gradient flows, that is curves of steepest descent in the Wasserstein space. A curve of steepest descent is an evolution that starting from an initial configuration evolves in time maximizing the rate of decrease of a given energy, i.e. decreasing it as fast as possible. Different forms of the energy generate different dynamics. The knowledge that a problem presents this particular structure can be beneficial for the design of a numerical scheme to approximate its solution. The objective of this work is indeed to propose approaches for solving these type of problems, preserving and exploiting their structure in order to have reliable and robust schemes.

The notion of gradient flow depends on the notion of metric space. The easiest example is the case of a finite dimensional gradient flow in $\mathbb{R}^d$, equipped with the euclidean distance, with respect to a real valued function $F : \mathbb{R}^d \to \mathbb{R}$. In this very simple setting, given a starting point $\boldsymbol{x}^0 \in \mathbb{R}^d$, a gradient flow is defined as the solution to the following Cauchy problem:

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = -\nabla F(\boldsymbol{x}(t))\,, \quad t \geq 0\,, \\ \boldsymbol{x}(0) = \boldsymbol{x}^0\,. \end{cases} \tag{1}$$

The vector field is given by the opposite of the gradient of the function F, which clearly motivates its name. The definition (1) of gradient flow makes perfectly sense also in the infinite dimensional case whenever the underlying space is of Hilbert type. Nevertheless, the notion of gradient flow can be extended to setting where the metric space does not have this geometrical structure, as it is the case for the Wasserstein space. Another definition is necessary however in these cases, where the time derivative and the gradient operator appearing in equation (1) do not make sense.

Although always sharing the same principle, there are different definitions of gradient flows, more or less suited depending on the notion of metric space or the specific energy considered. If the space has sufficient differential structure, as it is the case for the Wasserstein space, one can again resort to a characterization close to (1), which we could describe in some sense as the optimality conditions of the variational problem. Given a real valued energy functional $\mathcal{E}$ and an initial condition ρ^0, we can characterize a Wasserstein gradient flow with respect to $\mathcal{E}$ as the solution to the following partial differential equation:

$$\begin{cases} \partial_t \varrho - \nabla \cdot (\varrho \nabla \frac{\delta \mathcal{E}}{\delta \rho}[\varrho]) = 0 & \text{in } [0,T) \times \mathring{\Omega}\,, \\ \varrho \nabla \frac{\delta \mathcal{E}}{\delta \rho}[\varrho] \cdot \boldsymbol{n} = 0 & \text{on } [0,T) \times \partial\Omega\,, \\ \varrho(0,\cdot) = \rho^0 & \text{in } \Omega\,, \end{cases} \tag{2}$$

where $\Omega \subset \mathbb{R}^d$ is a convex and compact domain, $T \in \mathbb{R}_+$ is a time horizon. Equation (2) expresses the continuity equation for a time evolving density ϱ, starting from the initial condition ρ^0, convected by the velocity field $-\nabla \frac{\delta \mathcal{E}}{\delta \rho}[\varrho]$. It is complemented with a condition of zero flux across the boundary of the domain: the total mass, defined as the space integral of the density on the whole domain, is therefore preserved.

It is now well understood since the pioneering works of Otto [73, 107, 108] that equations of the form of (2) can be interpreted as the gradient flow in the Wasserstein space with respect to the energy $\mathcal{E}$. The first problem that has been shown having this structure is the linear Fokker-Planck equation [73, 25]. Beside this, many problems have been proven to exhibit the same variational structure. Porous media flows [108, 78, 34], magnetic fluids [107], superconductivity [5, 4], crowd motions [94], aggregation processes in biology [42, 23], semiconductor devices modelling [76], or multiphase mixtures [37, 72] are just few examples of problems that can be represented as gradient flows in the Wasserstein space. Designing efficient numerical schemes for approximating their solutions is therefore a major issue and our leading motivation.

There exist already numerous numerical schemes to solve problems of the form of (2). Nevertheless, the hidden variational structure is usually disregarded. The key principle that the evolution should decrease (as fast as possible) the energy is hardly enforced. The design of discretizations that, starting from the interpretation of problem (2) as steepest descent curve, preserve and exploit this structure can be extremely beneficial, both for robustness and reliability. Approaches in this direction have already been proposed recently (e.g., [17, 41]). However, the numerical analysis, the efficiency and the flexibility have not always been considered. The spirit of this work is also to push forward these aspects.

Content of this work

In this work we want to construct robust, reliable and efficient structure preserving discretizations of problems of the form of equation (2). The time discretization is based on variational approaches that mimic at the discrete (in time) level the behavior of steepest descent curves. The space discretization is based on the Finite Volume Method, a well-known methodology particularly suited for the discretization of partial differential equations that present a conservative structure as (2). We use in particular Two-Point Flux Approximation (TPFA) finite volumes, a simple, yet very flexible, discretization. We insist on designing schemes which preserve at the ultimate discrete level the variational structure of the problem: that is, we will always follow a first discretize then optimize approach. We will present first order and second order accurate schemes, in both time and space. The variational structure is linked with the notion of Wasserstein distance, a complex optimization problem. We will present also a deep analysis of the computation of this complex object.

In Chapter 1 we show in which sense problems of the form of (2) can be interpreted as steepest descent curves of the energy functional $\mathcal{E}$. This interpretation is based on the theory of optimal transport. We introduce the (quadratic) optimal transport problem and briefly characterize its solutions. We will insist in particular on its dynamical formulation, which is of major interest for us. We finally introduce the finite volume methodology.

In Chapter 2 we focus on the discretization with finite volumes of the quadratic optimal transport problem in the Benamou-Brenier dynamical form. We expose some stability issues related to this discretization of the problem and propose a possible solution based on nested

meshes. To validate our approach, we present some convergence results that will be verified numerically. We also introduce and analyze the interior point strategy we employ to solve the discrete optimization problem. In Appendix A we present another discretization, always in the framework of TPFA schemes, which enables to preserve at the discrete level the monotonicity of the continuous operators.

In Chapter 3 we present a first order accurate variational scheme to solve Wasserstein gradient flows. The time discretization is based on the Minimizing Movement Scheme (MMS) introduced by the De Giorgi [48]. In order to decrease the computational complexity, we approximate the complex Wasserstein distance involved in the scheme with a weighted H^{-1} norm. We show numerically that this strategy preserves the accuracy of the MMS. For the space discretization, we rely on upwind finite volumes, a specific instance of TPFA discretization. This choice allows to solve directly the problem with an efficient Newton scheme. To validate our approach, we show the convergence of the scheme to distributional solutions of the Fokker-Planck equation. We also present several numerical simulations of various problems that exhibit a Wasserstein gradient flow structure.

In Chapter 4 we deal with the problem of providing a second order accurate discretization in space for the approach we present in Chapter 3. We rely on the interior point technique we will present in Chapter 2 to solve the discrete problem. This allows us to be more flexible in the space discretization and use a more accurate TPFA strategy. We verify numerically the second order accuracy in space. We finally present and test a more precise, yet not second order in time, scheme based on a modification of the MMS time discretization which has been recently proposed in [41].

In Chapter 5 we construct a second order accurate scheme in both time and space. As the MMS scheme is an order one discretization in time, we introduce a new approach, a modification of the variational BDF2 scheme introduced in [93]. We prove that this new time discretization converges to distributional solutions of the Fokker-Planck equation, which shows its consistency. We show that it is also possible to recover gradient flows in the EVI sense. Thanks to this modified BDF2, we are able to propose a fully discrete scheme. We will rely again on the space discretization and the optimization strategy introduced in Chapter 4. We show numerically that this scheme is second order accurate.

Finally, in Appendix B, we present another possible space discretization strategy for the dynamical optimal transport problem based on conservative finite elements. This type of discretization is intimately related to the finite volume one. Moreover, it fits naturally the framework of the convergence proof designed in [79] for general discretization of optimal transport, as we will show. We use in this case a primal-dual technique to solve the discrete optimization problem, an optimization approach which is more common in the optimal transport community.

This work is mainly based on the three published papers [32, 104, 105], which are respectively presented in Chapter 3, Chapter 4 and Chapter 2. The presentation of these works is not chronological but follows a conceptual order. The latter two works in particular had been here extended. The content of Chapter 5 has not been, at the moment, submitted. Finally, the presentation in Appendix B is issued from [103] which is currently under review.

Acknowledgements

First of all, I want to thank Jean-David Benamou for accepting me in his team, the Mokaplan. Thanks to all its members for the welcoming atmosphere, the interesting discussions, the advices, the moments of relax and, of course, the beers we had together. Thanks to Derya Gök for the assistance she provided me. Thanks to the FSMP for the opportunity they gave me through the Cofund MathinParis progam, and in particular to Ariela Briani, who constantly took care of it and all of us candidates. Thanks also to the University of Paris Dauphine and Inria for hosting me during these three years. Thanks to all the members of the jury for accepting to read and evaluate my work.

Thanks to my supervisors, Clément Cancès and Thomas Gallouët, for guiding me in this journey. I hope in the future there will be the chance to continue working together. Clément, in spite of the distance you have always been there in the key moments. Thomas, I would have loved to learn more from you. Still, you taught me what I consider the most valuable lesson: work is not all that matters.

Thanks to my colleague Andrea Natale. I came to you with all my problems, you helped me out and decided to work with me. You have been essential in my thesis.

Thanks to my former supervisor Mario Putti, for hosting me during my visiting period in Padova but more importantly for his mentorship during the master in Mathematical Engineering. When I was struggling to find my way, you advised me and pushed me. I will always be grateful for this.

Thanks to all my fellow PhDs for sharing the difficulties of this journey, for the mathematical discussions and the comforting exchanges. I realized that a PhD is also a process of self-acceptance and self-confidence that everyone goes through, and it is simpler doing it together.

Thanks to all the friends that I found in all these years and that have been close to me. This PhD has been the most stressful and challenging period of my life, yet you made it also the happiest one.

Thanks to my family, for providing me with all the means and the support to pursue my objectives from the very beginning of my academic career, for the values they taught me, for their wisdom and love. I owe you everything.

Finally, thanks to the most important person, Silvia. In all these years you stayed by my side, comforting me and reassuring me, pushing me and cheering for me, sometimes also putting up with me. Our relationship is my biggest achievement. I love you.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 754362.



Finite volume notation

- K, L: cells, control volumes;
- σ: face;
- $K|L$: face common to K and L;
- $\boldsymbol{x}_K$: cell center of K;
- Δ_σ: diamond cell with vertices $\boldsymbol{x}_K, \boldsymbol{x}_L$ and the vertices of $\sigma = K|L$;
- $\mathcal{T}$: set of all cells;
- $\overline{\Sigma}$: set of all faces;
- $\big(\mathcal{T}, \overline{\Sigma}, (\boldsymbol{x}_K)_{K\in\mathcal{T}}\big)$: finite volume mesh;
- Σ: set of internal faces;
- Σ_{ext}: set of boundary faces, $\overline{\Sigma} \setminus \Sigma$;
- $\overline{\Sigma}_K$: set of faces of the cell K;
- Σ_K: set of internal faces of the cell K, $\overline{\Sigma}_K \cap \Sigma$;
- $\mathcal{N}_K$: neighboring cells of the cell K;
- m_K: Lebesgue measure of the cell K;
- m_σ: $(d-1)$-dimensional Lebesgue measure of the face σ;
- d_σ: Euclidean distance between $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ for $\sigma = K|L$;
- $d_{K,\sigma}$: Euclidean distance between $\boldsymbol{x}_K$ and σ;
- m_{Δ_σ}: Lebesgue measure of the diamond cell Δ_σ;
- a_σ: transmissivity of the face σ, $\frac{m_\sigma}{d_\sigma}$;
- $h_\mathcal{T}$: meshsize, maximum diamater among all cells K;

Contents

Preface i

Acknowledgements v

Finite volume notation vii

1 Prerequisites 1
- 1.1 Optimal transport and the Wasserstein distance . . . 1
  - 1.1.1 Generalities on optimal transport . . . 2
  - 1.1.2 Dynamical formulation . . . 6
  - 1.1.3 The Wasserstein space . . . 13
- 1.2 Wasserstein gradient flows . . . 16
  - 1.2.1 Generalized Minimizing Movement . . . 17
  - 1.2.2 Dynamical form of the JKO step . . . 20
- 1.3 Finite Volume Method . . . 23
  - 1.3.1 The discretization of Ω . . . 23
  - 1.3.2 Discrete continuity equation . . . 25
  - 1.3.3 Discrete spaces and operators . . . 26

2 Computation of optimal transport with finite volumes 29
- 2.1 Introduction . . . 29
  - 2.1.1 Discretization of dynamical optimal transport . . . 30
  - 2.1.2 Numerical solution . . . 31
- 2.2 Finite volume discretization . . . 32
  - 2.2.1 Nested meshes . . . 32
  - 2.2.2 Discrete spaces and operators . . . 33
- 2.3 Time discretization . . . 34
- 2.4 Discrete optimal transport problem . . . 36
- 2.5 Convergence to the continuous problem . . . 40
- 2.6 Primal-dual barrier method . . . 47
- 2.7 Numerical results . . . 52
  - 2.7.1 Oscillations . . . 53
  - 2.7.2 Convergence test . . . 55
  - 2.7.3 Geodesic . . . 58
- 2.8 Perspectives . . . 58

3 A variational finite volume scheme for Wasserstein gradient flows 63
3.1 Introduction . . . 63
3.1.1 JKO semi-discretization . . . 64
3.1.2 Implicit linearization of the Wasserstein distance and LJKO scheme . . . 65
3.1.3 Goal and organisation of the chapter . . . 67
3.2 A variational Finite Volume scheme . . . 67
3.2.1 Upstream weighted dissipation potentials . . . 68
3.2.2 A variational upstream mobility finite volume scheme . . . 70
3.2.3 Comparison with the classical backward Euler discretization . . . 74
3.3 Convergence in the Fokker-Planck case . . . 75
3.3.1 Some a priori estimates . . . 77
3.3.2 Compactness of the approximate solution . . . 81
3.3.3 Convergence towards a weak solution . . . 85
3.4 Numerical results . . . 87
3.4.1 Newton method . . . 87
3.4.2 Fokker-Planck equation . . . 88
3.4.3 Other examples of Wasserstein gradient flows . . . 91

4 Centered TPFA finite volume discretization for Wasserstein gradient flows 97
4.1 Introduction . . . 97
4.1.1 LJKO time discretization . . . 98
4.2 A second order discretization in space . . . 99
4.3 A different time approach . . . 103
4.4 Interior point strategy . . . 105
4.5 Numerical results . . . 106
4.6 Concluding remarks . . . 108

5 A modified BDF2 scheme for Wasserstein gradient flows 111
5.1 Introduction . . . 111
5.1.1 Existing variations of the JKO scheme . . . 111
5.1.2 A new formulation for the BDF2 . . . 113
5.1.3 Organization of the chapter . . . 114
5.2 Analysis of the modified BDF2 scheme . . . 115
5.2.1 Well-posedness and main properties of the scheme . . . 115
5.2.2 Convergence towards the Fokker-Planck equation . . . 118
5.2.3 Convergence in the EVI sense . . . 124
5.3 Finite volume discretization . . . 126
5.3.1 Extrapolation in the viscosity sense . . . 127
5.3.2 A second order finite volume scheme . . . 128
5.3.3 Other implementations . . . 131
5.4 Numerical validation of the modified BDF2 approach . . . 132
5.4.1 Comparison between the three approaches . . . 132
5.4.2 Convergence tests . . . 134

A A monotone discretization for the computation of geodesics **139**
A.1 Discrete setting . . . 139
A.2 Monotone discretization . . . 140

B A mixed finite element discretization of dynamical optimal transport **145**
B.1 Introduction . . . 145
B.1.1 Contributions and structure of the chapter . . . 147
B.2 Notation . . . 148
B.3 Dynamical formulation of optimal transport . . . 148
B.3.1 Hilbert space setting and proximal splitting . . . 149
B.4 Mixed finite element setting . . . 150
B.4.1 Finite element spaces on D . . . 150
B.4.2 Finite element spaces on Ω . . . 152
B.4.3 Discrete projection on the divergence-free subspace . . . 153
B.5 Discrete dynamical formulation and convergence . . . 154
B.6 The proximal splitting algorithm . . . 156
B.7 Regularization . . . 158
B.7.1 Mixed L^2-Wasserstein distance . . . 158
B.7.2 H^1 regularization . . . 159
B.8 Numerical results . . . 160
B.8.1 Qualitative behaviour and convergence of the proximal-splitting algorithm 160
B.8.2 Non-convex domain . . . 160
B.9 Proof of the convergence theorem . . . 162

Chapter 1

Prerequisites

In this chapter we will present the notions and the fundamental language needed to read and interpret this work.

1.1 Optimal transport and the Wasserstein distance

Optimal transport is a mathematical theory that deals with the very natural problem of finding the optimal way of reallocating an initial configuration of mass into a final one, minimizing a total cost of displacement. It is a very old problem originally introduced by Monge [100], who wondered what was the optimal way of reallocating sand piles. Starting from the work of Kantorovich [75], who recast it into a suitable mathematical framework, it developed considerably in the past decades. Despite the initial extremely practical taste, it became an incredible tool in many theoretical fields. In the last years optimal transport experienced a new growth particularly thanks to the development of numerical methods. The possibility to compute approximate solutions made this theory much more appealing. It finds nowadays application in many different fields, among which socio-economical problems, data science and machine learning, but also physics and other branches of mathematics, such as the theory of partial differential equations (PDEs), variational inequalities or probability theory.

Several features contribute to make optimal transport particularly useful. First of all, it can be applied to more general objects than simply mass distributions, as long as the reallocation problem can be formulated. Secondly, it provides a notion of distance as the cost for the optimal reallocation. In this way it is possible to conceive distances between objects that otherwise would be difficult to compare. Furthermore, if the transport cost takes into account the horizontal displacement, it resembles an intuitive notion of distance, in contrast to more classical L^p distances for example which measure the vertical displacement. Finally, it carries along a natural notion of interpolation, as we shall see later. These features explain in particular the many connections of optimal transport with physical problems, as it is the case for example of Wasserstein gradient flows that we are analyzing specifically in this work.

The most natural way of evaluating the total transport cost is to assign to each single displacement a unitary cost and then sum all the contributions. If this unitary cost is taken to be the squared euclidean distance between points in space, then the problem is referred to as the quadratic or the L^2 optimal transport problem. The square root of the total cost of displacement is called in this case the quadratic Wasserstein distance. In the following, we

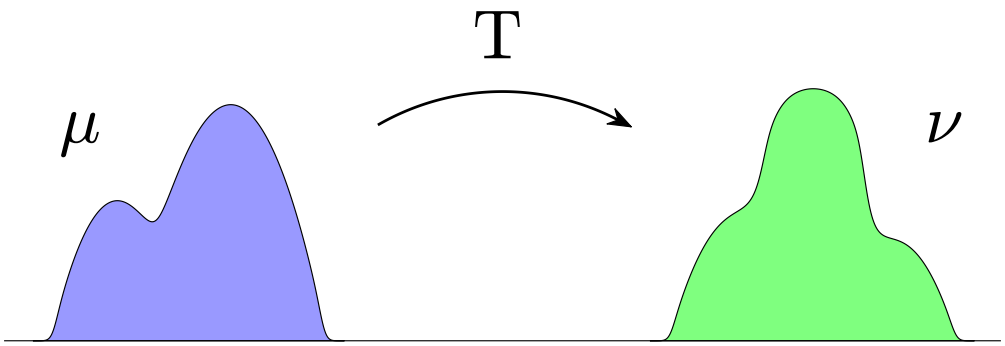


Figure 1.1: The problem of transporting an initial configuration of mass μ into a final one ν.

will quickly present the L^2 optimal transport problem. We will insist on its fluid dynamics formulation proposed by Benamou and Brenier, which is better suited for the purposes of this work, and the notion of geodesic interpolation. We will finally characterized what we refer to as the Wasserstein space.

Our aim is to introduce the reader to the concepts and the intuitions behind them and we will keep for this reason the presentation simple and formal. We will work in a very specific and simplified setting since the general theory will not be needed in this work. We refer to the monographs [115, 120, 121, 3, 111] for a general and precise presentation of the topics we will introduce and for a broad overview of the features and the advantages of using this theory in applications.

1.1.1 Generalities on optimal transport

Let us start by fixing the domain Ω to be a convex and compact subset of $\mathbb{R}^d$. We consider the space $\mathcal{P}(\Omega)$ of probability measures defined on Ω, subset of the topological dual space of the space of continuous function $C^0(\Omega)$ defined on Ω. We recall that probability measures are positive measures with fixed total mass equal to one, although the constant is not essential. When considering absolutely continuous measures in $\mathcal{P}(\Omega)$, that is measures that admit a density function $\rho \in L^1(\Omega; \mathbb{R}^+)$ for which it holds

$$\forall D \subset \Omega,\ \mu(D) = \int_D \rho(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}\,,$$

we will most of the time refer, throughout this work, directly to its density function, stating for example $\rho \in \mathcal{P}(\Omega)$, following a classical abuse of notation.

Given two measures $\mu, \nu \in \mathcal{P}(\Omega)$, we say that ν is the pushforward of μ through a map T if

$$\forall K \subset \Omega, \quad \nu(K) = \mathrm{T}_{\#}\mu(K) = \mu(\mathrm{T}^{-1}(K))\,, \tag{1.1}$$

or equivalently if

$$\int_\Omega f(\boldsymbol{y})\, \mathrm{d}\nu(\boldsymbol{y}) = \int_\Omega f(\mathrm{T}(\boldsymbol{x}))\, \mathrm{d}\mu(\boldsymbol{x}), \quad \forall f \in C^0(\Omega)\,. \tag{1.2}$$

With $\Pi(\mu,\nu)$ we denote the space of admissible couplings γ between μ and ν, that is the subset of probability measures defined on the product space $\Omega\times\Omega$ which verify the two marginal constraints

$$(\pi_1)_{\#}\gamma=\mu\,,\quad (\pi_2)_{\#}\gamma=\nu\,, \tag{1.3}$$

where π_1,π_2 stand for the canonical projections on the first and second coordinate of the product space $\Omega\times\Omega$. The two conditions can be equivalently stated as:

$$\int_{\Omega\times\Omega} f(\boldsymbol{x})\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})=\int_\Omega f(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x})\,,\quad \int_{\Omega\times\Omega} g(\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})=\int_\Omega g(\boldsymbol{y})\mathrm{d}\nu(\boldsymbol{y})\,,\quad \forall f,g\in C^0(\Omega)\,. \tag{1.4}$$

An element $\gamma\in\Pi(\mu,\nu)$ is called a coupling, or a transport plan, with respect to the two reference measures, as it prescribes for each couple of points $(\boldsymbol{x},\boldsymbol{y})$ how much mass contained in $\boldsymbol{x}$ of the measure μ should be assigned to $\boldsymbol{y}$, and vice versa.

The optimal transport problem and the Brenier solution

We are ready to state the optimal transport problem, in the Kantorovich formulation, between $\mu,\nu\in\mathcal{P}(\Omega)$. For a continuous cost function $c\in C^0(\Omega\times\Omega)$, which represents for each point $(\boldsymbol{x},\boldsymbol{y})$ the unitary cost of displacement from $\boldsymbol{x}$ to $\boldsymbol{y}$, the problem writes as

$$\inf_{\gamma\in\Pi(\mu,\nu)}\int_{\Omega\times\Omega} c(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})\,, \tag{1.5}$$

that is we want to find the transport plan between μ and ν that minimizes the total displacement cost. The cost function which has been considered and studied the most due the specific mathematical features it expresses, and to its many links and applications in physical problems and more, is the euclidean cost: $c(\boldsymbol{x},\boldsymbol{y})=|\boldsymbol{x}-\boldsymbol{y}|^2$. This is the cost we will consider. Originally, Monge considered the cost $c(\boldsymbol{x},\boldsymbol{y})=|\boldsymbol{x}-\boldsymbol{y}|$ which is theoretically much more involved.

Problem (1.5) is a well-posed convex optimization problem. It is more precisely a linear program. The set of optimal couplings $\Pi(\mu,\nu)$ is evidently convex. Existence of a solution is simple to establish using the direct method of the calculus of variations. The space of probability measures $\mathcal{P}(\Omega)$ (and equivalently $\mathcal{P}(\Omega\times\Omega)$) is indeed weakly-* compact. We recall that this means that for every sequence $(\mu_n)_{n\in\mathbb{N}}\subset\mathcal{P}(\Omega)$ there exists a subsequence μ_{n_k} and a probability measure μ such that

$$\int_\Omega f(\boldsymbol{x})\mathrm{d}\mu_{n_k}(\boldsymbol{x})\to\int_\Omega f(\boldsymbol{x})\mathrm{d}\mu(\boldsymbol{x}),\quad \forall f\in C^0(\Omega)\,.$$

The functional in (1.5) is by definition continuous in this topology and the constraints (1.4) clearly pass to the limit. The optimal transport plan γ then exists[1]. It is possible to prove this result under milder assumptions than the ones we considered, see [115, Section 1.1]. Uniqueness is a more delicate issue and is not guaranteed in general.

[1]Keep in mind that we assumed from the very beginning that the space Ω is bounded. The space of continuous functions $C(\Omega)$ coincides then with the spaces $C_0^0(\Omega)$ and $C_b^0(\Omega)$, namely the space of continuous functions vanishing at infinity and the space of bounded continuous functions. This implies that narrow convergence of measures, i.e. convergence in duality with $C_b^0(\Omega)$, coincides with the usual weak-* convergence. Relaxing the boundedness assumption on Ω can be done resorting to Prokhorov's theorem to recover compactness, see [115, Theorem 1.7].

In the case one of the two measures is absolutely continuous, let us say for example μ, Brenier [27] proved that it exists an optimal map $\mathrm{T} : \Omega \to \Omega$ sending μ to ν, in the sense that: ν is the pushforward of μ, $\nu = \mathrm{T}_{\#}\mu$, and the optimal transport plan in problem (1.5) has the form $\gamma = (\mathrm{Id}, \mathrm{T})_{\#}\mu$, where Id is the identity map. This means that the solution is not simply a measure but it has a precise analytical form: each infinitesimal quantity of mass is sent to a precise final location. This seems a desirable form for solutions to the transport problem. Therefore the optimal cost in problem (1.5) writes

$$\int_{\Omega} |\boldsymbol{x} - \mathrm{T}(\boldsymbol{x})|^2 \, \mathrm{d}\mu(\boldsymbol{x}) \,, \tag{1.6}$$

and the map T minimizes the functional (1.6) among all the maps that verifies the equivalent conditions (1.1)-(1.2) (otherwise there would exist a better transport plan). The map T can be written as the gradient of a convex function ψ, i.e. $\mathrm{T} = \nabla\psi$, and it is unique (on the support of μ). The function ψ is usually called the Brenier potential. Even more, if there exists a map T pushing μ to ν that can be written as the gradient of a convex function on the support of μ, then $\gamma = (\mathrm{Id}, \mathrm{T})_{\#}\mu$ is the optimal plan.

Minimizing the functional (1.6) over all maps T satisfying the constraint (1.1) was actually the original form of the optimal transport problem as it has been formulated by Monge. As the constraint expressed by the conditions (1.1)-(1.2) is neither linear nor convex, this form is not suitable. By allowing the mass to split, which is exactly what we do by optimizing over transport plans rather than transport maps, Kantorovich formulation simplifies enormously the problem. It can be shown that this latter formulation is indeed a convex relaxation in a specific sense of the original one (see [115, Section 1.5]). Nevertheless, if one of the two measures is absolutely continuous, the two problems coincide, in the sense that the two minima are the same and the optimal transport plan is concentrated on the graph of the optimal transport map. Mind of course the lack of symmetry whether the measure μ or ν is the absolutely continuous one: the Monge formulation introduces a direction in the transport which is not present in the Kantorovich one. If both measures are absolutely continuous, then there exist the optimal maps in the two directions and they are the inverse of one another.

If we assume both the measures μ, ν to be absolutely continuous, with respective densities f and g, we can explicitly characterize the map T. Performing the change of variables $\boldsymbol{y} = \mathrm{T}(\boldsymbol{x})$ in the left-hand side of (1.2), we obtain the following differential condition on the map T in order to push μ to ν:

$$g(\mathrm{T}(\boldsymbol{x})) \det\left(\nabla \mathrm{T}(\boldsymbol{x})\right) = f(\boldsymbol{x}) \,.$$

Considering then that we can write the optimal map T as the gradient of a convex function ψ, we obtain the following non-linear partial differential equation

$$g(\nabla\psi(\boldsymbol{x})) \det\left(\mathrm{Hess}(\psi)(\boldsymbol{x})\right) = f(\boldsymbol{x}) \,, \tag{1.7}$$

which is the Monge-Ampère equation. Solving (1.7) provides the Brenier potential and therefore the optimal transport map. It comes naturally that this simple idea inspired numerical approaches for solving the optimal transport problem (see [115, Section 6.3] and references therein). Nevertheless, tackling it numerically is a difficult task. On the other hand, from the analysis of this PDE we can get useful information. The regularity of solutions to this equation has been deeply studied by Caffarelli. See [120, Chapter 4] for a brief presentation

of the topic and in particular [120, Theorem 4.14] for a summary of Caffarelli's results. Essentially, if the density functions f and g are enough regular and bounded from below, then the Brenier potential is regular.

The dual problem

One of the fundamental tools to study a convex optimization problem is duality. A constrained minimization problem can be written as an infsup problem by simply augmenting the objective functional by adding a convex indicator function representing the constraints. With the problem at hand, using the constraints in the form (1.4), the infsup formulation of problem (1.5) writes:

$$\inf_{\gamma\in\Pi(\mu,\nu)}\sup_{\phi_0,\phi_1\in C^0(\Omega)}\int_{\Omega\times\Omega}|\boldsymbol{x}-\boldsymbol{y}|^2\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y}) + \int_\Omega(\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})-\mathrm{d}\mu(\boldsymbol{x}))\phi_0(\boldsymbol{x})+\int_\Omega(\mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})-\mathrm{d}\nu(\boldsymbol{y}))\phi_1(\boldsymbol{y})\,. \quad (1.8)$$

The two functions ϕ_0,ϕ_1 are called the potentials. The dual problem is obtained exchanging inf and sup and represents a lower bound for the primal one, i.e. the value attained by the former is always smaller or at most equal to the value attained by the latter. Minimizing the functional in (1.8) with resepct to $\gamma\in\Pi(\mu,\nu)$ we obtain the dual problem of (1.5):

$$\sup_{\phi_0,\phi_1\in C^0(\Omega)}\int_\Omega\phi_1(\boldsymbol{y})\,\mathrm{d}\nu(\boldsymbol{y})+\int_\Omega\phi_0(\boldsymbol{x})\,\mathrm{d}\mu(\boldsymbol{x})\,, \quad (1.9)$$

subject to the constraint:

$$\phi_0(\boldsymbol{x})+\phi_1(\boldsymbol{y})\le|\boldsymbol{x}-\boldsymbol{y}|^2\,. \quad (1.10)$$

In general, there could be a gap between the optimal primal and dual values. However, if the problem exhibits good properties, that is for example convexity and enough regularity, then it happens that the gap is zero and they are equivalent (meaning of course that the optimal values coincide, the problems are different). In this case, we say that strong duality holds. This is true for problems (1.5) and (1.9)-(1.10). The equivalence between them can be proven using for example the Fenchel-Rockafellar duality theorem [120, Theorem 1.3] (which also provides as a direct result the existence of the minimizer for problem (1.5)). Notice that the notation primal and dual should be swapped, as problem (1.5) is formulated on the space of measures, which is the dual space of the space $C^0(\Omega)$ where (1.9) is formulated. We will nevertheless stick to the standard notation in the optimal transport community, here and later on[2].

Recalling that the measures μ,ν are non-negative, one could intuitively argue that in order to find a good competitor (ϕ_0,ϕ_1), given for example ϕ_0, we could replace the other potential with the biggest function satisfying the constraint, i.e. with

$$\phi_0^c(\boldsymbol{y})=\inf_{\boldsymbol{x}}|\boldsymbol{x}-\boldsymbol{y}|^2-\phi_0(\boldsymbol{x})\,, \quad (1.11)$$

[2]But we will swap the order of primal and dual variables, as well as the order of the equations, when referring to the saddle-point problem.

which is the c-transform of the function ϕ_0. The c actually stands for cost function considered, the squared euclidean distance in this case. Then we could replace the function ϕ_0 by ϕ_0^{cc} in order to improve again the objective functional, and so on and so forth. However, $\phi_0^{ccc} = \phi^c$. In the case considered, that is for the cost function $c(\boldsymbol{x}, \boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|^2$, this coincides with the well known fact that the double Legendre transform of a convex and lower semi-continuous function is identical to itself. This means that the maximization in problem (1.9) can be restricted to functions that are the c-transforms of one another. This is the key argument which is used to prove existence of an optimal couple of potentials (ϕ_0, ϕ_1) for problem (1.9). First of all, as the c-transform has the same modulus of continuity of the cost function itself, we can expect the two optimal potentials to be more regular than continuous. Then, it is possible to prove the equicontinuity of any sequence of maximizer and, thanks to the compactness of Ω, also the equiboundedness. We can conclude thanks to the Ascoli-Arzelà theorem. See [115, Section 1.6] for details on c-transforms and more general existence arguments for problem (1.9)-(1.10). The two maximizers are often called the Kantorovich potentials.

Exploiting strong duality, we can relate the optimal dual potentials in (1.9)-(1.10) to the Brenier one. We know indeed that the values of the primal and the dual problem coincide, therefore it must hold

$$\phi_0(\boldsymbol{x}) + \phi_1(\boldsymbol{y}) = |\boldsymbol{x} - \boldsymbol{y}|^2 \,, \quad \gamma - \text{a.e.} \,, \tag{1.12}$$

for the optimal potentials ϕ_0, ϕ_1 and the optimal transport plan γ. To obtain (1.12), just rewrite the objective functional in (1.9) using $\gamma \in \Pi(\mu, \nu)$. If we disregard the fact that this equality is defined γ-a.e., we may differentiate it with respect to the $\boldsymbol{x}$ variable to obtain

$$\boldsymbol{y} = \boldsymbol{x} - \frac{1}{2}\nabla\phi_0(\boldsymbol{x}) = \mathrm{T}(\boldsymbol{x}) \,. \tag{1.13}$$

We have in this way an explicit relation between the optimal map and the optimal potential ϕ_0. Moreover,

$$\psi(\boldsymbol{x}) = \frac{|\boldsymbol{x}|^2}{2} - \frac{1}{2}\phi_0(\boldsymbol{x}) \,, \tag{1.14}$$

where we recall that ψ is the Brenier potential[3]. The same relations hold for more general cost functions c, although they may not be explicit. This was actually the original argument of Brenier. Handling carefully the support of the optimal transport plan when differentiating equality (1.13) provides the conditions for the existence of the optimal transport map. The condition μ absolutely continuous considered by Brenier is sufficient, but not necessary. See [95], or [120, Theorem 3.8], for a refined version of Brenier's theorem.

1.1.2 Dynamical formulation

A solution to problem (1.5) only provides the optimal reallocation of mass, but it doesn't specify how this reallocation should take place. In this sense, problem (1.5) is referred to as the static formulation of optimal transport. Benamou and Brenier introduced a dynamical formulation, which takes into account how the displacement has to be realized continuously in time [14]. It is especially thanks to this new formulation that optimal transport turned out to have incredible links with physical problems.

[3] Formulas (1.13)-(1.14) are usually written without the factor $\frac{1}{2}$ on the potential, which follows from considering $c(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}|\boldsymbol{x} - \boldsymbol{y}|^2$ as unitary cost. When dealing with Wasserstein gradient flows however, the factor $\frac{1}{2}$ is not considered as it is more natural in the formulation of the JKO scheme (see Section 1.2).

Benamou-Brenier formulation of optimal transport

We assume the two measures to be absolutely continuous, in order to simplify the presentation. Let us denote the initial and final densities as $\rho^{in}, \rho^f \in L^1(\Omega; \mathbb{R}_+)$, with the same total mass,

$$\int_\Omega \rho^{in}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \rho^f(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,,$$

not necessarily equal to one. We implicitly assume a positive time direction for the problem, whence the notation of initial and final density.

Benamou and Brenier showed that problem (1.5) can be formulated as finding a time-dependent density $\rho : [0,1]\times\Omega \to [0,+\infty)$ and a time-dependent velocity field $\boldsymbol{v} : [0,1]\times\Omega \to \mathbb{R}^d$ solving

$$\inf_{\substack{(\rho,\boldsymbol{v})\\ \rho\geq 0}} \int_0^1 \int_\Omega \rho|\boldsymbol{v}|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \tag{1.15}$$

where the two curves $(\rho, \boldsymbol{v})$ are subjected to the continuity equation constraint,

$$\begin{cases} \partial_t\rho + \nabla\cdot\rho\boldsymbol{v} = 0 & \text{in } [0,1]\times\Omega\,,\\ \rho\boldsymbol{v}\cdot\boldsymbol{n} = 0 & \text{on } [0,1]\times\partial\Omega\,, \end{cases} \tag{1.16}$$

which has to be satisfied in distributional sense, with the further initial and final conditions $\rho(0,\cdot) = \rho^{in}, \rho(1,\cdot) = \rho^f$. The optimal displacement is therefore the one that minimizes the total kinetic energy among all the admissible ones. It is reasonable that an admissible displacement should not create or destroy mass, neither disperse it outside the domain, whence the continuity constraint subjected to the condition of no flux across the boundary. For a precise statement of problem (1.15)-(1.16), see for example [120, Theorem 8.1].

Problem (1.15)-(1.16) is a non-linear and non-convex optimization problem. Nevertheless, as pointed out by Benamou and Brenier, it can be formulated as a convex optimization problem under linear constraints thanks to the change of variables $(\rho, \boldsymbol{v}) \mapsto (\rho, \boldsymbol{m} = \rho\boldsymbol{v})$ and defining $B : \mathbb{R}\times\mathbb{R}^d \to [0,+\infty]$,

$$B(p,\boldsymbol{Q}) \coloneqq \begin{cases} \frac{|\boldsymbol{Q}|^2}{2p} & \text{if } p > 0\,,\\ 0 & \text{if } p = 0,\ \boldsymbol{Q} = 0\,,\\ +\infty & \text{else}\,. \end{cases} \tag{1.17}$$

The function B is convex and lower semi-continuous as it can be written as the supremum of linear functions,

$$B(p,\boldsymbol{Q}) = \sup_{(a,\boldsymbol{b})\in K} ap + \langle \boldsymbol{b}, \boldsymbol{Q}\rangle\,,$$

where the convex set K is defined as: $K = \left\{(a,\boldsymbol{b}) \in \mathbb{R}\times\mathbb{R}^d : a + \frac{|\boldsymbol{b}|^2}{2} \leq 0\right\}$. Introducing the functional

$$\mathcal{B}(\rho,\boldsymbol{m}) = \int_0^1 \int_\Omega B(\rho(t,\boldsymbol{x}), \boldsymbol{m}(t,\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \tag{1.18}$$

the problem is then stated as:

$$\inf_{(\rho,\boldsymbol{m})} \mathcal{B}(\rho,\boldsymbol{m}) \tag{1.19}$$

where ρ and $\boldsymbol{m}$ satisfy in distributional sense the continuity equation

$$\begin{cases} \partial_t \rho + \nabla \cdot \boldsymbol{m} = 0 & \text{in } [0,1] \times \Omega\,, \\ \boldsymbol{m} \cdot \boldsymbol{n} = 0 & \text{on } [0,1] \times \partial\Omega\,, \end{cases} \tag{1.20}$$

with the further initial and final conditions $\rho(0,\cdot) = \rho^{in}, \rho(1,\cdot) = \rho^f$. The positivity constraint is now automatically enforced thanks to the definition of the functional $\mathcal{B}$ for a finite valued couple $(\rho, \boldsymbol{m})$. It is worth noticing that, whereas the change of variables solves the non-convexity and non-linearity issues, it introduces a regularity issue: as a results, finding solutions to problem (1.19)-(1.20) does not turn out to be easy. Pay attention that the functional (1.18) is defined as one half of the functional in (1.15)[4]. For a precise definition of the problem in the general setting of arbitrary initial and final probability measures, see for example [79]. Existence of a solution is easy to obtain (see Theorem 2.3 for the argument in the discrete setting) whereas uniqueness is more delicate. Again, it is guaranteed if one of the two measures is absolutely continuous [121, Corollary 7.23].

Also problem (1.19)-(1.20), as the original (1.9), admits a dual one for which strong duality holds. Let us derive it thanks to formal computations. Incorporating the constraint (1.20) in the objective functional, we can write the problem in the following saddle-point formulation:

$$\inf_{(\rho,\boldsymbol{m})} \sup_{\phi} \int_0^1 \int_\Omega B(\rho,\boldsymbol{m}) + \int_0^1 \int_\Omega (\partial_t \rho + \nabla \cdot \boldsymbol{m})\phi\,. \tag{1.21}$$

The function ϕ is now the time-space dependent potential. Integrating by parts, using the no flux boundary conditions and the initial and final conditions, we obtain

$$\inf_{(\rho,\boldsymbol{m})} \sup_{\phi} \int_0^1 \int_\Omega B(\rho,\boldsymbol{m}) - \int_0^1 \int_\Omega \rho \partial_t \phi - \int_0^1 \int_\Omega \boldsymbol{m} \cdot \nabla\phi + \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega \phi(0,\cdot)\rho^{in}\,.$$

Exchanging inf and sup we obtain the dual problem,

$$\sup_{\phi} \inf_{(\rho,\boldsymbol{m})} \int_0^1 \int_\Omega B(\rho,\boldsymbol{m}) - \int_0^1 \int_\Omega \rho \partial_t \phi - \int_0^1 \int_\Omega \boldsymbol{m} \cdot \nabla\phi + \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega \phi(0,\cdot)\rho^{in}\,,$$

which is equivalent to the primal one as we said, thanks to strong duality. The minimization in $\boldsymbol{m}$ then provides the optimality condition $\frac{\boldsymbol{m}}{\rho} = \nabla\phi$, which can be written in the form

$$\boldsymbol{m} = \rho \nabla \phi\,, \tag{1.22}$$

as we know that $\boldsymbol{m} = \boldsymbol{0}$ whenever $\rho = 0$, due to the definition of the function B. The definition of the function (1.17) imposes that for a momentum $\boldsymbol{m}$ to attain a finite value for the functional (1.18), it has to be absolutely continuous with respect to the density ρ. The optimal vector field $\boldsymbol{v}$ in (1.15) is therefore provided by $\nabla\phi$. Using this condition we can rewrite the saddle-point problem as

$$\sup_{\phi} \inf_{\rho \geq 0} -\int_0^1 \int_\Omega \frac{1}{2}\rho|\nabla\phi|^2 - \int_0^1 \int_\Omega \rho \partial_t \phi + \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega \phi(0,\cdot)\rho^{in}\,. \tag{1.23}$$

[4]This will turn out to be convenient notation-wise in the definition of the JKO scheme in section 1.2.

Minimizing now in ρ we finally obtain the form of the dual problem,

$$\sup_{\phi} \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega \phi(0,\cdot)\rho^{in}\,, \tag{1.24}$$

where the potential must satisfy the Hamilton-Jacobi equation:

$$\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 \le 0 \quad \text{in } [0,1]\times\Omega\,. \tag{1.25}$$

The inequality derives from the fact that the minimization in ρ is taken over non-negative densities, and the equality holds where ρ does not vanish, i.e. ρ-a.e.. Problem (1.24)-(1.25) is the corresponding dynamical formulation of the static problem (1.9)-(1.10). The objective functional in (1.24) can be identified with the one in (1.9) by changing sign to the potential ϕ_0. The strong duality between the two problems can be proven again using Fenchel-Rockafellar duality theory. See for example [39] for its application on a more general problem then the one considered here.

Thanks to strong duality, the primal and dual problem attain the same value. A primal-dual solution (ϕ,ρ) is a saddle point for the augmented functional in (1.23). It must then satisfy the system of primal-dual optimality conditions, or stationarity conditions,

$$\begin{cases} \partial_t\rho + \nabla\cdot(\rho\nabla\phi) = 0 & \text{in } [0,1]\times\Omega\,,\\ \partial_t\phi + \frac{1}{2}|\nabla\phi|^2 \le 0 & \text{in } [0,1]\times\Omega\,,\\ \partial_t\phi + \frac{1}{2}|\nabla\phi|^2 = 0 & \rho-\text{a.e.}\,, \end{cases} \tag{1.26}$$

complemented with the boundary conditions $\rho\nabla\phi\cdot\boldsymbol{n} = 0$ on $\partial\Omega$ and the initial/final conditions $\rho(0,\cdot)=\rho^{in}, \rho(1,\cdot)=\rho^f$. The optimal momentum is given by the optimality condition (1.22). A solution (ϕ,ρ) to system of equations (1.26) is necessarily a saddle-point of (1.23). Suppose indeed that $(\tilde\phi,\tilde\rho)$ solves system (1.26) but it is not a saddle-point. It holds

$$\begin{aligned} \int_\Omega \tilde\phi(1,\cdot)\rho^f - \int_\Omega \tilde\phi(0,\cdot)\rho^{in} &= \int_0^1\int_\Omega \partial_t(\tilde\phi\tilde\rho) = \int_0^1\int_\Omega \partial_t\tilde\phi\tilde\rho + \int_0^1\int_\Omega \tilde\phi\partial_t\tilde\rho\\ &= -\int_0^1\int_\Omega \frac{1}{2}\tilde\rho|\nabla\tilde\phi|^2 - \int_0^1\int_\Omega \tilde\phi\nabla\cdot(\tilde\rho\nabla\tilde\phi)\\ &= -\int_0^1\int_\Omega \frac{1}{2}\tilde\rho|\nabla\tilde\phi|^2 + \int_0^1\int_\Omega \tilde\rho|\nabla\tilde\phi|^2 = \int_0^1\int_\Omega \frac{1}{2}\tilde\rho|\nabla\tilde\phi|^2\,, \end{aligned}$$

providing

$$\begin{aligned} \sup_{\phi} \int_\Omega \phi(1)\rho^f - \int_\Omega \phi(0)\rho^{in} &\ge \int_\Omega \tilde\phi(1)\rho^f - \int_\Omega \tilde\phi(0)\rho^{in}\\ &= \int_0^1\int_\Omega \frac{1}{2}\tilde\rho|\nabla\tilde\phi|^2 \ge \inf_{(\rho,\boldsymbol{m})} \int_0^1\int_\Omega B(\rho,\boldsymbol{m})\,, \end{aligned}$$

(where the sup is taken on ϕ satisfying (1.25), the inf on $(\rho,\boldsymbol{m})$ satisfying (1.20)). One of the two inequalities is strict if $(\tilde\phi,\tilde\rho)$ is not a saddle-point, which is a contradiction to strong duality. Solutions to (1.26) are therefore saddle-points. As we said, uniqueness of solution to problem (1.19) and consequently system (1.26) is delicate to infer. Notice that problem

(1.23) is not strictly convex in the density. In any case, the potential is defined up to a global time-space constant, and it is not defined where the density vanishes.

We highlight that the Hamilton-Jacobi equation can be saturated thanks to the monotonicity of the Hamilton-Jacobi operator. For two potential curves ϕ_1, ϕ_2, such that $\phi_1(0) = \phi_2(0)$ and satisfying

$$\partial_t \phi_1 + \frac{1}{2}|\nabla \phi_1|^2 \leq \partial_t \phi_2 + \frac{1}{2}|\nabla \phi_2|^2 \,, \tag{1.27}$$

it holds $\phi_2(1) \geq \phi_1(1)$ (conversely, if $\phi_1(1) = \phi_2(1)$ and (1.27) holds then $\phi_2(0) \leq \phi_1(0)$). Saturating the equation then in (1.25) provides a better competitor for problem (1.24), i.e. the inequality can be replaced by the equality and consequently also in system (1.26) by strong duality. We will show in Section 1.2.2 how the argument works. Bear in mind that whereas the constraint (1.25) defines a convex set, the saturated equation does not. Solutions to the Hamilton-Jacobi equation

$$\partial_t \phi + \frac{1}{2}|\nabla \phi|^2 = 0 \quad \text{in } [0,1] \times \Omega \,,$$

are provided by the Hopf-Lax formula [11]. Given the initial condition $\phi(0, \cdot)$ (or equivalently the final condition $\phi(1, \cdot)$, by simply switching sign) we can explicitly write $\phi(t, \cdot)$ as

$$\phi(t, \boldsymbol{y}) = \inf_{\boldsymbol{x} \in \Omega} \frac{|\boldsymbol{x} - \boldsymbol{y}|^2}{t} + \phi(0, \boldsymbol{x}) \,, \quad \forall \boldsymbol{y} \in \Omega \,. \tag{1.28}$$

If we evaluate it at the final time, we recover the fact that the potentials are the c-transform of one another (up to the already mentioned sign change of ϕ_0). (1.28) is the time dependent version of (1.11). In the same way, it can provide useful information on the regularity of the potential. See [121, Chapter 7] for an introduction to the relation between the Hopf-Lax formula and the dual optimal transport problem.

Problem (1.19) does not guarantee any regularity on the interpolating density. For example, the interpolation between two delta measures is always a delta itself. We may wonder if there exist conditions on the initial and final measures that could ensure some regularity on the interpolation. As we already said, smoothness and strict positivity provides regularity on the Brenier potential, and consequently on the optimal dual potential ϕ_0. However we are not aware of results that extend to the whole time-dependent potential, nor on the density curve. Consider already that strict positivity does not necessarily hold on the whole curve even though the initial and final measures are strictly positive [117]. Nonetheless, smooth and strictly positive solutions to problem (1.26) exist and can be constructed (see Remark 2.11).

Lagrangian interpretation

Thanks to the Benamou-Brenier formulation, optimal transport defines a natural notion of interpolation between probability measures. Prior to their work, McCann introduced another idea [96]. Let us assume again that the measure μ is absolutely continuous, so that it exists the optimal map T transporting it to the final measure ν. Consider the map given by $\mathrm{T}_t = (1-t)\mathrm{Id} + t\mathrm{T}, \forall t \in [0,1]$. We can define the density curve $\mu_t = (\mathrm{T}_t)_{\#}\mu$, which is called the displacement interpolation. By the Brenier theorem, T_t is the optimal transport map between μ and μ_t, $\forall t \in [0,1]$. It holds indeed

$$\mathrm{T}_t(\boldsymbol{x}) = \boldsymbol{x} - \frac{1}{2}\nabla \phi_t(\boldsymbol{x}) = \boldsymbol{x} - \frac{t}{2}\nabla \phi_0(\boldsymbol{x}) \,, \tag{1.29}$$

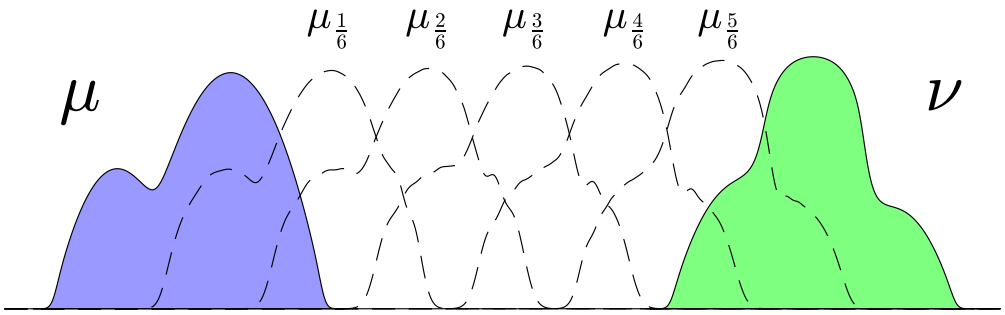


Figure 1.2: Interpolation between the two measures of the example in Figure 1.1.

and T_t is the gradient of the convex function

$$\psi_t = \frac{|\boldsymbol{x}|^2}{2} - \frac{t}{2}\phi_0\,, \tag{1.30}$$

where ϕ_0 is the optimal potential in the dual problem for the transport from μ to ν. At each time, the intermediate potential is simply given by the rescaling of ϕ_0. Notice that, as the Brenier potential ψ is convex, ψ_t is strongly convex, hence also strictly convex (see the discussion in Section 1.1.2). The cost increases quadratically in time along the curve,

$$\int_\Omega |\boldsymbol{x} - \mathrm{T}_t(\boldsymbol{x})|^2 \,\mathrm{d}\mu(\boldsymbol{x}) = t^2 \int_\Omega |\boldsymbol{x} - \mathrm{T}(\boldsymbol{x})|^2 \,\mathrm{d}\mu(\boldsymbol{x})\,.$$

Thanks to McCann's displacement interpolation, once we have found the optimal map T transporting the initial measure μ to ν, we can define the constant velocity $\boldsymbol{v} = \mathrm{T} - \mathrm{Id}$ and move the mass from the initial to the final position following straight trajectories, with constant velocity $\boldsymbol{v}$. The particles do not accelerate neither decelerate along the path. At each time, the trajectories remain optimal, meaning that there is no change of direction that could decrease further the total cost. Consequently, also the kinetic energy at each time remains constant and it is the minimal possible one. Intuitively, particles cannot bump into each other. More precisely, this is related to the fact that at each time the measure is the pushforward via the injective map T_t (as it is the gradient of a strongly convex function). It should not come as a surprise then that the interpolation μ_t coincides with the interpolation defined via the Benamou-Brenier formulation: the displacement interpolation solves the dynamical optimal transport problem, i.e. there exists a velocity field $\boldsymbol{v}_t$ such that the two curves $(\mu_t, \boldsymbol{v}_t)$ satisfy the continuity equation (1.16) and the total kinetic energy is minimal [121, Theorem 7.21]. The vector field $\boldsymbol{v}_t$ is given by

$$\boldsymbol{v}_t(t, \boldsymbol{x}) = \boldsymbol{v} \circ ((1-t)\boldsymbol{x} + t\mathrm{T}(\boldsymbol{x}))^{-1}\,,$$

that is, the velocity at time t in the point $\boldsymbol{x}$ is the velocity of the particle that will pass by $\boldsymbol{x}$ at time t. We highlight that the hypothesis of convexity of Ω is necessary in order to define the displacement interpolation, since otherwise $\mathrm{T}_t(\boldsymbol{x})$ may exit the domain. On the contrary,

problem (1.19)-(1.20) makes sense even in the case Ω is not convex but clearly particles cannot move on straight lines[5].

McCann's interpolation and the Benamou-Brenier formulation are two different description of the same continuous displacement. The notion of interpolation introduced by McCann is Lagrangian, which means that it prescribes the displacement and the velocity of each particle of the initial configuration. The notion introduced by Benamou and Brenier is instead Eulerian and it prescribes the evolution in time and space of the density and velocity field. The hypothesis that the measure μ is absolutely continuous is not necessary as the same concept can be formulated in terms of optimal transport plans. Recall that π_1 and π_2 represent the canonical projections on the first and second component of the product space $\Omega \times \Omega$, and let us define $\pi_t = (1-t)\pi_1 + t\pi_2$. Let γ be the optimal transport plan between μ and ν. We can define the displacement interpolation as the measure $\omega_t = (\pi_t)_{\#}\gamma$. If μ is absolutely continuous, we recover the previous definition. The curve ω_t is again optimal in problem (1.15) and all the considerations we have previously made, although less intuitive, hold true. The only difference is that in this case particles may split at the initial time.

Geodesic extrapolation

McCann defined an interpolation between two measures by linearly interpolating at the level of the transport maps (plans). It is possible to use the same idea in order to define an extrapolation. Again, let us start by the simpler case of μ absolutely continuous in order to ease the presentation. Given the optimal transport map T that pushes μ to ν, we define again the map $\mathrm{T}_t = (1-t)\mathrm{Id} + t\mathrm{T}$, this time for $t \geq 1$. We define the geodesic[6] t-extrapolation from μ to ν, at time $t \in [1, t^*]$, $t^* \geq 1$, as the measure $\mu_t = (\mathrm{T}_t)_{\#}\mu$. The value t^* is the maximum time for which this extrapolation is well defined, in the sense we specify below. Assume the Brenier potential ψ, relative to the transport from μ to ν, to be λ-convex (strongly convex with modulus λ), with $\lambda > 0$. This implies that

$$\frac{|\boldsymbol{x}|^2}{2} - \frac{t}{2}\phi_0 = t\Big(\frac{|\boldsymbol{x}|^2}{2t} - \frac{1}{2}\phi_0\Big) = t\Big(\frac{|\boldsymbol{x}|^2}{2t} - \Big(\frac{|\boldsymbol{x}|^2}{2} - \psi\Big)\Big) = t\Big(-\frac{(t-1)}{2t}|\boldsymbol{x}|^2 + \psi\Big)$$

is convex as long as $\frac{(t-1)}{2t} \leq \lambda$. This means that ψ_t defined as in (1.30) is the Brenier potential relative to the transport from μ to μ_t only if $t \leq \frac{1}{1-2\lambda} = t^*$, if $\lambda < \frac{1}{2}$, for any $t > 1$ if $\lambda \geq \frac{1}{2}$ ($t^* = +\infty$). Up until this time, particles keep moving straight beyond the final measure ν and the displacement is optimal between μ and μ_t. All the considerations we made in Section 1.1.2 hold true also in this case. After the time t^*, the Brenier potential, which has to be convex, is different from ψ_t and μ_t is not anymore the geodesic extrapolation.

Depending on the specific problem considered, after a certain point beyond the final measure ν particles may start to collide or reach the boundary of the domain, which is the reason why we cannot define the extrapolation for any time $t \geq 1$. If ν is not absolutely continuous, particles collide exactly at time $t = 1$, which means that the potential is 0-convex. The extrapolation does not exists at all in this case. Consider for example the transport between two measures uniformly distributed on two concentric annuli, where the optimal solution is

[5] Anyway, the unitary cost in the definition of problem (1.5) can be adapted to a non-convex domain Ω, passing from the straight euclidean distance to a curved one, in order to define again the displacement interpolation and state again its equivalence with the Benamou-Brenier one. See [121].

[6] The reason for the name geodesic extrapolation will be clear after Section 1.1.3.

clearly given by radial trajectories. If we consider the transport from the inner to the outer one, the particles can move straight for an indefinite time beyond the final measure and the extrapolation always exists. On the contrary, moving from the outer to the inner, they can only go as far as they reach the center and collapse into a delta. Nevertheless, the map T_t is always well defined even past the possible collision and we can always consider the measure $\mu_t = (\mathrm{T}_t)_{\#}\mu$, for all $t \geq 1$, provided of course we do not exit the domain. The measure μ_t defined in this way is the Lagrangian extrapolation. In the previous example, particles would start to escape again from the origin, keeping going straight. But the interpolation is not anymore the optimal one: there exists another map that pushes in an optimal way μ to μ_t, and the new optimal trajectories do not pass through ν.

If μ is not absolutely continuous, we can proceed as in the previous section. Denoting by γ the optimal transport plan between μ and ν, we can define the Lagrangian extrapolation as $\omega_t = (\pi_t)_{\#}\gamma$, where $\pi_t = (1-t)\pi_1 + t\pi_2$, this time for $t \geq 1$. If the optimal transport plan between μ and ω_t is given by $(\pi_1, \pi_t)_{\#}\gamma$, for all $t \in [1, t^*]$, then we say that ω_t is the geodesic t-extrapolation from μ to ν. Again, t^* is the maximum time for which this holds true, i.e. the time when a collision takes place. After t^* the Lagrangian extrapolation is well-defined but the trajectories are not the optimal ones.

From the initial time $t = 0$ till the time $t \leq t^*$, the interpolation defined between μ and μ_t is the same as the Benamou-Brenier one. It can be computed as solution to the system of equation (1.26). Recall that the optimal transport problem (1.19)-(1.20) is defined on the time interval $[0, 1]$. If we solve it for the measure μ and μ_t, $t \neq 1$, the solution $(\phi(s,\cdot), \rho(s,\cdot))$ we obtain coincides with the one defined via the McCann's interpolation only up to a rescale of the potential and the time variable by the factor t, $(t\phi(ts,\cdot), \rho(ts,\cdot))$. The fact that the particles do not collide along the measure curve interpolating μ and μ_t means that the solution to the Hamilton-Jacobi equation is a classical solution. A collision implies a loss of regularity: classical solutions cannot be considered in that event. However, we can consider viscosity solutions, which are solutions that dissipate the possible shock. This means that even though particles start to collide, we can keep integrating forward the system of equations (1.26), by considering viscosity solutions for the Hamilton-Jacobi equation [12]. The solution provided in this way is not the extrapolation (in the sense we defined it above), neither coincides with the pushforward of μ via the map T_t (after the shock). The viscosity solution dissipates in general the kinetic energy, which is on the contrary preserved if we simply assume that the particles continue to move straight even after the time t^*.

1.1.3 The Wasserstein space

We want to briefly characterize now what we refer to as Wasserstein space. First of all, given two measures $\mu, \nu \in \mathcal{P}(\Omega)$, we denote

$$\mathcal{W}_2^2(\mu,\nu) \coloneqq \min_{\gamma \in \Pi(\mu,\nu)} \int_{\Omega\times\Omega} |\boldsymbol{x}-\boldsymbol{y}|^2 \mathrm{d}\gamma(\boldsymbol{x},\boldsymbol{y})\,.$$

The mapping $\mathcal{W}_2(\cdot,\cdot) : \mathcal{P}(\Omega)\times\mathcal{P}(\Omega) \to \mathbb{R}_+$ can be proven to be a metric [115, Section 5.1]. It is called the quadratic Wasserstein distance. If a different power $p > 1$ is considered in the unitary cost, a different distance is obtained and it is denoted as $\mathcal{W}_p$. As we will only work with the quadratic case, we will omit to specify it from time to time. We call the Wasserstein space the metric space $(\mathcal{P}(\Omega), \mathcal{W}_2)$, obtained by equipping the space of probability measures with

the Wasserstein distance. The topology induced on $\mathcal{P}(\Omega)$ by $\mathcal{W}_2$ is the same one as the weak-* topology [115, Theorem 5.10][7]: given a sequence of probability measures $(\mu_n)_{n\in\mathbb{N}} \subset \mathcal{P}(\Omega)$ and a measure $\mu \in \mathcal{P}(\Omega)$, saying that

$$\mathcal{W}_2(\mu_n, \mu) \to 0$$

is therefore equivalent to

$$\int_\Omega \varphi \mathrm{d}\mu_n \to \int_\Omega \varphi \mathrm{d}\mu, \quad \forall \varphi \in C^0(\Omega)\,.$$

A direct consequence is that the functional $\mathcal{W}_2(\cdot, \nu) : \mathcal{P}(\Omega) \to \mathbb{R}_+$, for any fixed measure $\nu \in \mathcal{P}(\Omega)$, is continuous with respect to the weak-* convergence, which can be proven using the triangular inequality. This property will be fundamental in the study of gradient flows via minimizing movements (see Section 1.2 below).

From what we exposed in the previous sections, it appears that problem (1.19)-(1.20) selects a curve in $\mathcal{P}(\Omega)$ which minimizes the total distance in the Wasserstein sense. Something that resembles a geodesic curve on a Riemannian manifold. Ambrosio, Gigli and Savaré developed a whole formalism in order to justify this intuition, that can be applied to general metric spaces, not only the Wasserstein space. Let us just give simple definitions in order to clarify the previous statement. We refer to [3] for the detailed construction, or the more accessible [115, Chapter 5][8]. A curve in the space of probability measures ω is a continuous map defined on a time interval with values in $\mathcal{P}(\Omega)$, i.e. $\omega : [0,1] \to \mathcal{P}(\Omega)$. Up to a time rescaling, we can always consider the time interval to be $[0,1]$. The curve is said absolutely continuous if there exists $f \in L^1([0,1])$ such that

$$\mathcal{W}_2(\omega(t^0), \omega(t^1)) \leq \int_{t^0}^{t^1} f(s)\mathrm{d}s\,, \quad \forall\, t^0, t^1 \in [0,1],\, t^0 < t^1.$$

The velocity of the curve is not defined, as its time derivative has no meaning, but it is possible to give a meaning to its modulus:

$$|\omega'|(t) := \lim_{\Delta t \to 0} \frac{\mathcal{W}_2(\omega(t+\Delta t), \omega(t))}{|\Delta t|}\,.$$

The length of the curve is defined as

$$\mathrm{Length}(\omega) := \sup_N \left\{ \sum_{k=0}^{N-1} \mathcal{W}_2(\omega(t^k), \omega(t^{k+1})) \;:\; N \geq 1, 0 = t^0 < t^1 < \cdots < t^N = 1 \right\}, \quad (1.31)$$

and for an absolutely continuous curve it holds

$$\mathrm{Length}(\omega) = \int_0^1 |\omega'|(t)\mathrm{d}t\,.$$

A curve ω such that $\omega(0) = \mu, \omega(1) = \nu$ and minimizing (1.31) is called a geodesic curve between μ and ν.

[7] Recall that we are considering the case of $\Omega \subset \mathbb{R}^d$ compact and that in this case the weak-* convergence coincides with the narrow convergence. See [115, Theorem 5.11] for the general case of unbounded domain.

[8] We also suggest the survey [116] for a brief, yet clear, presentation.

In the Wasserstein space, absolutely continuous curves can be identified with curves that satisfy the continuity equation (1.16) in distributional sense for some vector field $\boldsymbol{v}$ and which have finite kinetic energy, i.e. with finite weighted norm $L^2_{\omega(t)}$ [115, Theorem 5.14][9]. Moreover, it holds

$$|\omega'|(t) = \left(\int_\Omega \omega(t) |\boldsymbol{v}(t)|^2 \right)^{\frac{1}{2}} .$$

Therefore, problem (1.15)-(1.16) selects an absolutely continuous curve which has the minimal length between two probability measures, i.e. a geodesic curve. The Wasserstein space is a geodesic space, which means that for any two points in the space, that is for any two measures, there exists a geodesic. The geodesic is moreover a constant speed curve, as we saw in Section 1.1.2, meaning that the distance along the interpolation is proportional to the travel time:

$$\mathcal{W}_2(\omega(t^0), \omega(t^1)) = |t^0 - t^1| \mathcal{W}_2(\omega(0), \omega(1)) , \quad \forall t^0, t^1 \in [0,1] .$$

Equivalently, the space integral of the density of kinetic energy is preserved along the curve.

Otto formalism

By building on this formal identification between the Wasserstein space and a Riemannian manifold, Otto [108] justified the link between gradient flows in this space and equations of the form:

$$\begin{cases} \partial_t \varrho - \nabla \cdot (\varrho \nabla \frac{\delta \mathcal{E}}{\delta \rho}[\varrho]) = 0 & \text{in } [0,T) \times \mathring{\Omega} , \\ \varrho \nabla \frac{\delta \mathcal{E}}{\delta \rho}[\varrho] \cdot \boldsymbol{n} = 0 & \text{on } [0,T) \times \partial\Omega , \\ \varrho(0,\cdot) = \rho^0 & \text{in } \Omega . \end{cases} \tag{1.32}$$

Let us briefly explain Otto's argument. This formal justification is nowadays called Otto calculus.

Firs of all, we need to give a definition of the tangent space at the point $\mu \in \mathcal{P}(\Omega)$ and endow it with an appropriate metric. Consider then measure curves $\rho : [-\varepsilon, \varepsilon] \to \mathcal{P}(\Omega)$ passing by μ at time zero, i.e. such that $\rho(0) = \mu$, and let us denote their velocity by $\partial_t \rho$. We want to give a representation of $\partial_t \rho|_{t=0}$. As these velocities should be seen as infinitesimal admissible displacements of the measure μ, we can intuitively identify them with elements of the form

$$\partial_t \rho|_{t=0} = -\nabla \cdot (\mu \boldsymbol{v}) , \tag{1.33}$$

for any vector field $\boldsymbol{v}$ which is tangent to the boundary of the domain, $\boldsymbol{v} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$, and which is compatible with the displacement. As there may be more then one compatible vector field, we select a specific one: the one associated with the minimal kinetic energy. That is, the one that can be represented as the gradient of a potential ϕ,

$$\partial_t \rho|_{t=0} = -\nabla \cdot (\mu \nabla \phi) , \tag{1.34}$$

where ϕ should also verify the boundary condition $\nabla\phi \cdot \boldsymbol{n} = 0$ on $\partial\Omega$. We can then formally identify the tangent space to the Riemannian manifold at the point μ as the set $H^1_\mu(\Omega)$ of potentials. Each potential is associated with a specific displacement, that is it verifies equation (1.34) for a specific curve ρ passing by μ. The next step is to define the notion of

[9] See also [3, Theorem 8.3.1].

metric we consider in the tangent space, i.e. in the space $H^1_\mu(\Omega)$ of potentials by our formal identification. This metric should be compatible with the formula for the total length of the geodesic in problem (1.19)-(1.20). Consider two curves ρ_1, ρ_2 and the respective potentials ϕ_1, ϕ_2 satisfying (1.34). We define the following metric on the tangent space:

$$\langle \partial_t \rho_1, \partial_t \rho_2 \rangle_\mu = \int_\Omega \nabla \phi_1 \cdot \nabla \phi_2 \, \mathrm{d}\mu \,.$$

We denote $|| \cdot ||_\mu$ the norm associated with this metric. We can then write the length of a curve as

$$\text{Length}(\rho)^2 = \int ||\partial_t \rho||^2_{\rho(t)} \mathrm{dt} = \int \int_\Omega \nabla \phi \cdot \nabla \phi \, \mathrm{d}\rho \mathrm{d}t = \int \int_\Omega |\nabla \phi|^2 \mathrm{d}\rho \mathrm{d}t \,.$$

We want now to identify the gradient operator. Consider a functional $\mathcal{E}$ and a curve ρ passing by μ at time zero. Using the chain rule, we know that

$$\frac{\mathrm{d}\mathcal{E}(\rho(t))}{\mathrm{d}t}|_{t=0} = \int_\Omega \frac{\delta \mathcal{E}}{\delta \rho}[\mu] \partial_t \rho \,.$$

On the other hand, we want to define the gradient such that it holds

$$\frac{\mathrm{d}\mathcal{E}(\rho(t))}{\mathrm{d}t}|_{t=0} = \langle \nabla_{\mathcal{W}_2} \mathcal{E}(\rho(t)), \partial_t \rho(t) \rangle_\mu |_{t=0} = \int_\Omega \nabla \psi \cdot \nabla \phi \, \mathrm{d}\mu \,, \tag{1.35}$$

for some potential $\psi \in H^1_\mu(\Omega)$. Using the identification (1.34) and integrating by parts, we can write

$$\int_\Omega \frac{\delta \mathcal{E}}{\delta \rho}[\mu] \partial_t \rho = -\int_\Omega \frac{\delta \mathcal{E}}{\delta \rho}[\mu] \nabla \cdot (\mu \nabla \phi) = \int_\Omega \nabla \frac{\delta \mathcal{E}}{\delta \rho}[\mu] \cdot \nabla \phi \mathrm{d}\mu \,. \tag{1.36}$$

Comparing (1.35) and (1.36), we can identify ψ with $\frac{\delta \mathcal{E}}{\delta \rho}[\mu]$ and define the gradient in the Wasserstein sense as

$$\nabla_{\mathcal{W}_2} \mathcal{E}(\mu) = -\nabla \cdot (\mu \nabla \frac{\delta \mathcal{E}}{\delta \rho}[\mu])$$

A gradient flow in the Wasserstein space can therefore be written as

$$\partial_t \rho = -\nabla_{\mathcal{W}_2} \mathcal{E}(\rho) \,,$$

complemented with the no flux boundary conditions, which is exactly equation (1.32).

1.2 Wasserstein gradient flows

A finite dimensional gradient flow in $\mathbb{R}^d$, equipped with the euclidean distance, with respect to a real valued function $F : \mathbb{R}^d \to \mathbb{R}$, coincides with solutions of the following Cauchy problem:

$$\begin{cases} \dfrac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} = -\nabla F(\boldsymbol{x}(t)) \,, \quad t \geq 0 \,, \\ \boldsymbol{x}(0) = \boldsymbol{x}^0 \,. \end{cases} \tag{1.37}$$

This definition relies on two things: the time derivative of the curve and the gradient operator. Both these tools can be defined in an infinite dimensional Hilbert space, but they make no

direct sense in the non-flat Wasserstein space, or a general metric space. In order to extend the notion of gradient flow to the Wasserstein space, its definition needs to be generalized.

Following [3], the reference book for gradient flows in metric spaces, a possible idea is to build a definition upon a property of finite dimensional gradient flows which does not require these notions, and extend it to the general setting[10]. Here is a possible construction. With reference to the gradient flow (1.37), assume the function F to be λ-convex and C^1. Therefore for any point $\boldsymbol{x} \in \mathbb{R}^d$ it holds:

$$F(\boldsymbol{x}) + \langle \nabla F(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{\lambda}{2}|\boldsymbol{y} - \boldsymbol{x}|^2 \leq F(\boldsymbol{y}), \quad \forall \boldsymbol{y} \in \mathbb{R}^d.$$

In case F is not C^1 we could write the same for each element in its subdifferential instead of considering the gradient (which is the only element in the subdifferential if F is differentiable). If we consider the quantity $\frac{1}{2}|\boldsymbol{x}(t) - \boldsymbol{y}|^2$, for $\boldsymbol{y} \in \mathbb{R}^d$, taking the time derivative we obtain:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}|\boldsymbol{x}(t) - \boldsymbol{y}|^2 &= \langle \boldsymbol{x}(t) - \boldsymbol{y}, \frac{\mathrm{d}\boldsymbol{x}(t)}{\mathrm{d}t} \rangle = -\langle \boldsymbol{x}(t) - \boldsymbol{y}, \nabla F(\boldsymbol{x}(t)) \rangle \\ &\leq F(\boldsymbol{y}) - F(\boldsymbol{x}(t)) - \frac{\lambda}{2}|\boldsymbol{y} - \boldsymbol{x}(t)|^2, \quad \forall t \geq 0. \end{aligned} \tag{1.38}$$

In the finite dimensional case, satisfying the inequality (1.38) is equivalent to satisfying (1.37), so that it is an equivalent definition of gradient flow. Inequality (1.38) is called Energy Variational Inequality (EVI). As this definition does not require any gradient or time derivative of the curve, it is possible to use it to generalize gradient flows to the infinite dimensional case. We say that a measure curve ϱ is a Wasserstein gradient flow in the EVI sense, with respect to the energy functional $\mathcal{E}$, if it satisfies for any given $\nu \in \mathcal{P}(\Omega)$ the following inequality:

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{W}_2^2(\varrho(t), \nu) \leq \mathcal{E}(\nu) - \mathcal{E}(\varrho(t)) - \frac{\lambda}{2}\mathcal{W}_2^2(\varrho(t), \nu), \quad \forall t \geq 0. \tag{1.39}$$

We just need to consider a proper notion of convexity, i.e. convexity along specific curves (Section 1.2.1). On the one hand, requiring the energy functional to be λ-convex may be too restrictive. On the other hand, it ensures uniqueness of the flow and also stability with respect to initial data [116, Section 3.3]. In Section 5.2.3, we will rely on this construction in order to prove convergence towards gradient flows of a time discretization of the flow. A weaker definition is possible, based again on another analogy with the finite dimensional case, but we prefer to follow another path. To define Wasserstein gradient flows, we rely on the construction based on the minimizing movements introduced by De Giorgi [48].

1.2.1 Generalized Minimizing Movement

Let us directly focus on the Wasserstein case for simplicity, although the idea is generalizable to any metric space. Given the measure $\rho^0 \in \mathcal{P}(\Omega)$ as initial condition and a real parameter $\tau > 0$, the Minimizing Movement Scheme (MMS) constructs a sequence of measures $(\rho^n)_{n\in\mathbb{N}} \subset \mathcal{P}(\Omega)$ recursively defined as

$$\rho^n \in \operatorname*{arginf}_{\rho \in \mathcal{P}(\Omega)} \frac{1}{2\tau}\mathcal{W}_2^2(\rho, \rho^{n-1}) + \mathcal{E}(\rho). \tag{1.40}$$

[10] As the latter monograph may be rather technical, we suggest also the expository paper [116], which provides a summary of the theory and focuses on the Wasserstein case.

The sequence of measures progressively minimizes the energy $\mathcal{E}$. Indeed, comparing the optimal measure ρ^n to the sub-optimal measure ρ^{n-1} for the objective functional, we deduce that at each step n it holds

$$\frac{1}{2\tau}\mathcal{W}_2^2(\rho^n, \rho^{n-1}) + \mathcal{E}(\rho^n) \leq \mathcal{E}(\rho^{n-1})\,. \tag{1.41}$$

The parameter τ, that we consider fixed but it may also vary at each iteration, plays the role of the time step and controls indeed the length of the step. For a big value the minimization of the energy functional prevails whereas for a small one the measure ρ^n will stick to the previous one. We can consider the gradient flow as the limit of this process that aims at sequentially minimizing its energy in smaller and smaller neighborhoods. For a time horizon $T > 0$, we can construct from the sequence $(\rho^n)_{n\in\mathbb{N}}$ a time dependent measure curve on $[0, T]$ by gluing them together in a piecewise continuous fashion:

$$\rho_\tau(t) = \sum_{n=1}^{N_\tau} \rho^n \mathbb{1}_{(t^{n-1}, t^n]}\,, \quad \rho_\tau(0) = \rho^0\,, \tag{1.42}$$

with $N_\tau = \frac{T}{\tau}$ total number of steps, $t^n = n\tau$ (assuming for simplicity the time step τ to be a divisor of T). We can define the Wasserstein gradient flow as the limit curve, if it exists, for $\tau \to 0$. This is the definition of Generalized Minimizing Movement (GMM) proposed by De Giorgi. The scheme is also called JKO scheme after Jordan, Kinderlehrer and Otto, who applied it in the Wasserstein setting [73]. Problem (1.40) is usually refer to as JKO step.

The Minimizing Movement Scheme is the variational generalization of the implicit Euler scheme. To see it, just write it in the finite dimensional case, considering the space $\mathbb{R}^d$ equipped with the euclidean distance:

$$\boldsymbol{x}^n \in \operatorname{argmin} \frac{1}{2\tau}\left|\boldsymbol{x} - \boldsymbol{x}^{n-1}\right|^2 + F(\boldsymbol{x})\,. \tag{1.43}$$

Taking the optimality conditions for this problem provides

$$\frac{\boldsymbol{x}^n - \boldsymbol{x}^{n-1}}{\tau} = -\nabla F(\boldsymbol{x}^n)\,,$$

which is the implicit Euler discretization of equation (1.37). The MMS then generalizes the Euler scheme to general metric spaces, as it is sufficient to replace the euclidean metric with any other metric. As it will be shown in Section 1.2.2, this time discretization differs from the implicit Euler discretization of the PDE (1.32) when applied in the Wasserstein setting. Defining the finite dimensional gradient flow as GMM, we can state its existence with milder hypotheses with respect to what the classical theory provide for the Cauchy problem (1.37), see [116, Section 2]. Although the MMS is usually thought just as a time discretization for a gradient flow, it can be taken as its very definition.

For a gradient flow to exist in the GMM sense, we need the scheme (1.40) to be well posed at each step n and we need to be able to extract a convergent subsequence. That is, we need sufficient compactness, which is the reason why it may be difficult to generalize this definition to other gradient flows. In the Wasserstein case, very few hypotheses on the energy functional are sufficient in order to find a minimizer in (1.40), lower semi-continuity with respect to the

weak-* convergence and a lower bound[11]. Summing the inequality (1.41) over n provides the classical estimate:

$$\frac{1}{2\tau}\sum_{n=0}^{N}\mathcal{W}_2^2(\rho^n,\rho^{n-1}) \leq \mathcal{E}(\rho^0) - \mathcal{E}(\rho^N)\,, \tag{1.44}$$

which is bounded if the energy is bounded from below and it is proper in the starting point ρ^0. With (1.44) it is straightforward to prove a uniform Hölder estimate for the sequence of curves $(\rho_\tau)_{\tau\in\mathbb{R}_+}$ defined in (1.42). Although the curves ρ_τ are not continuous, by taking advantage of the nice features of the Wasserstein space we can introduce a continuous analogous. As we said in section 1.1.3, given any two measures in the Wasserstein space we can find the geodesic curve joining them. We can then define a continuous curve $\tilde{\rho}_\tau$ by joining the geodesics between every two consecutive measures ρ^{n-1},ρ^n:

$$\tilde{\rho}_\tau = \sum_{n=1}^{N}\tilde{\rho}^n \mathbb{1}_{(t^{n-1},t^n]}\,, \quad \tilde{\rho}_\tau(0)=\rho^0\,, \tag{1.45}$$

with $\tilde{\rho}^n$ geodesic between ρ^{n-1} and ρ^n. The family of curves $(\tilde{\rho}_\tau)_{\tau\in\mathbb{R}_+}$ is again equicontinuous. We can therefore use the Ascoli-Arzelà compactness theorem and the compactness of the Wasserstein space to find a limit curve: the gradient flow. Using again the same estimate, we can further show that the family of piecewise continuous curves $(\rho_\tau)_{\tau\in\mathbb{R}_+}$ converges to the same limit. The previous argument is classical (see for example [115, Section 8.3] or Chapter 5) but we highlight that it is not necessarily true in other metric spaces.

Finally, the link between the GMM and solutions to equation (1.32) can be obtained passing to the limit the optimality conditions of problem (1.40). See [115, Proposition 7.17] and [73] for examples of possible variations that can be considered for this purpose. We remark that the construction we presented may be used, as we already pointed out for the finite dimensional case, as a strategy to study existence of (weak) solutions to problems of the form of equation (1.32). If we further assume the energy $\mathcal{E}$ to be λ-convex (in a specific sense, see Section 1.2.1 below), it is also possible to show that the limit measure of the family of curves $(\rho_\tau)_{\tau\in\mathbb{R}_+}$ satisfies the EVI inequality (1.39), see [3][12].

Convexity along generalized geodesics

We stress that convexity of the energy is not at all needed in general to define the gradient flow in this way, although it may help to provide uniqueness to problem (1.40) and characterize the limit curve. A straightforward consequence of the dual formulation (1.9) and strong duality is that the functional $\mathcal{W}_2^2(\cdot,\nu):\mathcal{P}(\Omega)\to\mathbb{R}$, which associates to a measure its optimal transport cost from itself and a fixed measure ν, is convex, as it can be expressed as the sup of linear functionals. However, a different notion of convexity may fit better the problem. Before introducing it, we stress that classical convexity is fundamental when tackling discretization and numerical solution of these problems, since all the nice mathematical structure is usually lost with the discretization. Concerning this point, see for example the discussion in Section 5.1.

[11] This last hypothesis can be relaxed if one can control the decay of the energy and settles for a condition on the time step τ [3].

[12] See also Section 5.2.3, where we will use this argument in order to prove convergence in the EVI sense of a discretization similar to the MMS.

The classical notion of convexity is linked to the standard convex interpolation $\rho(t) = (1-t)\rho_1 + t\rho_2$, for any $\rho_1, \rho_2 \in \mathcal{P}(\Omega)$. From the discussion in Section 1.1, we understood that for any two measures in $\mathcal{P}(\Omega)$ we can find another type of interpolation, which may be more suited for dealing with this type of problems: the geodesic. Convexity along this type of curves is referred to as displacement convexity and it has been introduced by McCann [96]. We say that a functional $\mathcal{F}$ is displacement convex if for any two measures $\rho_1, \rho_2 \in \mathcal{P}(\Omega)$ there exists a geodesic curve $\omega(t)$ such that $\omega(0) = \rho_1, \omega(1) = \rho_2$ and the function $\mathcal{F}(\omega(t))$ is convex on $[0,1]$[13]. A lot of classic functionals are displacement convex, but unfortunately not the Wasserstein distance: for a given measure $\nu \in \mathcal{P}(\Omega)$, the functional $\mathcal{W}_2^2(\omega(t), \nu)$ is not convex in general on $[0,1]$, unless either ρ_1 or ρ_2 coincides with ν (in which case we know it is quadratic). As we want to study variational problems involving the Wasserstein distance, this notion is not really interesting for us.

To overcome this issue, we consider a more general curve joining the two endpoints ρ_1, ρ_2. We fix a measure $\nu \in \mathcal{P}(\Omega)$, which is used to "center" the curve. To simplify the idea, we consider it to be absolutely continuous for now. There exist in this way the two maps T_1 and T_2, optimal transport maps from ν to ρ_1 and ρ_2, respectively. The generalized geodesic is defined as the curve

$$\omega(t) = ((1-t)\mathrm{T}_1 + t\mathrm{T}_2)_{\#}\nu\,,$$

pushforward of the linear interpolation of the two maps. The geodesic is recovered if ν coincides with either one of the two endpoints. The functional $\mathcal{F}$ is said generalized geodesically convex if for any three measures $\rho_1, \rho_2, \nu \in \mathcal{P}(\Omega)$, $\mathcal{F}(\omega(t))$ is convex on $[0,1]$. The functional $\mathcal{W}_2^2(\omega(t), \nu)$, where the second measure ν is also taken as the center of the generalized geodesic, is now convex. It is more precisely 2-convex [3, Section 9.2]. See [3, Section 9.3] for some examples of functionals that are (generalized) geodesically convex and the respective proofs.

This definition, as for the case of the displacement interpolation, can be generalized to the case where ν is not absolutely continuous. Consider the space $\Omega \times \Omega \times \Omega$ and recall that π_1 denotes the projection on the first component, π_2 the projection on the second one, and assume π_3 to be the projection on the third one. We denote $\pi_{1,3}$ and $\pi_{2,3}$ respectively the projection on the first and third components, the projection on the second and the third one. Then, if γ_1 is the optimal transport plan from ν to ρ_1, and γ_2 is the optimal transport plan from ν to ρ_2, there exists a measure $\sigma \in \mathcal{P}(\Omega \times \Omega \times \Omega)$ such that $(\pi_{1,3})_{\#}\sigma = \gamma_1$ and $(\pi_{2,3})_{\#}\sigma = \gamma_2$. The existence of such a measure σ is guaranteed by the gluing lemma [115, Lemma 5.5]. The generalized geodesic is the curve defined as: $\omega(t) = (\pi_t)_{\#}\sigma$, where $\pi_t = (1-t)\pi_1 + t\pi_2$.

1.2.2 Dynamical form of the JKO step

We show in this section how the presence of the energy in problem (1.40) modifies the optimality conditions of the dynamical transport problem, namely system of equations (1.26). Again, we will carry out formal calculations. Thanks to the Benamou-Brenier dynamic formulation of optimal transport, the problem can be written as

$$\inf_{\substack{(\rho,\boldsymbol{v})\\ \rho\geq 0}} \frac{1}{2}\int_{t^{n-1}}^{t^n}\int_\Omega \rho|\boldsymbol{v}|^2 + \mathcal{E}(\rho(t^n,\cdot)), \tag{1.46}$$

[13] Since in general there may exist more than one geodesic, the definition just requires that there exists at least one along which the functional is convex.

where the density and velocity curves satisfy weakly

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \boldsymbol{v}) = 0 & \text{in } [t^{n-1}, t^n] \times \Omega, \\ \rho \boldsymbol{v} \cdot \boldsymbol{n} = 0 & \text{on } [t^{n-1}, t^n] \times \partial\Omega, \\ \rho(t^{n-1}, \cdot) = \rho^{n-1} & \text{in } \Omega. \end{cases} \tag{1.47}$$

Notice that we rescaled the time with respect to problem (1.15)-(1.16), and consequently the time parameter τ disappeared. Rescaling the time simply leads to rescaling the potential ϕ, as already pointed out in Section 1.1.2. The next value ρ^n is chosen equal to $\rho(t^n, \cdot)$ for the optimal ρ in (1.46)-(1.47). Using the momentum $\boldsymbol{m} = \rho \boldsymbol{v}$ instead of $\boldsymbol{v}$ as a variable, and incorporating the constraint (1.47) in (1.46), thanks to the potential $-\phi$[14], yields the saddle-point problem

$$\inf_{(\rho, \boldsymbol{m})} \sup_{\phi} \int_{t^{n-1}}^{t^n} \int_\Omega B(\rho, \boldsymbol{m}) + \int_{t^{n-1}}^{t^n} \int_\Omega (\rho \partial_t \phi + \boldsymbol{m} \cdot \nabla \phi) + \int_\Omega [\phi(t^{n-1}, \cdot)\rho^{n-1} - \phi(t^n, \cdot)\rho(t^n, \cdot)] + \mathcal{E}(\rho(t^n, \cdot)) \,. \tag{1.48}$$

We refer again to (1.48) as the primal problem. The dual problem is obtained by exchanging inf and sup in (1.48):

$$\sup_{\phi} \inf_{(\rho, \boldsymbol{m})} \int_{t^{n-1}}^{t^n} \int_\Omega B(\rho, \boldsymbol{m}) + \int_{t^{n-1}}^{t^n} \int_\Omega (\rho \partial_t \phi + \boldsymbol{m} \cdot \nabla \phi) + \int_\Omega [\phi(t^{n-1}, \cdot)\rho^{n-1} - \phi(t^n, \cdot)\rho(t^n, \cdot)] + \mathcal{E}(\rho(t^n, \cdot)) \,. \tag{1.49}$$

Strong duality can be proven again and the problem hence does not change. Optimizing first with respect to $\boldsymbol{m}$ leads to $\boldsymbol{m} = -\rho \nabla \phi$, so that the dual problem reduces to

$$\sup_{\phi} \inf_{\rho \geq 0} \int_{t^{n-1}}^{t^n} \int_\Omega (\partial_t \phi - \frac{1}{2} |\nabla \phi|^2) \rho + \int_\Omega [\phi(t^{n-1}, \cdot)\rho^{n-1} - \phi(t^n, \cdot)\rho(t^n, \cdot)] + \mathcal{E}(\rho(t^n, \cdot)) \,. \tag{1.50}$$

Because of the first term in (1.50), the infimum is equal to $-\infty$ unless $-\partial_t \phi + \frac{1}{2}|\nabla \phi|^2 \leq 0$ in $(t^{n-1}, t^n) \times \Omega$ since $\rho \geq 0$, with equality ρ-a.e.. Moreover, optimizing with respect to $\rho(t^n, \cdot)$ provides that $\phi(t^n, \cdot) \leq \frac{\delta \mathcal{E}}{\delta \rho}[\rho(t^n, \cdot)]$ with equality $\rho(t^n, \cdot)$-a.e.. Hence the dual problem can be rewritten as

$$\sup_{\phi(t^{n-1}, \cdot)} \int_\Omega \phi(t^{n-1}, \cdot)\rho^{n-1} - \mathcal{E}^*(\phi(t^n, \cdot)) \,, \tag{1.51}$$

subject to the constraint

$$-\partial_t \phi + \frac{1}{2} |\nabla \phi|^2 \leq 0 \quad \text{in } [t^{n-1}, t^n] \times \Omega \,, \tag{1.52}$$

where $\mathcal{E}^*$ denotes the Legendre transform of the energy functional $\mathcal{E}$.

[14] We enforce the constraint using the potential $-\phi$, i.e. changing sign to the potential, differently from (1.21) in Section 1.1.2, in order to obtain the minus sign in the continuity equation in analogy with equation (1.32).

A couple (ϕ,ρ) is then a saddle point for problem (1.50) if it satifies the system of primal-dual optimality conditions:

$$\begin{cases} \partial_t\rho - \nabla\cdot(\rho\nabla\phi) = 0 & \text{in } [t^{n-1},t^n]\times\Omega, \\ -\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 \le 0 & \text{in } [t^{n-1},t^n]\times\Omega, \\ -\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 = 0 & \rho-\text{a.e.}, \end{cases} \quad \text{with} \quad \begin{cases} \rho(t^{n-1},\cdot) = \rho^{n-1} & \text{in } \Omega, \\ \phi(t^n,\cdot) \le \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n,\cdot)] & \text{in } \Omega, \\ \phi(t^n,\cdot) = \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n,\cdot)] & \rho(t^n,\cdot)-\text{a.e.}. \end{cases} \tag{1.53}$$

As we said in Section 1.1.2, the Hamilton-Jacobi equation in (1.26) can be saturated thanks to the monotonicity of the operator, preserving the optimality in problem (1.24). The same is true for system of equations (1.53) (for the two conditions in this case) and problem (1.51). We want to show here how the argument works. In Thereom 3.5 and Theorem A.1, we will transpose it to the discrete setting, where we will also justify the existence of a potential solving the Hamilton-Jacobi equation.

On the one hand, the monotonicity of the backward equation $-\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 = f$ with respect to its right-hand side $f \le 0$ implies that given $\phi(t^n,\cdot)$, the solution of $-\partial_t\phi+\frac{1}{2}|\nabla\phi|^2 = 0$ gives a bigger value at $\phi(t^{n-1},\cdot)$ and thus a better competitor for (1.51). On the other hand, in order to saturate the final time constraint we use the monotonicity of the equation with respect to its final time $\phi(t^n,\cdot)$. Indeed let $(\bar\phi,\bar\rho)$ be a saddle point of (1.50) and φ be the solution of $-\partial_t\varphi + \frac{1}{2}|\nabla\varphi|^2 = -\partial_t\bar\phi + \frac{1}{2}|\nabla\bar\phi|^2$ with $\varphi(t^n,\cdot) = \frac{\delta\mathcal{E}}{\delta\rho}[\bar\rho(t^n,\cdot)] \ge \bar\phi(t^n,\cdot)$. In particular (1.53) gives $\bar\phi(t^n,\cdot) = \varphi(t^n,\cdot)$ $\rho(t^n,\cdot)$-almost everywhere and the monotonicity implies $\varphi(t^{n-1},\cdot) \ge \bar\phi(t^{n-1},\cdot)$. All together these inequalities yield:

$$\begin{aligned} &\int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\varphi - \frac{1}{2}|\nabla\varphi|^2)\bar\rho + \int_\Omega [\phi(t^{n-1},\cdot)\rho^{n-1} - \varphi(t^n,\cdot)\bar\rho(t^n,\cdot)] + \mathcal{E}(\bar\rho(t^n,\cdot)) \\ &\ge \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\bar\phi - \frac{1}{2}|\nabla\bar\phi|^2)\bar\rho + \int_\Omega [\bar\phi(t^{n-1},\cdot)\rho^{n-1} - \bar\phi(t^n,\cdot)\bar\rho(t^n,\cdot)] + \mathcal{E}(\bar\rho(t^n,\cdot)) \\ &= \sup_\phi \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\phi - \frac{1}{2}|\nabla\phi|^2)\bar\rho + \int_\Omega [\phi(t^{n-1},\cdot)\rho^{n-1} - \phi(t^n,\cdot)\bar\rho(t^n,\cdot)] + \mathcal{E}(\bar\rho(t^n,\cdot)). \end{aligned}$$

Bearing in mind the optimality of $\bar\phi$, this last inequality is then an equality and φ is again optimal. Then, thanks to (1.53) and by convexity of the energy $\mathcal{E}$, we have

$$\begin{aligned} &\int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\phi - \frac{1}{2}|\nabla\phi|^2)\bar\rho + \int_\Omega [\phi(t^{n-1},\cdot)\rho^{n-1} - \phi(t^n,\cdot)\bar\rho(t^n,\cdot)] + \mathcal{E}(\bar\rho(t^n,\cdot)) \\ &\le \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\varphi - \frac{1}{2}|\nabla\varphi|^2)\bar\rho + \int_\Omega [\varphi(t^{n-1},\cdot)\rho^{n-1} - \varphi(t^n,\cdot)\bar\rho(t^n,\cdot)] + \mathcal{E}(\bar\rho(t^n,\cdot)) \\ &\le \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\varphi - \frac{1}{2}|\nabla\varphi|^2)\rho + \int_\Omega [\varphi(t^{n-1},\cdot)\rho^{n-1} - \varphi(t^n,\cdot)\rho(t^n,\cdot)] + \mathcal{E}(\rho(t^n,\cdot)) \end{aligned}$$

for all (ϕ,ρ), which means that $(\varphi,\bar\rho)$ is also a saddle point of (1.50). The second inequality derives from the fact that the Hamilton-Jacobi is satisfied everywhere and since $\varphi(t^n,\cdot) = \frac{\delta\mathcal{E}}{\delta\rho}[\bar\rho(t^n,\cdot)]$, by convexity of $\mathcal{E}$, it holds $\mathcal{E}(\bar\rho(t^n,\cdot)) + \varphi(t^n,\cdot)(\rho(t^n,\cdot) - \bar\rho(t^n,\cdot)) \le \mathcal{E}(\rho(t^n,\cdot)), \forall\rho$. At the end of the day, the primal-dual optimality conditions of problem (1.40) finally amounts to the mean field game

$$\begin{cases} \partial_t\rho - \nabla\cdot(\rho\nabla\phi) = 0, \\ \partial_t\phi - \frac{1}{2}|\nabla\phi|^2 = 0, \end{cases} \quad \text{in } [t^{n-1},t^n]\times\Omega, \quad \text{with} \quad \begin{cases} \rho(t^{n-1},\cdot) = \rho^{n-1}, \\ \phi(t^n,\cdot) = \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n,\cdot)], \end{cases} \quad \text{in } \Omega. \tag{1.54}$$

The optimal ρ^n of (1.40) is then equal to $\rho(t^n, \cdot)$. The no-flux boundary condition reduces to $\nabla\phi \cdot \boldsymbol{n} = 0$ on $[t^{n-1}, t^n] \times \partial\Omega$.

1.3 Finite Volume Method

Finite volumes are a mean of discretizing partial differential equations, particularly suited for conservative equations. The main idea is to discretize the integral version of a conservative equation on a given partitioning of the domain, relying on the divergence theorem in order to transform a differential operator into an integral one. The main technicality is then to consistently discretize the fluxes across the faces of the partitioning. The easiest idea, which we will rely on throughout this whole work, is to discretize vector fields on each face only along a specific direction, that is orthogonal to the face. In this way, the information of a vector field is only retained on one direction and the others are discarded. This type of discretization can be done in the framework of the so called Two-Point Flux Approximation (TPFA) and gives rise to the simplest finite volume scheme. We refer to [54] for an extensive introduction to finite volumes for partial differential equations. We shall here present as an example the discretization with this methodology of the continuity equation (1.16), our building block for the approximation of optimal transport related problems.

1.3.1 The discretization of Ω

In order to design TPFA finite volume schemes, we need sufficiently regular polygonal subdivisions of the domain. One simple possibility is to consider cartesian grids, a choice that can be beneficial both for the design of the scheme and its efficiency. On such partitioning, TPFA schemes coincide with centered finite differences. However, cartesian grids are not suited to discretize complex domains, which can be of interest for applications. We need to resort to more flexible partitionings.

Let us give the definition of the regular partitioning of the domain Ω we will use. These specifications are classical for TPFA Finite Volumes [54].

Definition 1.1 (Admissible mesh of Ω)**.** Assume the domain $\Omega \subset \mathbb{R}^d$ to be polygonal if $d = 2$ or polyhedral if $d = 3$. An admissible mesh of Ω is a triplet $\left(\mathcal{T}, \overline{\Sigma}, (\boldsymbol{x}_K)_{K\in\mathcal{T}}\right)$ such that the following conditions are fulfilled:

1. Each control volume (or cell) $K \in \mathcal{T}$ is non-empty, open, polyhedral and convex. We assume that $K \cap L = \emptyset$ if $K, L \in \mathcal{T}$ with $K \neq L$, while $\bigcup_{K\in\mathcal{T}} \overline{K} = \overline{\Omega}$. The Lebesgue measure of $K \in \mathcal{T}$ is denoted by $m_K > 0$.

2. Each face $\sigma \in \overline{\Sigma}$ is closed and is contained in a hyperplane of $\mathbb{R}^d$, with positive $(d-1)$-dimensional Hausdorff (or Lebesgue) measure denoted by $m_\sigma = \mathcal{H}^{d-1}(\sigma) > 0$. We assume that $\mathcal{H}^{d-1}(\sigma \cap \sigma') = 0$ for $\sigma, \sigma' \in \overline{\Sigma}$ unless $\sigma' = \sigma$. For all $K \in \mathcal{T}$, we assume that there exists a subset $\overline{\Sigma}_K$ of $\overline{\Sigma}$ such that $\partial K = \bigcup_{\sigma\in\overline{\Sigma}_K} \sigma$. Moreover, we suppose that $\bigcup_{K\in\mathcal{T}} \overline{\Sigma}_K = \overline{\Sigma}$. Given two distinct control volumes $K, L \in \mathcal{T}$, the intersection $\overline{K} \cap \overline{L}$ either reduces to a single face $\sigma \in \overline{\Sigma}$ denoted by $K|L$, or its $(d-1)$-dimensional Hausdorff measure is 0.

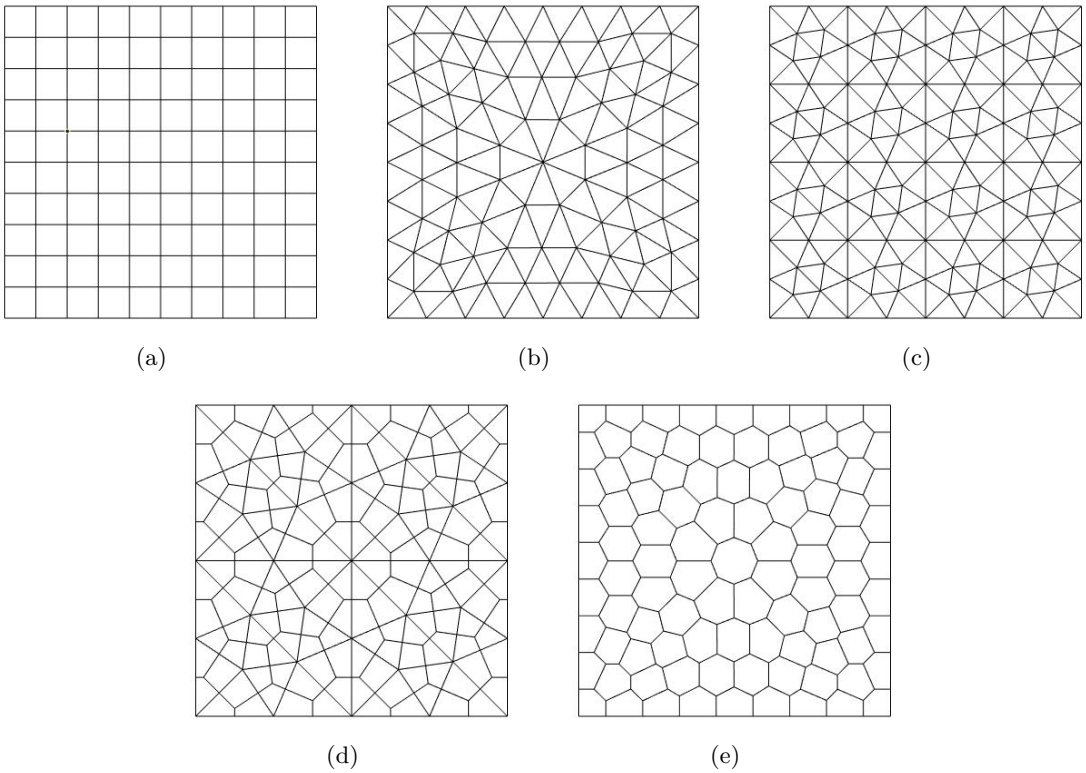


Figure 1.3: Different types of admissible mesh: cartesian grid (a), regular Delaunay triangulations (b,c), subdivision of a regular Delaunay triangulation (d), Voronoï tesselation (e).

3. The cell-centers $(\boldsymbol{x}_K)_{K\in\mathcal{T}} \subset \Omega$ are pairwise distinct and are such that, if $K, L \in \mathcal{T}$ share a face $K|L$, then the vector $\boldsymbol{x}_L - \boldsymbol{x}_K$ is orthogonal to $K|L$ and has the same orientation as the normal $\boldsymbol{n}_{K,\sigma}$ to $K|L$ outward with respect to K.

Cartesian grids, Delaunay triangulations or Voronoï tessellations are typical examples of admissible meshes in the above sense. In Figure 1.3 different type of admissible meshes are displayed for $\Omega = [0,1]^2$. We refer to [62] for a discussion on the need of such restrictive grids. Further hypotheses on the regularity of the mesh will be needed for specific convergence results and will be specified at a later time (see Sections 2.5 and 3.3).

In this work we will never consider three dimensional domains Ω and we will mainly restrict to the case $d = 2$. Nonetheless, we stress that the discretizations we will present, along with the theoretical results, can be straightforwardly recast in the three dimensional space, without any additional effort[15]. The only difference is the computational complexity for solving the discrete problems. It is of course possible to define a partioning of this type also for one dimensional domains, i.e. intervals $I \subset \mathbb{R}$. In this simplified setting, the domain can be divided in subintervals and edges collapse to points. The cell-centers can be taken in this case to be any point inside the cell, although the barycenter is the usual choice. Therefore, with the due care on the notation, also one dimensional cases will be covered by this work.

[15] For this reason, to be more general, we stuck to the three dimensional term of faces instead of edges.

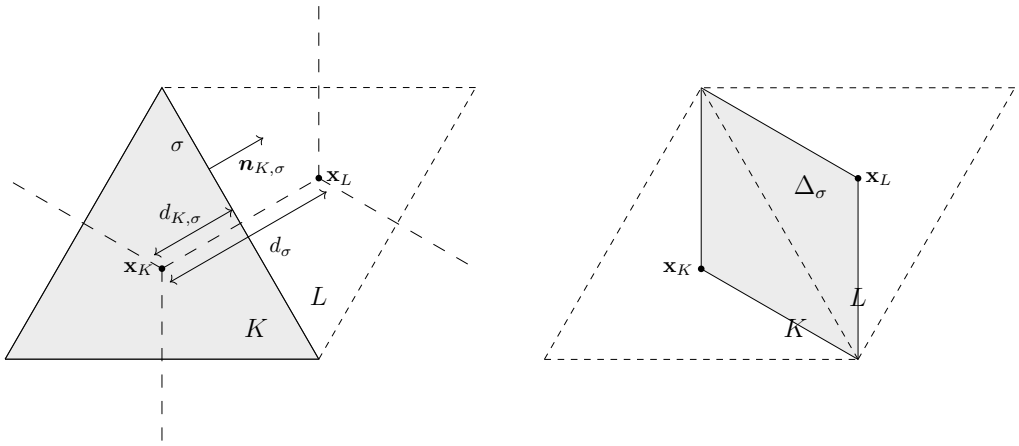


Figure 1.4: Exemplification of the notation for triangular cells.

We introduce some further notation. The set $\overline{\Sigma}$ consists of the two subsets of internal faces $\Sigma = \{\sigma \subset \Omega\}$ and external faces $\Sigma_{\text{ext}} = \{\sigma \subset \partial\Omega\} = \overline{\Sigma} \setminus \Sigma$. We denote by $\Sigma_K = \overline{\Sigma}_K \cap \Sigma$ the internal faces belonging to ∂K, and by $\mathcal{N}_K$ the neighboring cells of K, i.e., $\mathcal{N}_K = \{L \in \mathcal{T} \mid K|L \in \Sigma_K\}$. For each internal face $\sigma = K|L \in \Sigma$, we refer to the diamond cell Δ_σ as the polyhedron whose edges join $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ to the vertices of σ. The diamond cell Δ_σ is convex if $\boldsymbol{x}_K \in K$ and $\boldsymbol{x}_L \in L$. Denoting by $d_\sigma = |\boldsymbol{x}_K - \boldsymbol{x}_L|$, the measure m_{Δ_σ} of Δ_σ is then equal to $m_\sigma d_\sigma / d$, where d stands for the space dimension. We denote by $d_{K,\sigma}$ the Euclidean distance between the cell center $\mathbf{x}_K$ and the edge $\sigma \in \overline{\Sigma}_K$. The quantity $a_\sigma = m_\sigma / d_\sigma$ defines the transmissivity of the face $\sigma \in \Sigma$. In Figure 1.4 the notation is exemplified for a triangular cell.

1.3.2 Discrete continuity equation

In order to present how finite volumes work, let us consider the simple example of the continuity equation (1.16). Suppose for simplicity that the vector field $\boldsymbol{v}$ is given and that we want to approximate the density evolution ρ.

First of all, we discretize the time derivative with a simple (here implicit) Euler formula. For this purpose, consider a constant time step $\Delta t = \frac{1}{N+1}$, for some integer $N > 0$, and denote by $t^k = k\Delta t$, for $k \in \{0, .., N+1\}$. We can approximate the solution ρ at time t^k, $\forall k \in \{1, .., N+1\}$, by

$$\rho^k - \rho^{k-1} + \nabla \cdot (\rho^k \boldsymbol{v}(t^k)) = 0\,, \tag{1.55}$$

with the boundary conditions $\rho^0 = \rho^{in}, \rho^{N+1} = \rho^f$. Consider now a control volume $K \in \mathcal{T}$. Integrating at step k equation (1.55) over K and using the divergence theorem provides:

$$\int_K (\rho^k - \rho^{k-1}) + \int_{\partial K} \rho^k \boldsymbol{v}(t^k) \cdot \boldsymbol{n}_K = 0\,, \tag{1.56}$$

where $\boldsymbol{n}_K$ is the outer normal of K. Since we suppose the cell K to be polygonal, we can rewrite the second term in (1.56) as:

$$\int_{\partial K} \rho^k \boldsymbol{v}(t^k) \cdot \boldsymbol{n}_K = \sum_{\sigma \in \Sigma_K} \int_\sigma \rho^k \boldsymbol{v}(t^k) \cdot \boldsymbol{n}_{K,\sigma}\,. \tag{1.57}$$

Notice that we are automatically taking into account the boundary condition $\rho \boldsymbol{v} \cdot \boldsymbol{n} = 0$, which is the reason why we restrict the sum in (1.57) to the internal edges only of the cell K.

For each $K \in \mathcal{T}$ and $k \in \{0, .., N+1\}$ we associate a value ρ_K^k which is meant to approximate the mean value of ρ^k on the cell K,

$$\rho_K^k \approx \frac{1}{m_K} \int_K \rho^k \, .$$

The last step to have a fully discrete version of equation (1.16) is then to approximate the fluxes on each face $\sigma \in \Sigma$. That is, we have to choose how to reconstruct the mobility, the density associated with the flux across the face. For $\sigma = K|L$, the simplest idea is to use the discrete densities of the two neighboring cells, ρ_K^k and ρ_L^k. If we denote $r(\rho_K^k, \rho_L^k)$ such value, and we introduce

$$v_{K,\sigma}^k = \frac{1}{m_\sigma} \int_\sigma \boldsymbol{v}(t^k) \cdot \boldsymbol{n}_{K,\sigma} \, , \quad \forall K \in \mathcal{T} \, , \; \forall \sigma \in \Sigma_K \, , \; \forall k \in \{1, .., N+1\},$$

we can finally write equation (1.56) as:

$$(\rho_K^k - \rho_K^{k-1}) m_K + \sum_{\sigma \in \Sigma_K} r(\rho_K^k, \rho_L^k) v_{K,\sigma}^k m_\sigma = 0 \, . \tag{1.58}$$

Equation (1.58) is then complemented with the conditions $\rho_K^0 = \rho_K^{in}$, $\rho_K^0 = \rho_K^{in}$, $\forall K \in \mathcal{T}$, where $(\rho_K^{in})_{K \in \mathcal{T}}$ and $(\rho_K^f)_{K \in \mathcal{T}}$ are discrete approximations, on each cell, of the continuous counterparts ρ^{in}, ρ^f.

Different choices are possible for the function $r(\cdot, \cdot)$, which essentially consist in averaging the values of the two neighboring cells. We will consider several possibilities in the next chapters. If we denote $F_{K,\sigma}^k = r(\rho_K^k, \rho_L^k) v_{K,\sigma}^k$, we can further write the discrete continuity equation as

$$(\rho_K^k - \rho_K^{k-1}) m_K + \sum_{\sigma \in \Sigma_K} F_{K,\sigma}^k m_\sigma = 0 \, , \tag{1.59}$$

which is the discrete version of the equation in the density-momentum variables, namely equation (1.20)[16]. This type of finite volume discretization is called two-point flux because each discrete flux $F_{K,\sigma}^k$ in (1.59) depends only on the two values ρ_K^k and ρ_L^k.

1.3.3 Discrete spaces and operators

The Finite Volume Method, at least in the simple case we are considering, replaces the continuous solutions with discrete ones defined on the partitioning of the domain. Let us formalize better the approach by precisely introducing the discrete spaces and operators.

We denote by $\mathbb{R}^{\mathcal{T}}$ and $\mathbb{R}^{\Sigma}$ the two discrete spaces of discrete quantities defined on the control volumes and the diamond cells. They are respectively endowed with the weighted scalar products

$$\langle \cdot, \cdot \rangle_{\mathcal{T}} : (\boldsymbol{a}, \boldsymbol{b}) \in [\mathbb{R}^{\mathcal{T}}]^2 \mapsto \sum_{K \in \mathcal{T}} a_K b_K m_K \, ,$$
$$\langle \cdot, \cdot \rangle_{\Sigma} : (\boldsymbol{s}, \boldsymbol{u}) \in [\mathbb{R}^{\Sigma}]^2 \mapsto \sum_{\sigma \in \Sigma} s_\sigma u_\sigma m_\sigma d_\sigma \, .$$

[16] Notice that each $F_{K,\sigma}^k$ discretizes a flux according to (1.57) and not directly the momentum, whence our notation. We will sometimes refer anyway to it as discrete momentum.

For $\boldsymbol{a} \in \mathbb{R}^{\mathcal{T}}$, its mass is defined as $\langle \boldsymbol{a}, \mathbf{1} \rangle_{\mathcal{T}}$. We define the element-wise multiplication by $\odot$. We denote by $\mathbf{1}$ the constant element whose components are all equal to one, equivalently for $\mathbf{0}$. By convention, $\boldsymbol{a}^{-1}$ is the vector whose components are the reciprocal of the components of $\boldsymbol{a}$. The relation $\boldsymbol{a} \geq \boldsymbol{b}$ will always be intended componentwise.

We introduce also the space of discrete conservative fluxes,

$$\mathbb{F}_{\mathcal{T}} = \{ \boldsymbol{F} = (F_{K,\sigma}, F_{L,\sigma})_{\sigma \in \Sigma} \in \mathbb{R}^{2\Sigma} : F_{K,\sigma} + F_{L,\sigma} = 0 \} , \tag{1.60}$$

endowed with the scalar product

$$\langle \cdot, \cdot \rangle_{\mathbb{F}_{\mathcal{T}}} : (\boldsymbol{F}, \boldsymbol{G}) \in [\mathbb{F}_{\mathcal{T}}]^2 \mapsto \sum_{\sigma \in \Sigma} (F_{K,\sigma} G_{K,\sigma} + F_{L,\sigma} G_{L,\sigma}) \frac{m_\sigma d_\sigma}{2} .$$

We remark that the definition (1.60) does not take into account the boundary faces. Due to the no-flux boundary condition, there exist no fluxes on the boundary faces and we can discard them[17]. For the same reason, we do not need to reconstruct the mobility on the boundary and the space $\mathbb{R}^{\Sigma}$ only considers internal diamond cells.

We define now the discrete operators involved in equation (1.59). The discrete divergence $\mathrm{div}_{\mathcal{T}} : \mathbb{F}_{\mathcal{T}} \to \mathbb{R}^{\mathcal{T}}$ is defined by $(\mathrm{div}_{\mathcal{T}} \boldsymbol{F})_K \coloneqq \mathrm{div}_K(\boldsymbol{F})$, where

$$\mathrm{div}_K \boldsymbol{F} \coloneqq \frac{1}{m_K} \sum_{\sigma \in \Sigma_K} F_{K,\sigma} m_\sigma .$$

We can also define the discrete gradient $\nabla_\Sigma : \mathbb{R}^{\mathcal{T}} \to \mathbb{F}_{\mathcal{T}}$ by duality as $\langle \nabla_\Sigma \boldsymbol{a}, \boldsymbol{F} \rangle_{\mathbb{F}_{\mathcal{T}}} = -\langle \boldsymbol{a}, \mathrm{div}_{\mathcal{T}} \boldsymbol{F} \rangle_{\mathcal{T}}$. In particular we also have $(\nabla_\Sigma \boldsymbol{a})_{K,\sigma} = \nabla_{K,\sigma} \boldsymbol{a}$ where

$$\nabla_{K,\sigma} \boldsymbol{a} \coloneqq \frac{a_L - a_K}{d_\sigma} .$$

With these definitions at hand, we can rewrite equation (1.59). For $(\boldsymbol{\rho}^k)_{k=0}^{N+1} \subset \mathbb{R}^{\mathcal{T}}$ and $(\boldsymbol{F}^k)_{k=1}^{N+1} \subset \mathbb{F}_{\mathcal{T}}$, the discrete continuity equation writes as

$$\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} + \mathrm{div}_{\mathcal{T}} \boldsymbol{F}^k = 0 , \quad \forall k \in \{1, .., N+1\} ,$$

which is complemented by $\boldsymbol{\rho}^0 = \boldsymbol{\rho}^{in}, \boldsymbol{\rho}^{N+1} = \boldsymbol{\rho}^f$, where $\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f \in \mathbb{R}_+^{\mathcal{T}}$ are the discrete initial and final conditions. Due to the definition of the space of conservative fluxes (1.60), the equation is also automatically enforced with zero flux at the boundary. Then, owing to the conservation property $F_{K,\sigma} + F_{L,\sigma} = 0$, $\forall \sigma \in \mathbb{R}^{\Sigma}$, the total discrete mass is preserved at each iteration k, i.e.

$$\langle \frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} + \mathrm{div}_{\mathcal{T}} \boldsymbol{F}^k, \mathbf{1} \rangle_{\mathcal{T}} = \langle \frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}, \mathbf{1} \rangle_{\mathcal{T}} = 0,$$

which implies

$$\langle \boldsymbol{\rho}^k, \mathbf{1} \rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^{in}, \mathbf{1} \rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^f, \mathbf{1} \rangle_{\mathcal{T}} , \quad \forall k \in \{1, .., N\} ,$$

assuming of course the two discrete densities $\boldsymbol{\rho}^{in}$ and $\boldsymbol{\rho}^f$ to have the same total mass.

[17] This implies that, in the discrete optimization problems we will consider, the constraint on the no-flux condition will be explicitly enforced.

In the following chapters, we will also introduce for convenience the reconstruction operator from cells to diamond cells $\mathcal{R}_\Sigma : \mathbb{R}^\mathcal{T} \to \mathbb{R}^\Sigma$, which provides the definition for the mobility:

$$\left(\mathcal{R}_\Sigma(\boldsymbol{a})\right)_\sigma = r(a_K, a_L)\,, \quad \text{for } \boldsymbol{a} \in \mathbb{R}^\mathcal{T}\,.$$

Thanks to this operator, using the convention $\boldsymbol{u} \odot \boldsymbol{w} \in \mathbb{F}_\mathcal{T}, (\boldsymbol{u} \odot \boldsymbol{w})_{K,\sigma} = u_\sigma w_{K,\sigma}$, for $\boldsymbol{u} \in \mathbb{R}^\Sigma, \boldsymbol{w} \in \mathbb{F}_\mathcal{T}$, we can rewrite in vectorial form the change of variables

$$\boldsymbol{F}^k = \mathcal{R}_\Sigma(\boldsymbol{\rho}^k)\boldsymbol{v}^k\,, \quad \forall k \in \{1, .., N+1\}\,.$$

For the discretization of optimal transport problems we will need also another operator, $\mathcal{R}_\mathcal{T} : \mathbb{R}^\Sigma \to \mathbb{R}^\mathcal{T}$, which acts in the opposite way with respect to $\mathcal{R}_\Sigma$, reconstructing cell values from diamond cell ones. The form of $\mathcal{R}_\mathcal{T}$ depends on $\mathcal{R}_\Sigma$, in order to preserve the variational structure of the problem we will consider. The specific form of $\mathcal{R}_\Sigma$, and consequently of $\mathcal{R}_\mathcal{T}$, will be detailed later.

Chapter 2

Computation of optimal transport with finite volumes

This chapter contains an extended presentation of:

Andrea Natale and Gabriele Todeschi. Computation of optimal transport with finite volumes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(5):1847–1871, September 2021.

2.1 Introduction

We presented in the previous chapter the quadratic optimal transport problem, which provides the notion of distance we need to interpret problems of the form of equation (1.32) as gradient flows. We want to focus in this chapter on how we can compute numerically this distance. As we already said, optimal transport started to gain more and more attention recently thanks to the progresses in numerical methods to approximate it and nowadays several approaches are available. We suggest the interested reader to have a look at [115, Chapter 6],[98, 111, 13] for a complete introduction to the subject and the state of the art. For what concerns the quadratic problem, we could naively distinguish two categories: approaches for the static formulation (1.5) and for the dynamical one introduced by Benamou and Brenier. Let us recall this latter. Given two densities ρ^{in} and ρ^{f}, we aim at finding the two curves $(\rho, \boldsymbol{m}) : [0,1]\times\Omega \to \mathbb{R}_+ \times \mathbb{R}^d$, respectively the density displacement and the momentum, which solve

$$\inf_{(\rho,\boldsymbol{m})} \int_0^1 \int_\Omega B(\rho(t,\boldsymbol{x}), \boldsymbol{m}(t,\boldsymbol{x}))\, \mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \tag{2.1}$$

while satisfying in distributional sense the continuity equation

$$\begin{cases} \partial_t \rho + \nabla\cdot \boldsymbol{m} = 0 & \text{in } [0,1]\times\Omega\,, \\ \boldsymbol{m}\cdot\boldsymbol{n} = 0 & \text{on } [0,1]\times\partial\Omega\,, \end{cases} \tag{2.2}$$

with the further initial and final conditions $\rho(0,\cdot) = \rho^{in}, \rho(1,\cdot) = \rho^{f}$. The function $B : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty]$ represents the density of kinetic energy of the curve and is defined as

$$B(p, \boldsymbol{Q}) := \begin{cases} \frac{|\boldsymbol{Q}|^2}{2p} & \text{if } p > 0\,, \\ 0 & \text{if } p = 0,\ \boldsymbol{Q} = 0\,, \\ +\infty & \text{else}\,. \end{cases} \tag{2.3}$$

This dynamical formulation has inspired some of the first numerical methods for optimal transport. This form of the problem enables to compute not only the optimal reallocation but also the continuous in time displacement, which makes it more interesting for applications. In particular, it is easily generalizable to other problems, such as mean field games [18], the Schrödinger bridge problem [87], unbalanced optimal transport [46] or capacity constrained optimal transport [29], just to name a few. A lot of interest is still devoted for this reason on approaches to compute this dynamical formulation and several techniques are already available. However, only few of these can actually be generalized to more complex settings which are relevant for numerical modeling. Moreover, their numerical analysis is often neglected.

We consider here finite volume discretizations for the dynamical optimal transport problem, following the approach originally proposed by Benamou and Brenier [14]. We will focus on three main aspects. Firstly, we will expose some numerical issues related to the stability of finite volume methods that have been considered for this problem, and we propose a strategy based on nested meshes to overcome these. Secondly, we provide quantitative estimates on the convergence of the proposed methods to smooth solutions of the problem. Finally, we tackle the issue of the efficient computation of numerical solutions by applying and analyzing a classical interior point strategy adapted to our setting. We will provide extensive numerical results to validate our approach.

2.1.1 Discretization of dynamical optimal transport

In the original work of Benamou and Brenier [14] problem (2.1)-(2.2) was discretized on regular grids using centered finite differences. Later in [109] Papadakis, Peyré and Oudet introduced a finite difference discretization using staggered grids, which are better suited for the discretization of the continuity equation. Similar finite difference approaches have been used also in more recent works [41, 85]. Note that the use of regular grids can be beneficial for the efficient solution of the scheme, but it is not adapated to complex domains. Several finite element approaches have been considered in order to construct schemes able to handle more general unstructured grids [16, 17, 80]. In Appendix B we propose a $H(\mathrm{div})$-conforming finite element discretization that preserves at the discrete level the conservative form of the problem, in the same spirit of [109].

Another approach to discretize problem (2.1)-(2.2) is to use finite volumes, which is a natural choice given the conservative form of the constraint (2.2) and allows one to use unstructured grids. In [51], Erbar, Rumpf, Schmitzer and Simon considered a discretization of problem (2.1)-(2.2) on graphs, which can be written under the formalism of Two-Point Flux Approximation (TPFA) finite volumes [64]. They proved the Gamma-convergence of the discrete problem towards a semi-discrete version of (2.1)-(2.2), discrete in space and continuous in time. In [64], Gladbach, Kopfer and Maas proved a convergence result for this semi-discretization towards the continuous problem. Combining these two results, it is possible to obtain a global convergence result, under conditions on the ratio between the temporal

and spatial step sizes. Carrillo, Craig, Wang and Wei proved the Gamma-convergence without conditions on the step sizes but only for sufficiently regular and strictly positive solutions [41]. They used a centered finite difference discretization, which coincide with TPFA finite volumes on cartesian grids. Finally, in [79] Lavenant proved the weak convergence of discrete solutions (reconstructed as space-time measures) of a large class of time-space discretizations of (2.1)-(2.2), unconditionally with respect to time and space steps and without assuming any regularity. Lavenant applied this result to the discretization studied in [51] and the one proposed in [80]. We will apply it to the finite element discretization we present in Appendix B.

Our starting point in this work is the finite volume discretization presented in [79, 51]. We observe numerically that for this discretization the density interpolation can exhibit oscillations which prevent strong convergence of the numerical solution, even when the exact interpolation is smooth. The same phenomenon has been observed by Facca and coauthors in [56, 57] when dealing with finite element discretizations for the L^1 optimal transport problem, which is closely related to (2.1)-(2.2). Our strategy to overcome this issue is inspired by these last works and consists in enriching the space of discrete potentials. We will show numerically that such a modification attenuates the oscillations and favors a stronger convergence.

Note that with this modification, the convergence result in [79] cannot be applied straightforwardly. However, we will derive quantitative estimates for the convergence of the discrete Wasserstein distance and the discrete potential, which hold both in the enriched and original non-enriched case, in the case of smooth and strictly positive solutions. Even if such results are only partial as they do not apply to the density, they are still surprising given that the problem is not strictly convex. Moreover, we are not aware of similar estimates for the discretizations mentioned above. With these results at hand, it is possible to deduce again the weak convergence of the discrete density and momentum.

2.1.2 Numerical solution

A typical approach for solving discrete versions of the dynamical formulation (2.1)-(2.2) is to apply first order primal dual methods. This goes back to the original paper of Benamou and Brenier [14], who proposed to use an Alternating Direction Method of Multipliers (ADMM) approach applied to the augmented Lagrangian of the discrete saddle point problem. Later [109] considered different proximal splitting methods and recast the previous algorithm into the same framework. Nowadays, these approaches are frequently used [16, 17, 80, 41, 103]. In fact, they are robust and can take care automatically of the positivity of the density thanks to the definition of the objective functional and the function B (2.3). Nevertheless, they are not easy to apply to arbitrary discretizations of the problem (especially on unstructured grids). More importantly, they are efficient only as far as high accuracy is not mandatory and uniform grids are used.

In the present work, we apply the so called barrier method, an instance of the wider class of interior point methods [26, 65, 113, 58]. The problem is perturbed by adding to the functional a strictly convex barrier function which repulses the density away from zero. In this way it reduces to an equality-constrained minimization problem, where the minimizer is automatically greater than zero and the objective functional is locally smooth around it, and which can be effectively solved using a Newton scheme. The perturbation introduced by the barrier function can be tuned by multiplying it by a positive coefficient μ and the original

solution is recovered via a continuation method for μ going to zero. The final algorithm is robust and can be easily generalized to similar problems (see for example Chapter 4).

The idea of using a regularization term to deal with numerical solutions of optimal transport is not new. Among others, entropic regularization deserves a special mention. It has been introduced in [47] to approximate solutions of problem (1.5) and it has gained more and more attention thereafter. The idea is to add to the objective functional the entropy of the transport and it is particularly effective as it allows to use a robust algorithm to solve the discrete problem, the Sinkhorn algorithm. The approach is extremely simple, computationally cheap and particularly suited to high dimensions. However, the problem is solved for a finite value of the perturbation, as the complexity of the algorithm explodes otherwise, and it finds particular success for applications where some additional diffusion is tolerated. A modified version of the approach, able to recover the unpertubed solution via a continuation method, has been proposed only recently [19]. Notice that the same regularization technique can be applied to the dynamical formulation (2.1)-(2.2), by perturbing the problem with the entropy of the density curve. This entropic-dynamical formulation has gained a lot of attention recently for its relation with the Schrödinger problem [87]. However, we are not aware of numerical strategies employing this type of regularization for the dynamical optimal transport problem.

In [85], to regularize problem (2.1)-(2.2) and deal automatically with the positivity constraint, the authors proposed to use the Fisher information. However, they didn't consider a continuation method and the problem is solved for a fixed (small) value of the perturbation's parameter, leading to diffusive effects. A strategy similar to ours has been applied instead by Achdou and coauthors [1] (although in the context of mean field games), perturbing the Lagrangian associated to the problem with the Dirichlet energies of the density and the potential. Such a perturbation does not ensure the positivity of the solution and this forces the use of a monotone discretization. Using a barrier function allows us to consider more general discretizations, with higher accuracy in space.

2.2 Finite volume discretization

We introduced in section 1.3.3 the finite dimensional spaces and operators which are involved in the finite volume discretization. They rely on an admissible mesh discretizing the domain Ω according to Definition 1.1. We want to slightly generalize this construction.

2.2.1 Nested meshes

In order to be more flexible in the discretization of problem (2.1)-(2.2), we want to be able to decouple the discretization of the potential and the density. We introduce then two different (admissible) discretizations of Ω. In particular, we will take one as a subdivision of the other. We denote by $\left(\mathcal{T}', \overline{\Sigma'}, (\mathbf{x}_{K'})_{K' \in \mathcal{T}'}\right)$ the coarse mesh and by $\left(\mathcal{T}, \overline{\Sigma}, (\mathbf{x}_K)_{K \in \mathcal{T}}\right)$ the fine one, and we require that

$$\forall K \in \mathcal{T},\ \exists\, K' \in \mathcal{T}' \text{ such that } \overline{K} \subseteq \overline{K}'\,.$$

In practice we will consider two specific instances of this construction. The first is the trivial case where the two meshes coincide. The second holds at least in two dimensions and can be defined as follows. First, we take as coarse mesh a Delaunay triangulation, with cell centers $\boldsymbol{x}_{K'}$ the circumcenters of each cell K'. We further require that all the triangles are

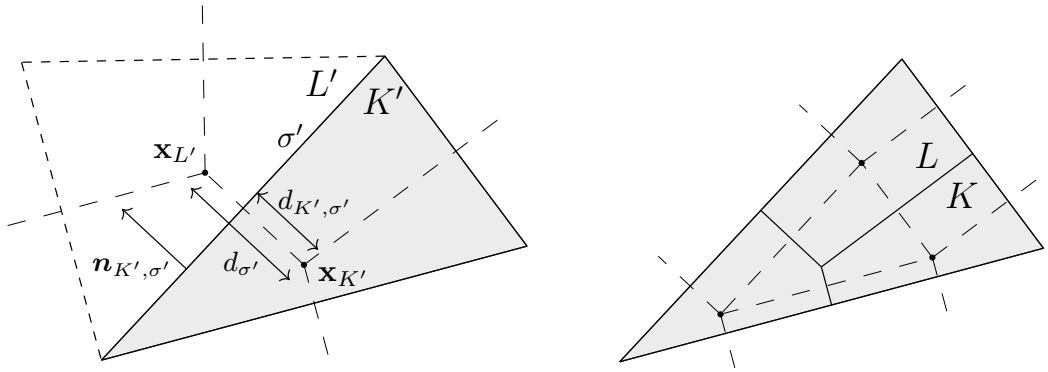


Figure 2.1: Exemplification of the notation of a triangular cell (left) and its subdivision (right).

acute, so that all the cell centers $\boldsymbol{x}_{K'}$ lie in the interior of the corresponding cell K'. Then, we define the fine mesh by dividing each triangular cell K' into three quadrilaterals by joining $\boldsymbol{x}_{K'}$ to the three midpoints of the edges $\sigma' \in \overline{\Sigma}'_{K'}$. We take again as cell centers $\boldsymbol{x}_K$ of the fine mesh the circumcenters of each cell K. This construction is illustrated in Figure 1.3d (which is the subdivision of a mesh of the type of Figure 1.3c) and in Figure 2.1. Note that the partition obtained in this way is indeed admissible. Other constructions are possible.

2.2.2 Discrete spaces and operators

We introduce two spaces of discrete variables defined on the two meshes, $\mathbb{R}^{\mathcal{T}}$ and $\mathbb{R}^{\mathcal{T}'}$, each one endowed with its own weighted scalar product,

$$\langle \cdot, \cdot \rangle_{\mathcal{T}} : (\boldsymbol{a}, \boldsymbol{b}) \in [\mathbb{R}^{\mathcal{T}}]^2 \mapsto \sum_{K \in \mathcal{T}} a_K b_K m_K \,,$$

and similarly for $\langle \cdot, \cdot \rangle_{\mathcal{T}'}$. Note that $\mathbb{R}^{\mathcal{T}'} \subseteq \mathbb{R}^{\mathcal{T}}$, and we denote by $\mathcal{I}$ the canonical injection operator, which is given explicitly by

$$\mathcal{I} : \mathbb{R}^{\mathcal{T}'} \to \mathbb{R}^{\mathcal{T}}, \quad (\mathcal{I}\rho)_K = \rho_{K'}, \quad \forall K \subset K'.$$

In the case where the two discretizations of Ω coincide, $\mathcal{I}$ is just the identity operator. We will denote by $\mathcal{I}^*$ the adjoint of $\mathcal{I}$, i.e. $\langle \mathcal{I}^* \cdot, \cdot \rangle_{\mathcal{T}'} = \langle \cdot, \mathcal{I} \cdot \rangle_{\mathcal{T}}$.

The space of discrete variables on diamond cells $\mathbb{R}^{\Sigma}$ and the space of conservative fluxes $\mathbb{F}_{\mathcal{T}}$,

$$\mathbb{F}_{\mathcal{T}} = \{ \boldsymbol{F} = (F_{K,\sigma}, F_{L,\sigma})_{\sigma \in \Sigma} \in \mathbb{R}^{2\Sigma} : F_{K,\sigma} + F_{L,\sigma} = 0 \} \,, \tag{2.4}$$

are both defined on the finer mesh only. They are respectively endowed with the scalar products:

$$\langle \cdot, \cdot \rangle_{\Sigma} : (\boldsymbol{u}, \boldsymbol{v}) \in [\mathbb{R}^{\Sigma}]^2 \mapsto \sum_{\sigma \in \Sigma} u_\sigma v_\sigma m_\sigma d_\sigma \,,$$

$$\langle \cdot, \cdot \rangle_{\mathbb{F}_{\mathcal{T}}} : (\boldsymbol{F}, \boldsymbol{G}) \in [\mathbb{F}_{\mathcal{T}}]^2 \mapsto \sum_{\sigma \in \Sigma} (F_{K,\sigma} G_{K,\sigma} + F_{L,\sigma} G_{L,\sigma}) \frac{m_\sigma d_\sigma}{2} \,.$$

We denote by $\| \cdot \|_{\mathcal{T}}$, $\| \cdot \|_{\mathcal{T}'}$, $\| \cdot \|_{\Sigma}$ and $\| \cdot \|_{\mathbb{F}_{\mathcal{T}}}$ the norms associated with the inner products defined above. We also denote $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$ and by convention $|\boldsymbol{F}| = (F_\sigma)_{\sigma \in \Sigma} \in \mathbb{R}^{\Sigma}$

and $(\boldsymbol{F})^2 = (F_\sigma^2)_{\sigma\in\Sigma} \in \mathbb{R}^\Sigma$, for $\boldsymbol{F} \in \mathbb{F}_\mathcal{T}$. Moreover, given $\boldsymbol{F}, \boldsymbol{G} \in \mathbb{F}_\mathcal{T}$ and $\boldsymbol{u} \in \mathbb{R}^\Sigma$, we define $\boldsymbol{F} \odot \boldsymbol{G}, \boldsymbol{u} \odot \boldsymbol{F} \in \mathbb{F}_\mathcal{T}$ by

$$[\boldsymbol{F} \odot \boldsymbol{G}]_{K,\sigma} \coloneqq F_{K,\sigma} G_{K,\sigma}\,, \quad [\boldsymbol{u} \odot \boldsymbol{F}]_{K,\sigma} \coloneqq u_\sigma F_{K,\sigma}\,.$$

The discrete divergence $\mathrm{div}_\mathcal{T} : \mathbb{F}_\mathcal{T} \to \mathbb{R}^\mathcal{T}$ and the discrete gradient $\nabla_\Sigma : \mathbb{R}^\mathcal{T} \to \mathbb{F}_\mathcal{T}$ are defined on the finer mesh and writes:

$$(\mathrm{div}_\mathcal{T} \boldsymbol{F})_K = \mathrm{div}_K \boldsymbol{F} = \frac{1}{m_K} \sum_{\sigma\in\Sigma_K} F_{K,\sigma} m_\sigma\,,$$

$$(\nabla_\Sigma \boldsymbol{a})_{K,\sigma} = \nabla_{K,\sigma} \boldsymbol{a} \coloneqq \frac{a_L - a_K}{d_\sigma}\,.$$

It holds $\langle \nabla_\Sigma \boldsymbol{a}, \boldsymbol{F}\rangle_{\mathbb{F}_\mathcal{T}} = -\langle \boldsymbol{a}, \mathrm{div}_\mathcal{T} \boldsymbol{F}\rangle_\mathcal{T}$. Moreover, as for the discrete conservative fluxes, we define $\nabla_\sigma \boldsymbol{a} \coloneqq |\nabla_{K,\sigma} \boldsymbol{a}|$. Finally, let us give the explicit form of the reconstruction operator $\mathcal{R}_\Sigma : \mathbb{R}^\mathcal{T} \to \mathbb{R}^\Sigma$, from cells to diamond cells of the finer grid, that we will use to discretize the mobility. We require the operator $\mathcal{R}_\Sigma$ to be a concave function (component-wise), positively 1-homogeneous and positivity preserving. We will consider two weighted means, $\mathcal{L}_\Sigma$ and $\mathcal{H}_\Sigma$, which correspond respectively to a linear and a harmonic mean, and are defined as follows:

$$(\mathcal{L}_\Sigma \boldsymbol{a})_\sigma = \frac{d_{K,\sigma}}{d_\sigma} a_K + \frac{d_{L,\sigma}}{d_\sigma} a_L\,, \quad (\mathcal{H}_\Sigma \boldsymbol{a})_\sigma = \frac{d_\sigma a_K a_L}{d_{K,\sigma} a_L + d_{L,\sigma} a_K}\,, \tag{2.5}$$

for any $\boldsymbol{a} \in \mathbb{R}^\mathcal{T}$. We denote by $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}] : \mathbb{R}^\mathcal{T} \to \mathbb{R}^\Sigma$ the differential of $\mathcal{R}_\Sigma$ with respect to $\boldsymbol{a}$, evaluated at a given $\boldsymbol{a} \in \mathbb{R}^\mathcal{T}$. Clearly, if $\mathcal{R}_\Sigma = \mathcal{L}_\Sigma$, we simply have $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}] = \mathcal{L}_\Sigma$. Moreover, we denote by $(\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}])^*$ the adjoint of $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}]$, with respect to the two different scalar products. For the two reconstructions we consider, this operator is given by either $\mathcal{L}_\Sigma^*$ or $(\mathrm{d}\mathcal{H}_\Sigma[\boldsymbol{a}])^*$, which are defined by

$$(\mathcal{L}_\Sigma^* \boldsymbol{u})_K = \sum_{\sigma\in\Sigma_K} u_\sigma \frac{m_\sigma d_{K,\sigma}}{m_K}\,, \quad ((\mathrm{d}\mathcal{H}_\Sigma[\boldsymbol{a}])^* \boldsymbol{u})_K = \sum_{\sigma\in\Sigma_K} \frac{(\mathcal{H}_\Sigma[\boldsymbol{a}])_\sigma^2}{a_K^2} u_\sigma \frac{m_\sigma d_{K,\sigma}}{m_K}\,, \tag{2.6}$$

for any $\boldsymbol{u} \in \mathbb{R}^\Sigma$.

2.3 Time discretization

The discretization in time of problem (2.1)-(2.2) deserves some comments. The continuous in time displacement of mass needs to be replaced with a sequence of finite steps. Consider an integer $N > 0$ and a discretization of the time interval $[0, 1]$ in $N + 1$ subintervals of constant length $\Delta t = \frac{1}{N+1}$, and let $t^k \coloneqq k\Delta t$ for all $k \in \{0, .., N + 1\}$. Consider a sequence of densities $(\rho^k)_{k=0}^{N+1} \subset \mathcal{P}(\Omega)$, such that $\rho^0 = \rho^{in}, \rho^{N+1} = \rho^f$. It is natural to discretize the time derivative in (2.2) with a simple Euler step with uniform time step Δt, introducing then a sequence of staggered momentum $(\boldsymbol{m}^k)_{k=1}^{N+1}$ that pushes the mass from one step to the other:

$$\frac{\rho^k - \rho^{k-1}}{\Delta t} + \nabla \cdot \boldsymbol{m}^k = 0, \quad \forall k \in \{1, .., N + 1\} \tag{2.7}$$

As the momentum are staggered in time with respect to the densities, we can see (2.7) as a midpoint discretization of equation (2.2). We can think of the time dependent density discretized with a piecewise linear one on each subinterval $[t^{k-1}, t^k]$, whereas the momentum with a piecewise constant one. On each subinterval the integral in time of the kinetic energy can be approximated for example with a left/right endpoint approximation or a midpoint rule. We want to comment on these choices thanks to the corresponding optimality conditions. The derivation of the discrete optimality conditions will be detailed later in Section 2.4, we want here to give just a quick look.

In order to have a finite kinetic energy, and therefore a finite candidate solution, the momentum has to be absolutely continuous with respect to the density, that is, loosely speaking, it has to be proportional to it (see Section 1.1.2). Consider a left endpoint approximation of the kinetic energy, that is

$$\int_0^1 \int_\Omega B(\rho, \boldsymbol{m}) \, \mathrm{d}\mathbf{x}\mathrm{d}t \approx \sum_{k=1}^{N+1} \Delta t \int_\Omega B(\rho^{k-1}, \boldsymbol{m}^k) \, \mathrm{d}\mathbf{x} \, .$$

Repeating the same steps as in Section 1.1.2, bearing in mind the optimality condition for the momentum (1.22), a left endpoint approximation would provide the following discrete continuity equation

$$\frac{\rho^k - \rho^{k-1}}{\Delta t} + \nabla \cdot (\rho^{k-1} \nabla \phi^k) = 0, \quad \forall k \in \{1, .., N+1\} \, ,$$

where $\boldsymbol{m}^k = \rho^{k-1} \nabla \phi^k$, ϕ^k being the Lagrange multiplier for the k-th equation. At each step, moving accordingly to the chosen time direction, the continuity equation is discretized with an explicit Euler step (considering the velocity field to be given). Assume the initial density $\rho^0 = \rho^{in}$ to have compact support. It is evident then that the mass cannot flow outside of its support, as the momentum is zero in this region and the density ρ^1 cannot have bigger support. Recursively, this is true at every step, therefore an admissible couple (ϕ, ρ) (or an admissible triplet $(\phi, \rho, \boldsymbol{F})$ with finite kinetic energy) can exist if and only if the support of the final density $\rho^{N+1} = \rho^f$ is not bigger than the support of ρ^{in}. The situation is the opposite for a right endpoint approximation of the kinetic energy. These choices appear too rigid.

Remark 2.1. *Due to the finite volumes discretization, as the momentum is defined on the edges and the mobility is averaged on adjacent cells (see Section 2.4), in the fully discrete scheme this issue may turn into a condition on a sufficiently high number of intermediate steps N for the existence of a candidate solution.*

Consider now a discretization of the kinetic energy with a midpoint rule:

$$\int_0^1 \int_\Omega B(\rho, \boldsymbol{m}) \, \mathrm{d}\mathbf{x}\mathrm{d}t \approx \sum_{k=1}^{N+1} \Delta t \int_\Omega B\Big(\frac{\rho^{k-1} + \rho^k}{2}, \boldsymbol{m}^k\Big) \, \mathrm{d}\mathbf{x} \, .$$

We obtain in this case the condition $\boldsymbol{m}^k = (\frac{\rho^k + \rho^{k-1}}{2}) \nabla \phi^k$ for the momentum, and the continuity equation is discretized with a midpoint rule as well (considering again the vector field $\nabla \phi^k$ to be given). The first advantage is that the discretization is now symmetric in time. As the optimal transport problem is symmetric, it would make sense to discretize it symmetrically.

One may also expect a higher precision for this discretization with respect to the previous one. More importantly, the previous issue with the supports of the final and initial conditions disappears as the mobility is not explicit in neither time directions. A candidate solution always exists, it suffices that the supports of two consecutive densities of the sequence $(\rho^k)_{k=0}^{N+1}$ always intersect. This of course turns into a high speed of propagation of the support of the densities if the number of intermediate steps N is not big.

Despite the several reasons for avoiding a left/right endpoint approximation of the kinetic energy, this choice presents an advantage with respect to the midpoint rule. Using a left (right) endpoint approximation leads to an implicit approximation of the Hamilton-Jacobi equation in the positive (negative) direction of time. The discrete equation in the first case is:

$$\frac{\phi^{k+1}-\phi^k}{\Delta t}+\frac{1}{2}|\nabla\phi^{k+1}|^2\leq 0,\quad \forall k\in\{1,..,N\}\,. \tag{2.8}$$

As mentioned in Section 1.1.2, the Hamilton-Jacobi equation can be saturated thanks to the monotonicity of the operator in both time directions. The discrete equation (2.8) is monotone as well, but only in one direction. We can show it thanks to formal computations. The following argument will be adjusted to the fully discrete case in Section 3.2 and appendix A. Consider indeed two potentials $(\phi_1^k)_{k=1}^{N+1}, (\phi_2^k)_{k=1}^{N+1}$ (sufficiently regular), verifying $\phi_1^1=\phi_2^1$ and

$$\frac{\phi_1^{k+1}-\phi_1^k}{\Delta t}+\frac{1}{2}|\nabla\phi_1^{k+1}|^2\leq\frac{\phi_2^{k+1}-\phi_2^k}{\Delta t}+\frac{1}{2}|\nabla\phi_2^{k+1}|^2\,,\quad \forall k\in\{1,..,N\}\,. \tag{2.9}$$

We define then by $(\boldsymbol{x}^{k+1})_{k=1}^N$ a sequence of points verifying $\boldsymbol{x}^{k+1}\in\operatorname{argmin}\phi_2^{k+1}-\phi_1^{k+1}$. For $k=1$, we have

$$\frac{\phi_1^2}{\Delta t}+\frac{1}{2}|\nabla\phi_1^2|^2\leq\frac{\phi_2^2}{\Delta t}+\frac{1}{2}|\nabla\phi_2^2|^2\,,$$

and since $\nabla\phi_1^2(\boldsymbol{x}^2)=\nabla\phi_2^2(\boldsymbol{x}^2)$ by definition of $\boldsymbol{x}^2$, this provides $\phi_1^2(\boldsymbol{x}^2)\leq\phi_2^2(\boldsymbol{x}^2)$ and therefore $\phi_1^2\leq\phi_2^2$. Assuming $\phi_1^k\leq\phi_2^k$, at the step $k+1$ it holds

$$\frac{\phi_1^{k+1}}{\Delta t}+\frac{1}{2}|\nabla\phi_1^{k+1}|^2\leq\frac{\phi_1^{k+1}}{\Delta t}+\frac{1}{2}|\nabla\phi_1^{k+1}|^2+\frac{\phi_2^k-\phi_1^k}{\Delta t}\leq\frac{\phi_2^{k+1}}{\Delta t}+\frac{1}{2}|\nabla\phi_2^{k+1}|^2\,,$$

so that $\phi_1^{k+1}(\boldsymbol{x}^{k+1})\leq\phi_2^{k+1}(\boldsymbol{x}^{k+1})$ and therefore $\phi_1^{k+1}\leq\phi_2^{k+1}$. By recurrence, this is true for all k and we have $\phi_1^{N+1}\leq\phi_2^{N+1}$.

The same reasoning cannot be applied to show that $\phi_2^1\leq\phi_1^1$ under the hypotheses $\phi_2^{N+1}=\phi_1^{N+1}$ and (2.9), and the result clearly does not hold. Therefore the discrete in time operator is monotone only in one direction, the positive direction of time. Approximating the kinetic energy with a right endpoint approximation would lead to a monotone operator in the negative direction of time. In this case it is possible to show that $\phi_2^1\leq\phi_1^1$ if $\phi_2^{N+1}=\phi_1^{N+1}$ and (2.9) holds, using the same argument (exchanging the min with a max). In either one direction or the other, it is possible to saturate the Hamilton-Jacobi equation in the discrete problem, as we did in the continuous case in Section 1.2.2 and as we shall see in the discrete setting in Appendix A.

2.4 Discrete optimal transport problem

We introduce now the fully discrete scheme. We will start from the convex formulation of the dynamical optimal transport problem in order to preserve the variational structure at the

discrete level. We will use here the midpoint discretization in time introduced in the previous section. Starting from the classical finite volume discretization of the continuity equation (2.2), we discretize the kinetic energy and we derive by duality the corresponding discrete form of the Hamilton-Jacobi equation (1.25), finally ending up with a discrete version of the system of optimality conditions (1.26).

We denote the time evolution of a discrete density by $\boldsymbol{\rho} := (\boldsymbol{\rho}^k)_{k=0}^{N+1}$, where $\boldsymbol{\rho}^k := (\rho^k_{K'})_{K'\in\mathcal{T}'}$. Similarly we denote by $\boldsymbol{F} := (\boldsymbol{F}^k)_{k=1}^{N+1}$ the time evolution of a discrete momentum, where $\boldsymbol{F}^k := (F^k_{K,\sigma}, F^k_{L,\sigma})_{\sigma\in\Sigma}$. Given a couple $(\boldsymbol{\rho}, \boldsymbol{F}) \in [\mathbb{R}^{\mathcal{T}'}]^{N+2} \times [\mathbb{F}_{\mathcal{T}}]^{N+1}$, we define the discrete equivalent of the objective functional in (2.1), $\mathcal{B}_{N,\mathcal{T}} : [\mathbb{R}^{\mathcal{T}}]^{N+2} \times [\mathbb{F}_{\mathcal{T}}]^{N+1} \to [0, +\infty]$, as follows:

$$\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) := \begin{cases} \sum_{k=1}^{N+1} \Delta t \sum_{\sigma\in\Sigma} B\big(((\mathcal{R}_\Sigma \circ \mathcal{I})(\frac{\boldsymbol{\rho}^k+\boldsymbol{\rho}^{k-1}}{2}))_\sigma, F^k_\sigma\big) m_\sigma d_\sigma & \text{if } \rho^k_{K'} \geq 0\,, \\ +\infty & \text{else}\,. \end{cases} \tag{2.10}$$

Since $\mathcal{R}_\Sigma$ is assumed to be concave, the function (2.10) is convex and lower semi-continuous.

At each time step, the kinetic energy is discretized on the diamond cells of the finer grid. As the momentum are defined on this grid, the density is first injected in the finer space and then reconstructed on the edges. Thanks to the discretization with the midpoint rule of the kinetic energy on each subinterval $[t^{k-1}, t^k] = [(k-1)\Delta t, k\Delta t]$, a given F^k_σ needs to vanish only if the reconstruction of $(\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1})/2$ on the same edge vanishes. Notice that the measure of each diamond cell is taken d times. This is done in order to compensate the unidirectional discretization of the vector field F and therefore obtain a consistent discretization (see, e.g., Lemma 2.6). Indeed, each F_σ is meant as an approximation of the modulus of the unitary flux, $|\boldsymbol{m} \cdot \boldsymbol{n}_{K,\sigma}|$, and encodes then the information of $\boldsymbol{m}$ only along the direction $\boldsymbol{n}_{K,\sigma}$. This choice is linked to the definition of inflated gradient (see [44, 55] for more details on this construction).

Remark 2.2. *Note that (2.10) is not simply the discretization of the objective functional in (2.1) on the diamond cells, in which case it would take the value $+\infty$ whenever the time-space reconstruction of the density is negative on some diamond cell. The functional in (2.10) takes the value $+\infty$ whenever the density is negative on some cell $K' \in \mathcal{T}'$, which is a stronger condition.*

Given two discrete densities $\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f \in \mathbb{R}_+^{\mathcal{T}'}$, with the same total discrete mass, $\langle \boldsymbol{\rho}^{in}, \mathbf{1}\rangle_{\mathcal{T}'} = \langle \boldsymbol{\rho}^f, \mathbf{1}\rangle_{\mathcal{T}'}$, we consider the following discrete version of problem (2.1)-(2.2):

$$\inf_{(\boldsymbol{\rho},\boldsymbol{F})\in\mathcal{C}_{N,\mathcal{T}}} \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F})\,, \tag{2.11}$$

where $\mathcal{C}_{N,\mathcal{T}} \subset [\mathbb{R}^{\mathcal{T}'}]^{N+2} \times [\mathbb{F}_{\mathcal{T}}]^{N+1}$ is the convex subset whose elements $(\boldsymbol{\rho}, \boldsymbol{F})$ satisfy both the discrete continuity equation

$$\mathcal{I}\Big(\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}\Big) + \mathrm{div}_{\mathcal{T}} \boldsymbol{F}^k = 0\,, \quad \forall k \in \{1, .., N+1\}, \tag{2.12}$$

and the initial and final conditions

$$\boldsymbol{\rho}^0 = \boldsymbol{\rho}^{in}, \quad \boldsymbol{\rho}^{N+1} = \boldsymbol{\rho}^f\,. \tag{2.13}$$

The continuity equation is discretized in time using the midpoint rule ($\boldsymbol{F}$ is indeed staggered in time with respect to $\boldsymbol{\rho}$). Moreover, given the definition of the discrete space of conservative fluxes and the operator $\mathrm{div}_{\mathcal{T}}$, (2.12) is to be understood with zero flux boundary conditions in space. Hence equations (2.12)-(2.13) imply that the total discrete mass is preserved at all times: $\langle \boldsymbol{\rho}^k, \mathbf{1}\rangle_{\mathcal{T}'} = \langle \boldsymbol{\rho}^{in}, \mathbf{1}\rangle_{\mathcal{T}'}, \forall k \in \{1, .., N\}$. In the following, we explicitly enforce the constraint (2.13), i.e. we identify $\boldsymbol{\rho}^0$ and $\boldsymbol{\rho}^{N+1}$ with $\boldsymbol{\rho}^{in}$ and $\boldsymbol{\rho}^f$, respectively.

Theorem 2.3. *Problem (2.11) admits a solution.*

Proof. First of all, notice that existence of a finite valued feasible point, i.e. an element $(\boldsymbol{\rho}, \boldsymbol{F})$ satisfying the constraint and with finite kinetic energy, is ensured thanks to the surjectivity of the divergence operator (to the space of discrete functions in $[\mathbb{R}^{\mathcal{T}}]^{N+1}$ with zero mean) and the time discretization of the functional (as explained in the previous section). Consider for example a density with mass everywhere.

Let us now show that a minimizing sequence is bounded. As the total mass is preserved, any density of the sequence lies in the compact set $||\boldsymbol{\rho}||_\infty \leq \frac{\langle \boldsymbol{\rho}^{in}, \mathbf{1}\rangle_{\mathcal{T}'}}{\min_{K' \in \mathcal{T}'}(m_{K'})}$. Thus we just need to show that also the momenta are bounded. For any $\boldsymbol{b} \in [\mathbb{F}_{\mathcal{T}}]^{N+1}$, with $|b^k_\sigma| \leq 1$ for all $\sigma \in \Sigma, k \in \{1, .., N+1\}$, denoting by $\boldsymbol{\rho}_s \in [\mathbb{R}^{\Sigma}]^{N+1}$ the term given by

$$\boldsymbol{\rho}^k_s = \sqrt{(\mathcal{R}_\Sigma \circ \mathcal{I})\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}}\,, \quad \forall k \in \{1, .., N+1\}\,,$$

it holds:

$$\begin{aligned}\sum_{k=1}^{N+1} \Delta t \langle \boldsymbol{F}^k, \boldsymbol{b}^k\rangle_{\mathbb{F}_{\mathcal{T}}} &= \sum_{k=1}^{N+1} \Delta t \langle \boldsymbol{F}^k \odot (\boldsymbol{\rho}^k_s)^{-1}, \boldsymbol{b}^k \odot \boldsymbol{\rho}^k_s\rangle_{\mathbb{F}_{\mathcal{T}}} \\ &\leq \sqrt{2\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F})}||\boldsymbol{b}||_{\boldsymbol{\rho}} \leq C\sqrt{2\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F})}\,.\end{aligned} \tag{2.14}$$

The weighted (semi-)norm $||\cdot||_{\boldsymbol{\rho}}$ is defined via (2.23). The first inequality derives applying Cauchy-Schwarz whereas the second one from the uniform bound on the density. Taking the sup with respect to $\boldsymbol{b}$ we obtain the bound on $\boldsymbol{F}$. Therefore a minimizing sequence is bounded. The existence of a minimizer follows from the lower semi-continuity of the function $\mathcal{B}_{N,\mathcal{T}}$ and the linearity of the constraint. □

By applying standard Lagrange duality for convex optimization problems (see for example [26]), we can observe that for problem (2.11) strong duality holds and that the dual problem attains its optimal value. Then we can equivalently say that a primal-dual couple $(\boldsymbol{\phi}, (\boldsymbol{\rho}, \boldsymbol{F}))$ of solutions to the primal and dual problems is the saddle point of the Lagrangian function

$$\mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{F}) = \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) + \sum_{k=1}^{N+1} \Delta t \langle \boldsymbol{\phi}^k, \mathcal{I}\Big(\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}\Big) + \mathrm{div}_{\mathcal{T}} \boldsymbol{F}^k\rangle_{\mathcal{T}}\,, \tag{2.15}$$

where the potential $\boldsymbol{\phi} \in [\mathbb{R}^{\mathcal{T}}]^{N+1}$ is the Lagrange multiplier for the continuity equation constraint.

We derive now the first order optimality conditions for problem (2.11), which are necessary and sufficient conditions for a solution. Equivalently, we want to derive the stationarity conditions of the Lagrangian $\mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{F})$. By definition of the discrete gradient operator

and thanks to the definition of conservative fluxes (2.4), the stationarity condition of $\mathcal{L}_{N,\mathcal{T}}$ with respect to $\boldsymbol{F}$ provides:

$$\boldsymbol{F}^k = (\mathcal{R}_\Sigma \circ \mathcal{I})\Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big) \odot \nabla_\Sigma \boldsymbol{\phi}^k\,, \quad \forall k \in \{1,..,N+1\}, \tag{2.16}$$

Plugging this condition in (2.15), the Lagrangian reduces to

$$\mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi},\boldsymbol{\rho}) = -\frac{\Delta t}{2}\sum_{k=1}^{N+1}\langle(\mathcal{R}_\Sigma \circ \mathcal{I})\Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big), (\nabla_\Sigma \boldsymbol{\phi}^k)^2\rangle_\Sigma + \sum_{k=1}^{N+1}\Delta t\langle \boldsymbol{\phi}^k, \mathcal{I}\Big(\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}\Big)\rangle_\mathcal{T}\,. \tag{2.17}$$

A stationary point of (2.17) must then satisfy the conditions:

$$\begin{cases} \mathcal{I}\Big(\dfrac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}\Big) + \mathrm{div}_\mathcal{T}\Big((\mathcal{R}_\Sigma \circ \mathcal{I})\Big(\dfrac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big) \odot \nabla_\Sigma \boldsymbol{\phi}^k\Big) = 0\,, \\ \mathcal{I}^*\Big(\dfrac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t}\Big) + \dfrac{1}{4}\mathcal{R}^k_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^k)^2 + \dfrac{1}{4}\mathcal{R}^{k+1}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^{k+1})^2 \leq 0\,, \end{cases} \tag{2.18}$$

where $k \in \{1,..,N+1\}$ for the discrete continuity equation, $k \in \{1,..,N\}$ for the discrete Hamilton-Jacobi equation. The linear operators $\mathcal{R}^k_{\mathcal{T}'} : \mathbb{R}^\Sigma \to \mathbb{R}^{\mathcal{T}'}$, for $k \in \{1,..,N+1\}$, are defined by

$$\langle \mathcal{R}^k_{\mathcal{T}'}\boldsymbol{u}, \boldsymbol{b}\rangle_{\mathcal{T}'} = \langle \boldsymbol{u}, (\mathrm{d}\mathcal{R}_\Sigma\Big[\mathcal{I}\Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big)\Big] \circ \mathcal{I})\,\boldsymbol{b}\rangle_\Sigma\,, \quad \text{for } \boldsymbol{u} \in \mathbb{R}^\Sigma\,,\, \boldsymbol{b} \in \mathbb{R}^{\mathcal{T}'}\,.$$

We recall that $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}] : \mathbb{R}^\mathcal{T} \to \mathbb{R}^\Sigma$ is the differential of $\mathcal{R}_\Sigma$ with respect to $\boldsymbol{a}$, evaluated at a given $\boldsymbol{a} \in \mathbb{R}^\mathcal{T}$. If $\mathcal{R}_\Sigma = \mathcal{L}_\Sigma$, then these operators do not depend on $\boldsymbol{\rho}$ and in particular we will drop such dependency in the notation by setting $\mathcal{R}^k_{\mathcal{T}'} = \mathcal{R}_{\mathcal{T}'} = \mathcal{I}^* \circ \mathcal{L}^*_\Sigma$.

The inequality in the second condition derives from the fact that the minimization in $\boldsymbol{\rho}$ is taken over non-negative values, and the equality holds where $\boldsymbol{\rho}^k$ does not vanish. Hence, we can write the full system of optimality conditions using a slack variable $\boldsymbol{\lambda} \in [\mathbb{R}^{\mathcal{T}'}_-]^N$:

$$\begin{cases} \mathcal{I}\Big(\dfrac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}\Big) + \mathrm{div}_\mathcal{T}\Big((\mathcal{R}_\Sigma \circ \mathcal{I})\Big(\dfrac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big) \odot \nabla_\Sigma \boldsymbol{\phi}^k\Big) = 0, \\ \mathcal{I}^*\Big(\dfrac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t}\Big) + \dfrac{1}{4}\mathcal{R}^k_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^k)^2 + \dfrac{1}{4}\mathcal{R}^{k+1}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^{k+1})^2 = \boldsymbol{\lambda}^k, \\ \boldsymbol{\rho}^k \geq 0,\ \boldsymbol{\lambda}^k \leq 0,\ \boldsymbol{\rho}^k \odot \boldsymbol{\lambda}^k = 0, \end{cases} \tag{2.19}$$

where $k \in \{1,..,N+1\}$ for the discrete continuity equation and $k \in \{1,..,N\}$ for the other conditions. Note that system (2.19) is a discrete version of the system of optimality conditions (1.26) holding at the continuous level. In particular, the continuity equation is discretized on the fine grid whereas the Hamilton-Jacobi equation on the coarse one. Using a discretization that preserves the monotonocity of the discrete Hamilton-Jacobi operator it is possible to show that the value zero for $\boldsymbol{\lambda}$ is optimal (see Appendix A), i.e. the discrete Hamilton-Jacobi equation can be saturated. However this is not the case for the discretizations we considered here since they do not preserve the monotonicity.

Remark 2.4. *If the two discretizations of Ω coincide, $\mathcal{I}$ becomes the identity and we recover the finite volumes discretization already considered in [51, 79], which is a fully discrete version of the continuous-time discrete optimal transport problem studied in [64].*

Remark 2.5. *Uniqueness of the density interpolation, which is guaranteed for the continuous problem (2.1)-(2.2) as soon as the initial (or final) measure is absolutely continuous with respect to the Lebesgue measure [121, Corollary 7.23], is not evident. System (2.19) is not guaranteed in general to have a unique solution. In particular, where the density vanishes, the potential and the positivity multiplier are clearly non unique. The potential is however uniquely defined, up to a global constant, if the density solution is unique and everywhere strictly positive.*

Given a solution $(\boldsymbol{\phi}, \boldsymbol{\rho})$ to system (2.19), we can construct the associated momentum $\boldsymbol{F}$ by equation (2.16) so that $(\boldsymbol{\rho}, \boldsymbol{F})$ is a minimizer of problem (2.11). Then, we define the discrete Wasserstein distance $W_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^{f})$ by

$$\frac{W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^{f})}{2} \coloneqq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}). \tag{2.20}$$

More precisely, replacing (2.16) in (2.20), the discrete Wasserstein distance can be computed using the following expression:

$$\frac{W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^{f})}{2} = \frac{\Delta t}{2} \sum_{k=1}^{N+1} \langle (\mathcal{R}_\Sigma \circ \mathcal{I}) \Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2} \Big), (\nabla_\Sigma \boldsymbol{\phi}^k)^2 \rangle_\Sigma \,. \tag{2.21}$$

In the case of the linear reconstruction, i.e. taking $\mathcal{R}_\Sigma = \mathcal{L}_\Sigma$, one can also easily express the dual to problem (2.11) in terms of the potential $\boldsymbol{\phi}$, as in the continuous case, i.e. problem (1.9). In this case in fact, maximizing the Lagrangian (2.17) over potentials $\boldsymbol{\phi}$ verifying the second condition of system (2.19) yields the following problem:

$$\sup_{\boldsymbol{\phi} \in \mathcal{K}_{N,\mathcal{T}}} \langle \mathcal{I}^* \boldsymbol{\phi}^{N+1} - \frac{\Delta t}{4} \mathcal{R}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^{N+1})^2, \boldsymbol{\rho}^f \rangle_{\mathcal{T}} - \langle \mathcal{I}^* \boldsymbol{\phi}^1 + \frac{\Delta t}{4} \mathcal{R}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^1)^2, \boldsymbol{\rho}^{in} \rangle_{\mathcal{T}} \tag{2.22}$$

where $\mathcal{R}_{\mathcal{T}'} = \mathcal{I}^* \circ \mathcal{L}^*_\Sigma$ and $\mathcal{K}_{N,\mathcal{T}} \subset [\mathbb{R}^{\mathcal{T}}]^{N+1}$ is the convex subset of potentials $\boldsymbol{\phi}$ verifying

$$\mathcal{I}^* \Big(\frac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t} \Big) + \frac{1}{4} \mathcal{R}_{\mathcal{T}'} \big((\nabla_\Sigma \boldsymbol{\phi}^k)^2 + (\nabla_\Sigma \boldsymbol{\phi}^{k+1})^2 \big) \leq 0 \,.$$

2.5 Convergence to the continuous problem

In this section, we provide quantitative estimates for the convergence of the action and the discrete potential $\boldsymbol{\phi}$ towards their continuous counterparts, in the case of solutions with smooth strictly positive densities. Note that we restrict ourselves to the case of the linear reconstruction operator, i.e. we take $\mathcal{R}_\Sigma = \mathcal{L}_\Sigma$. As a consequence of Remark 2.4, these results are also valid for the finite volume discretization considered in [79].

First of all, we introduce some additional notation. Let $\boldsymbol{F}, \boldsymbol{G} \in [\mathbb{F}_{\mathcal{T}}]^{N+1}$ and $\boldsymbol{\rho} \in [\mathbb{R}^{\mathcal{T}'}_+]^{N+2}$. We define the following weighted inner products:

$$\langle \boldsymbol{F}, \boldsymbol{G} \rangle_{\boldsymbol{\rho}} \coloneqq \Delta t \sum_{k=1}^{N+1} \langle \boldsymbol{F}^k, \boldsymbol{G}^k \rangle_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \,, \tag{2.23}$$

where

$$\langle \boldsymbol{F}^k, \boldsymbol{G}^k \rangle_{\boldsymbol{\rho}^k} \coloneqq \sum_{\sigma\in\Sigma} (F^k_{K,\sigma} G^k_{K,\sigma} + F^k_{L,\sigma} G^k_{L,\sigma})((\mathcal{R}_\Sigma \circ \mathcal{I})\boldsymbol{\rho}^k)_\sigma \frac{m_\sigma d_\sigma}{2} .$$

We will denote by $\|\cdot\|_{\boldsymbol{\rho}}$ and $\|\cdot\|_{\boldsymbol{\rho}^k}$ the (semi-)norms associated with these (semi-)inner products.

We will consider two sampling operators: one for the density $\Pi_{\mathcal{T}'}$, which performs an average on each cell, and one for the potential $\Pi_{\mathcal{T}}$, which evaluates the function at the cell centers. More precisely, given f and g sufficiently regular, we define

$$(\Pi_{\mathcal{T}'} f)_{K'} \coloneqq \frac{1}{m_{K'}} \int_{K'} f \,\mathrm{d}x , \quad (\Pi_{\mathcal{T}} g)_K \coloneqq g(\boldsymbol{x}_K) ,$$

for all $K' \in \mathcal{T}'$ and all $K \in \mathcal{T}$. For any time dependent functions ρ and ϕ sufficiently regular, we define $\overline{\Pi}_{\mathcal{T}'}\rho \coloneqq (\Pi_{\mathcal{T}'}\rho(t^k,\cdot))_{k=0}^{N+1}$ and

$$\overline{\Pi}_{\mathcal{T}}\phi \coloneqq \left(\frac{1}{\Delta t} \int_{t^{k-1}}^{t^k} \Pi_{\mathcal{T}}\phi(s,\cdot)\mathrm{d}s \right)_{k=1}^{N+1} .$$

We will denote by $h_{\mathcal{T}}$ the maximum cell diameter of the fine mesh, i.e. $h_{\mathcal{T}} \coloneqq \max_{K\in\mathcal{T}} \mathrm{diam}(K)$. We will assume two regularity conditions on the fine mesh. Firstly, there exists a constant ζ, which does not depend on $h_{\mathcal{T}}$, such that

$$\mathrm{diam}(K) \le \zeta d_\sigma \le \zeta^2 \mathrm{diam}(K), \quad \forall \sigma \in \Sigma_K,\ \forall K \in \mathcal{T} , \tag{2.24}$$

$$\mathrm{dist}(\boldsymbol{x}_K, K) \le \zeta\, \mathrm{diam}(K), \quad \forall K \in \mathcal{T} . \tag{2.25}$$

Secondly, there exists a constant $\eta_h > 0$ only depending on $h_{\mathcal{T}}$, with $\eta_h \to 0$ for $h_{\mathcal{T}} \to 0$, such that

$$\sum_{\sigma\in\Sigma_K} m_\sigma d_{K,\sigma} \boldsymbol{n}_{K,\sigma} \otimes \boldsymbol{n}_{K,\sigma} \le m_K(1+\eta_h)\mathrm{Id} , \quad \forall K \in \mathcal{T} . \tag{2.26}$$

The latter condition is essentially a specific instance of the asymptotic isotropy condition in [64] (see Definition 1.3). When the cell centers $\boldsymbol{x}_K$ are chosen as the circumcenters of the associated cell (as in the particular examples of meshes described in Section 2.2.1), a stronger property holds, which has been referred to as center of mass condition [64] or superadmissibility [53], and which reads as follows:

$$\sum_{\sigma\in\Sigma_K} m_\sigma d_{K,\sigma} \boldsymbol{n}_{K,\sigma} \otimes \boldsymbol{n}_{K,\sigma} = m_K \mathrm{Id} . \tag{2.27}$$

However, for generality of the discussion, in the following we will only require (2.26) and therefore we will keep the dependence on η_h explicit.

The following lemma collects some consistency properties of the projection $\Pi_{\mathcal{T}}$. In particular, point (3) below shows that the asymptotic isotropy condition implies the consistency of the quadratic term in the discrete Wasserstein distance (2.21), and justifies our discretization of the functional $\mathcal{B}_{N,\mathcal{T}}$.

Lemma 2.6. *The following properties hold:*

1. *for any* $\psi \in C^0(\Omega)$, $\max_{K\in\mathcal{T}} |(\Pi_\mathcal{T}\psi)_K| \le \|\psi\|_{C^0}$;
2. *for any* $\psi \in C^{0,1}(\Omega)$, *there exists a constant* $C>0$ *only depending on* ψ *and* ζ *such that*
$$\max_{K\in\mathcal{T}} \|(\Pi_\mathcal{T}\psi)_K - \psi\|_{C^0(\Omega)} \le Ch_\mathcal{T}\,;$$
3. *for any* $\psi \in C^{1,1}(\Omega)$, *there exists a constant* $C>0$ *only depending on* ψ *and* ζ *such that*
$$(\mathcal{L}_\Sigma^*|\nabla_\Sigma \Pi_\mathcal{T}\psi|^2)_K \le (\Pi_\mathcal{T}|\nabla\psi|^2)_K + C(h_\mathcal{T}+\eta_h)\,,$$
for all $K\in\mathcal{T}$, *where* $\mathcal{L}_\Sigma$ *is the linear reconstruction operator and* η_h *is defined as in* (2.26).

Proof. The first two points follow easily from the definition of $\Pi_\mathcal{T}$ and the regularity condition (2.25). For (3), observe that, by definition of the linear reconstruction operator,
$$(\mathcal{L}_\Sigma^*|\nabla_\sigma \Pi_\mathcal{T}\psi|^2)_K = \sum_{\sigma\in\Sigma_K} |\nabla_\sigma \Pi_\mathcal{T}\psi|^2 \frac{m_\sigma d_{K,\sigma}}{m_K}\,. \tag{2.28}$$
Then, using the definition of the operator $\Pi_\mathcal{T}$ and the regularity condition (2.24),
$$\nabla_\sigma \Pi_\mathcal{T}\psi = \left|\frac{\psi(\boldsymbol{x}_K)-\psi(\boldsymbol{x}_L)}{d_\sigma}\right| = \frac{1}{d_\sigma}\left|\int_0^1 \frac{\mathrm{d}}{\mathrm{d}s}\psi((1-s)\boldsymbol{x}_K + s\boldsymbol{x}_L)\mathrm{d}s\right| \le |\nabla\psi(x_K)\cdot\boldsymbol{n}_{K,\sigma}| + Ch_\mathcal{T}\,.$$
Replacing this into (2.28), neglecting higher order terms, and using the asymptotic isotropy assumption (2.26), we obtain the desired bound. □

Propostion 2.8 below is an adaptation to our setting of standard approximation results for elliptic problems. It quantifies the consistency of the projection $\overline{\Pi}_{\mathcal{T}'}$ in terms of the associated potential. As in [64], we will use it to construct an admissible competitor for the discrete problem. Before proving the result, we state the following classical finite-volume version of the Poincaré inequality.

Lemma 2.7 (Discrete mean Poincaré inequality, Lemma 10.2 in [54])**.** *There exists a constant* $C>0$, *only depending on* Ω, *such that for all admissible meshes* $\mathcal{T}$, *and for all* $\boldsymbol{\psi}\in\mathbb{R}^\mathcal{T}$, *the following inequality holds:*
$$\|\boldsymbol{\psi} - \frac{1}{|\Omega|}\sum_{K\in\mathcal{T}} \psi_K m_K\|_\mathcal{T} \le C\|\nabla_\Sigma\boldsymbol{\psi}\|_{\mathbb{F}_\mathcal{T}}\,.$$

Proposition 2.8. *Suppose that* $\rho, \partial_t\rho \in L^\infty([0,1];C^{0,1}(\Omega))$, *with* $\rho\ge\varepsilon>0$, *and let* $\phi\in L^\infty([0,1];C^{1,1}(\Omega))$ *be a solution of*
$$-\operatorname{div}(\rho\nabla\phi) = \partial_t\rho\,,\quad \nabla\phi\cdot n_{\partial\Omega} = 0 \ \text{ on } \partial\Omega\,. \tag{2.29}$$
Let $\boldsymbol{\rho} = \overline{\Pi}_{\mathcal{T}'}\rho$ *and let* $\boldsymbol{\phi}$ *be a solution of*
$$-\mathrm{div}_\mathcal{T}\Big((\mathcal{L}_\Sigma\circ\mathcal{I})\Big(\frac{\boldsymbol{\rho}^k+\boldsymbol{\rho}^{k-1}}{2}\Big)\odot\nabla_\Sigma\boldsymbol{\phi}^k\Big) = \mathcal{I}\Big(\frac{\boldsymbol{\rho}^k-\boldsymbol{\rho}^{k-1}}{\Delta t}\Big)\,.$$
Then, there exists a constant $C>0$ *depending only on* ϕ, ρ, ε, ζ *and* Ω, *such that*
$$\|\nabla_\Sigma\boldsymbol{\phi}\|_{\boldsymbol{\rho}}^2 \le \int_0^1\int_\Omega \rho|\nabla\phi|^2\,\mathrm{d}\boldsymbol{x}\mathrm{d}t + C(h_\mathcal{T}+\Delta t+\eta_h)\,, \tag{2.30}$$
with η_h *defined as in* (2.26).

Proof. First, we integrate equation (2.29) over the time-space cell $[t^{k-1}, t^k] \times K$ and divide it by $m_K \Delta t$. This yields

$$-\operatorname{div}_K \boldsymbol{u}^k = \frac{1}{m_K \Delta t} \int_K \int_{t^{k-1}}^{t^k} \partial_t \rho \, \mathrm{d}t \mathrm{d}\boldsymbol{x} \,. \tag{2.31}$$

where $\boldsymbol{u} \in [\mathbb{F}_{\mathcal{T}}]^{N+1}$ is defined by

$$u^k_{K,\sigma} \coloneqq \frac{1}{m_\sigma \Delta t} \int_\sigma \int_{t^{k-1}}^{t^k} (\rho \nabla \phi) \cdot \boldsymbol{n}_{K,\sigma} \, \mathrm{d}t \, \mathrm{d}s \,.$$

We define $\boldsymbol{e} \in [\mathbb{F}_{\mathcal{T}}]^{N+1}$ and $\boldsymbol{r} \in [\mathbb{R}^{\mathcal{T}}]^{N+1}$ by

$$e^k_{K,\sigma} = u^k_{K,\sigma} - \Big((\mathcal{L}_\Sigma \circ \mathcal{I}) \Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2} \Big) \Big)_\sigma \nabla_\sigma (\overline{\Pi}_{\mathcal{T}} \phi)^k \,,$$

and denoting by K' the cell in $\mathcal{T}'$ such that $K \subset K'$,

$$r^k_K \coloneqq \frac{1}{m_K \Delta t} \int_K \int_{t^{k-1}}^{t^k} \partial_t \rho \, \mathrm{d}t \mathrm{d}\boldsymbol{x} - \frac{1}{m_{K'} \Delta t} \int_{K'} \int_{t^{k-1}}^{t^k} \partial_t \rho \, \mathrm{d}t \mathrm{d}\boldsymbol{x} \,.$$

Then

$$-\operatorname{div}_{\mathcal{T}} \Big((\mathcal{L}_\Sigma \circ \mathcal{I}) \Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2} \Big) \odot \nabla_\Sigma (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \Big) = \boldsymbol{r}^k - \operatorname{div}_{\mathcal{T}} \boldsymbol{e}^k \,.$$

Multiplying both sides by $(\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k)$ we obtain

$$\| \nabla_\Sigma (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} = \langle \boldsymbol{r}^k - \operatorname{div}_{\mathcal{T}} \boldsymbol{e}^k, (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \rangle_{\mathcal{T}} \,.$$

Using the discrete Poincaré inequality of Lemma 2.7 and the lower bound on ρ, this implies

$$\| \nabla_\Sigma (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \|_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \leq C (\| \boldsymbol{r}^k \|_{\mathcal{T}} + \| \boldsymbol{e}^k \|_{\mathbb{F}_{\mathcal{T}}}) \,,$$

where $C > 0$ is a constant only depending on the lower bound ε and the domain. By the regularity of ϕ and ρ, and the estimate (2.24), we then obtain

$$\| \nabla_\Sigma (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \|_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \leq C (h_{\mathcal{T}} + \Delta t) \,, \tag{2.32}$$

where now C depends also on ρ and ϕ.

In order to get an estimate on the energy, we observe that $\boldsymbol{\phi}^k$ minimizes the functional

$$\boldsymbol{\psi} \in [\mathbb{R}^{\mathcal{T}}]^{N+1} \longmapsto \| \nabla_\Sigma \boldsymbol{\psi} \|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} - \langle \mathcal{I} \Big(\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} \Big), \boldsymbol{\psi} \rangle_{\mathcal{T}} \,,$$

which implies the inequality

$$\| \nabla_\Sigma \boldsymbol{\phi}^k \|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \leq \| \nabla_\Sigma (\overline{\Pi}_{\mathcal{T}} \phi)^k) \|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} + \langle \mathcal{I} \Big(\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} \Big), (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \rangle_{\mathcal{T}} \,.$$

Using again the discrete Poincaré inequality of Lemma 2.7 and the lower bound on ρ, as well as its regularity, we get

$$\| \nabla_\Sigma \boldsymbol{\phi}^k \|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \leq \| \nabla_\Sigma (\overline{\Pi}_{\mathcal{T}} \phi)^k) \|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} + C \| \nabla_\Sigma (\boldsymbol{\phi}^k - (\overline{\Pi}_{\mathcal{T}} \phi)^k) \|_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \,.$$

Hence, using (2.32), we obtain

$$\|\nabla_\Sigma \boldsymbol{\phi}\|_{\boldsymbol{\rho}}^2 \leq \|\nabla_\Sigma \overline{\Pi}_{\mathcal{T}} \phi)\|_{\boldsymbol{\rho}}^2 + C(h_{\mathcal{T}} + \Delta t)\,.$$

Finally, using Jensen's inequality and then Lemma 2.6, we find

$$\begin{aligned}
\|\nabla_\Sigma \overline{\Pi}_{\mathcal{T}} \phi\|_{\boldsymbol{\rho}}^2 &\leq \sum_{k=1}^{N+1} \int_{t^{k-1}}^{t^k} \|\nabla_\Sigma \Pi_{\mathcal{T}} \phi(t,\cdot))\|^2_{\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}} \,\mathrm{d}t \\
&\leq \sum_{k=1}^{N+1} \int_{t^{k-1}}^{t^k} \langle \mathcal{I}\Big(\frac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big), \Pi_{\mathcal{T}} |\nabla \phi(t,\cdot)|^2 \rangle_{\mathcal{T}} \mathrm{d}t + C(h_{\mathcal{T}} + \eta_h) \\
&= \sum_{k=1}^{N+1} \sum_{K \in \mathcal{T}} \int_{t^{k-1}}^{t^k} \int_K \frac{\rho(t^k,\cdot) + \rho(t^{k-1},\cdot)}{2} |\nabla \phi(t, \boldsymbol{x}_K)|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t + C(h_{\mathcal{T}} + \eta_h) \\
&\leq \int_0^1 \int_\Omega \rho |\nabla \phi|^2 \,\mathrm{d}\boldsymbol{x}\mathrm{d}t + C(h_{\mathcal{T}} + \Delta t + \eta_h)\,,
\end{aligned}$$

which concludes the proof. □

We are now ready to state the two main convergence results of this section, which provide quantitative estimates for the convergence rates of the discrete action and the discrete potential.

Theorem 2.9 (Convergence of the action)**.** *Suppose that $\phi : [0,1] \times \Omega \to \mathbb{R}$ is an optimal potential for the dual optimal transport problem (1.24)-(1.25) from ρ^{in} to ρ^f and that $\rho : [0,1] \times \Omega \to [0, +\infty)$ is the associated interpolation. Then, denoting $\boldsymbol{\rho}^{in} \coloneqq \Pi_{\mathcal{T}'} \rho^{in}$ and $\boldsymbol{\rho}^f \coloneqq \Pi_{\mathcal{T}'} \rho^f$, and taking η_h as in (2.26), the following holds:*

1. *if $\phi \in C^{1,1}([0,1] \times \Omega)$, there exists a constant $C > 0$ only dependent on ϕ and ζ such that*
$$W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f) \geq \mathcal{W}_2^2(\rho^{in}, \rho^f) - C(h_{\mathcal{T}} + \Delta t + \eta_h)\,;$$
2. *if $\phi \in L^\infty([0,1]; C^{1,1}(\Omega))$ and $\rho, \partial_t \rho \in L^\infty([0,1]; C^{0,1}(\Omega))$, with $\rho \geq \varepsilon > 0$, there exists a constant $C > 0$ depending only on ρ, ϕ, ε, ζ and Ω such that*
$$W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f) \leq \mathcal{W}_2^2(\rho^{in}, \rho^f) + C(h_{\mathcal{T}} + \Delta t + \eta_h)\,.$$

Proof. For the first point, we first observe that by Lemma 2.6 and the regularity of ϕ, $\overline{\Pi}_{\mathcal{T}} \phi$ verifies

$$\mathcal{I}^*\Big(\frac{(\overline{\Pi}_{\mathcal{T}}\phi)^{k+1} - (\overline{\Pi}_{\mathcal{T}}\phi)^k}{\Delta t}\Big) + \frac{\Delta t}{4} \mathcal{R}_{\mathcal{T}'}((\nabla_\Sigma (\overline{\Pi}_{\mathcal{T}}\phi)^k)^2 + (\nabla_\Sigma (\overline{\Pi}_{\mathcal{T}}\phi)^{k+1})^2) \leq C(h_{\mathcal{T}} + \Delta t + \eta_h)\,.$$

Define $\boldsymbol{\phi}$ by $\boldsymbol{\phi}^k \coloneqq (\overline{\Pi}_{\mathcal{T}}\phi)^k - C(t^k + t^{k-1})(h_{\mathcal{T}} + \Delta t + \eta_h)/2$, for $k \in \{1, \dots, N+1\}$. Then $\boldsymbol{\phi}$ is admissible for the dual problem (2.22), hence

$$\frac{W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f)}{2} \geq \langle \mathcal{I}^* \boldsymbol{\phi}^{N+1} - \frac{\Delta t}{4} \mathcal{R}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^{N+1})^2, \boldsymbol{\rho}^f \rangle_{\mathcal{T}'} - \langle \mathcal{I}^* \boldsymbol{\phi}^1 + \frac{\Delta t}{4} \mathcal{R}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^1)^2, \boldsymbol{\rho}^{in} \rangle_{\mathcal{T}'}\,.$$

Replacing back the definition of $\boldsymbol{\phi}$ and using the fact that $|\nabla_\sigma \boldsymbol{\phi}^1|$ and $|\nabla_\sigma \boldsymbol{\phi}^{N+1}|$ are uniformly bounded by a constant depending only on ϕ, we get

$$\frac{W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f)}{2} \geq \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega \phi(0,\cdot)\rho^{in} - C(h_\mathcal{T} + \Delta t + \eta_h)\,.$$

For the second point it suffices to observe that the couple (ρ, ϕ) satisfies (2.29). Then, defining $\boldsymbol{\rho}$ and $\boldsymbol{\phi}$ as in the statement of Proposition 2.8, we can construct an admissible competitor $(\boldsymbol{\rho}, \boldsymbol{F})$ for the discrete optimal transport problem by defining the momentum $\boldsymbol{F} \in [\mathbb{F}_\mathcal{T}]^{N+1}$ as in equation (2.16). Since, by definition,

$$W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f) \leq 2\,\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) = \|\nabla_\Sigma \boldsymbol{\phi}\|^2_{\boldsymbol{\rho}}\,,$$

we obtain the desired estimate using (2.30). □

The issue of convergence of the discrete solution $(\boldsymbol{\rho}, \boldsymbol{F})$ towards its continuous counterpart has been treated in detail in [79] for a general class of discretizations. These include the finite volume schemes considered here, in the case where the two domain decompositions coincide so that $\mathcal{I}$ is the identity operator (see Remark 2.4). For this case, one has that the discrete density $\boldsymbol{\rho}$ can be lifted to a measure on $[0,1] \times \Omega$ converging weakly to the exact optimal transport interpolation with mesh refinement.

It is not difficult to show that the second point of Theorem 2.9 implies a similar convergence result, for smooth positive solutions, also when the two discretizations of the domain do not coincide (e.g., this is a direct consequence of Theorem 2.18 in [79]). Besides this weak convergence result, Theorem 2.9 also implies the following quantitative estimate for the convergence of the potential, although in a norm dependent on the discrete solution itself.

Theorem 2.10 (Convergence of the potential)**.** *Suppose that $\phi : [0,1] \times \Omega \to \mathbb{R}$ is an optimal potential for the dual optimal transport problem (1.24)-(1.25) from ρ^{in} to ρ^f and that $\rho : [0,1] \times \Omega \to [0, +\infty)$ is the associated interpolation. Let $(\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\phi}})$ be the discrete solution associated with the boundary conditions $\boldsymbol{\rho}^{in} := \Pi_{\mathcal{T}'} \rho^{in}$ and $\boldsymbol{\rho}^f := \Pi_{\mathcal{T}'} \rho^f$. If $\phi \in C^{1,1}([0,1] \times \Omega)$ and $\rho, \partial_t \rho \in L^\infty([0,1]; C^{0,1}(\Omega))$, with $\rho \geq \varepsilon > 0$, there exists a constant $C > 0$ depending only on ρ, ϕ, ε, ζ and Ω, such that*

$$\|\nabla_\Sigma \tilde{\boldsymbol{\phi}} - \nabla_\Sigma \overline{\Pi}_\mathcal{T} \phi\|^2_{\tilde{\boldsymbol{\rho}}} \leq C(h_\mathcal{T} + \Delta t + \eta_h)\,,$$

for η_h defined as in (2.26).

Proof. Consider the quantity

$$\mathcal{E}_{N,\mathcal{T}}(\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\phi}} | \phi) := \frac{1}{2} \|\nabla_\Sigma \tilde{\boldsymbol{\phi}} - \nabla_\Sigma \overline{\Pi}_\mathcal{T} \phi\|^2_{\tilde{\boldsymbol{\rho}}}\,. \tag{2.33}$$

Expanding the square in (2.33) we obtain

$$\mathcal{E}_{N,\mathcal{T}}(\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\phi}} | \phi) = \mathcal{B}_{N,\mathcal{T}}(\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{F}}) + \frac{1}{2} \|\nabla_\Sigma \overline{\Pi}_\mathcal{T} \phi\|^2_{\tilde{\boldsymbol{\rho}}} - \langle \nabla_\Sigma \tilde{\boldsymbol{\phi}}, \nabla_\Sigma \overline{\Pi}_\mathcal{T} \phi \rangle_{\tilde{\boldsymbol{\rho}}}\,, \tag{2.34}$$

where $\tilde{\boldsymbol{F}}$ is given by equation (2.16). The second term in (2.34) can be written as

$$\begin{aligned}\frac{1}{2}\|\nabla_\Sigma \overline{\Pi}_{\mathcal{T}}\phi\|^2_{\tilde{\boldsymbol{\rho}}} &= \frac{1}{2}\sum_{k=1}^{N+1}\int_{t^{k-1}}^{t^k} \langle \mathcal{L}^*_\Sigma |\nabla_\Sigma \Pi_{\mathcal{T}}\phi(s,\cdot)|^2 - \Pi_{\mathcal{T}}|\nabla\phi(s,\cdot)|^2, \mathcal{I}\Big(\frac{\tilde{\boldsymbol{\rho}}^k + \tilde{\boldsymbol{\rho}}^{k-1}}{2}\Big)\rangle_{\mathcal{T}}\,\mathrm{d}s \\ &\quad - \sum_{k=1}^{N+1}\int_{t^{k-1}}^{t^k} \langle \Pi_{\mathcal{T}}\partial_t\phi(s,\cdot), \mathcal{I}\Big(\frac{\tilde{\boldsymbol{\rho}}^k + \tilde{\boldsymbol{\rho}}^{k-1}}{2}\Big)\rangle_{\mathcal{T}}\,\mathrm{d}s \\ &= I_1 - \sum_{k=1}^{N+1}\langle \Pi_{\mathcal{T}}\phi(t^k,\cdot) - \Pi_{\mathcal{T}}\phi(t^{k-1},\cdot), \mathcal{I}\Big(\frac{\tilde{\boldsymbol{\rho}}^k + \tilde{\boldsymbol{\rho}}^{k-1}}{2}\Big)\rangle_{\mathcal{T}}\,.\end{aligned} \tag{2.35}$$

The third term in (2.34) instead can be written as

$$\begin{aligned}-\langle \nabla_\Sigma\tilde{\boldsymbol{\phi}}, \nabla_\Sigma\overline{\Pi}_{\mathcal{T}}\phi\rangle_{\tilde{\boldsymbol{\rho}}} &= \sum_{k=1}^{N+1}\int_{t^{k-1}}^{t^k}\langle \mathrm{div}_{\mathcal{T}}\big((\mathcal{L}_\Sigma\circ\mathcal{I})\Big(\frac{\tilde{\boldsymbol{\rho}}^k + \tilde{\boldsymbol{\rho}}^{k-1}}{2}\Big)\odot\nabla_\Sigma\boldsymbol{\phi}^k\big), \Pi_{\mathcal{T}}\phi(s,\cdot)\rangle_{\mathcal{T}}\mathrm{d}s \\ &= -\sum_{k=1}^{N+1}\int_{t^{k-1}}^{t^k}\langle \mathcal{I}\Big(\frac{\tilde{\boldsymbol{\rho}}^k - \tilde{\boldsymbol{\rho}}^{k-1}}{\Delta t}\Big), \Pi_{\mathcal{T}}\phi(s,\cdot)\rangle_{\mathcal{T}}\,\mathrm{d}s \\ &= I_2 - \langle \mathcal{I}\tilde{\boldsymbol{\rho}}^{N+1}, \Pi_{\mathcal{T}}\phi(1,\cdot)\rangle_{\mathcal{T}} + \langle \mathcal{I}\tilde{\boldsymbol{\rho}}^0, \Pi_{\mathcal{T}}\phi(0,\cdot)\rangle_{\mathcal{T}} \\ &\quad + \sum_{k=1}^{N}\langle \Pi_{\mathcal{T}}\phi(t^k,\cdot) - \Pi_{\mathcal{T}}\phi(t^{k-1},\cdot), \mathcal{I}\Big(\frac{\tilde{\boldsymbol{\rho}}^k + \tilde{\boldsymbol{\rho}}^{k-1}}{2}\Big)\rangle_{\mathcal{T}}\,,\end{aligned} \tag{2.36}$$

where

$$I_2 \coloneqq \sum_{k=1}^{N}\int_{t^{k-1}}^{t^k}\langle \Pi_{\mathcal{T}}\partial_t\phi(s,\cdot) - \Pi_{\mathcal{T}}\Big(\frac{\phi(t^{k+1},\cdot) - \phi(t^k,\cdot)}{\Delta t}\Big), \mathcal{I}\tilde{\boldsymbol{\rho}}^{k-1,k}(s)\rangle_{\mathcal{T}}\,\mathrm{d}s$$

and $\tilde{\boldsymbol{\rho}}^{k-1,k}(s)$ is the linear interpolation between $\tilde{\boldsymbol{\rho}}^{k-1}$ and $\tilde{\boldsymbol{\rho}}^k$, i.e. $\tilde{\boldsymbol{\rho}}^{k-1,k}(s) \coloneqq \tilde{\boldsymbol{\rho}}^{k-1}(t^k - s)/\Delta t + \tilde{\boldsymbol{\rho}}^k(s - t_{k-1})/\Delta t$.

Adding and subtracting $\mathcal{W}_2^2(\rho^{in},\rho^f)/2 = \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega\phi(0,\cdot)\rho^{in}$ from the right-hand side of (2.34), substituting (2.35) and (2.36), and rearranging terms we obtain

$$\mathcal{E}_{N,\mathcal{T}}(\tilde{\boldsymbol{\rho}},\tilde{\boldsymbol{\phi}}|\phi) = \frac{W^2_{N,\mathcal{T}}(\boldsymbol{\rho}^{in},\boldsymbol{\rho}^f)}{2} - \frac{\mathcal{W}_2^2(\rho^{in},\rho^f)}{2} + I_1 + I_2 + I_3\,, \tag{2.37}$$

where

$$I_3 \coloneqq \int_\Omega \phi(1,\cdot)\rho^f - \int_\Omega\phi(0,\cdot)\rho^{in} - \langle \mathcal{I}^*\Pi_{\mathcal{T}}\phi(1,\cdot), \boldsymbol{\rho}^f\rangle + \langle \mathcal{I}^*\Pi_{\mathcal{T}}\phi(0,\cdot), \boldsymbol{\rho}^{in}\rangle\,,$$

since $\boldsymbol{\rho}^0 = \Pi_{\mathcal{T}'}\rho^{in}$ and $\boldsymbol{\rho}^{N+1} = \Pi_{\mathcal{T}'}\rho^f$. Finally, we estimate I_1 and I_3 using Lemma 2.6, I_2 using the regularity of ϕ, and the remaining term using the second point in Theorem 2.9. □

Remark 2.11. *It is easy to construct solutions to the optimality conditions (1.26), and therefore to problem (2.1)-(2.2), satisfying the assumptions of Theorem 2.9 or 2.10. In fact, given any smooth compactly-supported initial potential $\phi_0 : \Omega \to \mathbb{R}$, there exists $\delta > 0$ such that the map $x \mapsto T_t(x) \coloneqq x + t\nabla\phi_0(x)$ is a diffeomorphism for $t \in [0,\delta]$, and $\phi(t,\cdot) = \phi_0 \circ T_t^{-1}$ is a smooth solution to the Hamilton-Jacobi equation. Moreover, given a strictly postive and*

smooth initial density ρ_0, the density $\rho(t,\cdot) = (\rho_0/\det(\nabla T_t)) \circ T_t^{-1}$ solves the continuity equation with velocity $\nabla\phi(t,\cdot)$, and it is also smooth and strictly positive for $t \in [0,\delta]$. Then, the curve $t \mapsto (\delta\phi(\delta t,\cdot)), \rho(\delta t,\cdot))$ solves the optimality conditions (1.26) on the time interval $[0,1]$. On the other hand, even in the case where ρ_0 and ρ_1 are smooth and strictly positive the interpolation may not even be stricly positive as shown in [117].

Remark 2.12. *The quantity $\mathcal{E}_{N,\mathcal{T}}(\tilde{\boldsymbol{\rho}}, \tilde{\boldsymbol{\phi}}|\phi)$ defined in equation (2.33) is the discrete H^1 seminorm of the error weighted by the discrete solution $\tilde{\boldsymbol{\rho}}$. Note that this can also be seen as a discretization of the modulated energy (or relative entropy) of the kinetic energy, interpreted as a convex function of (ρ, F). In Section 2.7 we will use a similar quantity in order to evaluate numerically the convergence rate of the scheme.*

2.6 Primal-dual barrier method

We introduce now the primal-dual barrier method, the discrete optimization technique we use to deal with the uniqueness, smoothness and positivity issues and effectively solve problem (2.11). The method consists in perturbing the discrete problem with a barrier function which forces the density to be positive. Here we show that the solutions of such perturbed problem converge to the ones of the original problem, when the perturbation vanishes, therefore justifying the use of a continuation method. Finally, we will detail the implementation of the algorithm commenting on the choice of the parameters involved.

The most classical barrier function used when dealing with positivity constraints is the logarithmic barrier, $-\log\rho$. In order to write the perturbed problem, we first define precisely the barrier,

$$J(x) = \begin{cases} -\log(x) & \text{if } x > 0, \\ +\infty & \text{if } x \le 0, \end{cases}$$

so that it is convex and lower semi-continuous. We define the barrier function as

$$\mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho}) = \sum_{k=1}^{N} \Delta t \sum_{K' \in \mathcal{T}'} J(\rho^k_{K'}) m_{K'}\,,$$

and the perturbed version of problem (2.11) is therefore:

$$\inf_{(\boldsymbol{\rho},\boldsymbol{F}) \in \mathcal{C}_{N,\mathcal{T}}} \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho},\boldsymbol{F}) + \mu \mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho})\,. \tag{2.38}$$

Thanks to the strict convexity of the function $\mathcal{J}_{N,\mathcal{T}}$ on $[\mathbb{R}_+^{\mathcal{T}'} \setminus \{0\}]^N$, the solution $(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)$ is now unique. Proceeding as in Section 2.4, $\boldsymbol{\rho}^\mu$ can be characterized as solution to the system of optimality conditions

$$\begin{cases} \mathcal{I}\Big(\dfrac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t}\Big) + \operatorname{div}_{\mathcal{T}}\Big((\mathcal{R}_\Sigma \circ \mathcal{I})\Big(\dfrac{\boldsymbol{\rho}^k + \boldsymbol{\rho}^{k-1}}{2}\Big) \odot \nabla_\Sigma \boldsymbol{\phi}^k\Big) = 0, \\ \mathcal{I}^*\Big(\dfrac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t}\Big) + \dfrac{1}{4}\mathcal{R}^k_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^k)^2 + \dfrac{1}{4}\mathcal{R}^{k+1}_{\mathcal{T}'}(\nabla_\Sigma \boldsymbol{\phi}^{k+1})^2 = -\boldsymbol{s}^k, \\ \boldsymbol{\rho}^k \odot \boldsymbol{s}^k = \mu\,\mathbf{1}, \end{cases} \tag{2.39}$$

where $k \in \{1,..,N+1\}$ for the continuity equation and $k \in \{1,..,N\}$ for the other conditions. The variable $\boldsymbol{s} \in [\mathbb{R}^{\mathcal{T}'}]^N, (s^k)_{K'} = \mu\,(\rho^k_{K'})^{-1}$, has been introduced in order to decouple the optimization in $\boldsymbol{\rho}$ and $\boldsymbol{s}$, and it highlights the connection with system (2.19). In particular, system (2.39) can be seen as a perturbation of (2.19), where $\rho^k_{K'}$ and $s^k_{K'} = -\lambda^k_{K'}$ are automatically forced to be positive and the orthogonality is relaxed. In this way, the solution $(\boldsymbol{\phi}^\mu, \boldsymbol{\rho}^\mu, \boldsymbol{s}^\mu)$ is now unique, up to an additive constant for the potential, and the problem is smooth.

As it is classical in interior point methods (see, e.g., [26]), if we regard $(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)$ as an approximate solution to problem (2.11), we can derive an explicit estimate on how far it is from optimality. Given a solution $(\boldsymbol{\rho}, \boldsymbol{F})$ of the original problem, and defining $\tilde{\boldsymbol{\lambda}} \in [\mathbb{R}^{\mathcal{T}'}_-]^N$ by $(\tilde{\lambda}^k)_{K'} = -\frac{\mu}{(\rho^\mu)^k_{K'}}$, we have

$$\begin{aligned}
\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) &= \sup_{\boldsymbol{\phi}} \inf_{\boldsymbol{\rho}\geq 0, \boldsymbol{F}} \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{F}) \\
&\geq \inf_{\boldsymbol{\rho}\geq 0, \boldsymbol{F}} \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}^\mu, \boldsymbol{\rho}, \boldsymbol{F}) + \sum_{k=1}^N \Delta t \langle \tilde{\boldsymbol{\lambda}}^k, \boldsymbol{\rho}^k \rangle_{\mathcal{T}'} \\
&= \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}^\mu, \boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu) + \sum_{k=1}^N \Delta t \langle \tilde{\boldsymbol{\lambda}}^k, (\boldsymbol{\rho}^\mu)^k \rangle_{\mathcal{T}'} = \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu) - \mu \frac{N}{N+1} |\Omega|\,,
\end{aligned} \tag{2.40}$$

where we used the fact that $(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)$ is optimal for $\mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}^\mu, \boldsymbol{\rho}, \boldsymbol{F}) + \sum_{k=1}^N \Delta t \langle \tilde{\boldsymbol{\lambda}}^k, \boldsymbol{\rho}^k \rangle_{\mathcal{T}'}$, which can be easily verified by comparing the associated optimality conditions with (2.39). We have therefore

$$0 \leq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu) - \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) \leq \mu \frac{N}{N+1} |\Omega|\,. \tag{2.41}$$

As a consequence of (2.41), the smaller the parameter μ, the closer the perturbed solution is to the original one.

Theorem 2.13. *The solution $(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)$ of problem (2.38) converges up to extraction of a subsequence to $(\boldsymbol{\rho}, \boldsymbol{F})$ solution of (2.11) for $\mu \to 0$.*

Proof. Consider a sequence $(\mu_n)_n \subset \mathbb{R}_+$ converging to zero and the corresponding sequence $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ of solutions to problem (2.38). We first derive a bound on $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$, independent of μ. The bound on $\boldsymbol{\rho}^{\mu_n}$ derives easily from the conservation of mass. To obtain a bound for the momentum $\boldsymbol{F}^{\mu_n}$, for any $\boldsymbol{b} \in [\mathbb{F}_{\mathcal{T}}]^{N+1}$ with $|b^k_\sigma| \leq 1$ for all $\sigma \in \Sigma, k \in \{1,..,N+1\}$, we observe that there exists a constant $C > 0$ independent of μ such that

$$\sum_{k=1}^{N+1} \Delta t \langle (\boldsymbol{F}^{\mu_n})^k, \boldsymbol{b}^k \rangle_{\mathbb{F}_{\mathcal{T}}} \leq \sqrt{2\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})}\, ||\boldsymbol{b}||_{\boldsymbol{\rho}^\mu} \leq C\,, \tag{2.42}$$

where the weighted norm $||\cdot||_{\boldsymbol{\rho}^\mu}$ is defined via (2.23). The first inequality is the same as (2.14), applied to $\boldsymbol{\rho}^{\mu_n}$ and $\boldsymbol{F}^{\mu_n}$. The second one is obtained using the inequality (2.41). Taking the sup with respect to $\boldsymbol{b}$ in (2.42), we obtain the bound on $\boldsymbol{F}^{\mu_n}$.

The sequence $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ is bounded hence we can extract a converging subsequence (still labeled with μ_n for simplicity) $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n}) \to (\boldsymbol{\rho}^*, \boldsymbol{F}^*)$. Consider $(\boldsymbol{\rho}, \boldsymbol{F})$ minimizer of the unperturbed problem (2.11). From inequality (2.41), taking the limit for $n \to +\infty$, we obtain $\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^*, \boldsymbol{F}^*) = \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F})$, hence $(\boldsymbol{\rho}^*, \boldsymbol{F}^*)$ is a minimizer for problem (2.11). □

Remark 2.14. *If the solution $(\boldsymbol{\rho}, \boldsymbol{F})$ of the discrete problem (2.11) is unique, then the entire sequence $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ converges to it. In case it is not unique, due to*

$$0 \leq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n}) - \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) \leq \mu_n(\mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho}) - \mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho}^{\mu_n}))\,,$$

we know that $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ converges up to subsequence to a solution $(\boldsymbol{\rho}^*, \boldsymbol{F}^*)$ with minimal $\mathcal{J}_{N,\mathcal{T}}$. In case the solution $\boldsymbol{\rho}^*$ is strictly positive everywhere, the whole sequence $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ converges again.

The strict positivity derives automatically from the definition of the barrier function, which attains the value $+\infty$ in zero. As a consequence, for every value of $\mu > 0$ the objective function $\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) + \mu\mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho})$ is smooth in a neighborhood of the solution $(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)$, ensuring a good behavior of the Newton scheme for the solution of the system of equations (2.39). It is possible to derive a quantitative bound for the positivity of $\boldsymbol{\rho}^\mu$ as follows.

Proposition 2.15. *There exists a constant $C > 0$ independent of μ such that the density $\boldsymbol{\rho}^\mu$ solution to problem (2.38) satisfies the following bound:*

$$(\rho^\mu)^k_{K'} \geq C\mu, \quad \forall K' \in \mathcal{T}', \forall k\,. \tag{2.43}$$

Proof. Consider the solution $(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)$ to (2.38). We define the constant density $\boldsymbol{c} \in [\mathbb{R}^{\mathcal{T}'}_+]^N$, $c^k_{K'} = (\sum_{K\in\mathcal{T}} m_{K'})^{-1} = (|\Omega|)^{-1}$. It can be easily checked that $\boldsymbol{c}$ is the unique solution to

$$\min_{\boldsymbol{\rho}\in[\mathbb{R}^{\mathcal{T}'}]^N} \mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho}) \quad \text{such that} \quad \sum_{K'\in\mathcal{T}'} \rho^k_{K'} m_{K'} = 1,\ \forall k\,. \tag{2.44}$$

From now on, with a slight abuse of notation, we consider $\boldsymbol{c}$ to be complemented with the boundary conditions $\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f$. Thanks to the surjectivity of the divergence operator (to the space of discrete functions in $[\mathbb{R}^{\mathcal{T}}]^{N+1}$ with zero mean), we can find the unique momentum $\boldsymbol{F}^c$ with minimal $||\cdot||_{\boldsymbol{c}}$ norm (defined via equation (2.23)), such that $(\boldsymbol{c}, \boldsymbol{F}^c) \in \mathcal{C}_{N,\mathcal{T}}$:

$$\boldsymbol{F}^c = \operatorname*{argmin}_{\boldsymbol{F}\in\mathbb{F}_{\mathcal{T}}} \frac{1}{2}||\boldsymbol{F}||^2_{\boldsymbol{c}}, \quad \text{such that } (\boldsymbol{c}, \boldsymbol{F}^c) \in \mathcal{C}_{N,\mathcal{T}}\,.$$

Taking the admissible competitor $(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{F}}) = (\epsilon\boldsymbol{c} + (1-\epsilon)\boldsymbol{\rho}^\mu, \epsilon\boldsymbol{F}^c + (1-\epsilon)\boldsymbol{F}^\mu)$, $\epsilon \in [0,1]$, for problem (2.38), it holds

$$\mu\left(\mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu) - \mathcal{J}_{N,\mathcal{T}}(\hat{\boldsymbol{\rho}})\right) \leq \mathcal{B}_{N,\mathcal{T}}(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{F}}) - \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)\,. \tag{2.45}$$

The right-hand side of (2.45) is bounded: indeed, by convexity of $\mathcal{B}_{N,\mathcal{T}}$, it holds

$$\begin{aligned}\mathcal{B}_{N,\mathcal{T}}(\hat{\boldsymbol{\rho}}, \hat{\boldsymbol{F}}) - \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu) &\leq \epsilon\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^c, \boldsymbol{F}^c) + (1-\epsilon)\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu) - \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^\mu, \boldsymbol{F}^\mu)\\ &\leq C\epsilon\,.\end{aligned} \tag{2.46}$$

The left-hand side of (2.45) can be bounded from below thanks to the convexity of $\mathcal{J}_{N,\mathcal{T}}$, by the following quantity

$$\mu\sum_{k=1}^N \sum_{K'\in\mathcal{T}'} J'(\hat{\rho}^k_{K'})((\rho^\mu)^k_{K'} - \hat{\rho}^k_{K'})m_{K'}\Delta t = \mu\epsilon\sum_{k=1}^N \sum_{K'\in\mathcal{T}'} J'(\hat{\rho}^k_{K'})((\rho^\mu)^k_{K'} - c^k_{K'}))m_{K'}\Delta t\,.$$

Hence, we obtain

$$\mu\epsilon \sum_{k=1}^{N} \sum_{K' \in \mathcal{T}'} J'(\hat{\rho}^k_{K'})((\rho^\mu)^k_{K'} - c^k_{K'}))m_{K'}\Delta t \leq C\epsilon\,. \tag{2.47}$$

Simplifying ϵ in (2.47) and taking the limit for $\epsilon \to 0$, we obtain

$$\sum_{k=1}^{N} \sum_{K' \in \mathcal{T}'} \left(\frac{c^k_{K'}}{(\rho^\mu)^k_{K'}} - 1 \right) m_{K'}\Delta t \leq \frac{C}{\mu}\,,$$

and therefore

$$\min_{K'}(m_{K'})\Delta t \sum_{k=1}^{N} \sum_{K' \in \mathcal{T}'} \frac{c^k_{K'}}{(\rho^\mu)^k_{K'}} \leq \frac{C}{\mu} + |\Omega|T\,,$$

which implies the result. □

By Theorem 2.13 the solution of problem (2.38) provides an approximation to a solution $(\boldsymbol{\phi}, \boldsymbol{\rho})$ to problem (2.19), although the smaller the parameter the more difficult it is to solve the problem using a Newton method. The idea is then to use a continuation method, that is construct a sequence of solutions to problem (2.39) for a sequence of coefficients μ decreasing to zero, using each time the solution at the previous step as starting point for the Newton scheme. The resulting algorithm in shown in Algorithm 1. We denote by θ the rate of decay for μ; by ε_0 and ε_μ the tolerances for the solution to (2.19) and (2.39), respectively; and by δ_0 and δ_μ the error in the convergence towards solutions of the original and perturbed problem. The parameter δ_μ can be taken to be a norm of the residual of the system of equations (2.39) or of the Newton step $\boldsymbol{d}$. Concerning δ_0, it is either possible to choose a norm of the residual of the system of equations (2.19) or $\delta_0 = \mu\frac{N}{N+1}|\Omega|$, by virtue of (2.41), whether the proximity to the minimizer or to the minimum is preferred.

Algorithm 1:

Given the starting point $(\boldsymbol{\phi}_0, \boldsymbol{\rho}_0, \boldsymbol{s}_0)$ and the parameters $\mu_0 > 0, \theta \in (0,1), \varepsilon_0 > 0$;
while $\delta_0 > \varepsilon_0$ **do**
  $\mu = \theta\mu$;
  while $\delta_\mu > \varepsilon_\mu$ **do**
    compute Newton direction $\boldsymbol{d}$ for (2.39);
    compute $\alpha \in (0,1]$ such that $\boldsymbol{\rho} + \alpha\boldsymbol{d}_{\boldsymbol{\rho}} > 0$ and $\boldsymbol{s} + \alpha\boldsymbol{d}_{\boldsymbol{s}} > 0$;
    update: $(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{s}) = (\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{s}) + \alpha(\boldsymbol{d}_{\boldsymbol{\phi}}, \boldsymbol{d}_{\boldsymbol{\rho}}, \boldsymbol{d}_{\boldsymbol{s}})$;
    if $n > n_{max}$ **or** $\alpha < \alpha_{min}$ **then**
      increase μ and repeat from previous iteration;
    end
  end
end

Since any intermediate solution for $\mu \neq 0$ is not of interest, a very common approach in interior point methods is to set a relatively big tolerance ε_μ, or even to do just one Newton step per value of μ. Another possibility could be to consider a tolerance ε_μ decreasing to ε_0 as μ tends to zero. Nonetheless, a small tolerance ε_μ avoids the solution $\boldsymbol{\rho}^\mu$ of the pertubed

problem to get accidentally too close to the boundary of the feasibility domain, i.e. too close to zero. This would imply a drop in the regularity of the specific problem at hand, with evident consequences on the effectiveness of the Newton scheme. For this reason, we consider $\varepsilon_\mu = \varepsilon_0$. Notice that, in view of Proposition 2.15, it is possible to evaluate (underestimate) the constant for the lower bound on the density. It could be possible then try to employ the previous more effective strategies, taking care of controlling the proximity of the density to zero by means of bound (2.43).

A linesearch technique is typically employed in order to ensure global convergence of the Newton scheme. However, in many cases it leads to a non negligible cost by forcing the Newton scheme to do several steps before reaching convergence. Instead of modifying the step size α, we adaptively control θ in order to force the convergence. The Newton scheme is repeated with an increased θ (i.e. with an increased μ) if it is not able to converge in n_{max} steps. The step size α is chosen just to ensure that $\boldsymbol{\rho}$ and $\boldsymbol{s}$ do not become negative. Again, the Newton scheme is repeated if α needs to be smaller than α_{min}. In particular, taking $\alpha_{min} = 1$ only allows full Newton steps.

For the success of the algorithm, it is fundamental to have access to a good initial condition. Since the idea is to start from a relatively big value of the parameter μ, for which the problem is sufficiently regular, the common strategy is to use the starting point which minimizes the perturbation function.

Proposition 2.16. *The solution $(\rho^\mu, \boldsymbol{F}^\mu)$ of problem (2.38) converges to $(\boldsymbol{c}, \boldsymbol{F}^c)$, defined as in Proposition 2.15, for $\mu \to +\infty$.*

Proof. The proof is similar to the proof of Theorem 2.13. Consider a sequence $(\mu_n)_n \subset \mathbb{R}_+$, $\mu_n \to +\infty$ and the corresponding sequence $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ of solutions to problem (2.38). We already know from the proof of Theorem 2.13 that this latter is bounded and we can extract a convergent subsequence (without relabelling it) converging to $(\boldsymbol{\rho}^*, \boldsymbol{F}^*)$. Consider the couple $(\boldsymbol{c}, \boldsymbol{F}^c)$, defined as in the proof of Proposition 2.15. By optimality of $(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n})$ and recalling that $\boldsymbol{c}$ minimizes (2.44) , it holds:

$$0 \leq \mu_n(\mathcal{J}_{N,\mathcal{T}}(\boldsymbol{\rho}^{\mu_n}) - \mathcal{J}_{N,\mathcal{T}}(\boldsymbol{c})) \leq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{c}, \boldsymbol{F}^c) - \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n}) \leq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{c}, \boldsymbol{F}^c)\,. \tag{2.48}$$

We deduce that for $\boldsymbol{n} \to +\infty$, $\boldsymbol{\rho}^{\mu_n}$ converges to $\boldsymbol{\rho}^* = \boldsymbol{c}$, unique minimizer of $\mathcal{J}_{N,\mathcal{T}}$, and the whole sequence converges. From (2.48) we also know that

$$\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}^{\mu_n}, \boldsymbol{F}^{\mu_n}) \leq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{c}, \boldsymbol{F}^c), \quad \forall n\,,$$

so that, for $n \to +\infty$, thanks to the lower semi-continuity of $\mathcal{B}_{N,\mathcal{T}}$,

$$\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{c}, \boldsymbol{F}^*) \leq \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{c}, \boldsymbol{F}^c)\,.$$

Since $\boldsymbol{F}^c$ is the unique minimizer of $\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{c}, \boldsymbol{F})$, $\boldsymbol{F}^* = \boldsymbol{F}^c$ and the whole sequence converges. □

As shown in Proposition 2.16, for big values of the parameter μ the solution of the perturbed problem tends to $(\boldsymbol{c}, \boldsymbol{F}^c)$. This motivates the choice of the intial condition $\boldsymbol{\rho}_0 = \boldsymbol{c}$, $\boldsymbol{s}_0 = \mu(\boldsymbol{\rho}_0)^{-1}$ and $\boldsymbol{\phi}_0 = \boldsymbol{\phi}^c$, where $\boldsymbol{\phi}^c$ is the solution to

$$\mathrm{div}_{\mathcal{T}}\Big((\mathcal{R}_\Sigma \circ \mathcal{I})\Big(\frac{\boldsymbol{c}^k + \boldsymbol{c}^{k-1}}{2}\Big) \odot \nabla_\Sigma \boldsymbol{\phi}^k\Big) = \mathcal{I}\Big(\frac{\boldsymbol{c}^{k-1} - \boldsymbol{c}^k}{\Delta t}\Big), \quad \forall k \in \{1, .., N+1\}\,,$$

for which $(\boldsymbol{F}^c)^k = (\mathcal{R}_\Sigma \circ \mathcal{I})(\frac{\boldsymbol{c}^k+\boldsymbol{c}^{k-1}}{2}) \odot \nabla(\boldsymbol{\phi}^c)^k$. However, this initial condition may be not sufficiently good for starting the algorithm for relatively low values of μ. The issue is due to the presence of the boundary conditions on the density, which makes the kinetic energy contribution not negligible and the problem is not trivial to be solved. Taking $\boldsymbol{\rho}_0 = \boldsymbol{c}$ e $\boldsymbol{F} = \boldsymbol{0}$ is an alternative that performs well, as long as the time step Δt is not too small, since the boundary contributions become more important the smaller is Δt. For complex simulations, in one case or the other, one could be forced to consider high value for the initial parameter μ_0.

The issue with the choice of the starting point is inevitable due to the specific structure of the problem. In order to overcome it, one could devise a perturbation/smoothing of problem (2.11) which enables to state that the perturbed solution tends to the constant solution $(\boldsymbol{c}, \boldsymbol{0})$ for increasing values of μ. This could be obtained in different ways, for example relaxing the continuity equation or the boundary conditions, which would of course exit from the framework of the barrier method. A simple and effective strategy is to set the boundary conditions

$$\boldsymbol{\rho}_\mu^{in} = \frac{\boldsymbol{\rho}^{in} + \mu M}{1 + \mu|\Omega|}, \quad \boldsymbol{\rho}_\mu^{f} = \frac{\boldsymbol{\rho}^{f} + \mu M}{1 + \mu|\Omega|}, \tag{2.49}$$

for the perturbed problem, where M is the total discrete mass.

There exist of course several optimization solvers that could tackle the solution of problem (2.11), most of which are usually based on interior point strategies, especially for large scales. Nevertheless, the specificity of the problem at hand, its non-linearity of course but more importantly its lack of smoothness, led us to develop our own solver, in order to better handle it. Moreover, the solution of the sequence of linear systems requires an ad-hoc strategy, as mentioned in Sections 2.7-2.8. Finally, we remark that in the particular case of the linear reconstruction, the corresponding dual problem in (2.22) can be cast in the form of a second-order cone program, which can be solved again using an interior point method in polynomial time. This does not apply to the case of the harmonic reconstruction (or more general reconstructions) for which the dual problem has a more complex structure.

2.7 Numerical results

In this section we assess the performance of the scheme we presented in Section 2.4 using several two-dimensional numerical tests. In particular, we demonstrate the numerical implications of enriching the space of discrete potentials, both from a qualitative and quantitative point of view. As already noted in Remark 2.4, considering the two subdivisions of the domain to be the same and taking $\mathcal{I}$ to be the identity operator, we recover the discretization presented in [79]. We will refer to this case as the non-enriched scheme, to distinguish it from the enriched one. Needless to say, the greater is the richness of the space of discrete potentials the higher is the computational complexity. We will also show that the monotone scheme we introduce in Appendix A does not prevent the stability issues.

For the construction of the enriched scheme we use the nested meshes described in Section 2.2.1. In particular, the coarse mesh is given by a regular triangulation of the domain with only acute angles. Here, we will use the first family of grids provided in [60], which discretize the domain $\Omega = [0, 1]^2$. One of these grids is shown in Figure 1.3c. Unless specified differently, we will always refer to these grids throughout this section.

The code is implemented in MATLAB and is available online[1]. In particular, we exploit the built-in MATLAB direct solver to solve the sequence of linear systems generated by Algorithm 1. For $\mu \to 0$ the Jacobian matrix becomes ill-conditioned and the computation time, along with the memory consumption, rapidly increases for this solver. Using an iterative method could be extremely beneficial in this sense. However, the design of effective preconditioners is a delicate issue and should take into account the structure of the problem at hand (see, e.g., the general survey [20]). Therefore, we do not explore the use of such techniques in this article. We calibrated Algorithm 1 with the following parameters: $\theta = 0.2$, $\alpha_{min} = 0.1$, $\epsilon_0 = 10^{-6}$ ($\epsilon_0 = 10^{-8}$ for the convergence tests), $\mu_0 = 1, \boldsymbol{\phi}_0 = \mathbf{0}, \boldsymbol{\rho}_0 = \boldsymbol{c}$ ($\boldsymbol{c}$ defined as in 2.15) and $\boldsymbol{s}_0 = \mu(\boldsymbol{\rho}_0)^{-1}$. In all the simulations performed in this section, but also more generally, the algorithm proved to be extremely robust under this configuration. The Newton scheme rarely reaches a breakdown and, in case this happens, the adaptive strategy on the parameter θ overcomes the issue. Notice only that for complex simulations the value μ_0 may be increased to ease the start of the Newton scheme. On the other hand, the algorithm greatly benefits from perturbing the boundary conditions as in (2.49). Finally, we stress that all the results in this section are presented in their piecewise-constant form on the grid, without any kind of interpolation.

2.7.1 Oscillations

In this section we show that the discrete density obtained by using the non-enriched scheme can be very oscillatory. We observed numerically that the oscillations are more severe in cases where there is high compression of mass, i.e. when the corresponding continuous velocity field is not divergence free, and also more persistent with refinement (this is also confirmed by the convergence tests shown below in Section 2.7.2). On the other hand, this type of instability can be prevented using either cartesian grids or the enriched scheme, which eliminates the oscillations almost entirely. The monotone scheme we present in Appendix A on the contrary does not solve the issue.

In order to illustrate this phenomenon, consider the interpolation between two gaussian densities:

$$\rho^{in}(x,y) = \exp^{-3|\boldsymbol{x}-\boldsymbol{x}_1|^2}, \quad \rho^{f}(x,y) = \exp^{-3|\boldsymbol{x}-\boldsymbol{x}_2|^2},$$

where $\boldsymbol{x}_1 = (\frac{3}{10}, \frac{3}{10}), \boldsymbol{x}_2 = (\frac{7}{10}, \frac{7}{10})$. We compute the approximate solution between the discrete densities $\boldsymbol{\rho}^{in} = \left(\rho^{in}(\boldsymbol{x}_K)\right)_{K' \in \mathcal{T}'}$ and $\boldsymbol{\rho}^{f} = \left(\rho^{f}(\boldsymbol{x}_K)\right)_{K' \in \mathcal{T}'}$. We use both the enriched and the non-enriched scheme with linear reconstruction, for $h'_{\mathcal{T}} = 0.0625$ and $\#\mathcal{T}' = 896$, and for a number $N + 1 = 8$ of time steps. On the same grid and with the same number of time steps, we compute the solution also with the monotone discretization (i.e. the scheme introduced in Appendix A). Finally, we also use a different grid for the non-enriched scheme with linear reconstruction, a cartesian grid with $h'_{\mathcal{T}} = 0.0707$, that is with edges length 0.05 (using the same number of time steps). In Figure 2.2 we represented the four different midpoints. The non-enriched scheme with linear reconstruction exhibits severe oscillations on the unstructured grid. With the monotone scheme, the oscillations seems milder but the result is not satisfactory. Preserving the monotonicity of the Hamilton-Jacobi equation does not prevent this issue. Using the enriched scheme instead we obtain a good approximation

[1] https://github.com/gptod/OT-FV

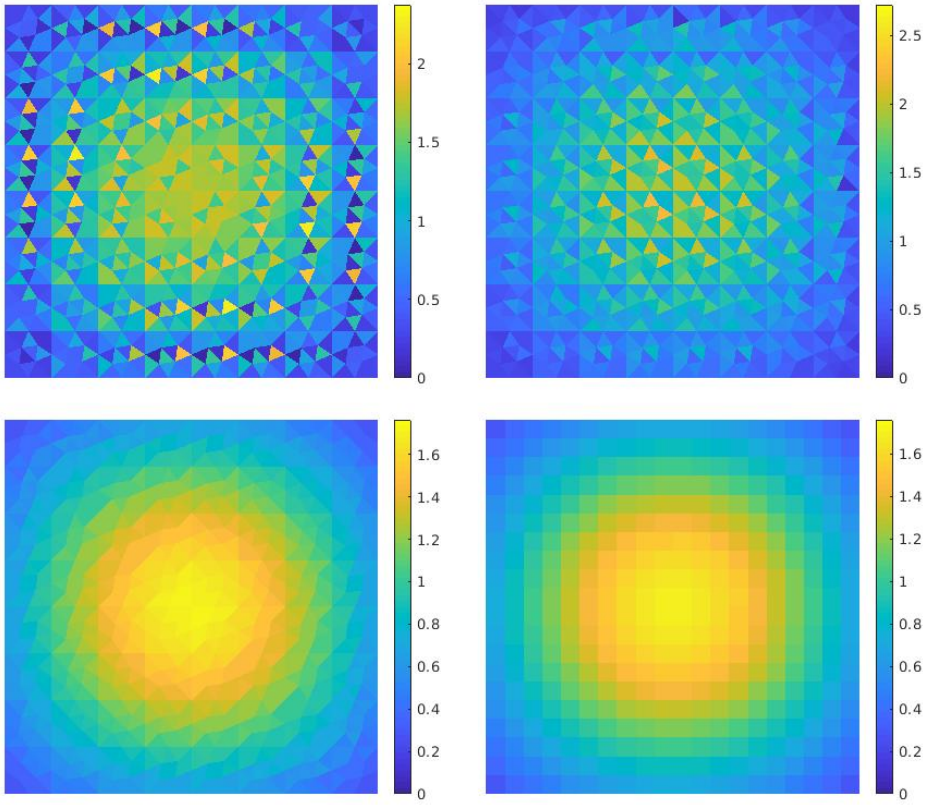


Figure 2.2: Midpoint between two gaussian functions. Non-enriched scheme with linear reconstruction (top-left), monotone discretization (top-right), enriched scheme with linear reconstruction (bottom-left), non-enriched scheme with linear reconstruction with cartesian grid (bottom-right).

of the expected result, as well as using a cartesian grid with the non-enriched one. The non-enriched scheme does not exhibit oscillations when using cartesian grids. Indeed, oscillations do not appear either in other works based on finite differences [41, 109, 85], which coincide with finite volumes on such simple grids.

The instability does not depend on the time refinement, which seems to exclude a possible condition on the time and space step sizes. We repeated the test with the non-enriched scheme with linear reconstruction using the same unstructured grid and with number of time steps $N+1 = 16, 32, 64$. The oscillations do not disappear as it is noticeable from the computed midpoints (Figure 2.3).

Let us consider another example, the interpolation between the two densities

$$\rho^{in}(x,y) = \cos\left(2\pi\left|\boldsymbol{x}-\boldsymbol{x}_0\right|\right) + \frac{3}{2}, \quad \rho^{f}(x,y) = M\Big(-\cos\left(2\pi\left|\boldsymbol{x}-\boldsymbol{x}_0\right|\right) + \frac{3}{2}\Big),$$

where $\boldsymbol{x}_0 = (\frac{1}{2}, \frac{1}{2})$ and M is chosen such that they have the same total mass. For $h'_{\mathcal{T}} = 0.0625$ and $\#\mathcal{T}' = 896$, and for a number $N+1 = 8$ of time steps, we compute the approximate Wasserstein interpolation between $\boldsymbol{\rho}^{in} = \big(\rho^{in}(\boldsymbol{x}_K)\big)_{K'\in\mathcal{T}'}$ and $\boldsymbol{\rho}^{f} = \big(\rho^{f}(\boldsymbol{x}_K)\big)_{K'\in\mathcal{T}'}$, by solv-

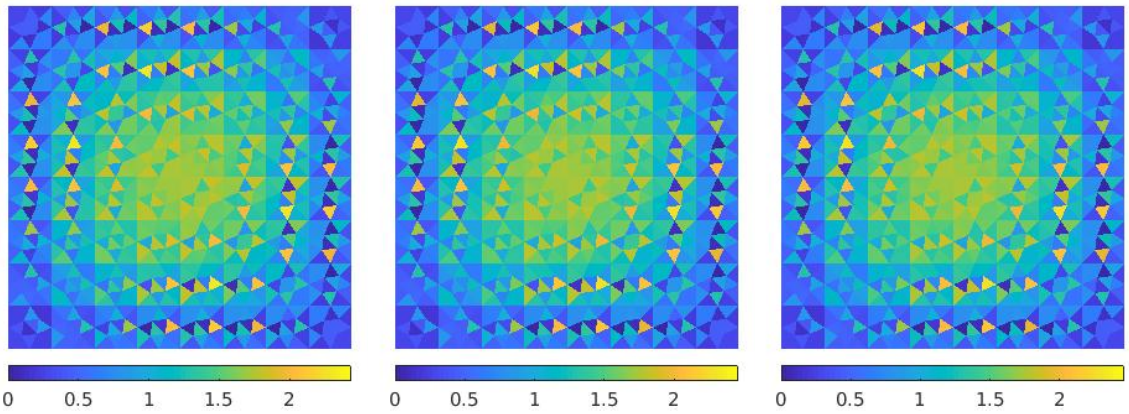


Figure 2.3: Midpoint between two gaussian functions computed with the non-enriched scheme with linear reconstruction. From left to right, number of time steps $N+1 = 16, 32, 64$.

ing problem (2.11) in four different ways: with the enriched and the non-enriched scheme, both with linear and harmonic reconstruction. The results are shown in Figure 2.4. Again, the severe oscillations which appear using the non-enriched scheme with linear reconstruction disappear using the enriched one. Oscillations are evident also using the harmonic reconstruction, although milder. The enriched scheme with harmonic reconstruction provides the smoothest solution.

2.7.2 Convergence test

We now quantify numerically the convergence rate for the potential, the Wasserstein distance and the density, by considering specific smooth solutions (ϕ, ρ) to (1.26) with compact support, and with smooth initial and final densities ρ^{in} and ρ^f. Note, however, that the convergence results of Section 2.5 are less general, since they require strictly positive densities, and only apply to the linear reconstruction.

We compute the solutions to problem (2.11), with $\boldsymbol{\rho}^{in} = (\rho^{in}(\boldsymbol{x}_{K'}))_{K'\in\mathcal{T}'}$ and $\boldsymbol{\rho}^f = (\rho^f(\boldsymbol{x}'_K))_{K'\in\mathcal{T}'}$, on a sequence of admissible meshes $(\mathcal{T}', \overline{\Sigma}', (\mathbf{x}_{K'})_{K'\in\mathcal{T}})$, and with an increasing number of time steps. We consider four type of errors: the error on the distance, the L^1 error on the density curve, the weighted L^2 error on the potential and on its gradient on the whole trajectory. We define a discrete potential $\boldsymbol{\phi} \in [\mathbb{R}^{\mathcal{T}}]^{N+1}$ by sampling the continuous solution, i.e. $\boldsymbol{\phi}_K^k = \phi(t^{k-1} + \frac{\Delta t}{2}, \boldsymbol{x}_K)$, for $k \in \{1, .., N+1\}$, and similarly for the density we introduce $\boldsymbol{\rho} \in [\mathbb{R}^{\mathcal{T}'}]^{N+1}$, with $\boldsymbol{\rho}_{K'} = \rho(t^{k-1} + \frac{\Delta t}{2}, \boldsymbol{x}_{K'})$, for $k \in \{1, .., N+1\}$. Given the discrete solution $(\tilde{\boldsymbol{\phi}}, \tilde{\boldsymbol{\rho}})$, the four errors are then computed as follows:

$$\epsilon_{\mathcal{W}_2} = |\mathcal{W}_2(\rho^{in}, \rho^f) - W_{N,\mathcal{T}}(\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f)|\,, \qquad \epsilon_\phi = \sum_{k=1}^{N+1} \Delta t \langle (\tilde{\phi}_K^k - \phi_K^k)^2, \mathcal{I}\Big(\frac{\tilde{\boldsymbol{\rho}}^k + \tilde{\boldsymbol{\rho}}^{k-1}}{2}\Big)\rangle_{\mathcal{T}}\,,$$

$$\epsilon_{\nabla\phi} = \|\nabla_\Sigma \tilde{\boldsymbol{\phi}} - \nabla_\Sigma \boldsymbol{\phi}\|_{\tilde{\boldsymbol{\rho}}}\,, \qquad \epsilon_\rho = \sum_{k=1}^{N+1} \Delta t \sum_{K'\in\mathcal{T}'} \Big|\boldsymbol{\rho}_{K'}^k - \frac{\tilde{\boldsymbol{\rho}}_{K'}^k + \tilde{\boldsymbol{\rho}}_{K'}^{k-1}}{2}\Big| m_{K'}\,,$$

where the weighted (semi-)norm $||\cdot||_{\tilde{\boldsymbol{\rho}}}$ is defined via (2.23). The rate of convergence is

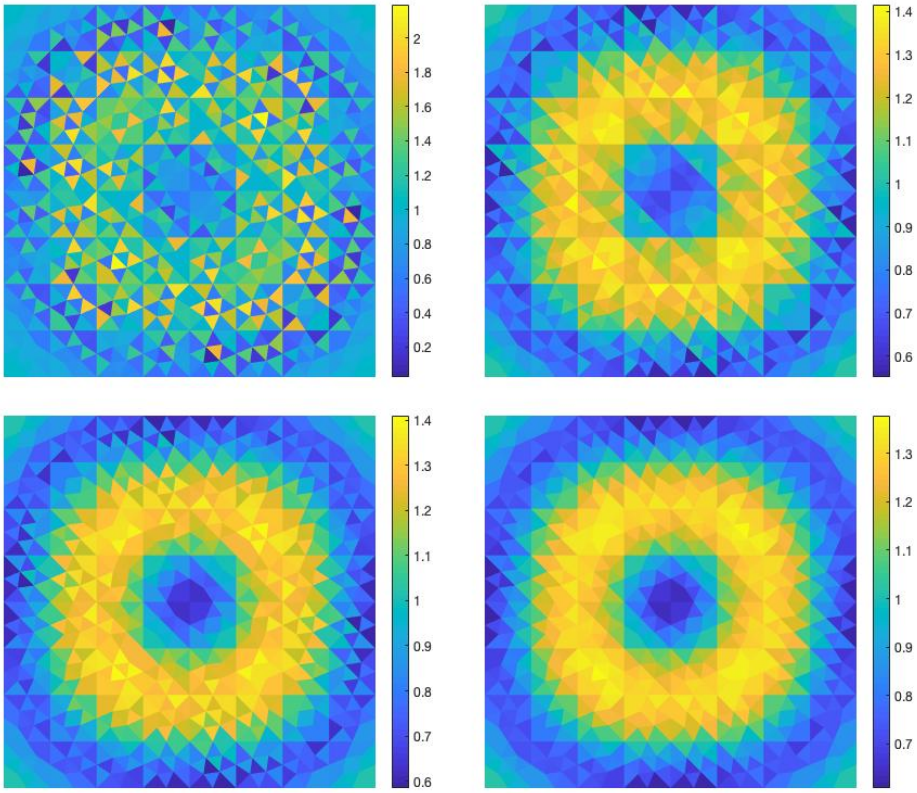


Figure 2.4: Midpoint between two sinusoidal functions. Non-enriched scheme in the top row, enriched scheme in the bottom one. Linear reconstruction on the left, harmonic reconstruction on the right.

evaluated as

$$\frac{\log(\epsilon_{m-1}) - \log(\epsilon_m)}{\log((h'_{\mathcal{T}})_{m-1}) - \log((h'_{\mathcal{T}})_m)}$$

for two consecutive values of errors ϵ and meshsizes $h'_{\mathcal{T}}$.

We first consider the simple case of a pure translation. We consider the optimal transport problem between the two following densities:

$$\rho^{in}(x,y) = \left(1 + \cos\left(\frac{10^2\pi}{3^2}\left|\boldsymbol{x} - \boldsymbol{x}_1\right|^2\right)\right)\mathbf{1}_{|\boldsymbol{x}-\boldsymbol{x}_1|\leq\frac{3}{10}}\,,$$

$$\rho^{f}(x,y) = \left(1 + \cos\left(\frac{10^2\pi}{3^2}\left|\boldsymbol{x} - \boldsymbol{x}_2\right|^2\right)\right)\mathbf{1}_{|\boldsymbol{x}-\boldsymbol{x}_2|\leq\frac{3}{10}}\,,$$

where $\boldsymbol{x}_1 = (\frac{3}{10}, \frac{3}{10}), \boldsymbol{x}_2 = (\frac{7}{10}, \frac{7}{10})$. The density interpolation and the potential are simply given by

$$\rho(t,x,y) = \left(1 + \cos\left(\frac{10^2\pi}{3^2}\left|\boldsymbol{x} - \boldsymbol{x}_t\right|^2\right)\right)\mathbf{1}_{|\boldsymbol{x}-\boldsymbol{x}_t|\leq\frac{3}{10}}\,,$$

$$\phi(t,x,y) = \frac{2}{5}x + \frac{2}{5}y - \frac{4}{25}t\,,$$

Table 2.1: Convergence test on the translation.

$h'_{\mathcal{T}}$	N	$\epsilon_{\mathcal{W}_2}$	rate	ϵ_ϕ	rate	$\epsilon_{\nabla\phi}$	rate	ϵ_ρ	rate
Non-enriched scheme with linear reconstruction									
0.250	1	3.109e-02	/	1.802e-02	/	2.153e-01	/	6.092e-01	/
0.125	3	3.375e-03	3.204	4.857e-03	1.892	9.574e-02	1.169	2.779e-01	1.132
0.062	7	1.190e-03	1.504	1.442e-03	1.752	3.947e-02	1.278	1.431e-01	0.958
0.031	15	2.351e-04	2.339	4.105e-04	1.813	1.550e-02	1.348	7.115e-02	1.008
0.016	31	2.874e-05	3.032	1.086e-04	1.919	5.708e-03	1.442	3.110e-02	1.194
Non-enriched scheme with harmonic reconstruction									
0.250	1	4.897e-02	/	2.382e-02	/	1.825e-01	/	5.870e-01	/
0.125	3	9.950e-03	2.299	5.635e-03	2.080	7.503e-02	1.282	2.535e-01	1.211
0.062	7	4.009e-03	1.311	1.751e-03	1.686	3.393e-02	1.145	1.172e-01	1.114
0.031	15	1.168e-03	1.780	5.055e-04	1.792	1.433e-02	1.243	4.907e-02	1.256
0.016	31	3.074e-04	1.925	1.409e-04	1.843	6.040e-03	1.247	2.057e-02	1.254
Enriched scheme with linear reconstruction									
0.250	1	3.880e-02	/	2.084e-02	/	2.231e-01	/	5.774e-01	/
0.125	3	3.714e-03	3.385	5.129e-03	2.023	9.375e-02	1.251	2.343e-01	1.301
0.062	7	1.457e-03	1.350	1.568e-03	1.710	4.303e-02	1.124	9.481e-02	1.305
0.031	15	3.551e-04	2.037	4.391e-04	1.836	1.935e-02	1.153	3.233e-02	1.552
0.016	31	6.712e-05	2.403	1.145e-04	1.939	8.719e-03	1.150	1.228e-02	1.397
Enriched scheme with harmonic reconstruction									
0.250	1	4.512e-02	/	2.240e-02	/	1.999e-01	/	5.740e-01	/
0.125	3	6.907e-03	2.708	5.187e-03	2.111	8.270e-02	1.273	2.370e-01	1.276
0.062	7	2.852e-03	1.276	1.597e-03	1.699	3.975e-02	1.057	1.036e-01	1.193
0.031	15	8.292e-04	1.782	4.521e-04	1.821	1.857e-02	1.098	4.014e-02	1.369
0.016	31	2.116e-04	1.970	1.221e-04	1.889	8.802e-03	1.077	1.668e-02	1.266

where $\boldsymbol{x}_t = (1-t)\boldsymbol{x}_1 + t\boldsymbol{x}_2 = (\frac{3}{10} + \frac{2}{5}t, \frac{3}{10} + \frac{2}{5}t)$, and the Wasserstein distance is $\mathcal{W}_2(\rho^{in}, \rho^f) = \frac{2\sqrt{2}}{5}$. Note in particular that the associated velocity field is constant in space. The errors defined above and the respective rates of convergence are shown in Table 2.1. In this case, all the considered errors converge with a rate of at least one for both the enriched and non-enriched scheme and both type of reconstructions.

We now consider a more challenging test, the optimal transport problem between the two densities

$$\rho^{in}(x,y) = \Big(1 + \cos\Big(2\pi\Big(x - \frac{1}{2}\Big)\Big)\Big),$$
$$\rho^{f}(x,y) = \frac{1}{c}\Big(1 + \cos\Big(\frac{2\pi}{c}\Big(x - \frac{1}{2}\Big)\Big)\Big)\mathbf{1}_{|x-\frac{1}{2}|\leq\frac{c}{2}},$$

where ρ^f is the compression of a factor c of ρ^{in}. The exact expression of the density interpo-

lation is

$$\rho(t,x,y) = \frac{1}{t(c-1)+1}\Big(1+\cos\Big(\frac{2\pi}{t(c-1)+1}\Big(x-\frac{1}{2}\Big)\Big)\Big)\mathbf{1}_{|x-\frac{1}{2}|\leq\frac{t(c-1)+1}{2}},$$

whereas the exact potential is

$$\phi(t,x,y) = \frac{1}{2}\frac{c-1}{t(c-1)+1}\Big(x-\frac{1}{2}\Big)^2.$$

The Wasserstein distance between the two densities is

$$\mathcal{W}_2(\rho^{in},\rho^f) = \sqrt{\frac{(\pi^2-6)(c-1)^2}{12\pi^2}}.$$

The numerical results for $c = 0.3$ are shown in Table 2.2. Again, in all the four cases, the Wasserstein distance and the gradient of the potential converge, with the errors exhibiting at least a linear rate of convergence. However, the density does not seem to converge in the non-enriched scheme with linear reconstruction, whereas it converges in the other cases.

It is noticeable from the convergence tests we performed how in the case of a pure translation the instability tends to disappear with refinement, whereas with compression this depends on the reconstruction used: the harmonic reconstruction seems to prevent the issue, the linear one does not. Our strategy of enriching the discrete space of potentials alleviates the problem and enables to recover the convergence of the density.

We performed the same convergence tests with cartesian grids, using the non-enriched scheme with linear reconstruction. Since the scheme does not oscillate with these meshes, it is interesting to see how it performs in this case in order to compare with the case of unstructured grids. The results are presented in Table 2.3. In this case, the density converges with an order of accuracy higher than one.

2.7.3 Geodesic

To conclude, we consider the transport problem between a cross distributed density and its rotation by 45 degrees. We compute the discrete solution with the enriched scheme, using the harmonic reconstruction, with $h'_{\mathcal{T}} = 0.0156, \#\mathcal{T}' = 14336$ and $N+1 = 32$ time steps. The approximate density interpolation is displayed in Figure 2.5: as expected, each branch of the cross splits symmetrically in two parts which move towards the two opposite branches of the rotated cross.

2.8 Perspectives

We considered TPFA discretizations of the dynamical formulation of the quadratic optimal transport problem. In particular, we proposed a method based on nested meshes to deal with numerical instabilities that occur when using this type of techniques. We also proved quantitative convergence estimates in the case of smooth solutions and proposed the use of interior point techniques for the efficient numerical solutions of the scheme. Several interesting questions remain open on all the three aspects of the problem we considered:

Table 2.2: Convergence test on the compression.

$h'_{\mathcal{T}}$	N	$\epsilon_{\mathcal{W}_2}$	rate	ϵ_ϕ	rate	$\epsilon_{\nabla\phi}$	rate	ϵ_ρ	rate
Non-enriched scheme with linear reconstruction									
0.250	1	1.653e-02	/	4.734e-03	/	6.903e-02	/	2.288e-01	/
0.125	3	1.421e-03	3.540	1.471e-03	1.687	3.301e-02	1.064	1.285e-01	0.832
0.062	7	2.978e-04	2.255	4.651e-04	1.661	1.729e-02	0.933	1.859e-01	-0.532
0.031	15	4.850e-04	-0.704	1.466e-04	1.666	1.038e-02	0.736	2.193e-01	-0.238
0.016	31	2.030e-04	1.257	4.491e-05	1.706	6.351e-03	0.709	2.378e-01	-0.117
Non-enriched scheme with harmonic reconstruction									
0.250	1	2.380e-03	/	2.785e-03	/	3.954e-02	/	2.666e-01	/
0.125	3	8.112e-03	-1.769	1.403e-03	0.989	2.384e-02	0.730	7.503e-02	1.829
0.062	7	2.805e-03	1.532	4.851e-04	1.532	1.162e-02	1.037	7.046e-02	0.091
0.031	15	6.207e-04	2.176	1.242e-04	1.966	5.419e-03	1.100	4.919e-02	0.518
0.016	31	1.652e-04	1.910	3.574e-05	1.797	2.690e-03	1.011	3.393e-02	0.536
Enriched scheme with linear reconstruction									
0.250	1	1.746e-02	/	4.130e-03	/	6.212e-02	/	2.333e-01	/
0.125	3	2.093e-03	3.060	9.486e-04	2.122	2.725e-02	1.189	7.694e-02	1.600
0.062	7	2.436e-04	3.103	2.827e-04	1.747	1.274e-02	1.097	5.805e-02	0.406
0.031	15	1.538e-04	0.664	7.698e-05	1.876	5.834e-03	1.127	3.551e-02	0.709
0.016	31	5.447e-05	1.497	1.932e-05	1.994	2.751e-03	1.085	2.325e-02	0.611
Enriched scheme with harmonic reconstruction									
0.250	1	7.281e-03	/	3.069e-03	/	4.756e-02	/	2.606e-01	/
0.125	3	2.609e-03	1.480	7.574e-04	2.019	2.332e-02	1.028	5.786e-02	2.171
0.062	7	1.626e-03	0.682	2.984e-04	1.344	1.112e-02	1.069	4.280e-02	0.435
0.031	15	2.752e-04	2.563	7.551e-05	1.983	5.378e-03	1.048	2.409e-02	0.829
0.016	31	6.788e-05	2.020	2.166e-05	1.802	2.700e-03	0.994	1.537e-02	0.648

1. As for the issue of the numerical instabilities, the origin of these remains unclear, although their appearance is not surprising since the optimal transport interpolation does not imply any direct regularizing effect (e.g., the interpolation between two Dirac masses stays a Dirac). Our approach (together with previous works on the L^1 optimal transport problem [56, 57]) points towards the existence of a hidden inf-sup type of condition, analogous to the well-known ones for linear saddle point problems, which guaranties some regularity in the interpolation.

2. The convergence results we proposed are only partial as they require that the density is strictly positive and also they do not apply to the density itself. Note, however, that the positivity requirement is only needed for the approximation result on the continuity equation in Proposition 2.8, and this could be avoided using for example the regularization technique used by Lavenant in [79]. Note also that the same type of inf-sup condition needed for stability could also be used to derive convergence rates for the

Table 2.3: Convergence test for the non-enriched scheme with linear reconstruction on cartesian grids.

$h'_{\mathcal{T}}$	N	$\epsilon_{\mathcal{W}_2}$	rate	ϵ_ϕ	rate	$\epsilon_{\nabla\phi}$	rate	ϵ_ρ	rate
Translation									
0.283	1	6.324e-02	/	2.978e-02	/	2.396e-01	/	8.181e-01	/
0.141	3	1.392e-02	2.184	7.749e-03	1.942	1.207e-01	0.990	2.907e-01	1.493
0.071	7	2.753e-03	2.338	2.389e-03	1.698	5.520e-02	1.128	1.270e-01	1.195
0.035	15	4.681e-04	2.556	7.085e-04	1.754	2.294e-02	1.267	4.361e-02	1.542
0.018	31	6.781e-05	2.787	1.932e-04	1.875	8.750e-03	1.390	1.385e-02	1.655
Compression									
0.283	1	1.080e-02	/	7.364e-03	/	9.636e-02	/	1.500e-01	/
0.141	3	1.956e-03	2.465	2.754e-03	1.419	5.187e-02	0.894	1.007e-01	0.575
0.071	7	2.283e-03	-0.224	1.005e-03	1.455	2.410e-02	1.106	3.968e-02	1.344
0.035	15	7.757e-04	1.558	2.874e-04	1.805	9.323e-03	1.370	1.115e-02	1.831
0.018	31	2.323e-04	1.740	7.498e-05	1.939	3.402e-03	1.454	3.135e-03	1.831

density.

3. The interior point technique we proposed for the solutions of the discrete system of optimality conditions can be made even more effective by using iterative methods for the solution of the linearized system. However, this is possible only once appropriate preconditioners are available. The challenging nature of the problem, which is mostly due to the interplay of the time and space discretization, implies that the design of such preconditioners requires a dedicated study and must be adapted to the discrete problem itself.

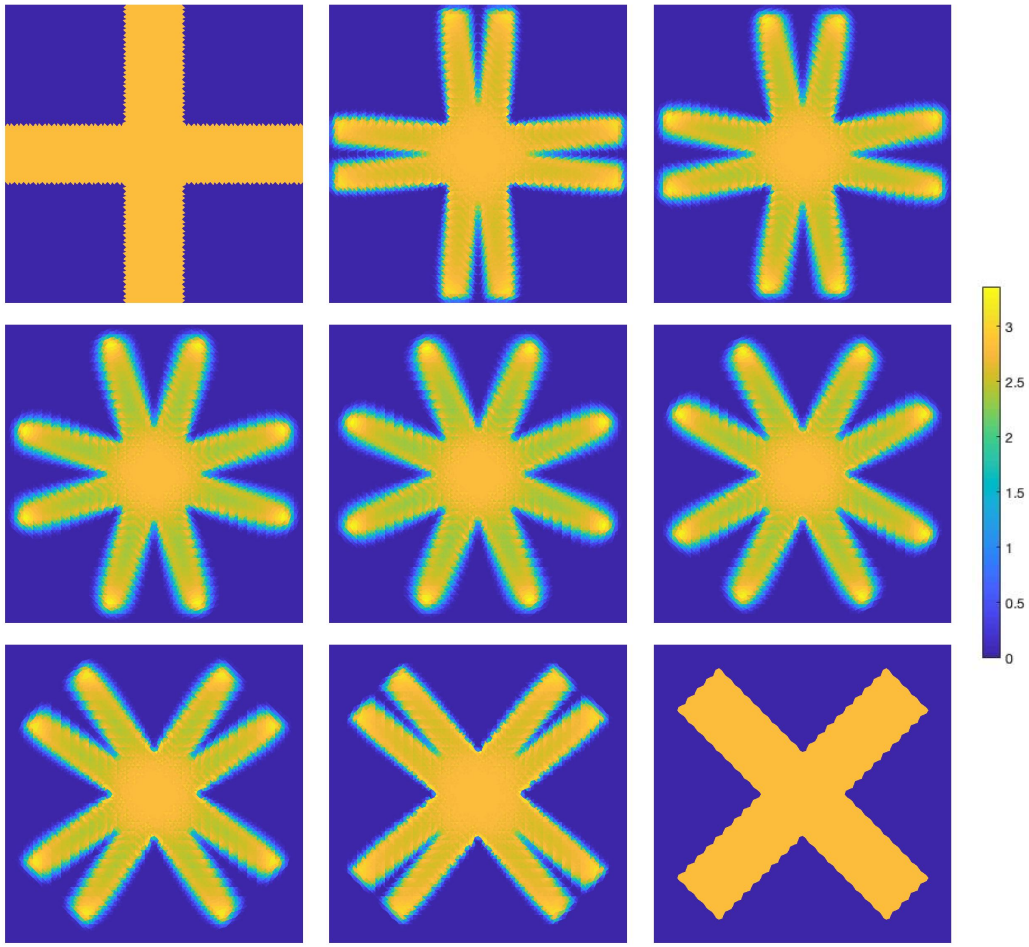


Figure 2.5: Wasserstein interpolation between a cross distributed density and its rotation by 45 degrees. Time increases from left to right, from top to bottom.

Chapter 3

A variational finite volume scheme for Wasserstein gradient flows

This chapter is issued from:

Clément Cancès, Thomas Galloët, and Gabriele Todeschi. A variational finite volume scheme for wasserstein gradient flows. *Numerische Mathematik*, 146:437–480, 10 2020.

3.1 Introduction

Our leading objective is to develop numerical schemes for the solution of Wasserstein gradient flows. Let us recall the form of the problem we want to tackle numerically. Given a convex and compact subset Ω of $\mathbb{R}^d$, a strictly convex and proper energy functional $\mathcal{E} : L^1(\Omega;\mathbb{R}_+) \to [0,+\infty]$, and given an initial density $\rho^0 \in L^1(\Omega;\mathbb{R}_+)$ with finite energy, i.e. such that $\mathcal{E}(\rho^0) < +\infty$, we want to solve problems of the form:

$$\begin{cases} \partial_t \varrho - \nabla\cdot(\rho\nabla\frac{\delta\mathcal{E}}{\delta\rho}[\rho]) = 0 & \text{in } Q_T = [0,T)\times\mathring{\Omega}\,, \\ \rho\nabla\frac{\delta\mathcal{E}}{\delta\rho}[\rho]\cdot\boldsymbol{n} = 0 & \text{on } \Sigma_T = [0,T)\times\partial\Omega\,, \\ \rho(0,\cdot) = \rho^0 & \text{in } \Omega\,. \end{cases} \tag{3.1}$$

As we explained in Chapter 1, a problem of this form can be interpreted as a gradient flow in the Wasserstein space with respect to the energy $\mathcal{E}$, i.e. a process that starting from ρ^0 evolves following the steepest decreasing direction of $\mathcal{E}$. A typical example of problem entering this framework is the linear Fokker-Planck equation [73]:

$$\partial_t\rho = \Delta\rho + \nabla\cdot(\rho\nabla V) \quad \text{in } Q_T, \tag{3.2}$$

complemented with no-flux boundary conditions and an initial condition. In (3.2), $V \in W^{1,\infty}(\Omega)$ denotes a Lipschitz continuous exterior potential. In this case, the energy functional is

$$\mathcal{E}(\rho) = \int_\Omega [\rho\log\frac{\rho}{e^{-V}} - \rho + e^{-V}]\mathrm{d}\boldsymbol{x}. \tag{3.3}$$

The potential V is defined up to an additive constant, which can be adjusted so that the densities e^{-V} and ρ^0 have the same mass.

3.1.1 JKO semi-discretization

Problem (3.1) can of course be directly discretized and solved using one of the many tools available nowadays for the numerical approximation of partial differential equations. The development of energy diminishing numerical methods based on classical ODE solvers for the march in time has been the purpose of many contributions in the recent past, see for instance [22, 35, 36, 31, 119, 106, 38]. Nevertheless, the aforementioned methods disregard the fact that the trajectory aims at optimizing the energy decay, in opposition to methods based on the JKO scheme. We recall that this scheme can be thought as a generalization to the space $\mathcal{P}(\Omega)$ (the mass being defined by the initial data ρ^0) equipped with the metric $\mathcal{W}_2$ of the backward Euler scheme. It defines recursively a sequence of measures $(\rho^n)_{n\in\mathbb{N}} \subset \mathcal{P}(\Omega)$, approximating the continuous curve, as

$$\rho^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{2\tau}\mathcal{W}_2^2(\rho,\rho^{n-1}) + \mathcal{E}(\rho)\,, \tag{3.4}$$

starting from the initial condition ρ^0. The parameter τ is the time discretization step. See Section 1.2.1 for more details.

Lagrangian numerical methods appear to be very natural (especially in dimension one) to approximate the Wasserstein distance and thus the solution to (3.4). This was already noticed in [77], and motivated numerous contributions, see for instance [91, 30, 92, 74, 43, 40, 81]. In our approach, we rather consider an Eulerian method based on finite volumes for the space discretization. The link between monotone finite volumes and optimal transportation was simultaneously highlighted by Mielke [99] and Maas [88, 63, 50, 89, 64]. But these works only focuses on the space discretization, whereas we are interested in the fully discrete setting. Moreover, the approximation based on upstream mobility we propose in Section 3.2.2 does not enter their framework. Last but not least, let us mention the so-called ALG2-JKO scheme [17, 33] where the optimization problem (3.4) is solved thanks to an augmented Lagrangian iterative method. Our approach is close to the one of [17], with the goal to obtain a faster numerical solver.

As we explained in Section 1.2.2, the primal-dual optimality conditions of problem (3.4) amounts to the mean field game

$$\begin{cases} \partial_t\rho - \nabla\cdot(\rho\nabla\phi) = 0\,, \\ \partial_t\phi - \frac{1}{2}|\nabla\phi|^2 = 0\,, \end{cases} \quad \text{in } [t^{n-1},t^n]\times\Omega, \text{ with} \quad \begin{cases} \rho(t^{n-1},\cdot) = \rho^{n-1}\,, \\ \phi(t^n,\cdot) = \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n,\cdot)]\,, \end{cases} \quad \text{in } \Omega\,, \tag{3.5}$$

complemented with the no-flux boundary condition $\nabla\phi\cdot\boldsymbol{n} = 0$ on $[t^{n-1},t^n]\times\partial\Omega$. The optimal ρ^n of (3.4) is then equal to $\rho(t^n,\cdot)$. The approximation of the system (3.5) is a natural strategy to approximate the solution to (3.1). Of course, its direct discretization would in general lose the variational structure of problem (3.4), disregarding the advantage of this semi-discretization in time. In [17, 41] two variational approaches have been proposed to solve problem (3.4), which essentially lead to find solutions to the system of optimality conditions (3.5) (more precisely to the non saturated version, see Section 1.2.2). Notice however that in [17] the authors did not discretize directly the optimization problem, contrary to [41], where the problem has been first discretized then optimized. These methods require a sub-time stepping to solve system (3.5) on each interval $[t^{n-1},t^n]$, yielding a possibly important computational cost. The avoidance of this sub-time stepping is the main motivation of the time discretization we propose now.

3.1.2 Implicit linearization of the Wasserstein distance and LJKO scheme

Let us introduce in the semi-discrete in time setting the time discretization to be used in the fully discrete setting later on. The following ansatz is at the basis of our approach: when τ is small, ρ^n is close to ρ^{n-1}. Then owing to [120, Section 7.6] (see also [110]), the Wasserstein distance between two densities ρ and μ of $\mathcal{P}(\Omega)$ is close to some weighted H^{-1} distance, namely

$$\|\rho-\mu\|_{\dot{H}^{-1}_\rho} = \mathcal{W}_2(\rho,\mu) + o(\mathcal{W}_2(\rho,\mu)), \quad \forall \rho,\mu \in \mathcal{P}(\Omega)\,. \tag{3.6}$$

In the above formula, we denoted by

$$\|h\|_{\dot{H}^{-1}_\rho} = \left\{ \sup_\varphi \int_\Omega h\varphi \,\mathrm{d}\boldsymbol{x} \;\middle|\; \|\varphi\|_{\dot{H}^1_\rho} \le 1 \right\}, \quad \text{with } \|\varphi\|_{\dot{H}^1_\rho} = \left(\int_\Omega \rho |\nabla\varphi|^2 \mathrm{d}\boldsymbol{x} \right)^{1/2}, \tag{3.7}$$

so that $\|\rho-\mu\|_{\dot{H}^{-1}_\rho} = \|\psi\|_{\dot{H}^1_\rho}$ with ψ solution to

$$\begin{cases} \rho-\mu-\nabla\cdot(\rho\nabla\psi) = 0 & \text{in } \Omega\,, \\ \nabla\psi\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega\,. \end{cases} \tag{3.8}$$

Indeed, in view of (3.7)-(3.8), there holds

$$\int_\Omega (\rho-\mu)\varphi \,\mathrm{d}\boldsymbol{x} = -\int_\Omega \nabla\cdot(\rho\nabla\psi)\varphi \,\mathrm{d}\boldsymbol{x} = \int_\Omega \rho\nabla\psi\cdot\nabla\varphi \,\mathrm{d}\boldsymbol{x} \le \|\psi\|_{\dot{H}^1_\rho}\|\varphi\|_{\dot{H}^1_\rho}\,,$$

with equality if $\varphi = \psi/\|\psi\|_{\dot{H}^1_\rho}$. Equation (3.8) can be thought as a linearization of the Monge-Ampère equation (1.7) ([120, Exercise 4.1]).

As the JKO scheme is an order one discretization in time of the continuous flow, it is reasonable to approximate the computation of the complex Wasserstein distance. In view of (3.6), a natural idea is to replace it by the weighted $\dot{H}^{-1}_\rho$ norm in (3.4), leading to what we call the implicitly linearized JKO (or LJKO) scheme:

$$\rho^n \in \operatorname*{argmin}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{2\tau}\left\|\rho-\rho^{n-1}\right\|^2_{\dot{H}^{-1}_\rho(\Omega)} + \mathcal{E}(\rho)\,. \tag{3.9}$$

We will show numerically that this approximation preserves the order one accuracy in time (Section 3.4). We stress that, as the $\dot{H}^{-1}_\rho$ norm is changing at every step n, scheme (3.9) cannot be considered as a minimizing movement in a metric space. The choice of an implicit weight ρ in (3.9) appears to be particularly important when $\{\rho^{n-1}=0\}$ has a non-empty interior set, which cannot be properly invaded by ρ^n if one chooses the explicit (but computationally cheaper) weight ρ^{n-1} as in [102]. Our time discretization is close to the one that was proposed very recently in [84] where the introduction of inner time stepping was also avoided. In [84], the authors introduce a regularization term based on Fisher information, which mainly amounts to stabilize the scheme thanks to some additional non-degenerate diffusion. In our approach, we manage to avoid this additional stabilization term by taking advantage of the monotonicity of the involved operators.

At each step $n \ge 1$, (3.9) can be formulated as a constrained optimization problem. To highlight its convexity, we perform the change of variables $(\rho,\psi) \mapsto (\rho, \boldsymbol{m} = -\rho\nabla\psi)$, in

analogy with the change of variables we performed in Section 1.2.2. Recalling the definition of the density of kinetic energy function $B : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty]$,

$$B(p, \boldsymbol{Q}) := \begin{cases} \frac{|\boldsymbol{Q}|^2}{2p} & \text{if } p > 0\,, \\ 0 & \text{if } p = 0,\ \boldsymbol{Q} = 0\,, \\ +\infty & \text{else}\,, \end{cases} \tag{3.10}$$

we can rewrite step n as:

$$\inf_{(\rho,\boldsymbol{m})} \frac{1}{\tau} \int_\Omega B(\rho, \boldsymbol{m}) \mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho), \quad \text{subject to: } \begin{cases} \rho - \rho^{n-1} + \nabla \cdot \boldsymbol{m} = 0 & \text{in } \Omega\,, \\ \boldsymbol{m} \cdot \boldsymbol{n} = 0 & \text{on } \partial\Omega\,. \end{cases} \tag{3.11}$$

Incorporating the constraint in the above formulation yields the following inf-sup problem:

$$\inf_{(\rho,\boldsymbol{m})} \sup_{\phi} \frac{1}{\tau} \int_\Omega B(\rho, \boldsymbol{m}) \mathrm{d}\boldsymbol{x} - \int_\Omega (\rho - \rho^{n-1})\phi \, \mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{m} \cdot \nabla\phi \, \mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho)\,, \tag{3.12}$$

the supremum with respect to ϕ being $+\infty$ unless the constraint is satisfied. Problem (3.12) is strictly convex in $(\rho, \boldsymbol{m})$ and concave (since linear) in ϕ. Exploiting Fenchel-Rockafellar duality theory it is possible to show that strong duality holds, so that (3.12) is equivalent to its dual problem where the inf and the sup have been swapped. Optimizing with respect to $\boldsymbol{m}$ yields the optimality condition $\boldsymbol{m} = -\tau\rho\nabla\phi$, hence the problem reduces to

$$\sup_{\phi} \int_\Omega \rho^{n-1}\phi \, \mathrm{d}\boldsymbol{x} + \inf_{\rho} \int_\Omega (-\phi - \frac{\tau}{2}|\nabla\phi|^2)\rho \, \mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho)\,. \tag{3.13}$$

The problem is now strictly convex in ρ and concave in ϕ. Optimizing with respect to ρ leads to the optimality condition

$$\phi^n + \frac{\tau}{2}|\nabla\phi^n|^2 \leq \frac{\delta\mathcal{E}}{\delta\rho}[\rho^n]\,, \tag{3.14}$$

with equality on $\{\rho^n > 0\}$. In the above formula, ϕ^n denotes the optimal ϕ realizing the sup in (3.13). Similarly to what has been done in Section 1.2.2 for the JKO scheme, it is possible to show again that saturating inequality (3.14) on $\{\rho^n = 0\}$ is optimal since the mapping $f \mapsto \phi$ solution to $\phi + \frac{\tau}{2}|\nabla\phi|^2 = f$ is monotone. Finally, the optimality conditions for the LJKO problem (3.9) write

$$\begin{cases} \dfrac{\rho^n - \rho^{n-1}}{\tau} - \nabla \cdot (\rho^n \nabla\phi^n) = 0\,, \\ \phi^n + \dfrac{\tau}{2}|\nabla\phi^n|^2 = \dfrac{\delta\mathcal{E}}{\delta\rho}[\rho^n]\,, \end{cases} \tag{3.15}$$

set on Ω, complemented with the homogeneous Neumann boundary condition $\nabla\phi^n \cdot \boldsymbol{n} = 0$ on $\partial\Omega$. We can interpret (3.15) as the one step resolvent of the mean-field game (3.5). Both the forward in time continuity equation and the backward in time Hamilton-Jacobi equation are discretized thanks to one step of backward Euler scheme.

Remark 3.1. *From the computations we just showed, our approximation of the Wasserstein distance in problem (3.11) coincides with a right endpoint approximation of the kinetic energy on the whole interval $[t^{n-1}, t^n]$ (which is the rescaled interval $[0, 1]$ in the definition of the dynamical optimal transport problem (1.19)-(1.20)) and a single Euler step for the continuity equation. Notice that differently from the computation of geodesics, the final density is not fixed here but to be chosen, which is the reason why this operation is not restrictive in this case. See Section 2.3 for more details.*

3.1.3 Goal and organisation of the chapter

As already noted, most of the numerical methods based on the backward Euler scheme disregard the optimal character of the trajectory $t \mapsto \varrho(t)$ of the exact solution to (3.1). Rather than discretizing directly the equation (3.1), which can be thought as the Euler-Lagrange equation for the steepest descent of the energy, we propose to first discretize with respect to space the functional appearing in the optimization problem (3.9), and then to optimize. The corresponding Euler-Lagrange equations will then encode the optimality of the trajectory. The choice of the LJKO scheme (3.9), rather than the classical JKO scheme (3.4), is motivated by the fact that solving (3.15) is computationally affordable. Indeed, it merely demands to approximate two functions, ρ^n and ϕ^n, rather than time depending trajectories as for the JKO scheme (3.5). This allows in particular to avoid inner time stepping as in [17, 41], making our approach much more tractable to solve complex problems.

Two-Point Flux Approximation (TPFA) finite volumes are a natural solution for the space discretization. They are naturally locally conservative thus well-suited to approximate conservation laws. Moreover, they naturally transpose to the discrete setting the monotonicity properties of the continuous operators. Monotonicity was crucial in the derivation of the optimality conditions (3.15), as it will also be the case in the fully discrete framework later on. This led us to use upstream mobilities in the definition of the discrete counterpart of the squared $\dot{H}_\rho^{-1}$ norm. The system (3.15) thus admits a discrete counterpart (3.30). The derivation of the fully discrete finite volume scheme based on the LJKO time discretization is performed in Section 3.2, where we also establish the well-posedness of the scheme, as well as the preservation at the discrete level of fundamental properties of the continuous model, namely the non-negativity of the densities and the decay of the energy along time. In Section 3.3, we show that our scheme converges in the case of the Fokker-Planck equation (3.2) under the assumption that the initial density is bounded from below by a positive constant. Even though we do not treat problem (3.1) in its full generality, this result shows the consistency of the scheme. Finally, Section 3.4 is devoted to numerical results, where our scheme is tested on several problems, including systems of equations of the type of (3.1).

3.2 A variational Finite Volume scheme

The goal of this section is to define the fully discrete scheme to solve (3.1), and to exhibit some important properties it has. We will rely on the notion of admissible mesh $(\mathcal{T}, \overline{\Sigma}, (\boldsymbol{x}_K)_{K\in\mathcal{T}})$, Definition 1.1, we gave in Section 1.3 for TPFA finite volumes, along with the notation introduced therein. We recall that the discrete space $\mathbb{R}^{\mathcal{T}}$, the space of discrete quantities defined on the control volumes of the partitioning[1], is equipped with the scalar product

$$\langle \boldsymbol{h}, \boldsymbol{\phi} \rangle_{\mathcal{T}} = \sum_{K\in\mathcal{T}} h_K \phi_K m_K \,, \quad \forall\, \boldsymbol{h} = (h_K)_{K\in\mathcal{T}} \,, \boldsymbol{\phi} = (\phi_K)_{K\in\mathcal{T}} \,,$$

which mimics the usual scalar product on $L^2(\Omega)$.

[1] We won't consider here a discretization based on nested meshes. The potential and the density variables will be discretized in the same space, since we did not experience any stability issue in this case. Indeed, the Wasserstein distance is discretized in time with a single step and the (final) density is regularized thanks to the energy functional $\mathcal{E}$.

3.2.1 Upstream weighted dissipation potentials

Since the LJKO time discretization presented in Section 3.1.2 relies on weighted $\dot{H}^1_\rho$ and $\dot{H}^{-1}_\rho$ norms, we introduce the discrete counterparts to be used in the sequel. As it will be explained in what follows, the upwinding yields problems to introduce discrete counterparts to the norms. To bypass this difficulty, we adopt a formalism based on dissipation potentials inspired from the one of generalized gradient flows introduced by Mielke in [99]. This framework was used for instance to study the convergence of the semi-discrete in space squareroot finite volume approximation of the Fokker-Planck equation, see [67].

Let $\boldsymbol{\rho} = (\rho_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}_+$, and let $\boldsymbol{\phi} = (\phi_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$, then we define the upstream weighted discrete counterpart of $\frac{1}{2}\|\phi\|^2_{\dot{H}^1_\rho}$ by

$$\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi}) = \frac{1}{2}\sum_{\sigma\in\Sigma} a_\sigma \rho_\sigma \left(\phi_K - \phi_L\right)^2 \geq 0\,, \tag{3.16}$$

where ρ_σ denotes the upwind value of $\boldsymbol{\rho}$ on $\sigma \in \Sigma$:

$$\rho_\sigma = \begin{cases} \rho_K & \text{if } \phi_K > \phi_L\,, \\ \rho_L & \text{if } \phi_K < \phi_L\,, \end{cases} \qquad \forall \sigma = K|L \in \Sigma\,. \tag{3.17}$$

Because of the upwind choice of the mobility (3.17), the functional (3.16) is not symmetric, i.e., $\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi}) \neq \mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};-\boldsymbol{\phi})$ in general, which prohibits to define a semi-norm from $\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\cdot)$. But one easily checks that $\boldsymbol{\phi} \mapsto \mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho},\boldsymbol{\phi})$ is convex, continuous thus lower semi-continuous (l.s.c.) and proper.

Let us now turn to the definition of the discrete counterpart of $\frac{1}{2}\|\cdot\|^2_{\dot{H}^{-1}_\rho}$. To this end, we recall the definition of the space of conservative fluxes $\mathbb{F}_{\mathcal{T}} \subset \mathbb{R}^{2\Sigma}$ we gave in Section 1.3.3. An element $\boldsymbol{F}$ of $\mathbb{F}_{\mathcal{T}}$ is made of two outward fluxes $F_{K,\sigma}, F_{L,\sigma}$ for each $\sigma = K|L \in \Sigma$. We impose the conservativity across each internal face

$$F_{K,\sigma} + F_{L,\sigma} = 0\,, \quad \forall \sigma = K|L \in \Sigma\,. \tag{3.18}$$

In what follows, we denote by $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$. There are no fluxes across the boundary faces. The space $\mathbb{F}_{\mathcal{T}}$ is then defined as

$$\mathbb{F}_{\mathcal{T}} = \left\{ \boldsymbol{F} = (F_{K,\sigma}, F_{L,\sigma})_{\sigma=K|L\in\Sigma} \in \mathbb{R}^{2\Sigma} \;\middle|\; (3.18) \text{ holds} \right\}.$$

Now, we define the subspace

$$\mathbb{R}^{\mathcal{T}}_0 = \left\{ \boldsymbol{h} = (h_K)_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}} \;\middle|\; \langle \boldsymbol{h}, \mathbf{1}\rangle_{\mathcal{T}} = 0 \right\}$$

and

$$\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{h}) = \inf_{\boldsymbol{F}} \sum_{\sigma\in\Sigma} B(\rho_\sigma, F_\sigma)\, d_\sigma m_\sigma \geq 0\,, \qquad \forall \boldsymbol{h} \in \mathbb{R}^{\mathcal{T}}_0\,, \tag{3.19}$$

where the minimization over $\boldsymbol{F}$ is restricted to the linear subspace of $\mathbb{F}_{\mathcal{T}}$ such that

$$h_K m_K = \sum_{\sigma\in\Sigma_K} m_\sigma F_{K,\sigma}, \qquad \forall K \in \mathcal{T}. \tag{3.20}$$

In (3.19), ρ_σ denotes the upwind value with respect to $\boldsymbol{F}$, i.e.,

$$\rho_\sigma = \begin{cases} \rho_K & \text{if } F_{K,\sigma} > 0\,, \\ \rho_L & \text{if } F_{L,\sigma} > 0\,, \end{cases} \qquad \forall \sigma = K|L \in \Sigma\,. \tag{3.21}$$

Summing (3.20) over $K \in \mathcal{T}$ and using the conservativity across the edges (3.18), one notices that there is no $\boldsymbol{F} \in \mathbb{F}_\mathcal{T}$ satisfying (3.20) unless $\boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}$. But when $\boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}$, the minimization set in (3.19) is never empty. Note that $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h})$ may take infinite values when $\boldsymbol{\rho}$ vanishes on some cells, for instance $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h}) = +\infty$ if $h_K > 0$ and $\rho_K = 0$ for some $K \in \mathcal{T}$.

Formula (3.19) deserves some comments. This sum is built to approximate $\int_\Omega \frac{|\boldsymbol{m}|^2}{2\rho} \mathrm{d}\boldsymbol{x}$. The flux F_σ approximates $|\boldsymbol{m} \cdot \boldsymbol{n}_\sigma|$, and thus encodes the information on $\boldsymbol{m}$ only in one direction (normal to the face σ) over d. But on the other hand, the volume $d_\sigma m_\sigma$ is equal to dm_{Δ_σ} which allows to hope that the sum is a consistent approximation of the integral. This remark has a strong link with the notion of inflated gradients introduced in [44, 55]. The convergence proof carried out in Section 3.3 somehow shows the non-obvious consistency of this formula.

At the continuous level, the norms $\|\cdot\|_{\dot{H}^1_\rho}$ and $\|\cdot\|_{\dot{H}^{-1}_\rho}$ are in duality. This property is transposed to the discrete level in the following sense.

Lemma 3.2. *Given $\boldsymbol{\rho} \geq \mathbf{0}$, the functionals $\boldsymbol{h} \mapsto \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h})$ and $\boldsymbol{\phi} \mapsto \mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{\phi})$ are one another Legendre transforms in the sense that*

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h}) = \sup_{\boldsymbol{\phi}} \langle \boldsymbol{h}, \boldsymbol{\phi} \rangle_\mathcal{T} - \mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{\phi})\,, \quad \forall \boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}\,. \tag{3.22}$$

In particular, both are proper convex l.s.c. functionals. Moreover, if $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h})$ is finite, then there exists a discrete Kantorovitch potential $\boldsymbol{\phi}$ solving

$$h_K m_K = \sum_{\sigma \in \Sigma} a_\sigma \rho_\sigma (\phi_K - \phi_L)\,, \quad \forall K \in \mathcal{T}\,, \tag{3.23}$$

such that

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h}) = \mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{\phi}) = \frac{1}{2} \langle \boldsymbol{h}, \boldsymbol{\phi} \rangle_\mathcal{T}\,. \tag{3.24}$$

Proof. Let $\boldsymbol{\rho} \geq \mathbf{0}$ be fixed. Incorporating the constraint (3.20) in (3.19), and using the definition of ρ_σ and the twice conservativity constraint (3.18), we obtain the saddle point primal problem

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h}) = \inf_{\boldsymbol{F}} \sup_{\boldsymbol{\phi}} \sum_{\sigma \in \Sigma} \Big[B(\rho_K, F^+_{K,\sigma}) + B(\rho_L, F^-_{K,\sigma}) \Big] m_\sigma d_\sigma \\ + \sum_{K \in \mathcal{T}} h_K \phi_K m_K - \sum_{\sigma \in \Sigma} m_\sigma F_{K,\sigma} (\phi_K - \phi_L).$$

The functional in the right-hand side is convex and coercive with respect to $\boldsymbol{F}$ and linear with respect to $\boldsymbol{\phi}$, so that strong duality holds. We can exchange the sup and the inf in the above formula to obtain the dual problem, and we minimize first with respect to $\boldsymbol{F}$, leading to [2]

$$F_{K,\sigma} = \rho_\sigma \Big(\frac{\phi_K - \phi_L}{d_\sigma} \Big)\,, \quad \forall \sigma \in \Sigma\,.$$

[2] Notice that $\boldsymbol{F}$ has opposite sign with respect to the gradient of $\boldsymbol{\phi}$.

Substituting $F_{K,\sigma}$ by $\rho_\sigma(\frac{\phi_K-\phi_L}{d_\sigma})$ in the dual problem leads to (3.22), while the constraint (3.20) turns to (3.23). The fact that $\mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho},\cdot)$ is also the Legendre transform of $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho},\cdot)$ follows from the fact that it is convex l.s.c., hence equal to its relaxation.

When $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{h})$ is finite, then the supremum in (3.22) is achieved, ensuring the existence of the corresponding discrete Kantorovitch potentials $\boldsymbol{\phi}$. Finally, multiplying (3.23) by the optimal ϕ_K and summing over $K\in\mathcal{T}$ yields $\langle \boldsymbol{h},\boldsymbol{\phi}\rangle_\mathcal{T} = 2\mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\phi})$. Substituting this relation in (3.22) shows the relation $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{h}) = \mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\phi})$. □

Our next lemma can be seen as an adaptation to our setting of a well known properties of optimal transportation, namely $\rho\mapsto \frac12\mathcal{W}_2^2(\rho,\mu)$ is convex, which is key in the study of Wasserstein gradient flows.

Lemma 3.3. *Let $\boldsymbol{\mu}\in\mathbb{R}_+^\mathcal{T}$, the function $\boldsymbol{\rho}\mapsto\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\mu}-\boldsymbol{\rho})$ is proper and convex on $(\boldsymbol{\mu}+\mathbb{R}_0^\mathcal{T})\cap\mathbb{R}_+^\mathcal{T}$.*

Proof. The function $\boldsymbol{\rho}\mapsto\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\mu}-\boldsymbol{\rho})$ is proper since it is equal to 0 at $\boldsymbol{\rho}=\boldsymbol{\mu}$. Then it follows from (3.22) that

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\mu}-\boldsymbol{\rho}) = \sup_{\boldsymbol{\phi}} \langle \boldsymbol{\mu}-\boldsymbol{\rho},\boldsymbol{\phi}\rangle_\mathcal{T} - \mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\phi})\,. \tag{3.25}$$

Since $\boldsymbol{\rho}\mapsto\mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\phi})$ is linear, $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\mu}-\boldsymbol{\rho})$ is defined as the supremum of linear functions, hence it is convex. □

3.2.2 A variational upstream mobility finite volume scheme

Given $\boldsymbol{\rho}^0\in\mathbb{R}_+^\mathcal{T}$, the space $\mathbb{P}_\mathcal{T}$, which is the discrete counterpart of $\mathcal{P}(\Omega)$, is defined by

$$\mathbb{P}_\mathcal{T} = \left\{\boldsymbol{\rho}\in\mathbb{R}_+^\mathcal{T} \;\middle|\; \langle\boldsymbol{\rho},\mathbf{1}\rangle_\mathcal{T} = \langle\boldsymbol{\rho}^0,\mathbf{1}\rangle_\mathcal{T}\right\} = (\boldsymbol{\rho}^0+\mathbb{R}_0^\mathcal{T})\cap\mathbb{R}_+^\mathcal{T}\,.$$

It is compact. The energy $\mathcal{E}$ is discretized into a strictly convex functional $\mathcal{E}_\mathcal{T}\in C^1(\mathbb{R}_+^\mathcal{T};\mathbb{R}_+)$ that we do not specify yet. We refer to Sections 3.3 and 3.4 for explicit examples.

We have introduced all the necessary material to introduce our numerical scheme, which combines upstream weighted finite volumes for the space discretization and the LJKO time discretization:

$$\boldsymbol{\rho}^n \in \operatorname*{arginf}_{\boldsymbol{\rho}\in\mathbb{P}_\mathcal{T}} \frac{1}{\tau}\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}) + \mathcal{E}_\mathcal{T}(\boldsymbol{\rho})\,,\quad n\geq 1\,. \tag{3.26}$$

A further characterization of the scheme is needed for its practical implementation, but the condensed expression (3.26) already provides crucial informations gathered in the following theorem. Note in particular that our scheme automatically preserves the mass and the positivity since the solutions $(\boldsymbol{\rho}^n)_{n\geq1}$ belong to $\mathbb{P}_\mathcal{T}$.

Theorem 3.4. *For all $n\geq1$, there exists a unique solution $\boldsymbol{\rho}^n\in\mathbb{P}_\mathcal{T}$ to (3.26). Moreover, energy is dissipated along the time steps. More precisely,*

$$\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^n) \leq \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^n) + \frac{1}{\tau}\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}^n;\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}^n) \leq \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^{n-1})\,,\quad \forall n\geq1\,. \tag{3.27}$$

Proof. The functional $\boldsymbol{\rho} \mapsto \frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})$ is lower semi-continuous and strictly convex on the compact set $\mathbb{P}_{\mathcal{T}}$ in view of Lemma 3.3 and of the assumptions on $\mathcal{E}_{\mathcal{T}}$. Moreover, it is proper since $\boldsymbol{\rho}^{n-1}$ belongs to its domain. Therefore, it admits a unique minimum on $\mathbb{P}_{\mathcal{T}}$. The energy / energy dissipation estimate (3.27) is obtained by choosing $\boldsymbol{\rho} = \boldsymbol{\rho}^{n-1}$ as a competitor in (3.26). □

In view of (3.25), and after rescaling the dual variable $\boldsymbol{\phi} \leftarrow \frac{\boldsymbol{\phi}}{\tau}$, solving (3.26) amounts to solve the saddle point problem

$$\inf_{\boldsymbol{\rho}\geq \mathbf{0}} \sup_{\boldsymbol{\phi}} \left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{\sigma\in\Sigma} a_\sigma \rho_\sigma (\phi_K - \phi_L)^2 + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) \,. \tag{3.28}$$

which is equivalent to its dual problem

$$\sup_{\boldsymbol{\phi}} \inf_{\boldsymbol{\rho}\geq \mathbf{0}} \left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{\sigma\in\Sigma} a_\sigma \rho_\sigma (\phi_K - \phi_L)^2 + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) \,. \tag{3.29}$$

Our strategy for the practical computation of the solution to (3.26) is to solve the system corresponding to the optimality conditions of (3.29). So far, we did not take advantage of the upwind choice of the mobility (3.17) (we only used the linearity of $(\boldsymbol{\rho}, \boldsymbol{\phi}) \mapsto (\rho_\sigma)_{\sigma\in\Sigma}$ in the proofs of Lemmas 3.2 and 3.3, which also holds true for a centered choice of the mobilities). The upwinding will be key in the proof of the following theorem, which, roughly speaking, states that there is no need of a Lagrange multiplier for the constraint $\boldsymbol{\rho} \geq \mathbf{0}$.

Theorem 3.5. *The unique solution $(\boldsymbol{\phi}^n, \boldsymbol{\rho}^n)$ to system*

$$\begin{cases} (\rho_K^n - \rho_K^{n-1}) m_K + \tau \displaystyle\sum_{\sigma\in\Sigma_K} a_\sigma \rho_\sigma^n (\phi_K^n - \phi_L^n) = 0 \,, \\ \phi_K^n m_K + \dfrac{\tau}{2} \displaystyle\sum_{\sigma\in\Sigma_K} a_\sigma \big((\phi_K^n - \phi_L^n)^+\big)^2 = \dfrac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n) \,, \end{cases} \qquad \forall K \in \mathcal{T} \,, \tag{3.30}$$

where ρ_σ^n denotes the upwind value, i.e.,

$$\rho_\sigma^n = \begin{cases} \rho_K^n & \text{if } \phi_K^n > \phi_L^n \,, \\ \rho_L^n & \text{if } \phi_K^n < \phi_L^n \,, \end{cases} \quad \forall \sigma = K|L \in \Sigma \,,$$

is a saddle point of (3.29).

System (3.30) is the discrete counterpart of (3.15), whose derivation relied on the monotonicity of the inverse of the operator $\phi \mapsto \phi + \frac{\tau}{2}|\nabla\phi|^2$. Before proving Theorem 3.5, let us show that the space discretization preserves this property at the discrete level. To this end, we introduce the functional $\boldsymbol{\mathcal{H}} = (\mathcal{H}_K)_K \in C^1(\mathbb{R}^{\mathcal{T}}; \mathbb{R}^{\mathcal{T}})$ defined by

$$\mathcal{H}_K(\boldsymbol{\phi}) := \phi_K + \frac{\tau}{2m_K} \sum_{\sigma\in\Sigma_K} a_\sigma \left((\phi_K - \phi_L)^+\right)^2 \,, \qquad \forall K \in \mathcal{T} \,.$$

Lemma 3.6. *Given $\boldsymbol{f} \in \mathbb{R}^{\mathcal{T}}$, there exists a unique solution to $\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}) = \boldsymbol{f}$, and it satisfies*

$$\min \boldsymbol{f} \leq \boldsymbol{\phi} \leq \max \boldsymbol{f}\,. \tag{3.31}$$

Moreover, let $\boldsymbol{\phi}, \widetilde{\boldsymbol{\phi}}$ be the solutions corresponding to $\boldsymbol{f}$ and $\widetilde{\boldsymbol{f}}$ respectively, then

$$\boldsymbol{f} \geq \widetilde{\boldsymbol{f}} \quad \Longrightarrow \quad \boldsymbol{\phi} \geq \widetilde{\boldsymbol{\phi}}\,. \tag{3.32}$$

Proof. Given $\boldsymbol{f} \geq \widetilde{\boldsymbol{f}}$ and $\boldsymbol{\phi}, \widetilde{\boldsymbol{\phi}}$ corresponding solutions, let K^* be the cell such that

$$\phi_{K^*} - \tilde{\phi}_{K^*} = \min_{K \in \mathcal{T}} \left(\phi_K - \tilde{\phi}_K\right).$$

Then, for all the neighboring cells L of K^*, it holds $\phi_{K^*} - \tilde{\phi}_{K^*} \leq \phi_L - \tilde{\phi}_L$ and therefore $\phi_{K^*} - \phi_L \leq \tilde{\phi}_{K^*} - \tilde{\phi}_L$ which implies

$$\frac{\tau}{2m_K} \sum_{\sigma \in \Sigma_{K^*}} a_\sigma \left((\phi_{K^*} - \phi_L)^+\right)^2 \leq \frac{\tau}{2m_K} \sum_{\sigma \in \Sigma_{K^*}} a_\sigma \left((\widetilde{\phi}_{K^*} - \widetilde{\phi}_L)^+\right)^2 . \tag{3.33}$$

Recall $\boldsymbol{f} \geq \widetilde{\boldsymbol{f}}$, so $\mathcal{H}_{K^*}(\boldsymbol{\phi}) \geq \mathcal{H}_{K^*}(\tilde{\boldsymbol{\phi}})$ together with (3.33) yield $\phi_{K^*} \geq \tilde{\phi}_{K^*}$. As in K^* the difference $\phi_K - \tilde{\phi}_K$ is minimal, we obtain $\phi_K \geq \tilde{\phi}_K$ for all $K \in \mathcal{T}$. The uniqueness of the solution $\boldsymbol{\phi}$ of $\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}) = \boldsymbol{f}$ follows directly. The maximum principle (3.31) is also a straightforward consequence of (3.32) as one can compare $\boldsymbol{\phi}$ to $(\min \boldsymbol{f})\mathbf{1}$ and $(\max \boldsymbol{f})\mathbf{1}$ which are fixed points of $\boldsymbol{\mathcal{H}}$. Finally, existence follows from Leray-Schauder fixed-point theorem [83] as the bounds (3.31) are uniform whatever $\tau \geq 0$. □

With Lemma 3.6 at hand, we can now prove Theorem 3.5.

Proof of Theorem 3.5. Uniqueness of the solution $\boldsymbol{\rho}^n$ to (3.26) was already proved in Theorem 3.4. Owing to (3.27), $\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n)$ is finite. So Lemma 3.2 ensures the existence of a discrete Kantorovitch potential $\widetilde{\boldsymbol{\phi}}^n$ satisfying (after a suitable rescaling by τ^{-1})

$$(\rho_K^n - \rho_K^{n-1})m_K + \tau \sum_{\sigma \in \Sigma_K} a_\sigma \rho_\sigma^n (\widetilde{\phi}_K^n - \widetilde{\phi}_L^n) = 0\,, \qquad \forall K \in \mathcal{T}\,. \tag{3.34}$$

The above condition is the optimality condition with respect to $\boldsymbol{\phi}$ in (3.29). To compute the optimality condition with respect to $\boldsymbol{\rho}$ in (3.29), let us rewrite the objective using the definition of ρ_σ and $\boldsymbol{\mathcal{H}}$:

$$\begin{aligned}
&\left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{\sigma \in \Sigma} a_\sigma \rho_\sigma (\phi_K - \phi_L)^2 + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) = \\
&= \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{\sigma \in \Sigma} \left[a_\sigma \rho_K \left((\phi_K - \phi_L)^+\right)^2 + a_\sigma \rho_L \left((\phi_L - \phi_K)^+\right)^2 \right] \\
&= \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{K} \sum_{\sigma \in \Sigma_K} a_\sigma \rho_K \left((\phi_K - \phi_L)^+\right)^2 \\
&= \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \langle \boldsymbol{\rho}, \boldsymbol{\phi} \rangle_{\mathcal{T}} - \sum_K m_K \rho_K \Big[\frac{\tau}{2m_K} \sum_{\sigma \in \Sigma_K} a_\sigma \left((\phi_K - \phi_L)^+\right)^2 \Big] \\
&= \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \langle \boldsymbol{\rho}, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}) \rangle_{\mathcal{T}}\,.
\end{aligned}$$

Thus (3.29) rewrites

$$\sup_{\boldsymbol{\phi}} \inf_{\boldsymbol{\rho}\geq \mathbf{0}} \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \left\langle \boldsymbol{\rho}, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}) \right\rangle_{\mathcal{T}} . \tag{3.35}$$

Denote by

$$\mathcal{Z}^n = \{K \in \mathcal{T} \mid \rho_K^n = 0\}, \quad \mathcal{P}^n = \{K \in \mathcal{T} \mid \rho_K^n > 0\} = (\mathcal{Z}^n)^c,$$

Using (3.35) the optimality conditions with respect to $\boldsymbol{\rho}$ of (3.29) thus read

$$\widetilde{\phi}_K^n m_K + \frac{\tau}{2} \sum_{\sigma \in \Sigma_{0,K}} a_\sigma \big((\widetilde{\phi}_K^n - \widetilde{\phi}_L^n)^+ \big)^2 = \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n), \quad \forall K \in \mathcal{P}^n \tag{3.36}$$

and

$$\widetilde{\phi}_K^n m_K + \frac{\tau}{2} \sum_{\sigma \in \Sigma_{0,K}} a_\sigma \big((\widetilde{\phi}_K^n - \widetilde{\phi}_L^n)^+ \big)^2 \leq \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n), \quad \forall K \in \mathcal{Z}^n . \tag{3.37}$$

By definition, $(\widetilde{\boldsymbol{\phi}}^n, \boldsymbol{\rho}^n)$ is a saddle point of (3.29), so equivalently of (3.35) and by strong duality it is also a saddle point of

$$\inf_{\boldsymbol{\rho}\geq \mathbf{0}} \sup_{\boldsymbol{\phi}} \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \left\langle \boldsymbol{\rho}, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}) \right\rangle_{\mathcal{T}} . \tag{3.38}$$

In particular $\widetilde{\boldsymbol{\phi}}^n$ is optimal in

$$\sup_{\boldsymbol{\phi}} \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \left\langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}) \right\rangle_{\mathcal{T}} . \tag{3.39}$$

To prove Theorem 3.5, we have to prove that, given $\boldsymbol{\rho}^n$, we can saturate the inequality in both (3.36) and (3.37) while preserving the optimality in (3.39). Lemma 3.6 gives the existence of a solution $\boldsymbol{\phi}^n \in \mathbb{R}^{\mathcal{T}}$ to

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^n) = \left(\frac{1}{m_K} \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n) \right)_{K \in \mathcal{T}} . \tag{3.40}$$

Note that (3.36) implies

$$\mathcal{H}_K(\boldsymbol{\phi}^n) = \mathcal{H}_K(\widetilde{\boldsymbol{\phi}}^n), \quad \forall K \in \mathcal{P}^n,$$

so

$$\left\langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^n) \right\rangle_{\mathcal{T}} = \left\langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{H}}(\widetilde{\boldsymbol{\phi}}^n) \right\rangle_{\mathcal{T}} . \tag{3.41}$$

The combination of (3.36) and (3.37) is exactly $\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^n) \geq \boldsymbol{\mathcal{H}}(\widetilde{\boldsymbol{\phi}}^n)$, thus Lemma 3.6 gives $\boldsymbol{\phi}^n \geq \widetilde{\boldsymbol{\phi}}^n$. Consequently,

$$\left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi}^n \right\rangle_{\mathcal{T}} \geq \left\langle \boldsymbol{\rho}^{n-1}, \widetilde{\boldsymbol{\phi}}^n \right\rangle_{\mathcal{T}} \tag{3.42}$$

since $\boldsymbol{\rho}^{n-1} \geq \mathbf{0}$. Incorporating (3.41) and (3.42) in (3.39) shows that $\boldsymbol{\phi}^n$ is a better competitor than $\widetilde{\boldsymbol{\phi}}^n$, and therefore again optimal. Thanks to the convexity of $\mathcal{E}_{\mathcal{T}}$ and (3.40), it also holds

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi}^n \right\rangle_{\mathcal{T}} - \left\langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^n) \right\rangle_{\mathcal{T}} \leq \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi}^n \right\rangle_{\mathcal{T}} - \left\langle \boldsymbol{\rho}, \boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^n) \right\rangle_{\mathcal{T}}$$

for every $\boldsymbol{\rho}$, which means that $(\boldsymbol{\phi}^n, \boldsymbol{\rho}^n)$ is again a saddle point of (3.29) and satisfies (3.30). Finally, owing to Lemma 3.6, the solution $\boldsymbol{\phi}^n$ to (3.40) is unique, concluding the proof of Theorem 3.5. □

3.2.3 Comparison with the classical backward Euler discretization

The scheme (3.26) is based on a "first discretize then optimize" approach. We have built a discrete counterpart of $\frac{1}{2}\mathcal{W}_2^2$ and a discrete energy $\mathcal{E}_\mathcal{T}$, then the discrete dynamics is chosen in an optimal way by (3.26). In opposition, the continuous equation (3.1) can be thought as the Euler-Lagrange optimality condition for the steepest descent of the energy. A classical approach to approximate the optimal dynamics is to discretize directly (3.1), leading to what we call a "first optimize then discretize" approach. It is classical for the semi-discretization in time of (3.1) to use a backward Euler scheme. If one combines this technique with upstream weighted finite volumes, we obtain the following fully discrete scheme:

$$(\check\rho_K^n-\rho_K^{n-1})m_K+\tau\sum_{\sigma\in\Sigma_K} a_\sigma\check\rho_\sigma^n(\check\phi_K^n-\check\phi_L^n)=0\,,\quad\text{with}\quad \check\phi_K^n=\frac{1}{m_K}\frac{\partial\mathcal{E}_\mathcal{T}}{\partial\rho_K}(\check{\boldsymbol{\rho}}^n)\,,\quad\forall K\in\mathcal{T}\,. \tag{3.43}$$

This scheme has no clear variational structure in the sense that, to our knowledge, $\check{\boldsymbol{\rho}}^n$ is no longer the solution to an optimization problem. However, it shares some common features with our scheme (3.26): it is mass and positivity preserving as well as energy diminishing.

Proposition 3.7. *Given $\boldsymbol{\rho}^{n-1}\in\mathbb{P}_\mathcal{T}$, there exists at least one solution $(\check{\boldsymbol{\rho}}^n,\check{\boldsymbol{\phi}}^n)\in\mathbb{P}_\mathcal{T}\times\mathbb{R}^\mathcal{T}$ to system (3.43), which satisfies*

$$\mathcal{E}_\mathcal{T}(\check{\boldsymbol{\rho}}^n)+\frac{1}{\tau}\mathcal{A}_\mathcal{T}(\check{\boldsymbol{\rho}}^n;\boldsymbol{\rho}^{n-1}-\check{\boldsymbol{\rho}}^n)+\tau\mathcal{A}_\mathcal{T}^*(\check{\boldsymbol{\rho}}^n;\check{\boldsymbol{\phi}}^n)\le\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^{n-1})\,. \tag{3.44}$$

Proof. Summing (3.43) over $K\in\mathcal{T}$ provides directly the conservation of mass, i.e., $\langle\check{\boldsymbol{\rho}}^n,\mathbf{1}\rangle_\mathcal{T}=\langle\boldsymbol{\rho}^{n-1},\mathbf{1}\rangle_\mathcal{T}$. Assume for contradiction that $\mathcal{K}^n=\{K\in\mathcal{T}\mid\check\rho_K^n<0\}\neq\emptyset$, then choose $K^\star\in\mathcal{K}^n$ such that $\check\phi_{K^\star}^n\ge\check\phi_K^n$ for all $K\in\mathcal{K}^n$. Then it follows from the upwind choice of the mobility in (3.43) that

$$\sum_{\sigma\in\Sigma_{K^\star}} a_\sigma\check\rho_\sigma^n(\check\phi_{K^\star}^n-\check\phi_L^n)\le 0\,,$$

so that $\check\rho_{K^\star}^n\ge\rho_{K^\star}^{n-1}\ge 0$, showing a contradiction. Therefore, $\mathcal{K}^n=\emptyset$ and $\check{\boldsymbol{\rho}}^n\ge\mathbf{0}$. These two *a priori* estimates (mass and positivity preservation) are uniform with respect to $\tau\ge 0$, thus they are sufficient to prove the existence of a solution $(\check{\boldsymbol{\rho}}^n,\check{\boldsymbol{\phi}}^n)$ to (3.43) thanks to a topological degree argument [83].

Let us now turn to the derivation of the energy / energy dissipation inequality (3.44). Multiplying (3.43) by $\check\phi_K^n$ and summing over $K\in\mathcal{T}$ provides

$$\langle\check{\boldsymbol{\rho}}^n-\boldsymbol{\rho}^{n-1},\check{\boldsymbol{\phi}}^n\rangle_\mathcal{T}+2\tau\mathcal{A}_\mathcal{T}^*(\check{\boldsymbol{\rho}}^n;\check{\boldsymbol{\phi}}^n)=0\,.$$

The definition of $\check{\boldsymbol{\phi}}^n$ and the convexity of $\mathcal{E}_\mathcal{T}$ yield $\langle\check{\boldsymbol{\rho}}^n-\boldsymbol{\rho}^{n-1},\check{\boldsymbol{\phi}}^n\rangle_\mathcal{T}\ge\mathcal{E}_\mathcal{T}(\check{\boldsymbol{\rho}}^n)-\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^{n-1})$. Thus to prove (3.44), it remains to check that

$$\frac{1}{\tau}\mathcal{A}_\mathcal{T}(\check{\boldsymbol{\rho}}^n;\boldsymbol{\rho}^{n-1}-\check{\boldsymbol{\rho}}^n)=\tau\mathcal{A}_\mathcal{T}^*(\check{\boldsymbol{\rho}}^n;\check{\boldsymbol{\phi}}^n)=\frac{1}{\tau}\mathcal{A}_\mathcal{T}^*(\check{\boldsymbol{\rho}}^n;\tau\check{\boldsymbol{\phi}}^n)\,. \tag{3.45}$$

In view of (3.23), $\tau\check{\boldsymbol{\phi}}^n$ is a discrete Kantorovitch potential sending $\boldsymbol{\rho}^{n-1}$ on $\check{\boldsymbol{\rho}}^n$ for the mobility corresponding to $\check{\boldsymbol{\rho}}^n$. Therefore (3.45) holds as a consequence of (3.24). □

Next proposition provides a finer energy / energy dissipation estimate than (3.27), which can be thought as discrete counterpart to the energy / energy dissipation inequality (EDI) which is a characterization of generalized gradient flows [3, 99].

Proposition 3.8. *Given $\boldsymbol{\rho}^{n-1} \in \mathbb{P}_{\mathcal{T}}$, let $\boldsymbol{\rho}^n$ be the unique solution to (3.26) and let $\check{\boldsymbol{\rho}}^n$ be a solution to (3.43), then*

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \tau \mathcal{A}^*_{\mathcal{T}}\left(\boldsymbol{\rho}^n; \boldsymbol{\phi}^n\right) + \tau \mathcal{A}^*_{\mathcal{T}}\left(\check{\boldsymbol{\rho}}^n; \check{\boldsymbol{\phi}}^n\right) \leq \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^{n-1}),$$

where $\check{\boldsymbol{\phi}}^n$ is defined by $m_K \check{\phi}^n_K = \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\check{\boldsymbol{\rho}}^n)$ for all $K \in \mathcal{T}$.

Proof. Since $\check{\boldsymbol{\rho}}^n$ belongs to $\mathbb{P}_{\mathcal{T}}$, it is an admissible competitor for (3.26), thus

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) \leq \mathcal{E}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n) + \frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n; \boldsymbol{\rho}^{n-1} - \check{\boldsymbol{\rho}}^n). \tag{3.46}$$

Combining this with (3.44) and bearing in mind that $\frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) = \tau \mathcal{A}^*_{\mathcal{T}}\left(\boldsymbol{\rho}^n; \boldsymbol{\phi}^n\right)$ thanks to (3.24), we obtain the desired inequality (3.46). □

3.3 Convergence in the Fokker-Planck case

In this section, we investigate the limit of the scheme when the time step τ and the size of the mesh $h_{\mathcal{T}}$ tend to 0 in the specific case of the Fokker-Planck equation (3.2). We recall that the size of the mesh is defined by $h_{\mathcal{T}} = \max_{K\in\mathcal{T}} h_K$ with $h_K = \operatorname{diam}(K)$. To this end, we consider a sequence $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m}\right)_{m\geq 1}$ of admissible discretizations of Ω in the sense of Definition 1.1 and a sequence $(\tau_m)_{m\geq 1}$ of time steps such that $\lim_{m\to\infty} \tau_m = \lim_{m\to\infty} h_{\mathcal{T}_m} = 0$. We also make the further assumptions on the mesh sequence: there exists $\zeta > 0$ such that, for all $m \geq 1$,

$$h_K \leq \zeta d_\sigma \leq \zeta^2 h_K, \quad \forall \sigma \in \Sigma_K,\ \forall K \in \mathcal{T}_m, \tag{3.47a}$$

$$\operatorname{dist}(\boldsymbol{x}_K, \overline{K}) \leq \zeta h_K, \quad \forall K \in \mathcal{T}_m, \tag{3.47b}$$

and

$$\sum_{\sigma\in\sigma_K} m_{\Delta_\sigma} \leq \zeta m_K, \quad \forall K \in \mathcal{T}_m. \tag{3.47c}$$

Let $T > 0$ be an arbitrary finite time horizon, then we assume for the sake of simplicity that $\tau_m = T/N_m$ for some integer N_m tending to $+\infty$ with m. For the ease of reading, we remove the subscript $m \geq 1$ when it appears to be unnecessary for understanding.

Given $V \in C^2(\Omega)$, we define the discrete counterpart of the energy (3.3) by

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) = \sum_{K\in\mathcal{T}} m_K \left[\rho_K \log \frac{\rho_K}{e^{-V_K}} - \rho_K + e^{-V_K}\right], \quad \forall \boldsymbol{\rho} \in \mathbb{R}_+^{\mathcal{T}},$$

where $V_K = V(\boldsymbol{x}_K)$ for all $K \in \mathcal{T}$. In view of the above formula, there holds

$$\frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}) = m_K(\log(\rho_K) + V_K), \quad \forall K \in \mathcal{T}. \tag{3.48}$$

Given an initial condition $\varrho^0 \in \mathcal{P}(\Omega)$ with positive mass, i.e. $\int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x} > 0$, and such that $\mathcal{E}(\varrho^0) < \infty$, it is discretized into $\boldsymbol{\rho}^0 = \left(\rho^0_K\right)_{K\in\mathcal{T}}$ defined by

$$\rho^0_K = \frac{1}{m_K}\int_K \varrho^0 \mathrm{d}\boldsymbol{x} \geq 0, \quad \forall K \in \mathcal{T}. \tag{3.49}$$

Note that the energy $\mathcal{E}_\mathcal{T}$ is not in $C^1(\mathbb{R}_+^\mathcal{T})$ since its gradient blows up on $\partial\mathbb{R}_+^\mathcal{T}$. However, the functional $\mathcal{E}_\mathcal{T}$ is continuous and strictly convex on $\mathbb{R}_+^\mathcal{T}$, hence the scheme (3.26) still admits a unique solution $\boldsymbol{\rho}^n$ for all $n \geq 1$ thanks to Theorem 3.4, since its proof does not use the differentiability of the energy. Thanks to the conservativity of the scheme and definition (3.49) of $\boldsymbol{\rho}^0$, one has

$$\langle \boldsymbol{\rho}^n, \mathbf{1}\rangle_\mathcal{T} = \langle \boldsymbol{\rho}^0, \mathbf{1}\rangle_\mathcal{T} = \int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x} > 0\,, \quad \forall n \geq 1.$$

Let us show that $\boldsymbol{\rho}^n > \mathbf{0}$ for all $n \geq 1$. To this end, we proceed as in [115, Lemma 8.6].

Lemma 3.9. *Assume that ϱ^0 has positive mass, then the iterated solutions $(\boldsymbol{\rho}^n)_{n\geq 1}$ to scheme (3.26) satisfy $\boldsymbol{\rho}^n > \mathbf{0}$ for all $n \geq 1$. Moreover, there exists a unique sequence $(\boldsymbol{\phi}^n)_{\geq 1}$ of discrete Kantorovitch potentials such that the following optimality conditions are satisfied for all $K \in \mathcal{T}$ and all $n \geq 1$:*

$$(\rho_K^n - \rho_K^{n-1})m_K + \tau \sum_{\sigma\in\Sigma} a_\sigma \rho_\sigma^n(\phi_K^n - \phi_L^n) = 0\,, \tag{3.50}$$

$$\phi_K^n + \frac{\tau}{2m_K} \sum_{\sigma=K|L\in\Sigma_K} a_\sigma \left((\phi_K^n - \phi_L^n)^+\right)^2 = \log(\rho_K^n) + V_K\,, \tag{3.51}$$

Proof. Define $\overline{\rho} = \frac{1}{|\Omega|}\int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x}$ and $\overline{\boldsymbol{\rho}} = \overline{\rho}\mathbf{1} \in \mathbb{P}_\mathcal{T}$, and by $\boldsymbol{\rho}_\epsilon^n = \left(\rho_{K,\epsilon}^n\right)_{K\in\mathcal{T}} = \epsilon\overline{\boldsymbol{\rho}} + (1-\epsilon)\boldsymbol{\rho}^n \in \mathbb{P}_\mathcal{T}$ for some arbitrary $\epsilon \in (0,1)$. Since $\boldsymbol{\rho}^n$ is optimal in (3.26), there holds

$$\begin{aligned}\sum_{K\in\mathcal{T}} m_K\left[\rho_K^n \log \rho_K^n - \rho_{K,\epsilon}^n \log \rho_{K,\epsilon}^n\right] \leq \sum_{K\in\mathcal{T}} m_K\left(\rho_{K,\epsilon}^n - \rho_K^n\right) V_K \\ + \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}_\epsilon^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}_\epsilon^n) - \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n).\end{aligned} \tag{3.52}$$

The convexity of $\boldsymbol{\rho} \mapsto \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}, \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho})$ implies that

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}_\epsilon^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}_\epsilon^n) \leq \epsilon\mathcal{A}_\mathcal{T}(\overline{\boldsymbol{\rho}}; \boldsymbol{\rho}^{n-1} - \overline{\boldsymbol{\rho}}) + (1-\epsilon)\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n)\,,$$

while the boundedness of V provides

$$\sum_{K\in\mathcal{T}} m_K\left(\rho_{K,\epsilon}^n - \rho_K^n\right) V_K \leq \epsilon\|V\|_{L^\infty(\Omega)}\|\varrho^0\|_{L^1(\Omega)}\,.$$

Therefore, the right-hand side in (3.52) can be overestimated by

$$\sum_{K\in\mathcal{T}} m_K\left[\rho_K^n \log \rho_K^n - \rho_{K,\epsilon}^n \log \rho_{K,\epsilon}^n\right] \leq C\epsilon$$

for some C depending on $\boldsymbol{\rho}^n, \boldsymbol{\rho}^{n-1}$ and V but not on ϵ. Setting $\mathcal{Z}^n = \{K \in \mathcal{T} \mid \rho_K^n = 0\}$ and $\mathcal{P}^n = \{K \in \mathcal{T} \mid \rho_K^n > 0\} = (\mathcal{Z}^n)^c$, we have

$$\sum_{K\in\mathcal{Z}^n} m_K\left[\rho_K^n \log \rho_K^n - \rho_{K,\epsilon}^n \log \rho_{K,\epsilon}^n\right] = \epsilon \sum_{K\in\mathcal{Z}^n} m_K\overline{\rho}\log\epsilon\overline{\rho}\,,$$

and, thanks to the convexity of $\rho \mapsto \rho\log\rho$ and to the monotonicity of $\rho \mapsto \log\rho$,

$$\begin{aligned}\sum_{K\in\mathcal{P}^n} m_K\left[\rho_K^n \log \rho_K^n - \rho_{K,\epsilon}^n \log \rho_{K,\epsilon}^n\right] &\geq \epsilon \sum_{K\in\mathcal{P}^n} m_K(\rho_K^n - \overline{\rho})(1 + \log(\rho_{K,\epsilon}^n)) \\ &\geq \epsilon \sum_{K\in\mathcal{P}^n} m_K(\rho_K^n - \overline{\rho})(1 + \log(\overline{\rho})) \geq -C\epsilon\,.\end{aligned}$$

Then dividing by ϵ and letting ϵ tend to 0, we obtain that

$$\limsup_{\epsilon\to 0} \sum_{K\in\mathcal{Z}^n} m_K \overline{\rho} \log \epsilon\overline{\rho} \le C\,,$$

which is only possible if $\mathcal{Z}^n = \emptyset$, i.e., $\boldsymbol{\rho}^n > \mathbf{0}$. This implies that $\mathcal{E}_{\mathcal{T}}$ is differentiable at $\boldsymbol{\rho}^n$, hence the optimality conditions (3.30) hold, which rewrites as (3.51)–(3.50) thanks to (3.48). The uniqueness of the discrete Kantorovitch potential $\boldsymbol{\phi}^n$ for all $n \ge 1$ is then provided by Theorem 3.5. □

Lemma 3.9 allows to define two functions $\rho_{\mathcal{T},\tau}$ and $\phi_{\mathcal{T},\tau}$ by setting

$$\rho_{\mathcal{T},\tau}(t,\boldsymbol{x}) = \rho_K^n, \quad \phi_{\mathcal{T},\tau}(t,\boldsymbol{x}) = \phi_K^n \quad \text{if } (t,\boldsymbol{x}) \in (t^{n-1},t^n]\times K\,.$$

It follows from the conservativity of the scheme and definition (3.49) of $\boldsymbol{\rho}^0$ that

$$\int_\Omega \rho_{\mathcal{T},\tau}(t^n,\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \langle \boldsymbol{\rho}^n, \mathbf{1}\rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^0, \mathbf{1}\rangle_{\mathcal{T}} = \int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x} > 0\,,$$

so that $\rho_{\mathcal{T},\tau}(t,\cdot)$ belongs to $\mathcal{P}(\Omega)$ for all $t \in (0,T)$.

The goal of this section is to prove the following theorem.

Theorem 3.10. *Assume that $\varrho^0 \ge \rho_\star$ for some $\rho_\star \in (0,+\infty)$ and that $\mathcal{E}(\varrho^0) < +\infty$, and let $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m}\right)_{m\ge 1}$ be a sequence of admissible discretizations of Ω such that $h_{\mathcal{T}_m}$ and τ_m tend to 0 while conditions (3.47) hold. Then up to a subsequence, $(\rho_{\mathcal{T}_m,\tau_m})_{m\ge1}$ tends in $L^1(Q_T)$ towards a weak solution $\varrho \in L^\infty((0,T);L^1(\Omega)) \cap L^2((0,T);W^{1,1}(\Omega))$ of (3.2) corresponding to the initial data ϱ^0.*

The proof is based on compactness arguments. At first in Section 3.3.1, we derive some a priori estimates on the discrete solution. These estimates will be used to obtain some compactness on $\rho_{\mathcal{T}_m,\tau_m}$ and $\phi_{\mathcal{T}_m,\tau_m}$ in Section 3.3.2. Finally, we identify the limit value as a weak solution in Section 3.3.3.

Remark 3.11. *We restrict our attention to the case of the linear Fokker-Planck equation for simplicity. The linearity of the continuous equation plays no role in our study. What is important is the fact that the discrete and continuous solutions are uniformly bounded away from 0 so that the weighted $\dot{H}^1_\rho$ norm controls the non-weighted $\dot{H}^1$ norm. Such a uniform lower bound can also be derived for the porous medium equation without drift.*

3.3.1 Some a priori estimates

First, let us show that if the continuous initial energy $\mathcal{E}(\varrho^0)$ is bounded, then so does its discrete counterpart $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^0)$.

Lemma 3.12. *Given $\varrho^0 \in \mathcal{P}(\Omega)$ such that $\mathcal{E}(\varrho^0) < +\infty$, and let $\boldsymbol{\rho}^0$ be defined by (3.49), then there exists C_1 depending only on Ω, V and ϱ^0 (but not on $\mathcal{T}$) such that $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) \le C_1$ for all $n \ge 0$.*

Proof. It follows from (3.27) that $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) \le \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^0)$ for all $n \ge 1$. Rewriting $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^0)$ as

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^0) = T_1 + T_2 + T_3 \tag{3.53}$$

with

$$T_1 = \sum_{K\in\mathcal{T}} m_K[\rho_K^0 \log \rho_K^0 - \rho_K^0]\,, \quad T_2 = \sum_{K\in\mathcal{T}} m_K \rho_K^0 V_K\,, \quad \text{and} \quad T_3 = \sum_{K\in\mathcal{T}} m_K e^{-V_K}\,,$$

we deduce from the definition (3.49) of $\boldsymbol{\rho}^0$ and Jensen's inequality that

$$T_1 \le \int_\Omega [\varrho^0 \log \varrho^0 - \varrho^0]\mathrm{d}\boldsymbol{x}\,. \tag{3.54}$$

Since V is continuous, there exists $\widetilde{\boldsymbol{x}}_K \in K$ such that $\int_K e^{-V}\mathrm{d}\boldsymbol{x} = m_K e^{-V(\widetilde{\boldsymbol{x}}_K)}$. Therefore,

$$T_3 = \int_\Omega e^{-V}\mathrm{d}\boldsymbol{x} + \sum_{K\in\mathcal{T}} m_K[e^{-V(\boldsymbol{x}_K)} - e^{-V(\widetilde{\boldsymbol{x}}_K)}] \le \int_\Omega e^{-V}\mathrm{d}\boldsymbol{x} + e^{\|V^-\|_\infty}\|\nabla V\|_\infty \mathrm{diam}(\Omega)\,. \tag{3.55}$$

Similarly, it follows from the mean value theorem that there exists $\check{\boldsymbol{x}}_K \in K$ such that $m_K V(\check{\boldsymbol{x}}_K)\rho_K^0 = \int_K \varrho^0 V\mathrm{d}\boldsymbol{x}$. Hence,

$$T_2 = \int_\Omega \varrho^0 V\mathrm{d}\boldsymbol{x} + \sum_{K\in\mathcal{T}} m_K \rho_K^0 [V(\boldsymbol{x}_K) - V(\check{\boldsymbol{x}}_K)] \le \int_\Omega \varrho^0 V\mathrm{d}\boldsymbol{x} + \|\nabla V\|_\infty \mathrm{diam}(\Omega)\int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x}\,. \tag{3.56}$$

Combining (3.54)–(3.56) in (3.53) shows that $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^0) \le \mathcal{E}(\varrho^0) + C$ for some C depending only on V, Ω and ϱ^0. □

Our next lemma shows that if ϱ^0 is bounded away from 0, then so does $\rho_{\mathcal{T},\tau}$.

Lemma 3.13. *Using the convention* $\log(0) = -\infty$*, one has*

$$\min_{K\in\mathcal{T}} [\log(\rho_K^n) + V_K] \ge \min_{K\in\mathcal{T}} \left[\log(\rho_K^{n-1}) + V_K\right], \quad \forall n \ge 1.$$

In particular, if $\varrho^0 \ge \rho_\star$ *for some* $\rho_\star \in (0, +\infty)$*, then there exists* $\alpha > 0$ *depending only on* V *and* $\rho_\star$ *(but not on* $\mathcal{T}, \tau$ *and* n*) such that* $\boldsymbol{\rho}^n \ge \alpha \mathbf{1}$ *for all* $n \ge 1$*.*

Proof. It follows directly from (3.51) that $\log(\rho_K^n) + V_K \ge \phi_K^n$ for all $K \in \mathcal{T}$. Let $K_\star \in \mathcal{T}$ be such that $\phi_{K_\star}^n \le \phi_K^n$ for all $K \in \mathcal{T}$, then the conservation equation (3.50) ensures that $\rho_{K_\star}^n \ge \rho_{K_\star}^{n-1}$. On the other hand, since

$$\sum_{\sigma\in\Sigma_{K_\star}} a_\sigma \left((\phi_{K_\star}^n - \phi_L^n)^+\right)^2 = 0\,,$$

the discrete HJ equation (3.51) provides that

$$\phi_{K^\star}^n = \log(\rho_{K_\star}^n) + V_{K_\star} = \min_{K\in\mathcal{T}} [\log(\rho_K^n) + V_K] \ge \log(\rho_{K_\star}^{n-1}) + V_{K_\star} \ge \min_{K\in\mathcal{T}} \left[\log(\rho_K^{n-1}) + V_K\right].$$

Assume now that $\varrho^0 \ge \rho_\star$, then for all $K \in \mathcal{T}$ and all $n \ge 0$,

$$\log(\rho_K^n) \ge \min_{L\in\mathcal{T}} [\log(\rho_L^0) + V_L] - V_K \ge \min_{L\in\mathcal{T}} \log(\rho_L^0) - 2\|V\|_\infty \ge \log(\rho_\star) - \|V^+\|_\infty - \|V^-\|_\infty\,.$$

Therefore, we obtain the desired inequality with $\alpha = \rho_\star e^{-\|V^+\|_\infty - \|V^-\|_\infty}$. □

Our third lemma deals with some estimates on the discrete gradient of the discrete Kantorovitch potentials $(\boldsymbol{\phi}^n)_n$.

Lemma 3.14. *Let $(\boldsymbol{\phi}^n, \boldsymbol{\rho}^n)$ be the iterated solution to (3.30), then*

$$\sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} a_\sigma \rho_\sigma^n (\phi_K^n - \phi_L^n)^2 \leq C_1 \,. \tag{3.57}$$

Moreover, if $\varrho^0 \geq \rho_\star \in (0, +\infty)$, then there exists C_2 (depending on Ω, V and ϱ^0) such that

$$\sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} a_\sigma (\phi_K^n - \phi_L^n)^2 \leq C_2 \,. \tag{3.58}$$

Proof. Since $\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) \geq 0$ for all $\boldsymbol{\rho} \in \mathbb{P}_\mathcal{T}$, summing (3.27) over $n \in \{1, \dots, N\}$ yields

$$\sum_{n=1}^{N} \frac{1}{\tau} \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) \leq \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^0) \,.$$

Thanks to (3.24), the left-hand side rewrites

$$\sum_{n=1}^{N} \frac{1}{\tau} \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) = \sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} a_\sigma \rho_\sigma^n (\phi_K^n - \phi_L^n)^2,$$

so that it only remains to use Lemma 3.12 to recover (3.57).

Finally, if ϱ^0 is bounded from below by some $\rho_\star > 0$, then Lemma 3.13 shows that $\rho_K^n \geq \alpha$ for some α depending only on $\rho_\star$ and V. Therefore, since ρ_σ^n is either equal to ρ_K^n or to ρ_L^n for $\sigma = K|L \in \Sigma$, then (3.58) holds with $C_2 = \frac{C_1}{\alpha}$. □

The discrete solution $\rho_{\mathcal{T},\tau}$ is piecewise constant on the cells. To study the convergence of the scheme, we also need a second reconstruction $\rho_{\Sigma,\tau}$ of the density corresponding to the edge mobilities. It is defined by

$$\rho_{\Sigma,\tau}(t, \boldsymbol{x}) = \begin{cases} \rho_\sigma^n & \text{if } (t, \boldsymbol{x}) \in (t^{n-1}, t^n] \times \Delta_\sigma \,, \quad \sigma \in \Sigma \,, \\ \rho_K^n & \text{if } (t, \boldsymbol{x}) \in (t^{n-1}, t^n] \times K \setminus \left(\bigcup_{\sigma\in\Sigma_K} \Delta_\sigma \right) \,, \quad K \in \mathcal{T} \,. \end{cases} \tag{3.59}$$

Lemma 3.15. *There exists C_3 depending only on ζ and ϱ^0 such that*

$$\int_\Omega \rho_{\Sigma,\tau}(t, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq C_3 \,, \quad \forall t > 0 \,. \tag{3.60}$$

Moreover, there exists C_4 depending only on ζ, V and ϱ^0 such that

$$\int_\Omega \rho_{\Sigma,\tau}(t, \boldsymbol{x}) \log \rho_{\Sigma,\tau}(t, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq C_4 \,, \quad \forall t > 0 \,. \tag{3.61}$$

Proof. Since $t \mapsto \rho_{\Sigma,\tau}(t,\cdot)$ is piecewise constant, it suffices to check that the above properties at each t^n, $1 \le n \le N$. In view of the definition of $\rho_{\Sigma,\tau}$, one has

$$\int_\Omega \rho_{\Sigma,\tau}(t^n,\boldsymbol{x})\mathrm{d}\boldsymbol{x} \le \sum_{K\in\mathcal{T}}\sum_{\sigma\in\Sigma_K} \rho_K^n m_K + \sum_{\sigma\in\Sigma}\rho_\sigma^n m_{\Delta_\sigma}\,.$$

The first term can easily be overestimated by $\int_\Omega \rho_{\mathcal{T},\tau}(t^n,\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int_\Omega \varrho^0\mathrm{d}\boldsymbol{x}$. Since $\rho_\sigma^n \le \rho_K^n + \rho_L^n$, the second term in the above expression can be overestimated by

$$\sum_{\sigma\in\Sigma}\rho_\sigma^n m_{\Delta_\sigma} \le \sum_{K\in\mathcal{T}}\rho_K^n\left(\sum_{\sigma\in\Sigma_K} m_{\Delta_\sigma}\right).$$

Using the regularity property of the mesh (3.47c), we obtain that

$$\sum_{\sigma\in\Sigma}\rho_\sigma^n m_{\Delta_\sigma} \le \zeta\int_\Omega \varrho^0\mathrm{d}\boldsymbol{x}\,,$$

so that (3.60) holds with $C_3 = (1+\zeta)\int_\Omega\varrho^0\mathrm{d}\boldsymbol{x}$.

Reproducing the above calculations, one gets that

$$\begin{aligned}\int_\Omega \rho_{\Sigma,\tau}(t,\boldsymbol{x})\log\rho_{\Sigma,\tau}(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x} &\le (1+\zeta)\int_\Omega \rho_{\mathcal{T},\tau}(t,\boldsymbol{x})\log\rho_{\mathcal{T},\tau}(t,\boldsymbol{x})\mathrm{d}\boldsymbol{x}\\ &= (1+\zeta)\Big(\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^n) + \sum_{K\in\mathcal{T}} m_K[\rho_K^n(1-V_K) - e^{-V_K}]\Big).\end{aligned}$$

Since $\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^n) \le \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^0) \le C_1$ and since V is uniformly bounded, we obtain that (3.61) holds with $C_4 = (1+\zeta)\,(C_1 + \|(1-V)^+\|_\infty)$. □

The last lemma of this section can be thought as a discrete $\big(L^\infty((0,T);W^{1,\infty}(\Omega))\big)'$ estimate on $\partial_t\rho_{\mathcal{T},\tau}$. This estimate will be used to apply a discrete nonlinear Aubin-Simon lemma [6] in the next section.

Lemma 3.16. *Let $\varphi \in C_c^\infty(Q_T)$, then define $\varphi_K^n = \frac{1}{m_K}\int_K \varphi(t^n,\boldsymbol{x})\mathrm{d}\boldsymbol{x}$ for all $K\in\mathcal{T}$. There exists C_5 depending only on ζ, T, ϱ^0, d, such that*

$$\sum_{n=1}^N\sum_{K\in\mathcal{T}} m_K(\rho_K^n - \rho_K^{n-1})\varphi_K \le C_5\|\nabla\varphi\|_{L^\infty(Q_T)}\,.$$

Proof. Multiplying (3.50) by φ_K^n and summing over $K\in\mathcal{T}$ and $n\in\{1,\dots,N\}$ yields

$$A := \sum_{n=1}^N\sum_{K\in\mathcal{T}} m_K(\rho_K^n-\rho_K^{n-1})\varphi_K = -\sum_{n=1}^N\tau\sum_{\sigma\in\Sigma} a_\sigma\rho_\sigma^n(\phi_K^n-\phi_L^n)(\varphi_K^n-\varphi_L^n)\,.$$

Applying Cauchy-Schwarz inequality on the right-hand side then provides

$$A^2 \le \left(\sum_{n=1}^N\tau\sum_{\sigma\in\Sigma} a_\sigma\rho_\sigma^n(\phi_K^n-\phi_L^n)^2\right)\left(\sum_{n=1}^N\tau\sum_{\sigma\in\Sigma} a_\sigma\rho_\sigma^n(\varphi_K^n-\varphi_L^n)^2\right). \tag{3.62}$$

The first term in the right-hand side is bounded thanks to Lemma 3.14. On the other hand, the regularity of φ ensures that there exists $\widetilde{\boldsymbol{x}}_K \in K$ such that $\varphi(t^n, \boldsymbol{x}_K) = \varphi_K^n$ for all $K \in \mathcal{T}$. Thanks to the regularity assumptions (3.47a)–(3.47b) on the mesh, there holds

$$|\varphi_K^n - \varphi_L^n| \le \|\nabla\varphi\|_\infty |\widetilde{\boldsymbol{x}}_K - \widetilde{\boldsymbol{x}}_L| \le (1 + 2\zeta(1+\zeta))\|\nabla\varphi\|_\infty d_\sigma\,, \quad \sigma = K|L\,.$$

Hence, the second term of the right-hand side in (3.62) can be overestimated by

$$\begin{aligned}\sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n (\varphi_K^n - \varphi_L^n)^2 &\le (1+2\zeta(1+\zeta))^2 \|\nabla\varphi\|_\infty^2 \sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} m_\sigma d_\sigma \rho_\sigma^n \\ &\le (1+2\zeta(1+\zeta))^2 d \|\nabla\varphi\|_\infty^2 \iint_{Q_T} \rho_{\Sigma,\tau} \mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\le (1+2\zeta(1+\zeta))^2 C_3 T d \|\nabla\varphi\|_\infty^2\,,\end{aligned}$$

the last inequality being a consequence of Lemma 3.15. Combining all this material in (3.62) shows the desired estimate with $C_5 = (1+2\zeta(1+\zeta))\sqrt{C_1 C_3 T d}$. □

3.3.2 Compactness of the approximate solution

The goal of this section is to show enough compactness in order to be able to pass to the limit $m \to \infty$. For the sake of readability, we remove the subscript m unless necessary.

Owing to Lemma 3.12, one has $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) \le C_1$ for all $n \in \{1, \dots, N\}$. Proceeding as in the proof of Lemma 3.15, this allows to show that

$$\int_\Omega \rho_{\mathcal{T},\tau}(t,\boldsymbol{x}) \log \rho_{\mathcal{T},\tau}(t,\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \le C_6\,, \quad \forall t \in (0,T]\,, \tag{3.63}$$

for some C_6 depending only on ϱ^0, ζ and V. Combining de La Vallée Poussin's theorem with Dunford-Pettis' one [122, Ch. XI, Theorem 3.6], there exists $\varrho \in L^\infty((0,T); L^1(\Omega))$ such that, up to a subsequence,

$$\rho_{\mathcal{T}_m,\tau_m} \text{ tends to } \varrho \text{ weakly in } L^1(Q_T) \text{ as } m \text{ tends to } +\infty\,. \tag{3.64}$$

Since $\rho \mapsto \rho\log\rho$ is convex, $f \mapsto \iint_{Q_T} f\log f \mathrm{d}\boldsymbol{x}\mathrm{d}t$ is l.s.c. for the weak convergence in $L^1(Q_T)$ (see for instance [28, Corollary 3.9]), so that (3.63) yields

$$\iint_{Q_T} \varrho \log \varrho \mathrm{d}\boldsymbol{x}\mathrm{d}t \le C_6 T\,. \tag{3.65}$$

Moreover, since $\rho_{\mathcal{T},\tau} \ge \alpha$ thanks to Lemma 3.13, then $\varrho \ge \alpha$ too.

Our goal is to show that ϱ is a weak solution to the Fokker-Planck equation (3.2) corresponding to the initial data ϱ^0. Even though the continuous problem is linear, (3.64) is not enough to pass to the limit in our nonlinear scheme. Refined compactness have to be derived in this section so that one can identify ϱ as the solution to (3.2) in the next section. To show enhanced compactness (and most of all the consistency of the scheme in the next section), we have to assume that the initial data is bounded away from 0.

Proposition 3.17. *Assume that $\varrho^0 \geq \rho_\star \in (0,+\infty)$, then, up to a subsequence,*

$$\rho_{\mathcal{T}_m,\tau_m} \underset{m\to\infty}{\longrightarrow} \varrho \qquad \text{strongly in } L^1(Q_T), \tag{3.66}$$

$$\log \rho_{\mathcal{T}_m,\tau_m} \underset{m\to\infty}{\longrightarrow} \log \varrho \qquad \text{strongly in } L^1(Q_T), \tag{3.67}$$

$$\phi_{\mathcal{T}_m,\tau_m} \underset{m\to\infty}{\longrightarrow} \log \varrho + V \qquad \text{strongly in } L^1(Q_T). \tag{3.68}$$

Proof. Our proof of (3.66)–(3.67) relies on ideas introduced in [101] that were adapted to the discrete setting in [6]. Define the two convex and increasing conjugated functions defined on $\mathbb{R}_+$:

$$\Upsilon : x \mapsto e^x - x - 1 \quad \text{and} \quad \Upsilon^* : y \mapsto (1+y)\log(1+y) - y\,,$$

then the following inequality holds for any measurable functions $f, g : Q_T \to \mathbb{R}$:

$$\iint_{Q_T} |fg| \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \iint_{Q_T} \Upsilon(|f|)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \iint_{Q_T} \Upsilon^*(|g|)\mathrm{d}\boldsymbol{x}\mathrm{d}t\,. \tag{3.69}$$

Now, notice that since $\rho_{\mathcal{T},\tau}$ is bounded from below thanks to Lemma 3.13 and bounded in $L^1(Q_T)$, then $\log \rho_{\mathcal{T},\tau}$ is bounded in $L^p(Q_T)$ for all $p \in [1,\infty)$ and $\Upsilon(|\log(\rho_{\mathcal{T},\tau})|)$ is bounded in $L^1(Q_T)$. As a consequence, there exists $\ell \in L^\infty((0,T);L^p(\Omega))$ such that

$$\log \rho_{\mathcal{T}_m,\tau_m} \underset{m\to\infty}{\longrightarrow} \ell \quad \text{weakly in } L^1(Q_T)\,. \tag{3.70}$$

Since $f \mapsto \iint_{Q_T} \Upsilon(|f|)$ is convex thus l.s.c. for the weak convergence, we infer that $\Upsilon(|\ell|)$ belongs to $L^1(Q_T)$. Moreover, in view of (3.65), $\Upsilon^*(\varrho)$ belongs also to $L^1(Q_T)$. Therefore, thanks to (3.69), the function $\varrho\ell$ is in $L^1(Q_T)$.

Define the quantities

$$r_K^n = \frac{\tau}{2m_K} a_\sigma \sum_{\sigma\in\Sigma_K} \left((\phi_K^n - \phi_L^n)^+\right)^2 \geq 0,\ \forall K \in \mathcal{T},\ \forall n \in \{1,\dots,N\}\,,$$

and by $r_{\mathcal{T},\tau} \in L^1(Q_T)$ the function defined

$$r_{\mathcal{T},\tau}(t,\boldsymbol{x}) = r_K^n \quad \text{if } (t,\boldsymbol{x}) \in (t^{n-1},t^n]\times K\,,$$

Thanks to Lemma 3.14, $\|r_{\mathcal{T},\tau}\|_{L^1(Q_T)} \leq \frac{1}{2}C_2\tau$. As a consequence, $r_{\mathcal{T}_m,\tau_m}$ tends to 0 in $L^1(Q_T)$ as m tends to $+\infty$.

Let $\boldsymbol{\xi} \in \mathbb{R}^d$ be arbitrary, we denote by $\Omega_{\boldsymbol{\xi}} = \{\boldsymbol{x} \in \Omega \mid \boldsymbol{x}+\boldsymbol{\xi} \in \Omega\}$. Then using (3.51) and the triangle inequality, we obtain that for all $m \geq 1$, there holds

$$\int_0^T \int_{\Omega_{\boldsymbol{\xi}}} |\log \rho_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x}+\boldsymbol{\xi}) - \log \rho_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x})|\,\mathrm{d}\boldsymbol{x}\mathrm{d}t \leq A_{1,m}(\boldsymbol{\xi}) + A_{2,m}(\boldsymbol{\xi}) + A_{3,m}(\boldsymbol{\xi})\,,$$

where, denoting by $V_{\mathcal{T}}(\boldsymbol{x}) = V_K$ if $\boldsymbol{x} \in K$, we have set

$$A_{1,m}(\boldsymbol{\xi}) = \int_0^T \int_{\Omega_{\boldsymbol{\xi}}} |r_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x}+\boldsymbol{\xi}) - r_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\mathrm{d}t\,,$$

$$A_{2,m}(\boldsymbol{\xi}) = \int_0^T \int_{\Omega_{\boldsymbol{\xi}}} |\phi_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x}+\boldsymbol{\xi}) - \phi_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\mathrm{d}t\,,$$

$$A_{3,m}(\boldsymbol{\xi}) = T\int_{\Omega_{\boldsymbol{\xi}}} |V_{\mathcal{T}_m}(\boldsymbol{x}+\boldsymbol{\xi}) - V_{\mathcal{T}_m}(\boldsymbol{x})|\mathrm{d}\boldsymbol{x}\,.$$

Since $(r_{\mathcal{T}_m,\tau_m})_{m\geq 1}$ and $(V_{\mathcal{T}_m})_{m\geq 1}$ are compact in $L^1(Q_T)$ and $L^1(\Omega)$ respectively, it follows from the Riesz-Frechet-Kolmogorov theorem (see for instance [28, Exercise 4.34]) that there exists $\omega \in C(\mathbb{R}_+;\mathbb{R}_+)$ with $\omega(0)=0$ such that

$$A_{1,m}(\boldsymbol{\xi}) + A_{3,m}(\boldsymbol{\xi}) \leq \omega(|\boldsymbol{\xi}|), \quad \forall \boldsymbol{\xi}\in\mathbb{R}^d,\ \forall m\geq 0\,. \tag{3.71}$$

On the other hand, the function $\phi_{\mathcal{T},\tau}$ belongs to $L^1((0,T);BV(\Omega))$ and the integral in time of its total variation in space can be estimated as follows:

$$\begin{aligned}\iint_{Q_T} |\nabla \phi_{\mathcal{T}_m,\tau_m}| &= \sum_{n=1}^N \tau \sum_{\sigma\in\Sigma} m_\sigma |\phi_K^n - \phi_L^n| \\ &\leq \left(d|\Omega| T \sum_{n=1}^N \tau \sum_{\sigma\in\Sigma} m_\sigma (\phi_K^n - \phi_L^n)^2 \right)^{1/2} \leq C_7\,.\end{aligned}$$

with $C_7 = \sqrt{d|\Omega|TC_2}$. This implies in particular that $A_{2,m}(\boldsymbol{\xi}) \leq C_7|\boldsymbol{\xi}|$ for all $m\geq 1$. Combining this estimate with (3.71) in (3.51) yields

$$\sup_{m\geq 1} \int_0^T \int_{\Omega_\xi} |\log \rho_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x}+\boldsymbol{\xi}) - \log \rho_{\mathcal{T}_m,\tau_m}(t,\boldsymbol{x})| \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{|\boldsymbol{\xi}|\to 0}{\longrightarrow} 0\,. \tag{3.72}$$

The combination of (3.72) with Lemma 3.16 is exactly what one needs to reproduce the proof of [6, Proposition 3.8], which shows that the product of the weakly convergent sequences $(\rho_{\mathcal{T}_m,\tau_m})_m$ and $(\log \rho_{\mathcal{T}_m,\tau_m})_m$ converges towards the product of their weak limits:

$$\iint_{Q_T} \rho_{\mathcal{T}_m,\tau_m} \log \rho_{\mathcal{T}_m,\tau_m} \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} \varrho\ell\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t\,, \quad \forall \varphi\in C_c^\infty(Q_T)\,. \tag{3.73}$$

Let us now identify ℓ as $\log(\varrho)$ thanks to Minty's trick. Let $\kappa>0$ and $\varphi\in C_c^\infty(Q_T;\mathbb{R}_+)$ be arbitrary, then thanks to (3.73),

$$0 \leq \iint_{Q_T} (\rho_{\mathcal{T}_m,\tau_m} - \kappa)(\log \rho_{\mathcal{T}_m,\tau_m} - \log\kappa)\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} (\varrho-\kappa)(\ell-\log\kappa)\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t\,.$$

As a consequence, $(\varrho-\kappa)(\ell-\log\kappa)\geq 0$ a.e. in Q_T for all $\kappa>0$, which holds if and only if $\ell=\log\varrho$. To finalize the proof of (3.66)–(3.67), define

$$c_m = (\rho_{\mathcal{T}_m,\tau_m} - \varrho)(\log \rho_{\mathcal{T}_m,\tau_m} - \log\varrho) \in L^1(Q_T;\mathbb{R}_+)\,, \quad \forall m\geq 1\,.$$

Then (3.73) implies that

$$\iint_{Q_T} c_m \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{m\to\infty}{\longrightarrow} 0, \quad \forall \varphi\in C_c^\infty(Q_T),\ \varphi\geq 0\,.$$

As a consequence, c_m tends to 0 almost everywhere in Q_T, which implies that $\rho_{\mathcal{T}_m,\tau_m}$ tends almost everywhere towards ϱ (up to a subsequence). Then (3.66)–(3.67) follow from Vitali's convergence theorem (see for instance [122, Chap. XI, Theorem 3.9]).

Finally, one has $\phi_{\mathcal{T},\tau} = \log\rho_{\mathcal{T},\tau} + V_{\mathcal{T}} - r_{\mathcal{T},\tau}$. In view of the above discussion, the right-hand side converges strongly in $L^1(Q_T)$ up to a subsequence towards $\log\varrho + V$, then so does the left-hand side. This provides (3.68) and concludes the proof of Proposition 3.17. □

Next lemma shows that $\rho_{\Sigma,\tau}$ shares the same limit ϱ as $\rho_{\mathcal{T},\tau}$.

Lemma 3.18. *Assume that $\varrho^0 \geq \rho_\star \in (0,+\infty)$, then*

$$\|\rho_{\Sigma_m,\tau_m} - \rho_{\mathcal{T}_m,\tau_m}\|_{L^1(Q_T)} \underset{m\to\infty}{\longrightarrow} 0\,.$$

Proof. Thanks to Lemma 3.15, it follows from the de La Vallée-Poussin and Dunford Pettis theorems that $(\rho_{\Sigma_m,\tau_m})_{m\geq 1}$ is relatively compact for the weak topology of $L^1(Q_T)$. Combining this with (3.64), we infer that, up to a subsequence, $(\rho_{\Sigma_m,\tau_m} - \rho_{\mathcal{T}_m,\tau_m})_{m\geq 1}$ converges towards some w weakly in $L^1(Q_T)$. Thanks to Vitali's convergence theorem, it suffices to show that from any subsequence of $(\rho_{\Sigma_m,\tau_m} - \rho_{\mathcal{T}_m,\tau_m})_{m\geq 1}$, one can extract a subsequence that tends to 0 a.e. in Q_T (so that the whole sequence converges towards $w = 0$), or equivalently

$$\|\log \rho_{\Sigma_m,\tau_m} - \log \rho_{\mathcal{T}_m,\tau_m}\|_{L^1(Q_T)} \underset{m\to\infty}{\longrightarrow} 0\,, \tag{3.74}$$

since both $(\rho_{\Sigma_m,\tau_m})_{m\geq 1}$ and $(\rho_{\mathcal{T}_m,\tau_m})_{m\geq 1}$ are bounded away from 0 thanks to Lemma 3.13. Bearing in mind the definition (3.59) of ρ_{Σ_m,τ_m}, one has

$$\|\log \rho_{\Sigma,\tau} - \log \rho_{\mathcal{T},\tau}\|_{L^1(Q_T)} \leq \sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} m_{\Delta_\sigma} |\log \rho_K^n - \log \rho_L^n|\,.$$

Using (3.51) and the triangle inequality, one gets that

$$\|\log \rho_{\Sigma,\tau} - \log \rho_{\mathcal{T},\tau}\|_{L^1(Q_T)} \leq R_1 + R_2 + TR_3\,,$$

with

$$R_1 = \sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} m_{\Delta_\sigma} |\phi_K^n - \phi_L^n|\,, \qquad R_2 = \sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} m_{\Delta_\sigma} |r_K^n - r_L^n|\,,$$

and

$$R_3 = \sum_{\sigma\in\Sigma} m_{\Delta_\sigma} |V_K - V_L|\,.$$

Using again that $dm_{\Delta_\sigma} = d_\sigma m_\sigma \leq \zeta h_{\mathcal{T}} m_\sigma$ thanks to (3.47a), one has

$$R_1 \leq \frac{\zeta}{d} h_{\mathcal{T}} \sum_{n=1}^{N} \tau \sum_{\sigma\in\Sigma} m_\sigma |\phi_K^n - \phi_L^n| \leq \frac{C_7\zeta}{d} h_{\mathcal{T}} \underset{m\to\infty}{\longrightarrow} 0\,.$$

Since $|r_K^n - r_L^n| \leq r_K^n + r_L^n$, the regularity assumption (3.47c) on the mesh implies that

$$R_2 \leq \sum_{n=1}^{N} \tau \sum_{K\in\mathcal{T}} \sum_{\sigma\in\Sigma_K} m_{\Delta_\sigma} r_K^n \leq \zeta \|r_{\mathcal{T},\tau}\|_{L^1(Q_T)} \underset{m\to\infty}{\longrightarrow} 0\,.$$

Since V is Lipschitz continuous, $|V_K - V_L| \leq \|\nabla V\|_\infty d_\sigma \leq \zeta \|\nabla V\|_\infty h_{\mathcal{T}}$ for all $\sigma = K|L \in \Sigma$ thanks to (3.47a). Therefore,

$$R_3 \leq \zeta \|\nabla V\|_\infty |\Omega| h_{\mathcal{T}} \underset{m\to\infty}{\longrightarrow} 0\,,$$

so that (3.74) holds, concluding the proof of Lemma 3.18. □

3.3.3 Convergence towards a weak solution

Our next lemma is an important step towards the identification of the limit ϱ as a weak solution to the continuous Fokker-Planck equation (3.2). Define the vector field $\boldsymbol{F}_{\Sigma,\tau} : Q_T \to \mathbb{R}^d$ by

$$\boldsymbol{F}_{\Sigma,\tau}(t,\boldsymbol{x}) = \begin{cases} d\rho_\sigma^n \big(\frac{\phi_K^n - \phi_L^n}{d_\sigma}\big)\boldsymbol{n}_{K\sigma} & \text{if } (t,\boldsymbol{x}) \in (t^{n-1},t^n] \times \Delta_\sigma\,, \\ 0 & \text{otherwise}\,. \end{cases}$$

Lemma 3.19. *Assume that $\varrho^0 \geq \rho_\star \in (0,+\infty)$, then, up to a subsequence, the vector field $\boldsymbol{F}_{\Sigma_m,\tau_m}$ converges weakly in $L^1(Q_T)^d$ towards $-\nabla\varrho - \varrho\nabla V$ as m tends to $+\infty$. Moreover, $\sqrt{\varrho}$ belongs to $L^2((0,T);H^1(\Omega))$, while ϱ belongs to $L^2((0,T);W^{1,1}(\Omega))$.*

Proof. Let us introduce the inflated discrete gradient $\boldsymbol{G}_{\Sigma,\tau}$ of $\phi_{\mathcal{T},\tau}$ defined by

$$\boldsymbol{G}_{\Sigma,\tau}(t,\boldsymbol{x}) = \begin{cases} d\big(\frac{\phi_L^n - \phi_K^n}{d_\sigma}\big)\boldsymbol{n}_{K\sigma} & \text{if } (t,\boldsymbol{x}) \in (t^{n-1},t^n] \times \Delta_\sigma\,, \\ 0 & \text{otherwise}\,, \end{cases}$$

so that $\boldsymbol{F}_{\Sigma,\tau} = -\rho_{\Sigma,\tau}\boldsymbol{G}_{\Sigma,\tau}$. Thanks to Lemma 3.14,

$$\|\boldsymbol{G}_{\Sigma,\tau}\|^2_{L^2(Q_T)^d} = d\sum_{n=1}^N \tau \sum_{\sigma\in\Sigma} a_\sigma(\phi_K^n - \phi_L^n)^2 \leq dC_2\,,$$

thus we know that, up to a subsequence, $\boldsymbol{G}_{\Sigma,\tau}$ converges weakly towards some $\boldsymbol{G}$ in $L^2(Q_T)^d$ as m tends to $+\infty$. Since $\phi_{\mathcal{T},\tau}$ tends to $\log\varrho + V$, cf. (3.68), then the weak consistency of the inflated gradient [44, 55] implies that $\boldsymbol{G} = \nabla(\log\varrho + V)$.

Define now $\boldsymbol{H}_{\Sigma,\tau} = \sqrt{\rho_{\Sigma,\tau}}\boldsymbol{G}_{\Sigma,\tau}$, then using again Lemma 3.14,

$$\|\boldsymbol{H}_{\Sigma,\tau}\|^2_{L^2(Q_T)^d} = d\sum_{n=1}^N \tau \sum_{\sigma\in\Sigma} a_\sigma\rho_\sigma^n(\phi_K^n - \phi_L^n)^2 \leq dC_1\,,$$

so that there exists $\boldsymbol{H} \in L^2(Q_T)^d$ such that, up to a subsequence, $\boldsymbol{H}_{\Sigma,\tau}$ tends to $\boldsymbol{H}$ weakly in $L^2(Q_T)^d$. But since $\sqrt{\rho_{\Sigma,\tau}}$ converges strongly towards $\sqrt{\varrho}$ in $L^2(Q_T)$, cf. Lemma 3.15, and since $\boldsymbol{G}_{\Sigma,\tau}$ tends weakly towards $\nabla(\log\varrho + V)$ in $L^2(Q_T)^d$, we deduce that $\boldsymbol{H}_{\Sigma,\tau}$ tends weakly in $L^1(Q_T)^d$ towards $\sqrt{\varrho}\nabla(\log\varrho + V) = 2\nabla\sqrt{\varrho} + \sqrt{\varrho}\nabla V = \boldsymbol{H}$. In particular, $\sqrt{\varrho}$ belongs to $L^2((0,T);H^1(\Omega))$. Now, we can pass in the limit $m \to +\infty$ in $\boldsymbol{F}_{\Sigma,\tau} = -\sqrt{\rho_{\Sigma,\tau}}\boldsymbol{H}_{\Sigma,\tau}$, leading to the desired result. □

In order to conclude the proof of Theorem 3.10, it remains to check that any limit value ϱ of the scheme is a solution to the Fokker-Planck equation (3.2) in the distributional sense.

Proposition 3.20. *Let ϱ be a limit value of $(\rho_{\mathcal{T}_m,\tau_m})_{m\geq 1}$ as described in Section 3.3.2, then for all $\varphi \in C_c^\infty([0,T)\times\Omega)$, one has*

$$\iint_{Q_T} \varrho\partial_t\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega \varrho^0\varphi(0,\cdot)\mathrm{d}\boldsymbol{x} - \iint_{Q_T} (\varrho\nabla V + \nabla\varrho)\cdot\nabla\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0\,. \tag{3.75}$$

Proof. Given $\varphi \in C_c^\infty([0,T)\times\Omega)$, we denote by $\varphi_K^n = \varphi(t^n, \boldsymbol{x}_K)$. Then multipying (3.50) by $-\varphi_K^{n-1}$ and summing over $K\in\mathcal{T}$ and $n\in\{1,\dots,N\}$ leads to

$$B_1 + B_2 + B_3 = 0\,,$$

where we have set

$$B_1 = \sum_{n=1}^{N}\tau\sum_{K\in\mathcal{T}} m_K \frac{\varphi_K^n-\varphi_K^{n-1}}{\tau}\rho_K^n\,, \qquad B_2 = \sum_{K\in\mathcal{T}} m_K\varphi_K^0\rho_K^0\,,$$

and

$$B_3 = -\sum_{n=1}^{N}\tau\sum_{\sigma=K|L\in\Sigma} a_\sigma\rho_\sigma^n\left(\phi_K^n-\phi_L^n\right)\left(\varphi_K^{n-1}-\varphi_L^{n-1}\right)\,.$$

Since $\rho_{\mathcal{T},\tau}$ converges in $L^1(Q_T)$ towards ϱ, cf. Proposition 3.17, and since φ is smooth,

$$B_1 \underset{m\to\infty}{\longrightarrow} \iint_{Q_T}\varrho\partial_t\varphi\mathrm{d}\boldsymbol{x}\mathrm{d}t\,.$$

It follows from the definition (3.49) of ρ_K^0 that the piecewise constant function $\rho_{\mathcal{T}}^0$, defined by $\rho_{\mathcal{T}}^0(\boldsymbol{x}) = \rho_K^0$ if $\boldsymbol{x}\in\mathcal{T}$, converges in $L^1(\Omega)$ towards ϱ^0. Therefore, since φ is smooth,

$$B_2 \underset{m\to\infty}{\longrightarrow} \int_\Omega \varrho^0\varphi(0,\cdot)\mathrm{d}\boldsymbol{x}\,.$$

Let us define

$$B_3' = \iint_{Q_T}\boldsymbol{F}_{\Sigma,\tau}\cdot\nabla\varphi\mathrm{d}\boldsymbol{x}\mathrm{d}t\,.$$

Then it follows from Lemma 3.19 that

$$B_3' \underset{m\to\infty}{\longrightarrow} -\iint_{Q_T}(\varrho\nabla V+\nabla\varrho)\cdot\nabla\varphi\mathrm{d}\boldsymbol{x}\mathrm{d}t\,.$$

To conclude the proof of Proposition 3.20, it only remains to check that

$$\left|B_3-B_3'\right| \le \sum_{n=1}^{N}\tau\sum_{\sigma=K|L\in\Sigma} a_\sigma\rho_\sigma^n\left|\phi_K^n-\phi_L^n\right|\left|\varphi_K^{n-1}-\varphi_L^{n-1}+\frac{1}{\tau m_{\Delta_\sigma}}\int_{t^{n-1}}^{t^n}\int_{\Delta_\sigma} d_\sigma\nabla\varphi\cdot\boldsymbol{n}_{KL}\mathrm{d}\boldsymbol{x}\mathrm{d}t\right|.$$

Since φ is smooth and since $d_\sigma\boldsymbol{n}_{KL} = \boldsymbol{x}_K-\boldsymbol{x}_L$ thanks to the orthogonality condition satisfied by the mesh,

$$\left|\varphi_K^{n-1}-\varphi_L^{n-1}+\frac{1}{\tau m_{\Delta_\sigma}}\int_{t^{n-1}}^{t^n}\int_{\Delta_\sigma} d_\sigma\nabla\varphi\cdot\boldsymbol{n}_{KL}\mathrm{d}\boldsymbol{x}\mathrm{d}t\right| \le C_\varphi d_\sigma(\tau+d_\sigma)$$

for some C_φ depending only on φ. Therefore,

$$\left|B_3-B_3'\right| \le C_\varphi(\tau+d_\sigma)\sum_{n=1}^{N}\tau\sum_{\sigma\in\Sigma} m_\sigma\rho_\sigma^n\left|\phi_K^n-\phi_L^n\right|\,.$$

Applying Cauchy-Schwarz inequality, one gets that

$$\left|B_3-B_3'\right| \le C_\varphi(\tau+d_\sigma)C_1 d\left\|\rho_{\Sigma,\tau}\right\|_{L^1(Q_T)} \underset{m\to\infty}{\longrightarrow} 0$$

thanks to Lemma 3.15. □

3.4 Numerical results

To check the correctness and reliability of our formulation we performed some numerical tests. Before that, we are going to present some details on the solution of the nonlinear system involved in the scheme.

3.4.1 Newton method

Due to the explicit formulation of the optimality conditions of the saddle point problem (3.29), it appears extremely convenient to use a Newton method for their solution. Given $\boldsymbol{u}^{n-1} = (\boldsymbol{\phi}^{n-1}, \boldsymbol{\rho}^{n-1}) \in \mathbb{R}^{2\mathcal{T}}$ solution of the scheme at the time step $n-1$, the Newton method aims at constructing a sequence of approximations of $\boldsymbol{u}^n$ as $\boldsymbol{u}^{n,k+1} = \boldsymbol{u}^{n,k} - \boldsymbol{d}^k$, $\boldsymbol{d}^k = (\boldsymbol{d}^k_{\boldsymbol{\phi}}, \boldsymbol{d}^k_{\boldsymbol{\rho}})$ being the Newton direction, solution to the block-structured system of equations

$$\boldsymbol{J}^k \boldsymbol{d}^k = \begin{bmatrix} \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}} & \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \\ \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}} & \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}} \end{bmatrix} \begin{bmatrix} \boldsymbol{d}^k_{\boldsymbol{\phi}} \\ \boldsymbol{d}^k_{\boldsymbol{\rho}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}^k_{\boldsymbol{\phi}} \\ \boldsymbol{f}^k_{\boldsymbol{\rho}} \end{bmatrix}. \tag{3.76}$$

In the above linear system, $\boldsymbol{f}^k_{\boldsymbol{\phi}}$ and $\boldsymbol{f}^k_{\boldsymbol{\rho}}$ are the discrete continuity and HJ equations evaluated in $\boldsymbol{u}^{n,k}$, and $\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}}$, $\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}}$, $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}$ and $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ are the four blocks of the Hessian matrix $\boldsymbol{J}^k$ of the discrete functional in (3.29) evaluated in $\boldsymbol{u}^{n,k}$. The sequence converges to the unique solution $\boldsymbol{u}^n$ as soon as the initial guess is sufficiently close to it, which is ensured for a sufficiently small time step by taking $\boldsymbol{u}^{n,0} = \boldsymbol{u}^{n-1}$. The algorithm stops when the ℓ^∞ norm of the discrete equations is smaller than a prescribed tolerance or if the maximum number of iterations is reached. It is possible to implement an adaptative time stepping: if the Newton method converges in few iterations the time step τ increases; if it reaches the maximum number of iterations the time step is decreased and the method restarted. Issues could arise if the iterate $\boldsymbol{u}^{n,k}$ reaches negative values, especially if the energy is not defined for negative densities. Two possible strategies may be implemented to avoid this problem: the iterate may be projected on the set of positive measure by taking $\boldsymbol{u}^{n,k} = (\boldsymbol{u}^{n,k})^+$; the method may be restarted with a smaller time step.

In case of a local energy functional, as it is the case for the Fokker-Planck energy and many more examples, the block $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ is diagonal and therefore straightforward to invert. System (3.76) can be rewritten in term of the Schur complement and solved for $\boldsymbol{d}^k_{\boldsymbol{\phi}}$ as

$$\left[\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}} - \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}\right)^{-1} \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}\right] \boldsymbol{d}^k_{\boldsymbol{\phi}} = \boldsymbol{f}^k_{\boldsymbol{\phi}} - \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}\right)^{-1} \boldsymbol{f}^k_{\boldsymbol{\rho}}, \tag{3.77}$$

while $\boldsymbol{d}^k_{\boldsymbol{\rho}} = \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}\right)^{-1} \left(\boldsymbol{f}^k_{\boldsymbol{\rho}} - \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}\, \boldsymbol{d}^k_{\boldsymbol{\phi}}\right)$.

Proposition 3.21. *The Schur complement $\boldsymbol{S}^k = \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}} - \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}\right)^{-1} \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}$ is symmetric and negative definite.*

Proof. $\boldsymbol{S}^k$ is symmetric since $\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}}$ and $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ are, while $\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} = \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}\right)^T$. The matrix $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ is positive definite since the problem is strictly convex, whereas $\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}}$ is negative definite if $\rho^{n,k}_K > 0, \forall K \in \mathcal{T}$, since the problem is strictly concave, but it is semi-negative definite if the density vanishes somewhere. Therefore, it is sufficient to show that the matrix $\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} = \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}\right)^T = \boldsymbol{M} + \boldsymbol{A}^k$ is invertible. $\boldsymbol{M}$ is a diagonal matrix such that $(\boldsymbol{M})_{K,K} = m_K$, whereas

$$(\boldsymbol{A}^k)_{K,K} = \tau \sum_{\sigma \in \Sigma_K} a_\sigma (\phi^{n,k}_K - \phi^{n,k}_L)^+ \geq 0\,,$$

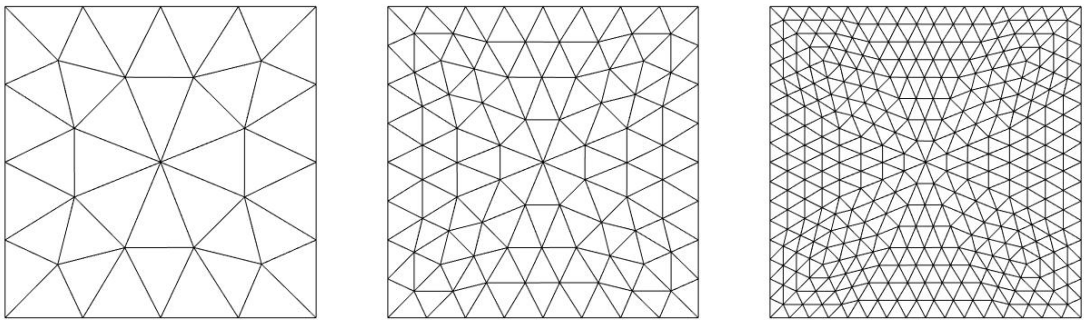


Figure 3.1: Sequence of regular triangular meshes.

and, for $L \neq K$,

$$(\boldsymbol{A}^k)_{K,L} = -\tau a_\sigma (\phi_L^{n,k} - \phi_K^n)^+ \leq 0 \quad \text{if } \sigma = K|L, \quad (\boldsymbol{A}^k)_{K,L} = 0 \quad \text{otherwise.}$$

Therefore the columns of $\boldsymbol{A}^k$ sum up to 0, so that $(\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}})$ is a column M-matrix [59] and thus invertible. □

In case the matrix $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ is simple to invert it is then possible to decrease the computational complexity of the solution of system (3.76). Moreover, it is possible to exploit for the solution of system (3.77) solvers which are computationally more efficient, since the system is symmetric and negative definite.

3.4.2 Fokker-Planck equation

We first tackle the gradient flow of the Fokker-Planck energy, namely equation (3.2). In section 3.3 we showed the L^1 convergence of the scheme. Consider the specific potential $V(\mathbf{x}) = -gx$: for this case it is possible to design an analytical solution and test numerically the convergence of the scheme. Consider the domain $\Omega = [0,1]^2$, the time interval $[0, 0.25]$ and the following analytical solution of the Fokker-Planck equation (built from a one-dimensional one):

$$\varrho(t,x,y) = \exp\left(-\alpha t + \frac{g}{2}x\right)\left(\pi\cos(\pi x) + \frac{g}{2}\sin(\pi x)\right) + \pi\exp\left(g\left(x - \frac{1}{2}\right)\right),$$

where $\alpha = \pi^2 + \frac{g^2}{4}$. On the domain $\Omega = [0,1]^2$, the function $\varrho(t,x,y)$ is positive and satisfies the mixed boundary conditions $(\nabla\varrho + \varrho\nabla V)\cdot\boldsymbol{n}|_{\partial\Omega} = 0$. We consider $g = 1$. We want to exploit the knowledge of this exact solution to compute the error we commit in the spatial and time integration. Consider a sequence of meshes $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m}\right)$ with decreasing mesh size $h_{\mathcal{T}_m}$ and a sequence of decreasing time steps τ_m such that $\frac{h_{\mathcal{T}_{m+1}}}{h_{\mathcal{T}_m}} = \frac{\tau_{m+1}}{\tau_m}$. In particular, we used a sequence of Delaunay triangular meshes such that the mesh size halves at each step, obtained subdividing at each step each triangle into four using the edges midpoints. Three subsequent partitioning of the domain are shown in Figure 3.1.

Table 3.1: Time-space convergence for the two schemes. Integration on the interval $[0, 0.25]$.

		FV				LJKO			
h	dt	ϵ_{L^∞}	r	ϵ_{L^1}	r	ϵ_{L^∞}	r	ϵ_{L^1}	r
0.2986	0.0500	0.1634	/	0.0350	/	0.1463	/	0.0334	/
0.1493	0.0250	0.0856	0.932	0.0176	0.997	0.0651	1.169	0.0145	1.120
0.0747	0.0125	0.0434	0.979	0.0087	1.015	0.0449	0.535	0.0066	1.134
0.0373	0.0063	0.0218	0.996	0.0043	1.009	0.0297	0.598	0.0033	1.007
0.0187	0.0031	0.0109	0.999	0.0022	1.004	0.0174	0.770	0.0017	0.943
0.0093	0.0016	0.0054	1.000	0.0011	1.001	0.0095	0.870	0.0009	0.947

Let us introduce the following mesh-dependent errors:

$$\begin{aligned}
\epsilon_1^n &= \sum_{K\in\mathcal{T}_m} |\rho_K^n - \varrho(n\tau, \boldsymbol{x}_K)| m_K \quad \to \quad \text{discrete } L^1 \text{ error}\,,\\
\epsilon_{L^\infty} &= \max_n(\epsilon_n^1) \quad \to \quad \text{discrete } L^\infty((0,T);L^1(\Omega)) \text{ error}\,,\\
\epsilon_{L^1} &= \sum_n \tau\, \epsilon_1^n \quad \to \quad \text{discrete } L^1((0,T);L^1(\Omega)) \text{ error}\,,
\end{aligned}$$

where $\varrho(n\tau_m, \boldsymbol{x}_K)$ is the value in the cell center of the triangle K of the analytical solution at time $n\tau_m$, n running from 1 to the total number of time steps N_m. The upstream Finite Volume scheme with backward Euler discretization of the temporal derivative, namely scheme (3.43), is known to exhibit order one of convergence applied to this problem, both in time and space. This means that the $L^\infty((0,T);L^1(\Omega))$ and $L^1((0,T);L^1(\Omega))$ errors halve whenever $h_\mathcal{T}$ and τ halve. We want to inspect whether scheme (3.30) recovers the same behavior.

For the sequence of meshes and time steps, for m going from one to the total number of meshes, we computed the solution to the linear Fokker-Planck equations and the errors, using both schemes (3.43) and (3.30). The results are shown in Table 3.1. For each mesh size and time step m, it is represented the error together with the rate with respect to the previous one. Scheme (3.30) exhibits the same order of convergence of scheme (3.43). It is noticeable that the rate of convergence of the former scheme senses a big drop and then recovers order one, especially in the $L^\infty((0,T);L^1(\Omega))$ error. This is due to the fact that the initial condition $\varrho(0, \boldsymbol{x}_K)$ is too close to zero, and in particular equal to zero on the set $1 \times [0,1]$, and scheme (3.30) tends to be repulsed away from zero due to the singularity of the gradient of the first variation of the energy. In Table 3.2 we repeated the convergence test for the time interval $[0.05, 0.25]$: the convergence profile sensibly improves.

To further investigate and compare the behavior of the two schemes, we computed also the energy decay along the trajectory. We call dissipation the difference $\mathcal{E}(\varrho) - \mathcal{E}(\varrho^\infty)$, where ϱ^∞ is the final equilibrium condition, the long time behavior. Since we are discretizing a gradient flow, its dissipation is a useful criteria to assess the goodness of the scheme. The long time value of the energy is equal to:

$$\begin{aligned}
\mathcal{E}\left(\lim_{t\to\infty}\varrho\right) &= \int_\Omega \lim_{t\to\infty} (\varrho\log\varrho - \varrho g x)\,\mathrm{d}\boldsymbol{x}\\
&= \exp\left(\frac{g}{2}\right)\left(\frac{\pi\log(\pi)}{g} + \frac{\pi}{2} - \frac{\pi}{g}\right) + \exp\left(-\frac{g}{2}\right)\left(-\frac{\pi\log(\pi)}{g} - \frac{\pi}{2} + \frac{\pi}{g}\right).
\end{aligned}$$

Table 3.2: Time-space convergence for scheme (3.30). Integration on the interval $[0.5, 0.25]$.

		LJKO			
h	dt	ϵ_{L^∞}	r	ϵ_{L^1}	r
0.2986	0.0500	0.1186	/	0.0216	/
0.1493	0.0250	0.0618	0.9411	0.0109	0.9857
0.0747	0.0125	0.0307	1.0110	0.0053	1.0311
0.0373	0.0063	0.0152	1.0116	0.0026	1.0213
0.0187	0.0031	0.0076	1.0078	0.0013	1.0119
0.0093	0.0016	0.0038	1.0042	0.0006	1.0062

It is possible to define the equilibrium solution also on the discrete dynamics on the grid. Namely, the equilibrium solution $\boldsymbol{\rho}^\infty$ for the discrete dynamics is

$$\rho_K^\infty = M \exp(-V_K),\ V_K = V(\boldsymbol{x}_K),\ \forall K \in \mathcal{T},$$

which is the unique minimizer of the discrete energy $\mathcal{E}_\mathcal{T} = \sum_{K\in\mathcal{T}} \big(\rho_K \log \rho_K + \rho_K V(\boldsymbol{x}_K)\big) m_K$ subject to the constraint of the conservation of the mass. The optimality conditions for this problem provide indeed:

$$\begin{aligned} &\frac{\partial}{\partial \rho_K}\big(\mathcal{E}_\mathcal{T} + \lambda \sum_{K\in\mathcal{T}} (\rho_K - \rho_K^0) m_K\big)|_{\rho_K^\infty} = \big(\log \rho_K^\infty + 1 + V_K + \lambda\big) m_K = 0, \quad \forall K \in \mathcal{T} \\ &\implies \quad \rho_K^\infty = \exp(-(1+\lambda) - V_K) = M \exp(-V_K), \quad \forall K \in \mathcal{T}, \end{aligned}$$

with λ lagrange multiplier associated with the constraint. M is the constant that makes $\boldsymbol{\rho}^\infty$ have the same total mass:

$$M = \frac{\sum_{K\in\mathcal{T}} \rho_K^0 m_K}{\sum_{K\in\mathcal{T}} \exp^{-V_K} m_K}.$$

It is immediate to observe that this is indeed the equilibrium solution for scheme (3.43), since with such density the potential is constant:

$$\phi_K = \frac{\delta \mathcal{E}_\mathcal{T}(\boldsymbol{\rho})}{\delta \rho_K}|_{\rho_K^\infty} = \log \rho_K^\infty + 1 + V_K = \log M - V_K + 1 + V_K = \log M + 1, \quad \forall K \in \mathcal{T}.$$

For the scheme (3.30) instead, as it appears clear from Lemma 3.2, whenever $\rho_K^n = \rho_K^{n-1}, \forall K \in \mathcal{T}$, as it is the case for an equilibrium solution, the potential is constant. From the potential equation one gets again

$$\phi_K = \frac{\delta \mathcal{E}_\mathcal{T}(\boldsymbol{\rho})}{\delta \rho_K}|_{\rho_K^\infty} = \log M + 1, \forall K \in \mathcal{T}.$$

In Figure 3.2 it is represented the semilog plot of the dissipation of the system in the time interval $[0, 3]$, computed for the two schemes, $\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) - \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^\infty)$, and the real solution, $\mathcal{E}(\varrho) - \mathcal{E}(\varrho^\infty)$. In Figure 3.2a it is noticeable that scheme (3.30) dissipates the energy faster than the other, being indeed a bit more diffusive. This is an expected behavior since the scheme is built to maximize the decrease of the energy and this is actually one of the main

strength of the approach. In Figure 3.2b, one can see that the two dissipations tend to the real one when a finer mesh and a smaller time step are used, for both schemes, despite the fact that (3.30) still dissipates faster. In the end, in Figure 3.2c it is possible to remark that for a very small time step the dissipations tend to coincide, as it is expected. For the time parameter going to zero the two schemes coincide.

3.4.3 Other examples of Wasserstein gradient flows

Let us consider now more qualitative tests on other examples of Wasserstein gradient flows, in order to show the general validity of our approach.

Porous medium equation

The porous medium equation,

$$\partial_t \varrho = \Delta \varrho^\delta + \nabla \cdot (\varrho \nabla V)\,,$$

has been proven in [108] to be a gradient flow in Wasserstein space with respect to the energy

$$\mathcal{E}(\rho) = \int_\Omega \frac{1}{\delta - 1} \rho^\delta \mathrm{d}\boldsymbol{x} + \int_\Omega \rho V \mathrm{d}\boldsymbol{x}\,, \tag{3.78}$$

for a given δ strictly greater than one. Our aim is to show that scheme (3.30) works regardless of the uniform bound from below on the density. For this reason, we use an initial density ρ^0 with compact support and a confining potential $V(\boldsymbol{x}) = \frac{1}{2}|\boldsymbol{x} - \boldsymbol{x}_0|^2$, where $\boldsymbol{x}_0 = (\frac{1}{2}, \frac{1}{2})$. The equilibrium solution of the gradient flow should then be the Barenblatt profile

$$\varrho^\infty(\boldsymbol{x}) = \max\left(\left(\frac{M}{2\pi}\right)^{\frac{\delta-1}{\delta}} - \frac{\delta - 1}{2\delta}\,|\boldsymbol{x} - \boldsymbol{x}_0|^2\,, 0\right)^{\frac{1}{\delta-1}}\,,$$

with M total mass of the initial condition.

In Figure 3.3 the evolution of an initial density close to a dirac in the center of the domain $\Omega = [0,1]^2$ is shown for the case $m = 4$. In Figure 3.4 it is represented the dissipation of the energy, $\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) - \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^\infty)$, in semi-logarithmic scale, where $\boldsymbol{\rho}^\infty_K = \varrho^\infty(\boldsymbol{x}_K), \forall K \in \mathcal{T}$. The energy $\mathcal{E}_\mathcal{T}$ is the straightforward discretization of (3.78), as it has been done for the Fokker-Planck energy. As expected, the solution converges towards the Barenblatt profile.

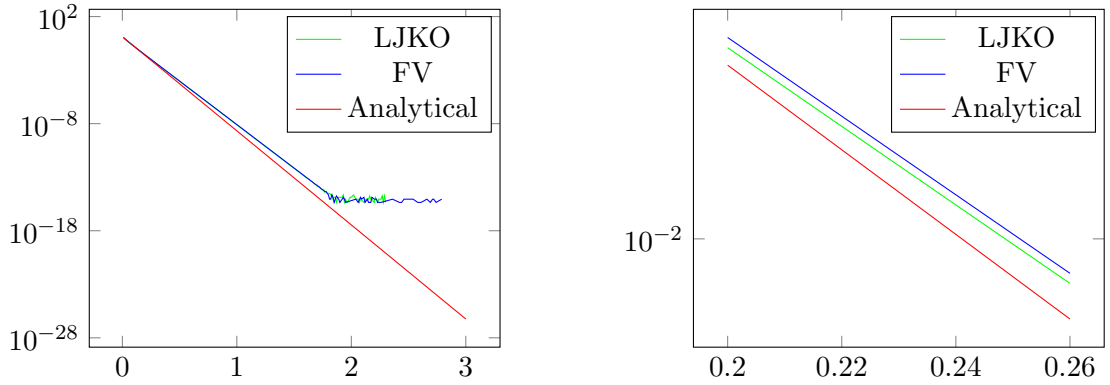
Thin film equation

In order to show that scheme (3.30) can be employed also on more complex problems, we consider the Wasserstein gradient flow with respect to the energy

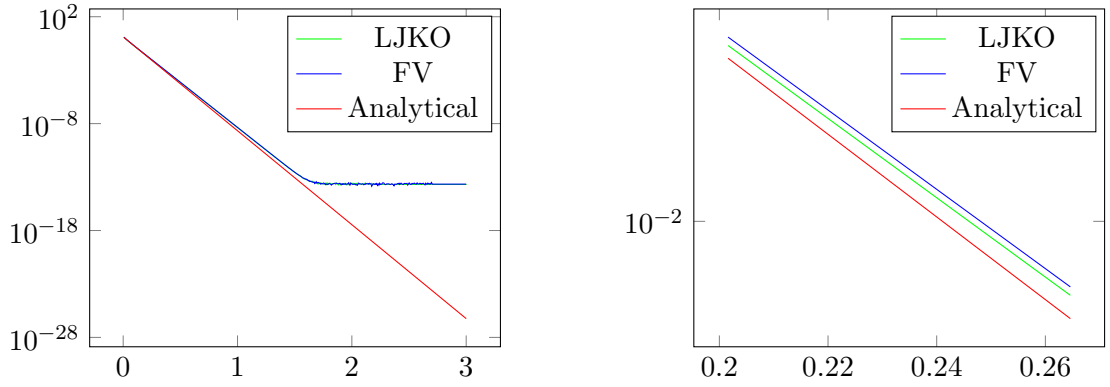
$$\mathcal{E}(\rho) = \frac{1}{2}\int_\Omega |\nabla \rho|^2 \mathrm{d}\boldsymbol{x} + \int_\Omega \rho V \mathrm{d}\boldsymbol{x}\,,$$

which gives rise to a phenomenon modeled by the thin film equation

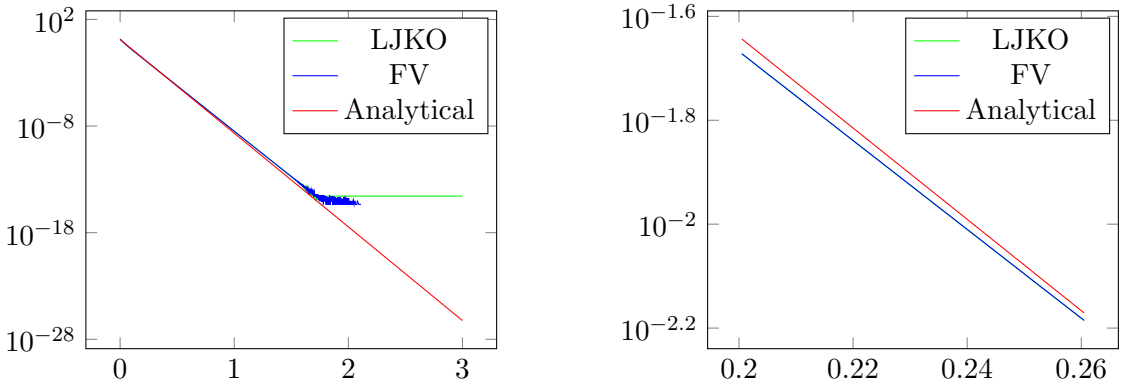
$$\partial_t \varrho = -\nabla \cdot (\varrho \nabla(\Delta \varrho)) + \nabla \cdot (\varrho \nabla V)\,,$$



(a) $\tau_m = 0.01,\ h_{\mathcal{T}_m} = 0.1493.$



(b) $\tau_m = 0.0063,\ h_{\mathcal{T}_m} = 0.0373.$



(c) $\tau_m = 0.0001,\ h_{\mathcal{T}_m} = 0.1493.$

Figure 3.2: Comparison of the dissipation along time of the system computed with the two numerical schemes (3.30) (LJKO) and (3.43) (FV), and in the real case. Semi-logarithmic plot.

a particular case of a family of nonlinear fourth order equations [90]. The energy $\mathcal{E}(\rho)$ is discretized as

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) = \frac{1}{2}\sum_{\sigma\in\Sigma}\left(\frac{\rho_L - \rho_K}{d_\sigma}\right)^2 d_\sigma m_\sigma + \sum_{K\in\mathcal{T}} \rho_K V(\boldsymbol{x}_K) m_K\,,$$

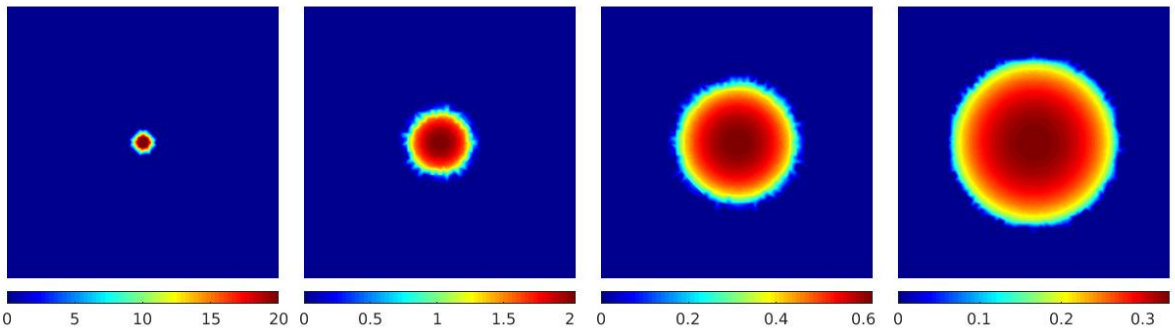


Figure 3.3: Evolution of an initial density close to a dirac according to the porous medium equation. From left to right, $t = 0, 0.0001, 0.01, 1$. In each picture the scaling is different for the sake of the representation.

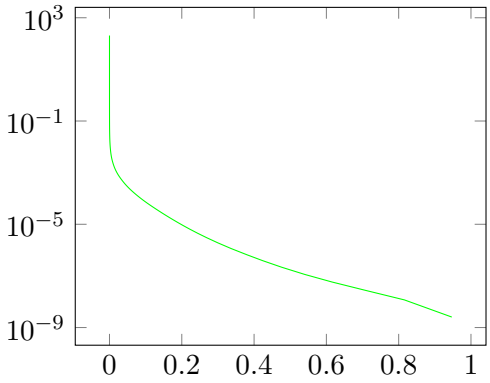


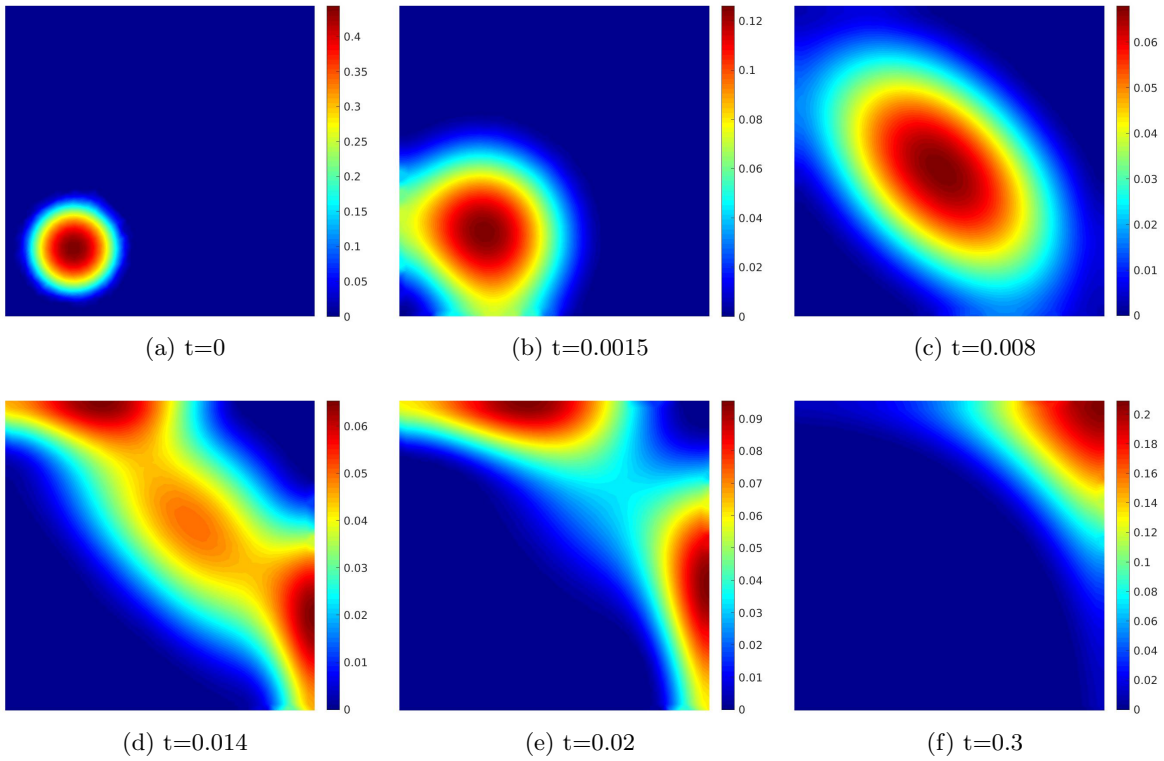
Figure 3.4: Dissipation along time of the energy for the porous medium equation. Semi-logarithmic plot.

where again we made use of the inflated gradient definition for the discretization of the Dirichlet energy. Notice that even though the continuous energy functional $\mathcal{E}(\rho)$ is local, the discrete counterpart is not. The matrix $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ in (3.77) is not diagonal and the Schur complement technique for the solution of the linear system (3.76) is not necessarily convenient anymore. In Figure 3.5 it is represented the evolution of an initial density with quadratic profile and compact support in the domain $\Omega = [0,1]^2$. The potential is $V(\boldsymbol{x}) = (x-1)(y-1)$.

Salinity intrusion problem

We want to show now that scheme (3.30) can be used for the solutions of systems of equations of the type of (3.1). We consider the problem of salinity intrusion in an unconfined aquifer. Under the assumption that the two fluids, the fresh and the salt water, are immiscible and the domains occupied by each fluid are separated by a sharp interface, the problem can be modeled via the system of equations

$$\begin{cases} \partial_t f - \nabla \cdot (\nu f \nabla (f + g + b)) = 0\,, \\ \partial_t g - \nabla \cdot (g \nabla (\nu f + g + b)) = 0\,. \end{cases} \tag{3.79}$$



(a) t=0 (b) t=0.0015 (c) t=0.008

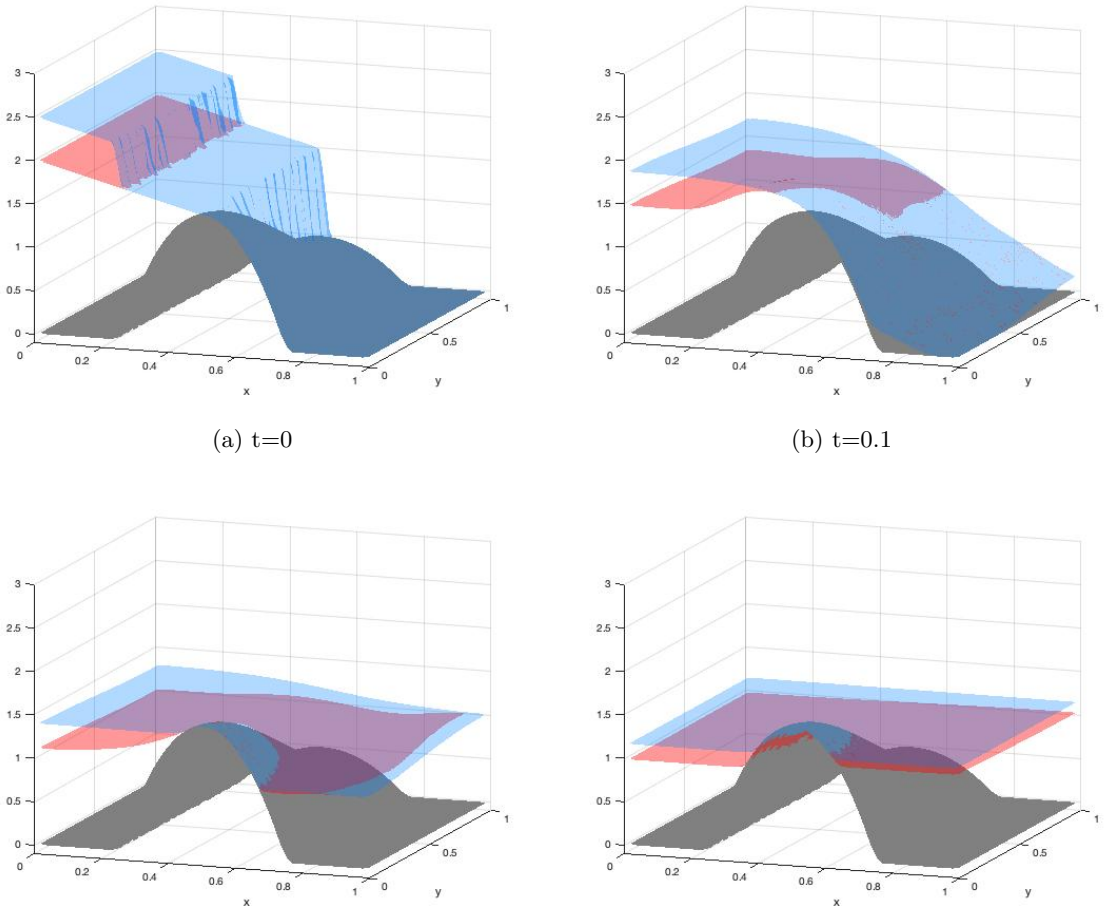
(d) t=0.014 (e) t=0.02 (f) t=0.3

Figure 3.5: Evolution of an initial quadratic density according to the thin film equation. In each picture the scaling is different for the sake of the representation.

System (3.79) is completed with the no-flux boundary conditions $\nabla f \cdot \boldsymbol{n} = \nabla g \cdot \boldsymbol{n} = 0$ on $\partial\Omega, \forall t \geq 0$, and the initial conditions $f(0,\cdot) = f_0, g(0,\cdot) = g_0$, with $f_0, g_0 \in L^\infty(\Omega)$, $f_0, g_0 \geq 0$. The quantities f, g, and b represent respectively the thickness of the fresh water layer, the thickness of the salt water layer and the height of the bedrock. Therefore the quantity $b+g$ represents the height of the sharp interface separating the two fluids. The parameter $\nu = \frac{\rho_f}{\rho_s}$ is the ratio between the constant mass densities of the fresh and salt water. Equation (3.79) has been proven in [78] to be a Wasserstein gradient flow with respect to the energy

$$\mathcal{E}(f,g) = \int_\Omega \Big(\frac{\nu}{2}(b+g+f)^2 + \frac{1-\nu}{2}(b+g)^2\Big)\mathrm{d}\boldsymbol{x}\,. \tag{3.80}$$

The discretization of (3.80) is again straightforward. In Figure 3.6 it is represented an evolution of the two surfaces of salt and fresh water (see [2] for a full description of the test case). Given the particular configuration of the bedrock b, the two surfaces are represented respectively by $b+g$ and $b+g+f$. Also this case is not covered from the theoretical analysis we performed on the convergence of the scheme but still scheme (3.30) works. As already said, from numerical evidences the scheme works under much more general and mild hypotheses.



(a) t=0

(b) t=0.1

(c) t=0.5

(d) t=10

Figure 3.6: Evolution of the two interfaces of salt (red) and fresh (blue) water.

Chapter 4

Centered TPFA finite volume discretization for Wasserstein gradient flows

This chapter contains an extended presentation of:

Andrea Natale and Gabriele Todeschi. TPFA Finite Volume Approximation of Wasserstein Gradient Flows. In *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, pages 193–201. Springer International Publishing, 2020.

4.1 Introduction

In the previous chapter we proposed a discretization for Wasserstein gradient flows. By taking advantage of the variational structure of the problem, i.e. the fact that the flow can be seen as the limit curve generated via the JKO scheme (1.40), we derived a non-trivial discretization of the partial differential equation representing it, namely equation (1.32). For this discretization, the energy decay and the positivity of the discrete density can be ensured. Furthermore, the scheme is intrinsically robust, as it is formulated as the solution of a well-posed convex optimization problem. Preserving at the discrete level the monotonicity of the Hamilton-Jacobi operator, at each step of the scheme the solution can be computed solving the system of non-linear equations (3.30). However, the discretization is of order one, in time by the very nature of the JKO scheme, and in space due to the choice of the upwind reconstruction. Moreover, the mere application of the Newton scheme we considered for solving the equations is not a robust optimization technique. The convergence is indeed ensured by reducing, whenever necessary, the time step τ (see Section 3.4.1). The objective of this chapter is to improve these two aspects.

Improving the accuracy in time is the objective of the next chapter, even though a first try inspired by [41] will be discussed here. We will mainly focus here on the space discretization. In the previous chapter we strongly relied on the upwind discretization in order to preserve the monotonocity and write the discrete system of equations (3.30). Without it, one has to deal with the more general problem (4.22), see below. We follow the path we set in Chapter 2 and employ the same interior point strategy. In this way we can be more flexible in the choice

of the discretization: we will use in particular a centered reconstruction for the mobility as we did in Chapter 2, which typically leads to second order accuracy in space. Furthermore, we can obtain in this way a more robust solver by taking advantage of the smoothing effect and the continuation method.

4.1.1 LJKO time discretization

As we already said, we can see a gradient flow as the limit process of a discrete evolution that minimizes at each step the sum of an energy term and a penalization on the Wasserstein distance. This is the so called JKO scheme, which we recall here. For an intial condition $\rho^0 \in \mathcal{P}(\Omega)$ and an increasing sequence $(t^n)_{n\in\mathbb{N}} \subset \mathbb{R}$ of time steps such that $\cup_n [t^{n-1}, t^n] = [0, T]$, $[t^{n-1}, t^n] = \tau$ (not necessarily constant), the JKO scheme constructs a sequence $(\rho^n)_{n\in\mathbb{N}} \subset \mathcal{P}(\Omega)$ as follows: given the measure ρ^{n-1}, compute $\rho^n = \rho(t^n)$, where $(\rho, \boldsymbol{m}) : [t^{n-1}, t^n] \times \Omega \to \mathbb{R}^+ \times \mathbb{R}^d$ solve

$$\inf_{(\rho,\boldsymbol{m})} \int_{t^{n-1}}^{t^n} \int_\Omega B(\rho, \boldsymbol{m}) \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mathcal{E}(\rho(t^n, \cdot)) \,, \tag{4.1}$$

among all $(\rho, \boldsymbol{m})$ belonging to the convex subset of distributional solutions to the continuity equation:

$$\begin{cases} \partial_t \rho + \nabla \cdot \boldsymbol{m} = 0 & \text{in } [t^{n-1}, t^n] \times \Omega \,, \\ \boldsymbol{m} \cdot \boldsymbol{n} = 0 & \text{on } [t^{n-1}, t^n] \times \partial\Omega \,, \\ \rho(t^{n-1}, \cdot) = \rho^{n-1} \,. \end{cases} \tag{4.2}$$

We used the Benamou-Brenier formulation of optimal transport. The function $B : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty]$ is the density of kinetic energy,

$$B(p, \boldsymbol{Q}) \coloneqq \begin{cases} \frac{|\boldsymbol{Q}|^2}{2p} & \text{if } p > 0 \,, \\ 0 & \text{if } p = 0,\ \boldsymbol{Q} = 0 \,, \\ +\infty & \text{else} \,, \end{cases} \tag{4.3}$$

and the minimal total kinetic energy of the curve $(\rho, \boldsymbol{m})$ satisfying the continuity equation constraint (4.2) represents the squared Wasserstein distance between ρ^{n-1} and ρ^n. The sequence computed in this way is then an approximation in time of the continuous flow. However, the Wasserstein distance involved in (4.1) is a complex time dependent problem and needs to be further discretized. Since the JKO scheme is naturally of order one, a first order time discretization is sufficient (as we have shown numerically in Chapter 3) and leads to a reasonable computational complexity. As the Wasserstein distance is closed to a weighted $\dot{H}_\rho^{-1}$ norm for two measures that are sufficiently close to each other ([120, Section 7.6],[110]), namely

$$\|\rho - \nu\|_{\dot{H}_\rho^{-1}} = \mathcal{W}_2(\rho, \nu) + o(\mathcal{W}_2(\rho, \nu)), \quad \forall \rho, \nu \in \mathcal{P}(\Omega),$$

we can enormously simplify the problem by replacing it in (4.1)-(4.2). We obtain in this way the following scheme:

$$\rho^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{2\tau} \left\|\rho - \rho^{n-1}\right\|_{\dot{H}_\rho^{-1}}^2 + \mathcal{E}(\rho), \quad \forall n \geq 1 \,. \tag{4.4}$$

As we have shown in Chapter 3, by the very same change of variables of the Benamou-Brenier formulation, the weighted $\dot{H}^{-1}_\rho$ norm squared can be written as a convex optimization problem. Indeed, it holds

$$\frac{1}{2}\left\|\rho-\rho^{n-1}\right\|^2_{\dot{H}^{-1}_\rho} = \inf_{(\rho,\boldsymbol{m})}\left\{\int_\Omega B(\rho,\boldsymbol{m})\mathrm{d}\boldsymbol{x}\,:\,\begin{cases}\rho-\rho^{n-1}+\nabla\cdot\boldsymbol{m}=0 & \text{in }\Omega,\\ \boldsymbol{m}\cdot\boldsymbol{n}=0 & \text{on }\partial\Omega,\end{cases}\right\}, \tag{4.5}$$

see Section 3.1.2. Replacing problem (4.1)-(4.2) with problem (4.4) essentially consists, up to a time rescale by the factor τ of the momentum, in discretizing the continuity equation using a single implicit Euler step and the time integral using a right endpoint approximation. We refer to this time discretization as Linearized JKO scheme (LJKO).

4.2 A second order discretization in space

The space discretization is based again on TPFA finite volumes. Let us recall the main definitions and notation that we have set in Section 1.3. We denote by $\left(\mathcal{T},\overline{\Sigma},(\boldsymbol{x}_K)_{K\in\mathcal{T}}\right)$ an admissible mesh according to Definition 1.1, namely the triplet of the set of polyhedral control volumes, the set of faces and the set of cell centers. The space of discrete conservative fluxes is given by

$$\mathbb{F}_\mathcal{T} = \{\boldsymbol{F}=(F_{K,\sigma},F_{L,\sigma})_{\sigma\in\Sigma}\in\mathbb{R}^{2\Sigma}\,:\,F_{K,\sigma}+F_{L,\sigma}=0\}\,. \tag{4.6}$$

We denote $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$ and, by convention, $|\boldsymbol{F}| = (F_\sigma)_{\sigma\in\Sigma}\in\mathbb{R}^\Sigma$ and $(\boldsymbol{F})^2 = (F^2_\sigma)_{\sigma\in\Sigma}\in\mathbb{R}^\Sigma$, for $\boldsymbol{F}\in\mathbb{F}_\mathcal{T}$. The spaces of discrete variables are $\mathbb{R}^\mathcal{T}$ and $\mathbb{R}^\Sigma$, defined respectively on the cells and the diamond cells of the mesh, endowed with the two scalar products

$$\langle\cdot,\cdot\rangle_\mathcal{T} : (\boldsymbol{a},\boldsymbol{b})\in[\mathbb{R}^\mathcal{T}]^2\mapsto\sum_{K\in\mathcal{T}}a_Kb_Km_K\,,$$

$$\langle\cdot,\cdot\rangle_\Sigma : (\boldsymbol{u},\boldsymbol{v})\in[\mathbb{R}^\Sigma]^2\mapsto\sum_{\sigma\in\Sigma}u_\sigma v_\sigma m_\sigma d_\sigma\,.$$

To define the mobility on the diamond cells, we employ again the average operators $\mathcal{L}_\Sigma,\mathcal{H}_\Sigma:\mathbb{R}^\mathcal{T}\to\mathbb{R}^\Sigma$ we introduced in Section 2.2.2: for any $\boldsymbol{a}\in\mathbb{R}^\mathcal{T}$,

$$(\mathcal{L}_\Sigma\boldsymbol{a})_\sigma = \lambda_{K,\sigma}a_K+\lambda_{L,\sigma}a_L\,,\quad (\mathcal{H}_\Sigma\boldsymbol{a})_\sigma = \frac{a_Ka_L}{\lambda_{K,\sigma}a_L+\lambda_{L,\sigma}a_K}\,, \tag{4.7}$$

which correspond to weighted arithmetic and harmonic averages, where $\lambda_{K,\sigma},\lambda_{L,\sigma}\in[0,1]$, $\lambda_{K,\sigma}+\lambda_{L,\sigma}=1$. Other choices are possible, such as geometric or logarithmic averages [51, 64]. For both reconstructions we will consider the weights $(\lambda_{K,\sigma},\lambda_{L,\sigma}) = (\frac{d_{K,\sigma}}{d_\sigma},\frac{d_{L,\sigma}}{d_\sigma})$, which provide a mass weighted mean. For the linear reconstruction, we will also consider $(\frac{1}{2},\frac{1}{2})$, the standard arithmetic mean, and $(\frac{d_{L,\sigma}}{d_\sigma},\frac{d_{K,\sigma}}{d_\sigma})$, which provides a linear reconstruction of the density at the edge midpoint. Thanks to these reconstructions we expect to obtain second order accuracy for the space discretization.

We denote by $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}]:\mathbb{R}^\mathcal{T}\to\mathbb{R}^\Sigma$ the differential of $\mathcal{R}_\Sigma$ with respect to $\boldsymbol{a}$, evaluated at a given $\boldsymbol{a}\in\mathbb{R}^\mathcal{T}$. Clearly, if $\mathcal{R}_\Sigma=\mathcal{L}_\Sigma$, we simply have $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}]=\mathcal{L}_\Sigma$. Moreover, we denote by $(\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}])^*$ the adjoint of $\mathrm{d}\mathcal{R}_\Sigma[\boldsymbol{a}]$, with respect to the two different scalar products. For the

two reconstructions we consider, this operator is given by either $\mathcal{L}_\Sigma^*$ or $(\mathrm{d}\mathcal{H}_\Sigma[\boldsymbol{a}])^*$, which are defined by

$$(\mathcal{L}_\Sigma^*\boldsymbol{u})_K = \sum_{\sigma\in\Sigma_K} u_\sigma \lambda_{K,\sigma}\frac{m_\sigma d_\sigma}{m_K}\,, \quad ((\mathrm{d}\mathcal{H}_\Sigma[\boldsymbol{a}])^*\boldsymbol{u})_K = \sum_{\sigma\in\Sigma_K} \frac{(\mathcal{H}_\Sigma[\boldsymbol{a}])_\sigma^2}{a_K^2} u_\sigma \lambda_{K,\sigma}\frac{m_\sigma d_\sigma}{m_K}\,, \tag{4.8}$$

for any $\boldsymbol{u}\in\mathbb{R}^\Sigma$.

Centered finite volume scheme

The discrete scheme can be derived straightforwardly by following the same steps as in Section 3.2. We begin again by defining the discrete counterpart of the weighted norm (4.5). First, we introduce the discrete space of zero mass variables:

$$\mathbb{R}_0^\mathcal{T} = \left\{\boldsymbol{h}\in\mathbb{R}^\mathcal{T} : \langle \boldsymbol{h}, \mathbf{1}\rangle_\mathcal{T} = 0\right\}.$$

Then, given the discrete density $\boldsymbol{\rho}\in\mathbb{R}^\mathcal{T}$ and for any $\boldsymbol{h}\in\mathbb{R}_0^\mathcal{T}$, the discrete counterpart of the $\dot{H}_\rho^{-1}$ norm squared is

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{h}) = \begin{cases} \inf\limits_{\boldsymbol{F}\in\mathbb{F}_\mathcal{T}} \left\{ \sum\limits_{\sigma\in\Sigma} B\left((\mathcal{R}_\mathcal{T}(\boldsymbol{\rho}))_\sigma, F_\sigma\right) m_\sigma d_\sigma : h_K m_K = \sum\limits_{\sigma\in\Sigma_K} F_{K,\sigma} m_\sigma \right\} & \text{if } \langle \boldsymbol{h},\mathbf{1}\rangle_\mathcal{T} = 0\,, \\ +\infty & \text{else}\,. \end{cases} \tag{4.9}$$

The total kinetic energy is discretized on the diamond cells. Each flux F_σ is meant as an approximation of the quantity $|\boldsymbol{m}\cdot\boldsymbol{n}_{K,\sigma}|$ and the measure of each diamond cell is taken $m_\sigma d_\sigma = d m_{\Delta_\sigma}$, i.e. d times the actual measure, where d stands for the space dimension, in order to compensate this unidirectional discretization. The constraint $\boldsymbol{m}\cdot\boldsymbol{n} = 0$ on the boundary is automatically taken into account disregarding the flux on the boundary edges in the definition of the space of discrete conservative fluxes (4.6). For any $\boldsymbol{\rho}\in\mathbb{R}^\mathcal{T}$, the function $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\cdot)$ is proper ($\boldsymbol{h} = \mathbf{0}$ has always finite value), convex and lower semi-continuous as infimum of convex and lower semi-continuous functions. The function $B(\cdot, F_\sigma)$ is indeed convex and decreasing and the reconstructions $\mathcal{R}_\Sigma$ defined in (4.8) are component-wise concave. Existence of the optimal flux can be obtained easily thanks to the coercivity of the objective function with respect to $\boldsymbol{F}$. The function $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\cdot)$ is then equal to its double Legendre transform and it holds:

$$\begin{aligned} \mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{h}) &= \inf_{\boldsymbol{F}\in\mathbb{F}_\mathcal{T}} \sup_{\boldsymbol{\phi}\in\mathbb{R}^\mathcal{T}} \sum_{\sigma\in\Sigma} B\left((\mathcal{R}_\mathcal{T}(\boldsymbol{\rho}))_\sigma, F_\sigma\right) m_\sigma d_\sigma + \sum_{K\in\mathcal{T}} \phi_K h_K m_K - \sum_{\sigma\in\Sigma} F_{K,\sigma}(\phi_K - \phi_L) m_\sigma \\ &= \sup_{\boldsymbol{\phi}\in\mathbb{R}^\mathcal{T}} \inf_{\boldsymbol{F}\in\mathbb{F}_\mathcal{T}} \sum_{\sigma\in\Sigma} B\left((\mathcal{R}_\mathcal{T}(\boldsymbol{\rho}))_\sigma, F_\sigma\right) m_\sigma d_\sigma + \sum_{K\in\mathcal{T}} \phi_K h_K m_K - \sum_{\sigma\in\Sigma} F_{K,\sigma}(\phi_K - \phi_L) m_\sigma \\ &= \sup_{\boldsymbol{\phi}\in\mathbb{R}^\mathcal{T}} \langle \boldsymbol{\phi}, \boldsymbol{h}\rangle_\mathcal{T} - \frac{1}{2}\sum_{\sigma\in\Sigma} (\mathcal{R}_\Sigma(\boldsymbol{\rho}))_\sigma \Big(\frac{\phi_K - \phi_L}{d_\sigma}\Big)^2 m_\sigma d_\sigma \\ &= \sup_{\boldsymbol{\phi}\in\mathbb{R}^\mathcal{T}} \langle \boldsymbol{\phi}, \boldsymbol{h}\rangle_\mathcal{T} - \mathcal{A}_\mathcal{T}^*(\boldsymbol{\rho};\boldsymbol{\phi})\,. \end{aligned} \tag{4.10}$$

Let us comment on the above calculations. After incorporating the constraint via the Lagrange multiplier $\boldsymbol{\phi}$, we used the conservativity property of the space $\mathbb{F}_{\mathcal{T}}$ to write:

$$\sum_{K\in\mathcal{T}} \phi_K \sum_{\sigma\in\Sigma_K} F_{K,\sigma} m_\sigma = \sum_{\sigma\in\Sigma} F_{K,\sigma}(\phi_K - \phi_L) m_\sigma \,.$$

Then we swapped inf and sup thanks to strong duality, the objective function being convex in $\boldsymbol{F}$ and linear in $\boldsymbol{\phi}$. Finally, we used the optimality conditions with respect to $\boldsymbol{F}$ which provides

$$F_{K,\sigma} = (\mathcal{R}_\Sigma(\boldsymbol{\rho}))_\sigma \Big(\frac{\phi_K - \phi_L}{d_\sigma}\Big) \,.$$

Notice that differently from the upwind case presented in Chapter 3, which introduces an asymmetry in the problem, in this case both $\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{h})$ and $\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi})$ are symmetric, i.e. $\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};-\boldsymbol{h}) = \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{h})$ and $\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};-\boldsymbol{\phi}) = \mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi})$.

We can now formulate the discrete versions of problem (4.4). Assume the discrete energy $\boldsymbol{\rho} \mapsto \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})$ to be a proper, strictly convex and lower semi-continuous scalar function. Given the discrete initial density $\boldsymbol{\rho}^0 \in \mathbb{R}^{\mathcal{T}}_+$, define the discrete space

$$\mathbb{P}_{\mathcal{T}} = \{\boldsymbol{\rho} \in \mathbb{R}^{\mathcal{T}}_+ : \langle \boldsymbol{\rho}, \mathbf{1}\rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^0, \mathbf{1}\rangle_{\mathcal{T}}\} \,.$$

For a time step $\tau > 0$, the discrete analogous of the LJKO scheme (4.4) computes recursively the sequence of densities $(\boldsymbol{\rho}^n)_{n\in\mathbb{N}}$ as: given $\boldsymbol{\rho}^{n-1} \in \mathbb{P}_{\mathcal{T}}$, compute $\boldsymbol{\rho}^n$ solution to

$$\inf_{\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}} \frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) \,. \tag{4.11}$$

The function $\boldsymbol{\rho} \mapsto \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho})$ is convex and lower semi-continuous, as it can be written as the sup of convex and lower semi-continuous functions thanks to (4.10) and by concavity of $\mathcal{R}_\Sigma$. Existence of a solution $\boldsymbol{\rho}^n$ is then ensured as the whole objective function is lower semi-continuous and the minimization is carried out on the compact set $\mathbb{P}_{\mathcal{T}}$. Uniqueness is guaranteed by the strict convexity of the energy. Thanks to the duality formula (4.10), this ensures the existence also of the optimal potential $\boldsymbol{\phi}^n$. The conservation of mass is automatically enforced thanks to the conservativity of the finite volume discretization, by definition of the function $\mathcal{A}_{\mathcal{T}}$, and therefore

$$\langle \boldsymbol{\rho}^n, \mathbf{1}\rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^{n-1}, \mathbf{1}\rangle_{\mathcal{T}} \,, \quad \forall n \,.$$

The minimization in $\boldsymbol{\rho}$ can be performed on the whole subspace $\mathbb{R}^{\mathcal{T}}_+$. Furthermore, the scheme guarantees the discrete energy-dissipation property: given the solution $\boldsymbol{\rho}^n$ to (4.11), the competitor $\boldsymbol{\rho}^{n-1}$ provides

$$\frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n;\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}^n) + \mathcal{E}(\boldsymbol{\rho}^n) \le \mathcal{E}(\boldsymbol{\rho}^{n-1}) \,.$$

Thanks to (4.10), rescaling the potential as $\boldsymbol{\phi} \leftarrow \frac{\boldsymbol{\phi}}{\tau}$, problem (4.11) can be formulated as

$$\inf_{\boldsymbol{\rho}\in\mathbb{R}^{\mathcal{T}}_+} \sup_{\boldsymbol{\phi}\in\mathbb{R}^{\mathcal{T}}} \langle \boldsymbol{\phi}, \boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}\rangle_{\mathcal{T}} - \tau\,\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi}) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) \,. \tag{4.12}$$

We already know that a solution exists (the potential is not necessarily unique) and strong duality holds again. We can therefore characterize $(\boldsymbol{\phi}^n, \boldsymbol{\rho}^n)$ as solution to the following system of necessary and sufficient optimality conditions:

$$\begin{cases} (\rho_K^n - \rho_K^{n-1}) m_K + \tau \sum_{\sigma \in \Sigma_K} (\mathcal{R}_\Sigma(\boldsymbol{\rho}^n))_\sigma \Big(\dfrac{\phi_K^n - \phi_L^n}{d_\sigma} \Big) m_\sigma = 0 \,, \\ \phi_K^n m_K - \dfrac{\partial \mathcal{E}_\mathcal{T}}{\partial \rho_K}(\rho_K^n) + \dfrac{\tau}{2} \Big(\mathcal{R}_\mathcal{T} \Big(\dfrac{\phi_K^n - \phi_L^n}{d_\sigma} \Big)^2 \Big)_K m_K \leq 0 \,, \end{cases} \qquad \forall K \in \mathcal{T} \,. \tag{4.13}$$

The linear operator $\mathcal{R}_\mathcal{T} : \mathbb{R}^\Sigma \to \mathbb{R}^\mathcal{T}$ is given by either $\mathcal{L}_\Sigma^*$ or $(\mathrm{d}\mathcal{H}_\Sigma[\boldsymbol{\rho}^n])^*$, as defined in (4.8). The inequality derives from the fact that we are optimizing on the set of positive densities and the equality holds in the cells where $\rho_K > 0$. By introducing the auxiliary variable $\boldsymbol{\lambda}^n \in \mathbb{R}_-^\mathcal{T}$, we can rewrite system (4.13) as:

$$\begin{cases} (\rho_K^n - \rho_K^{n-1}) m_K + \tau \sum_{\sigma \in \Sigma_K} (\mathcal{R}_\Sigma(\rho^n))_\sigma \Big(\dfrac{\phi_K^n - \phi_L^n}{d_\sigma} \Big) m_\sigma = 0 \,, \\ (\phi_K^n - \lambda_K^n) m_K - \dfrac{\partial \mathcal{E}_\mathcal{T}}{\partial \rho_K}(\rho_K^n) + \dfrac{\tau}{2} \Big(\mathcal{R}_\mathcal{T} \Big(\dfrac{\phi_K^n - \phi_L^n}{d_\sigma} \Big)^2 \Big)_K m_K = 0 \,, \qquad \forall K \in \mathcal{T} \,. \\ \rho_K^n \geq 0, \ \lambda_K^n \leq 0, \ \rho_K^n \lambda_K^n = 0 \,, \end{cases} \tag{4.14}$$

The variable $\boldsymbol{\lambda}^n \in \mathbb{R}_-^\mathcal{T}$ acts as the Lagrange multiplier of the positivity constraint on $\boldsymbol{\rho}^n$.

Remark on the Hamilton-Jacobi equation

System (4.14) is not easy to solve, the major problem being the non-uniqueness of the multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\phi}$ whenever the density vanishes. The discretization we proposed in Chapter 3, based on the upwind reconstruction for the mobility (3.21), allowed us to consider $\boldsymbol{\lambda} = \mathbf{0}$ an admissible solution and discard it. That is, it allowed us to saturate the Hamilton-Jacobi equation, while preserving the optimality of the solution, Theorem 3.5. We relied on the monotonicity of the discrete Hamilton-Jacobi operator, a property that was essential in Lemma 3.6, the building block of Theorem 3.5. The upwind reconstruction has however another nice feature as shown in Section 3.2.3: the discrete continuity equation admits only non-negative solutions. We may think that, in order to consider $\boldsymbol{\lambda} = \mathbf{0}$ an admissible solution for system (4.14), ensuring the positivity is sufficient and that the monotonicity is not needed. Let us show that such positivity preservation property is not enough.

Consider the weighted harmonic reconstruction operator $\mathcal{H}_\Sigma : \mathbb{R}^\mathcal{T} \to \mathbb{R}^\Sigma$ and let us define it as:

$$(\mathcal{H}_\Sigma(\boldsymbol{\rho}))_\sigma = \begin{cases} \frac{d_\sigma \rho_K \rho_L}{d_{K,\sigma} \rho_L + d_{L,\sigma} \rho_K} & \text{if } \rho_K, \rho_L > 0 \,, \\ 0 & \text{else} \,, \end{cases} \tag{4.15}$$

so that it is continuous and defined everywhere. The discrete continuity equation in (4.13) is in this case:

$$(\rho_K^n - \rho_K^{n-1}) m_K + \tau \sum_{\sigma \in \Sigma_K} (\mathcal{H}_\Sigma(\rho^n))_\sigma \Big(\frac{\phi_K^n - \phi_L^n}{d_\sigma} \Big) m_\sigma = 0 \,, \quad \forall K \in \mathcal{T} \,. \tag{4.16}$$

Equation (4.16) admits only positive solutions as soon as $\boldsymbol{\rho}^{n-1}$ is positive. Suppose indeed it exists a solution $\boldsymbol{\rho}^n \in \mathbb{R}^{\mathcal{T}}$ with $\rho_K^n < 0$, for a control volume $K \in \mathcal{T}$. Then we obtain $\rho_K^n = \rho_K^{n-1} < 0$, which is a contradiction. With this reconstruction we could hope to saturate the Hamilton-Jacobi equation in system (4.14), i.e. take by default $\lambda = \mathbf{0}$ as we did with the upwind reconstruction, and compute the optimal solution to problem (4.11). This is not the case as the following counterexample shows. Consider the energy $\mathcal{E}(\rho) = \int_\Omega \frac{1}{2}\rho^2$, which gives rise to the porous medium equation, a nonlinear type of diffusion. The discrete counterpart of this energy is $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) = \sum_{K\in\mathcal{T}} \frac{1}{2}\rho_K^2 m_K$. Consider a trivial mesh made of two square cells of edge length one and cell centers $\boldsymbol{x}_1 = (-\frac{1}{2}, 0), \boldsymbol{x}_2 = (\frac{1}{2}, 0)$, so that there is only one internal edge $\sigma = 1|2$, $d_\sigma = 1$ and $|\boldsymbol{x}_1 - \boldsymbol{x}_2| = 1$. System of equations (4.13) writes in this case

$$\begin{cases} \rho_1^n - \rho_1^{n-1} + \tau(\phi_2^n - \phi_1^n)\frac{2\rho_1^n\rho_2^n}{\rho_1^n+\rho_2^n} = 0\,, \\ \rho_2^n - \rho_2^{n-1} + \tau(\phi_1^n - \phi_2^n)\frac{2\rho_1^n\rho_2^n}{\rho_1^n+\rho_2^n} = 0\,, \\ \phi_1^n - \rho_1^n + \frac{\tau}{2}(\phi_2^n - \phi_1^n)^2\frac{2(\rho_2^n)^2}{(\rho_1^n+\rho_2^n)^2} = 0\,, \\ \phi_2^n - \rho_2^n + \frac{\tau}{2}(\phi_1^n - \phi_2^n)^2\frac{2(\rho_1^n)^2}{(\rho_1^n+\rho_2^n)^2} = 0\,, \end{cases}$$

where we saturated the Hamilton-Jacobi equations. Take then as initial condition $\boldsymbol{\rho}^0 = (1, 0)$. We can check then that the density $\boldsymbol{\rho}^n = (1, 0), \forall n \geq 1$, is an admissible stationary solution for the system of equations as long as $\tau < \frac{1}{4}$, i.e. the density does not evolve. This is clearly not the correct dynamics, which should instead tend to diffuse the mass everywhere. If the discretization of the Hamilton-Jacobi operator is not monotone, saturating the equation does not necessarily provide an optimal solution in problem (4.12).

4.3 A different time approach

We want to propose an alternative, possibly more precise, time discretization to (4.4). It relies on two things: a variant of the JKO time discretization and a more precise approximation of the Wasserstein distance. We present it here, instead of postponing it to the next chapter, since the structure of problem does not change.

The accuracy in time is limited by the very nature of the JKO scheme. In [41] a simple modification has been proposed which seems to be more accurate. It is based on the analogy with the Crank-Nicolson time discretization and consists in replacing the energy $\mathcal{E}$ in the n-th JKO step with $\tilde{\mathcal{E}}^{n-1}$ defined as:

$$\widetilde{\mathcal{E}}^{n-1}(\rho) = \frac{1}{2}\left(\mathcal{E}(\rho) + \int_\Omega \frac{\delta\mathcal{E}}{\delta\rho}[\rho^{n-1}]\,\rho\right).$$

The energy $\widetilde{\mathcal{E}}^{n-1}$ is the average of the original $\mathcal{E}$ with its first order expansion in ρ^{n-1} (up to a constant term). In the finite dimensional case, given a function $F : \mathbb{R}^d \to \mathbb{R}$, the problem writes as

$$\boldsymbol{x}^n \in \operatorname*{arginf}_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2\tau}\left|\boldsymbol{x} - \boldsymbol{x}^{n-1}\right|^2 + \frac{1}{2}\left(F(\boldsymbol{x}) + \langle\nabla F(\boldsymbol{x}^{n-1}), \boldsymbol{x}\rangle\right).$$

The optimality conditions for the minimizer $\boldsymbol{x}^n$ provides the recurrence formula

$$\frac{\boldsymbol{x}^n - \boldsymbol{x}^{n-1}}{\tau} = -\frac{1}{2}\left(\nabla F(\boldsymbol{x}^n) + \nabla F(\boldsymbol{x}^{n-1})\right),$$

which is the Crank-Nicolson time discretization of the Cauchy problem (1.37), the gradient flow of F.

As the numerical experiments in Chapter 2 suggest, we may be able to obtain a more precise scheme by taking the arithmetic mean of the measures ρ and ρ^{n-1} as weight for the H^{-1} norm:

$$\frac{1}{2}\left\|\rho-\rho^{n-1}\right\|^2_{\dot{H}^{-1}_{\frac{\rho+\rho^{n-1}}{2}}} = \inf_{(\rho,\boldsymbol{m})}\left\{\int_\Omega B\Big(\frac{\rho+\rho^{n-1}}{2},\boldsymbol{m}\Big)\mathrm{d}\boldsymbol{x} : \begin{cases}\rho-\rho^{n-1}+\ \nabla\cdot\boldsymbol{m}=0, & \text{in }\Omega,\\ \boldsymbol{m}\cdot\boldsymbol{n}=0, & \text{on }\partial\Omega,\end{cases}\right\}. \tag{4.17}$$

In this way the approximation of the Wasserstein distance could be more accurate as the time integral of the kinetic energy is approximated with a midpoint rule.

Using these simple ideas we can try to devise a higher accurate in time scheme with respect to (4.4):

$$\rho^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{2\tau}\left\|\rho-\rho^{n-1}\right\|^2_{\dot{H}^{-1}_{\frac{\rho+\rho^{n-1}}{2}}} + \widetilde{\mathcal{E}}^{n-1}(\rho)\,, \quad \forall n\geq 1\,. \tag{4.18}$$

The discrete analogous of scheme (4.18) is derived in the same way. At each step n, the density $\boldsymbol{\rho}^n$ is computed as:

$$\inf_{\boldsymbol{\rho}\in\mathbb{P}_\mathcal{T}} \frac{1}{\tau}\mathcal{A}_\mathcal{T}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-1}}{2};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}\Big) + \widetilde{\mathcal{E}}^{n-1}_\mathcal{T}(\boldsymbol{\rho})\,, \tag{4.19}$$

where

$$\widetilde{\mathcal{E}}^{n-1}_\mathcal{T}(\boldsymbol{\rho}) = \frac{1}{2}\left(\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) + \langle\nabla_{\boldsymbol{\rho}}\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^{n-1}),\boldsymbol{\rho}\rangle_\mathcal{T}\right),$$

and $\nabla_{\boldsymbol{\rho}}\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^{n-1}) = (\frac{\partial\mathcal{E}_\mathcal{T}}{\partial\rho_K}(\boldsymbol{\rho}^{n-1}))_{K\in\mathcal{T}}\in\mathbb{R}^\mathcal{T}$. The analogous of formula (4.9) and (4.10) can be straighforwardly written. The optimality conditions are derived in the same way and writes in this case:

$$\begin{cases}(\rho^n_K-\rho^{n-1}_K)m_K + \tau\displaystyle\sum_{\sigma\in\Sigma_K}\Big(\mathcal{R}_\Sigma\Big(\frac{\boldsymbol{\rho}^n+\boldsymbol{\rho}^{n-1}}{2}\Big)\Big)_\sigma\Big(\frac{\phi^n_K-\phi^n_L}{d_\sigma}\Big)m_\sigma = 0\,,\\ \phi^n_K m_K - \dfrac{\partial\widetilde{\mathcal{E}}^{n-1}_\mathcal{T}}{\partial\rho_K}(\rho^n_K) + \dfrac{\tau}{4}\Big(\mathcal{R}_\mathcal{T}\Big(\dfrac{\phi^n_K-\phi^n_L}{d_\sigma}\Big)^2\Big)_K m_K \leq 0\,,\end{cases} \quad \forall K\in\mathcal{T}\,. \tag{4.20}$$

The linear operator $\mathcal{R}_\mathcal{T}:\mathbb{R}^\Sigma\to\mathbb{R}^\mathcal{T}$ is in this case either $\mathcal{L}^*_\Sigma$ or $(\mathrm{d}\mathcal{H}_\Sigma[\frac{\boldsymbol{\rho}^n+\boldsymbol{\rho}^{n-1}}{2}])^*$, as defined in (4.8). We stress that scheme (4.19) does not necessarily decrease the energy at each step. The decrease of $\tilde{\mathcal{E}}^{n-1}$ does not implies the decrease of $\mathcal{E}$. We will test this scheme and show that, despite being more precise, it is not second order accurate in time.

We conclude with another final remark on the Hamilton-Jacobi equation and its saturation. With the discretization we presented in Chapter 3, the functional in the dual problem, (3.39), is linear and increasing in $\boldsymbol{\phi}$. This is not the case using the arithmetic mean of the densities $\boldsymbol{\rho}^n$ and $\boldsymbol{\rho}^{n-1}$ in the definition of the function $\mathcal{A}_\mathcal{T}$ as in problem (4.19). Indeed, this term writes

$$\mathcal{A}_\mathcal{T}(\frac{\boldsymbol{\rho}^n+\boldsymbol{\rho}^{n-1}}{2};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}^n) = \langle\boldsymbol{\phi}^n,\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}^n\rangle_\mathcal{T} - \frac{1}{2}\sum_{\sigma\in\Sigma}(\mathcal{R}_\Sigma(\frac{\boldsymbol{\rho}^n+\boldsymbol{\rho}^{n-1}}{2}))_\sigma(\frac{\phi^n_K-\phi^n_L}{d_\sigma})^2 m_\sigma d_\sigma\,.$$

where $\boldsymbol{\phi}^n$ is an optimal potential. In the objective functional, and hence in the dual problem, there is a term that depends on the density $\boldsymbol{\rho}^{n-1}$ and on the gradient squared of the potential $\boldsymbol{\phi}^n$. Consequently, assuming the discrete Hamilton-Jacobi equation to be monotone, saturating it provides a bigger potential but it does not necessarily increase the value of the objective functional, as we do not control its gradient squared. The potential obtained by saturating the equation is not necessarily optimal in this case. For this reason, in the scheme we proposed in Chapter 3 we cannot consider an arithmetic average of the densities in the definition of $\mathcal{A}_\mathcal{T}$. The possibility of saturating the equation depends also on the form of the dual problem and not only on the monotonicity of the discrete operator.

4.4 Interior point strategy

In order to find solutions to problems (4.11) and (4.19) we use again the interior point strategy we devised in Section 2.6. As we pointed out already, this approach is very flexible and can be adapted easily to different problems. Consider for example problem (4.11). We introduce its perturbed version:

$$\inf_{\boldsymbol{\rho}\in\mathbb{R}^\mathcal{T}} \frac{1}{\tau}\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}) + \mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) - \mu\sum_{K\in\mathcal{T}}\log(\rho_K)m_K\,. \tag{4.21}$$

The addition of the barrier function forces the solution to be positive, taking care automatically of the positivity coinstraint, and regularizes the problem. The optimality conditions of the perturbed problem are in this case

$$\begin{cases} (\rho_K^n-\rho_K^{n-1})m_K + \tau\displaystyle\sum_{\sigma\in\Sigma_K}(\mathcal{R}_\Sigma(\rho^n))_\sigma\Big(\frac{\phi_K^n-\phi_L^n}{d_\sigma}\Big)m_\sigma = 0\,, \\ (\phi_K^n+s_K)m_K - \dfrac{\partial\mathcal{E}_\mathcal{T}}{\partial\rho_K}(\rho_K^n) + \dfrac{\tau}{2}\Big(R_\mathcal{T}\Big(\dfrac{\phi_K^n-\phi_L^n}{d_\sigma}\Big)^2\Big)_K m_K = 0\,, \qquad \forall K\in\mathcal{T}, \\ s_K\rho_K = \mu\,, \end{cases} \tag{4.22}$$

which can be seen as a pertubation of (4.14), where ρ_K and $s_K = -\lambda_K$ are automatically forced to be positive and the orthogonality is relaxed. For small value of the parameter μ we can then approximate the solution of problem (4.11) (the convergence proof for $\mu\to 0$ is identical to the proof of Theorem 2.13) and the true solution is recovered via a continuation method. The resulting algorithm is identical to Algorithm 1. See Section 2.6 for a careful description of the algorithm and all the considerations regarding its implementation, which apply also in this case.

We remark that solving the gradient flow with respect to an energy which is singular or has singular derivative in zero, as for example the entropy $\mathcal{E}(\rho) = \int_\Omega \rho\log(\rho)$, enforces automatically the strict positivity of the density. In system (4.14) the equality holds then on every cell and the problem can be solved with the Newton scheme again. However, one cannot control the magnitude of the energy and therefore the interior point method, even if not strictly necessary, helps to get a more robust solver.

4.5 Numerical results

We are going now to perform a convergence test to show that scheme (4.11) is second order accurate in space. We will also show that the new time discretization (4.19) does not attain second order accuracy in time. We will finally present a qualitative test, a porous medium flow.

Convergence test

In Section 3.4, we perfomed a convergence test in order to test the accuracy of the scheme. We considered an analytical solution to the Fokker-Planck equation. We want to repeat the same test in the present case, in order to show that the new discretization does attain a second order accuracy in space.

We recall that the model problem is:

$$\partial_t \varrho = \Delta \varrho + \nabla \cdot (\varrho \nabla V)\,, \tag{4.23}$$

complemented with no flux boundary condition and a positive initial condition, with $V \in W^{1,\infty}(\Omega)$ a Lipschitz continuous exterior potential. Equation (4.23) defines a gradient flow in the Wasserstein space with respect to the energy $\mathcal{E}(\rho) = \int_\Omega (\rho \log \rho + \rho V) \mathrm{d}\boldsymbol{x}$, and it has been one of the first problem to be recast in this way [73]. The analytical solution we consider is

$$\varrho(t, \boldsymbol{x}) = \exp\left(-\alpha t + \frac{g}{2} x\right)\left(\pi \cos(\pi x) + \frac{g}{2}\sin(\pi x)\right) + \pi \exp\left(g\left(x - \frac{1}{2}\right)\right),$$

where $\alpha = \pi^2 + \frac{g^2}{4}$, which solves equation (4.23) in the domain $[0, 0.25] \times [0, 1]^2$ with potential $V(\boldsymbol{x}) = -gx$. We take $g = 1$. Consider then a sequence of meshes $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m}\right)$ with decreasing mesh size $h_m = h_{\mathcal{T}_m}$, and a sequence of decreasing time steps τ_m such that $(\frac{\tau_{m+1}}{\tau_m}) = (\frac{h_{m+1}}{h_m})^2$. We use the same sequence of meshes as in Section 3.4 (see Figure 3.1). We solve problem (4.23) with scheme (4.11) using this sequence of meshes and using as discrete initial condition $\rho_K^0 = \varrho(0, \boldsymbol{x}_K)$. The continuous energy is straightforwardly discretized as

$$\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) = \sum_{K\in\mathcal{T}} (\rho_K \log \rho_K + \rho_K V(\boldsymbol{x}_K))\, m_K\,.$$

For each solution we compute the mesh-dependent $L^1((0,T); L^1(\Omega))$ error:

$$\epsilon_m = \sum_n \tau_m \sum_{K\in\mathcal{T}_m} |\rho_K^n - \rho_s(\boldsymbol{x}_K, n\tau_m)| m_K\,.$$

In Table 4.1 are listed the errors for each m together with the convergence rate,

$$\frac{\log(\epsilon_{m-1}) - \log(\epsilon_m)}{\log(h_{m-1}) - \log(h_m)}\,,$$

for the three different weighted arithmetic averages and the harmonic one. The scheme is first order accurate in time and second order accurate in space. The results can be compared with the results presented in Table 3.1.

We want to test now the scheme (4.19) to see if the strategy proposed in [41], together with the arithmetic average in time of the mobilities in the definition of function $\mathcal{A}_\mathcal{T}$, allows

Table 4.1: Time-space convergence for the scheme (4.11). Linear reconstruction for the mobility in the first three cases, harmonic in the last one.

		$\mathcal{L}_\Sigma$						$\mathcal{H}_\Sigma$	
h_m	τ_m	ϵ_m^a	rate	ϵ_m^b	rate	ϵ_m^c	rate	ϵ_m	rate
0.2986	0.0500	3.281e-02	/	3.336e-02	/	3.300e-02	/	3.238e-02	/
0.1493	0.0125	7.443e-03	2.140	7.609e-03	2.132	7.341e-03	2.168	7.479e-03	2.114
0.0747	0.0031	1.759e-03	2.081	1.788e-03	2.089	1.736e-03	2.080	1.792e-03	2.062
0.0373	0.0008	4.332e-04	2.021	4.389e-04	2.026	4.288e-04	2.018	4.434e-04	2.015
0.0187	0.0002	1.080e-04	2.004	1.092e-04	2.007	1.070e-04	2.002	1.106e-04	2.004

[a] Weights $(\frac{1}{2}, \frac{1}{2})$. [b] Weights $(\frac{d_L}{d_\sigma}, \frac{d_K}{d_\sigma})$. [c] Weights $(\frac{d_K}{d_\sigma}, \frac{d_L}{d_\sigma})$.

Table 4.2: Time-space convergence for the schemes (4.11) and (4.19).

		Scheme (4.11)				Scheme (4.19)			
		$\mathcal{L}_\Sigma$		$\mathcal{H}_\Sigma$		$\mathcal{L}_\Sigma$		$\mathcal{H}_\Sigma$	
h_m	τ_m	ϵ_m	rate	ϵ_m	rate	ϵ_m	rate	ϵ_m	rate
0.2986	0.0500	2.041e-02	/	2.043e-02	/	5.916e-03	/	6.047e-03	/
0.1493	0.0250	1.071e-02	0.930	1.073e-02	0.929	2.304e-03	1.361	2.367e-03	1.353
0.0747	0.0125	5.404e-03	0.987	5.411e-03	0.988	1.010e-03	1.190	1.025e-03	1.208
0.0373	0.0063	2.709e-03	0.996	2.712e-03	0.997	4.754e-04	1.087	4.769e-04	1.103
0.0187	0.0031	1.356e-03	0.999	1.357e-03	0.999	2.320e-04	1.035	2.320e-04	1.040

to obtain a higher accuracy in time. We repeat the same test using scheme (4.19), using the same sequence of grids but considering this time a linearly decreasing sequence of time steps τ_m, satisfying precisely: $(\frac{\tau_{m+1}}{\tau_m}) = (\frac{h_{m+1}}{h_m})$. We perform also the test with scheme (4.11) to compare the results. In both cases, we consider the weights $(\frac{d_K}{d_\sigma}, \frac{d_L}{d_\sigma})$ and both the linear and the harmonic reconstruction. We integrate the equation on the time interval $[0.05, 0.25]$ to avoid the effects of the singularity at time $t = 0$ in $1 \times [0, 1]$. The results are presented in Table 4.2. Although being more precise, the approach (4.19) is only first order accurate.

Porous medium flow

As second application, we consider a gradient flow of an energy which is not singular in zero. On the domain $\Omega = [-1.5, 1.5]^2$, for a time interval $[0, T]$, consider the porous medium equation,

$$\partial_t \varrho = \Delta \varrho^\delta + \nabla \cdot (\varrho \nabla V) \,,$$

which we recall is a gradient flow in the Wasserstein space with respect to the energy $\mathcal{E}(\rho) = \int_\Omega \frac{1}{\delta - 1} \rho^\delta + \rho V$, for a given δ strictly greater than one [108]. We consider the confining potential $V(\boldsymbol{x}) = \frac{1}{2} |\boldsymbol{x}|^2$, which forces the density to concentrate at the origin. We compute the discrete

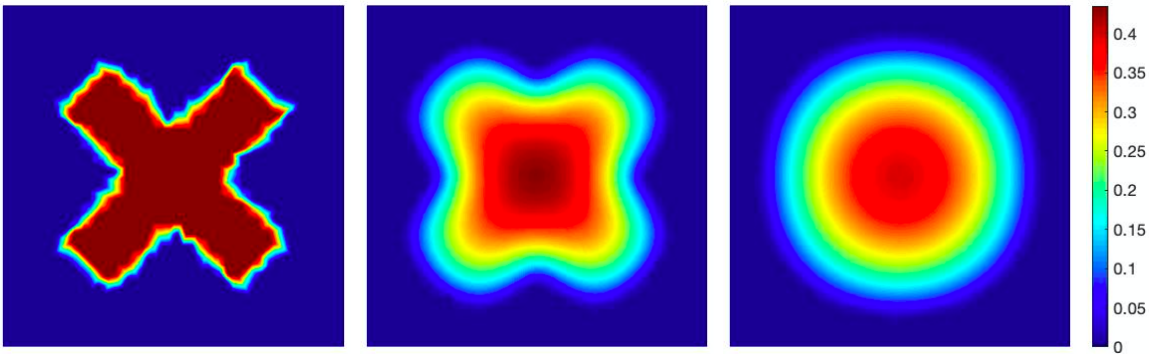


Figure 4.1: Convergence towards the Barenblatt solution ($\gamma = 2$). Time steps $t = 0$, $t = 0.1$ and $t = 0.7$.

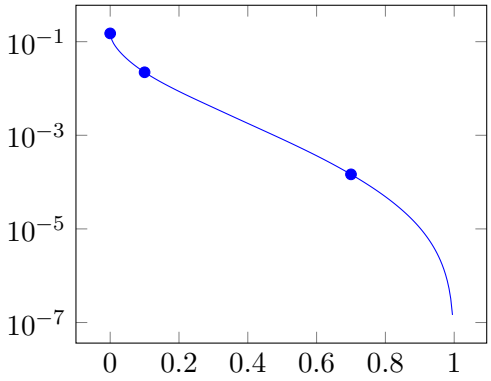


Figure 4.2: Exponential decay profile of the dissipation along time, with the three values corresponding to Figure 4.1. Semi-logarithmic plot.

flow using scheme (4.11) with linear reconstruction, for the discrete energy

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) = \sum_{K\in\mathcal{T}} \left(\frac{1}{\delta - 1} \rho_K^{\delta} + \rho_K V(\boldsymbol{x}_K) \right) m_K \,.$$

In (4.1) the evolution of an initial cross shaped density is shown for the case $\delta = 2$. As expected, the solution converges towards the Barenblatt profile

$$\rho^{\infty}(\boldsymbol{x}) = \max\left(\left(\frac{M}{2\pi} \right)^{\frac{\delta-1}{\delta}} - \frac{\delta - 1}{2\delta} |\boldsymbol{x}|^2 , 0 \right)^{\frac{1}{\delta-1}} ,$$

M being the total mass of the initial condition (Figure 4.1). In Figure 4.2, it is represented the dissipation profile of the energy, $\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) - \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^{\infty})$, which converges exponentially towards the energy of the Barenblatt equilibrium solution.

4.6 Concluding remarks

In order to be able to discretize Wasserstein gradient flows with second order accuracy in space, we presented a strategy based again on TPFA finite volumes. Instead of considering

the upwind reconstruction, we used in this case a centered reconstruction for the density, which enabled to reach the second order accuracy in space. The time discretization is again based on the LJKO scheme, which limits the accuracy in time and therefore the overall accuracy of the scheme. The modification based on the strategy proposed in [41] does not provide a second order accurate scheme. We will consider a second order time discretization in the next chapter.

The great advantage of using the upwind reconstruction is to be able to solve the system of optimality conditions directly, resulting in an extremely efficient approach. In comparison, the interior point strategy we proposed for solving the problem with the new discretization is evidently more involved, as it requires the solution of several non-linear system of equations. Concerning specifically the solution of the sequence of linear systems, all the considerations we made in Section 3.4.1 are again valid: the barrier function has a local structure (i.e. its hessian is diagonal) and if the discrete energy functional $\mathcal{E}_{\mathcal{T}}$ is local as well, the system of equations can be reduced to the strictly positive definite and symmetric Schur complement and solved with efficient techniques. Notice that the third equation in (4.22) has been added artificially in order to decouple the optimization in $\boldsymbol{\rho}$ and $\boldsymbol{s}$, but the linear system can be easily reduced to the first two equations.

On the other hand, the smoothing effect of the perturbation and the continuation method result in an extremely robust solver: no matter the time step τ chosen, the solution exists and the algorithm is able to compute it. In Chapter 3 we chose to apply directly the Newton scheme to solve the system of equations (3.30), without considering a globalization technique as for example the use of a linesearch. The convergence was there enforced by reducing the time step when necessary. Indeed, for the Newton scheme the natural starting point is the solution at step $n-1$. The closer this point is to the solution at step n, the easier it is to solve the problem. The strategy therefore was to link the accuracy of the solution to the possibility to solve the system of equations. In this case, the interior point method is sufficiently robust to allow us to solve the problem for any time step. The time step has therefore to be chosen in advance and a careful control, based for example on the (approximate) Wasserstein distance between consecutive densities, could improve the performance of the scheme. We did not explore this possibility in this work.

These last considerations deserve a final comment regarding the present approach. Due to the interior point strategy, the previous known solution is not a valid starting point because of the presence of the perturbation, especially if it has a compact support. The initialization has always to be the constant density in order to ease the start of the algorithm, see Section 2.6 for details. This implies that all the information we know thanks to the previous value of the solution is disregarded in the new step. This seems to be an unavoidable consequence of using an interior point strategy and it is particularly penalizing in the computation of gradient flows.

Chapter 5

A modified BDF2 scheme for Wasserstein gradient flows

5.1 Introduction

In Chapters 3 and 4 we presented strategies for discretizing Wasserstein gradient flows. Concerning the time discretization, we relied on the JKO scheme (1.40), variational generalization of the implicit Euler scheme. For the space discretization we followed two strategies based always on TPFA finite volumes. In Chapter 3 we used the upwind choice for the reconstruction of the mobility, which leads to a monotone discretization and enables to obtain a one step resolvant of the system of equations (1.54). The problem can then be solved using directly a Newton scheme. The drawback is that the accuracy in space of the scheme is limited to order one. In Chapter 4, we used instead a centered reconstruction which enables to obtain a second order accuracy in space but requires more effort for solving the discrete optimization problem. The objective is now to increase the accuracy in time and propose a second order in time and space scheme.

The JKO scheme is naturally an order one discretization in time. A new variational strategy is required if we want to increase the accuracy. New approaches can be designed generalizing second order schemes for ordinary differential equations, following the example of the JKO scheme. The simple strategy we considered in Chapter 4 which has been proposed in [41], based on the analogy with the Crank-Nicolson scheme, is not enough. Two other strategies have been proposed, which are not however particularly suited to be implemented: the variational implicit midpoint [82] and the variational backward differentiation formula of order two [93]. We will propose a modified version of the latter. The new strategy can be justified rigorously and we propose a possible non rigorous, yet effective, implementation. To the best of our knowledge, there exist no numerical approach able to compute with second order accuracy general Wasserstein gradient flows while preserving the variational structure.

5.1.1 Existing variations of the JKO scheme

Two well-known second order accurate schemes for ordinary differential equations are the midpoint scheme and the BDF2 (Backward Differentiation Formula of order 2) scheme. Let us consider the domain $\mathbb{R}^d$ endowed with the standard euclidean metric. We can cast these two schemes in a variational form as follows. Consider a real valued scalar function $F : \mathbb{R}^d \to \mathbb{R}$,

a starting point $\boldsymbol{x}^0 \in \mathbb{R}^d$ and a time step $\tau > 0$. The Variational Implicit Midpoint (VIM) scheme can be written as: at each step n, compute $\boldsymbol{x}^n$ as

$$\boldsymbol{x}^n \in \operatorname*{arginf}_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{2\tau}\left|\boldsymbol{x}-\boldsymbol{x}^{n-1}\right|^2 + 2F\Big(\frac{\boldsymbol{x}+\boldsymbol{x}^{n-1}}{2}\Big)\,. \tag{5.1}$$

The optimality conditions of (5.1) provide the recurrence formula

$$\frac{1}{\tau}(\boldsymbol{x}^n-\boldsymbol{x}^{n-1}) = -\nabla F\Big(\frac{\boldsymbol{x}^n+\boldsymbol{x}^{n-1}}{2}\Big)\,,$$

which is indeed the midpoint scheme for the Cauchy problem (1.37). If we consider instead two initial conditions $\boldsymbol{x}^0, \boldsymbol{x}^1 \in \mathbb{R}^d$, where $\boldsymbol{x}^1$ can be computed for example with a single Euler step (1.43), the variational BDF2 scheme can be written as: compute at each step n the point $\boldsymbol{x}^n$ as

$$\boldsymbol{x}^n \in \operatorname*{arginf}_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{\tau}\left(\alpha\left|\boldsymbol{x}-\boldsymbol{x}^{n-1}\right|^2 - \beta\left|\boldsymbol{x}-\boldsymbol{x}^{n-2}\right|^2\right) + F(\boldsymbol{x})\,, \tag{5.2}$$

with $\alpha = 1, \beta = \frac{1}{4}$. The optimality conditions of (5.2) are in this case

$$\frac{1}{\tau}\Big(\frac{3}{2}\boldsymbol{x}^n - 2\boldsymbol{x}^{n-1} + \frac{1}{2}\boldsymbol{x}^{n-2}\Big) = -\nabla F(\boldsymbol{x}^n)\,, \tag{5.3}$$

that is the Backward Differentiation Formula of order 2.

These two schemes have been recently extended to the computation of more general gradient flows, respectively by Legendre and Turinici in [82] and by Matthes and Plazotta in [93]. Both schemes have been formulated for general metric spaces, but we will focus on our case of interest. Let us place ourselves in the Wasserstein space and consider the energy functional $\mathcal{E} : \mathcal{P}(\Omega) \to \mathbb{R}$. As for the JKO, these two schemes generate a sequence of measures $(\rho^n)_{n\in\mathbb{N}} \subset \mathcal{P}(\Omega)$ which discretize the continuous flow. To define the VIM scheme in this case, the linear interpolation needs to be substituted with a more general curve, the geodesic. Given then $\rho^{n-1} \in \mathcal{P}(\Omega)$, the n-th VIM step writes:

$$\rho^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{2\tau}\mathcal{W}_2^2(\rho,\rho^{n-1}) + 2\mathcal{E}(\tilde{\rho}^n)\,, \tag{5.4}$$

where $\tilde{\rho}^n$ is the midpoint of the (not necessarily unique) geodesic between ρ and ρ^{n-1}. Given instead $\rho^{n-1}, \rho^{n-2} \in \mathcal{P}(\Omega)$, the n-th BDF2 step can be naturally written as:

$$\rho^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{\tau}\left(\alpha\mathcal{W}_2^2(\rho,\rho^{n-1}) - \beta\mathcal{W}_2^2(\rho,\rho^{n-2})\right) + \mathcal{E}(\rho)\,. \tag{5.5}$$

where again $\alpha = 1, \beta = \frac{1}{4}$. These two approaches are not however of immediate implementation.

Scheme (5.4) is not numerically feasible as it requires an explicit formula for the midpoint given the initial and final density. This may also lead to convexity issues. In the finite dimensional case, where geodesics are straight lines and the midpoint coincides with the simple arithmetic mean, the objective function is convex as soon as F is. In the case of problem (5.4), it is simple to see that the objective functional is generalized geodesically convex as soon as

the energy functional $\mathcal{E}$ is[1]. Nevertheless, at the discrete level obtaining a convex problem does not seem to be trivial, depending of course on how we represent the geodesic midpoint. Notice however that the VIM scheme can be formulated in another, equivalent way. In order to compute ρ^n at the n-th step, one can compute first $\tilde{\rho}^n$ with a JKO step with halved time step $\frac{\tau}{2}$,

$$\tilde{\rho}^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{\tau}\mathcal{W}_2^2(\rho,\rho^{n-1}) + \mathcal{E}(\rho)\,, \tag{5.6}$$

and then compute ρ^n as the 2-extrapolation[2] along the geodesic from ρ^{n-1} to $\tilde{\rho}^n$. The equivalence of the two formulations is immediate since for the three measures $\rho^{n-1}, \tilde{\rho}^n$ and ρ^n it holds

$$\frac{1}{2\tau}\mathcal{W}_2^2(\rho^n,\rho^{n-1}) + 2\mathcal{E}(\tilde{\rho}^n) = 2\Big(\frac{1}{\tau}\mathcal{W}_2^2(\tilde{\rho}^n,\rho^{n-1}) + \mathcal{E}(\tilde{\rho}^n)\Big)\,.$$

Problem (5.6) is classically (generalized geodesically) convex as soon as the functional $\mathcal{E}$ is classically (generalized geodesically) convex. Nevertheless, it requires to extrapolate the measure curve. Although numerically feasible (see Section 5.3), in this way there are no guarantees on the regularity of the final measure, as the extrapolation does not provide any regularity. We will expose this issue in Section 5.4.1.

The BDF2 scheme does not require to compute any extrapolation or geodesic. Problem (5.5) is not however a convex optimization problem in the classical sense. We know that the functional $\mathcal{W}_2^2(\cdot,\mu)$, for any fixed measure $\mu\in\mathcal{P}(\Omega)$, is convex. Despite the fact that $\alpha>\beta$, the difference of the two distances in (5.5) is not necessarily convex. Consider the following simple counterexample. Take $\rho^{n-1}=\delta_0$ and $\rho_1=\delta_{-\boldsymbol{x}}, \rho_2=\delta_{\boldsymbol{x}}$, three delta measures centered respectively in $0,\boldsymbol{x}$ and $-\boldsymbol{x}$. Along the interpolation $\rho_t=(1-t)\rho_1+t\rho_2$, the first term of the functional, $\mathcal{W}_2^2(\rho_t,\rho^{n-1})$, is constant whereas the other one is not for a general ρ^{n-2}. Then the second term is concave along the interpolation. The overall convexity of the functional in (5.5) depends on the energy $\mathcal{E}$. This lack of convexity inevitably leads to difficulties in its numerical implementation, see Sections 5.3.3 and 5.4.1.

5.1.2 A new formulation for the BDF2

The BDF2 in the euclidean setting does not suffer from this convexity issue. Problem (5.2) is indeed convex since it holds

$$\alpha\left|\boldsymbol{x}-\boldsymbol{x}^{n-1}\right|^2 - \beta\left|\boldsymbol{x}-\boldsymbol{x}^{n-2}\right|^2 = (\alpha-\beta)\left|\boldsymbol{x}-\boldsymbol{x}_e^{n-1}\right|^2 - \frac{\alpha\beta}{\alpha-\beta}\left|\boldsymbol{x}^{n-1}-\boldsymbol{x}^{n-2}\right|^2$$

and $\alpha>\beta$, where

$$\boldsymbol{x}_e^{n-1} = \frac{\alpha\boldsymbol{x}^{n-1}-\beta\boldsymbol{x}^{n-2}}{\alpha-\beta} = \boldsymbol{x}^{n-2} + \frac{\alpha}{\alpha-\beta}(\boldsymbol{x}^{n-1}-\boldsymbol{x}^{n-2}) = \boldsymbol{x}^{n-2} + \frac{4}{3}(\boldsymbol{x}^{n-1}-\boldsymbol{x}^{n-2}) \tag{5.7}$$

is the euclidean $\frac{4}{3}$-extrapolation from x^{n-2} to x^{n-1}. The point $\boldsymbol{x}_e^{n-1}$ is an approximation of the gradient flow at time $t^{n-1}+\frac{\tau}{3}$. It can be defined as the unique solution to the following

[1] Assume for simplicity that $\mathrm{T}^1,\mathrm{T}^2$ are the two optimal transport plans from ρ^{n-1} (or any other measure) to the two measures ρ^1,ρ^2. If we consider $\mathrm{T}_s^1=(1-s)\mathrm{Id}+s\mathrm{T}^1$, and identically T_s^2, then $\mu_t=((1-t)\mathrm{T}_s^1+t\mathrm{T}_s^2)_\#\rho^{n-1}$ is a generalized geodesic for any $s\in[0,1]$.

[2] Assuming for simplicity T to be the optimal transport maps from ρ^1 to ρ^2, we recall that the t-extrapolation is the measure $\mu_t=(\mathrm{T}_t)_\#\rho^1$, where $\mathrm{T}_t=(1-t)\mathrm{Id}+t\mathrm{T}$ for $t>1$.

problem:

$$\operatorname*{arginf}_{\boldsymbol{x}\in\mathbb{R}^d} \alpha \left|\boldsymbol{x}-\boldsymbol{x}^{n-1}\right|^2 - \beta \left|\boldsymbol{x}-\boldsymbol{x}^{n-2}\right|^2 . \tag{5.8}$$

Thanks to (5.7) we can rewrite the step (5.2) as

$$\boldsymbol{x}^n \in \operatorname*{arginf}_{\boldsymbol{x}\in\mathbb{R}^d} \frac{1}{\tau}(\alpha-\beta)\left|\boldsymbol{x}-\boldsymbol{x}_e^{n-1}\right|^2 + F(\boldsymbol{x}) , \tag{5.9}$$

where we neglect the constant term.

In analogy with (5.9), we propose the following modified version of the BDF2 scheme. Given a time step $\tau > 0$ and two initial conditions $\rho^0, \rho^1 \in \mathcal{P}(\Omega)$, where ρ^1 can be computed for example with a single JKO step (1.40), compute ρ^n as

$$\rho^n \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \frac{1}{\tau}(\alpha-\beta)\mathcal{W}_2^2(\rho,\rho_e^{n-1}) + \mathcal{E}(\rho) , \tag{5.10}$$

where the density ρ_e^{n-1} is defined as

$$\rho_e^{n-1} \in \operatorname*{arginf}_{\rho\in\mathcal{P}(\Omega)} \alpha\mathcal{W}_2^2(\rho,\rho^{n-1}) - \beta\mathcal{W}_2^2(\rho,\rho^{n-2}) . \tag{5.11}$$

In the same way as problem (5.8) recasts in a variational way the euclidean extrapolation, problem (5.11) provides a variational notion of Wasserstein extrapolation in a metric sense. At each step, instead of directly computing the new measure by solving problem (5.5), we first compute the extrapolation at time $t^{n-1}+\frac{\tau}{3}$ and then realize a JKO step of length $\frac{\tau}{2(\alpha-\beta)} = \frac{2\tau}{3}$ from this point in order to compute the measure ρ^n at time $t^n = t^{n-1}+\tau$.

This modified BDF2 scheme requires to compute at each step n the density ρ_e^{n-1}, which is to be computed in advance and does not depend on the unknown density ρ. The problem (5.5) is therefore convex as long as $\mathcal{E}$ is. Differently from the VIM scheme, in this case the extrapolation is computed before taking the JKO step and the energy is evaluated in the final density, providing regularity to the solution. Problem (5.11) suffers from the same convexity issues of the original problem (5.5). Nonetheless, a sufficiently accurate approximation to the solution ρ_e^{n-1} can be effectively computed, which motivates its use from the numerical point of view. We remark that the metric notion of Wasserstein extrapolation provided by problem (5.11) differs in general from the geodesic extrapolation we presented in Section 1.1.2. See the discussion in Section 5.2.1 below.

5.1.3 Organization of the chapter

In Section 5.2 we will analyze this new formulation of the BDF2 scheme. We will first state the well-posedness of the scheme and characterize its properties. We will then show the consistency of the approach: firstly, by proving the convergence of the scheme towards distributional solutions of the Fokker-Planck equation, in the same spirit of the original paper [73]; secondly, by showing that this time discretization recovers Wasserstein gradient flows defined in the EVI sense. Relying on the finite volumes techniques developed in the previous chapters, we will propose a numerical implementation of the scheme. Since (5.11) is not a convex optimization problem, we will propose another definition of extrapolation. We will verify the consistency and the second order accuracy in Section 5.4. We will also present simple implementations of the VIM and the original BDF2 schemes, test their accuracy and compare their solutions.

5.2 Analysis of the modified BDF2 scheme

In order to carry out the analysis of the scheme, let us make the following assumptions on the energy functional: $\mathcal{E}$ is lower-semicontinuous with respect to the weak-* topology, bounded from below and generalized geodesically convex (see Section 1.2.1). To simplify the presentation, we consider the time step τ to be fixed at each step n and to be always an integer divisor of the total integration time T, i.e. $N_\tau = \frac{T}{\tau} \in \mathbb{N}$. In the following, we will keep most of the time the explicit reference to the coefficients α, β in order to keep track of the computations. We recall that $\alpha = 1$ and $\beta = \frac{1}{4}$.

5.2.1 Well-posedness and main properties of the scheme

Given the two measures $\rho^{n-1}, \rho^{n-2} \in \mathcal{P}(\Omega)$ and the two real numbers α, β, let us define for compactness of notation the functional $\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho)$ as

$$\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho) := \alpha \mathcal{W}_2^2(\rho, \rho^{n-1}) - \beta \mathcal{W}_2^2(\rho, \rho^{n-2}) \,.$$

We first show that the minimizer of the functional $\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho)$ is always well-defined. In this sense, the following lemma, which is a particular case of [93, Theorem 3.4], is fundamental. It states that $\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \cdot)$ is convex along the generalized geodesic curve centered in ρ^{n-1} (see Section 1.2.1).

Lemma 5.1. *For any two given measures $\omega_0, \omega_1 \in \mathcal{P}(\Omega)$, denote by $\omega : [0,1] \to \mathcal{P}(\Omega), \omega(0) = \omega_0, \omega(1) = \omega_1$, the generalized geodesic curve joining them and centered in ρ^{n-1}. Then, $\forall t \in [0,1]$, it holds:*

$$\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \omega(t)) \leq (1-t)\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \omega_0) + t\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \omega_1) - (\alpha-\beta)t(1-t)\mathcal{W}_2^2(\omega_0, \omega_1). \tag{5.12}$$

Theorem 5.2 (Metric extrapolation)**.** *There exists a unique solution ρ_e^{n-1} to problem (5.11). Moreover, it holds*

$$(\alpha - \beta)\mathcal{W}_2^2(\rho, \rho_e^{n-1}) + \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho_e^{n-1}) \leq \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho) \,, \quad \forall \rho \in \mathcal{P}(\Omega) \,, \tag{5.13}$$

and

$$\mathcal{W}_2^2(\rho_e^{n-1}, \rho^{n-1}) \leq \Big(\frac{\beta}{\alpha - \beta}\Big)^2 \mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) \,, \tag{5.14}$$

$$\mathcal{W}_2^2(\rho_e^{n-1}, \rho^{n-2}) \leq \Big(\frac{\alpha}{\alpha - \beta}\Big)^2 \mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) \,. \tag{5.15}$$

Proof. First of all, the functional $\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho)$ is bounded from below: using indeed the inequality

$$\mathcal{W}_2^2(\rho, \rho^{n-2}) \leq \Big(1 + \frac{1}{c}\Big)\mathcal{W}_2^2(\rho, \rho^{n-1}) + (1+c)\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) \,,$$

(which derives from the triangular inequality and Young's inequality) with $c = \frac{\beta}{\alpha-\beta} > 0$, we get

$$\begin{aligned} \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho) &\geq \Big(\alpha - \beta\Big(1 + \frac{1}{c}\Big)\Big)\mathcal{W}_2^2(\rho, \rho^{n-1}) - \beta(1+c)\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) \\ &= -\frac{\alpha\beta}{\alpha - \beta}\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) \,. \end{aligned} \tag{5.16}$$

Existence of the solution is then immediate, as the space $\mathcal{P}(\Omega)$ is compact in the weak-* topology and the functional is continuous for the same topology. Uniqueness derives immediately from Lemma 5.1.

Inequality (5.13) derives again from Lemma 5.1. For $\rho \in \mathcal{P}(\Omega)$, consider a measure curve ω as in Lemma 5.1, with $\omega_0 = \rho_e^{n-1}$ and $\omega_1 = \rho$. By optimality of ρ_e^{n-1}, it holds

$$\begin{aligned} 0 &\leq \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \omega(t)) - \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho_e^{n-1}) \\ &= t\big(\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho) - \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho_e^{n-1})\big) - (\alpha - \beta)t(1-t)\mathcal{W}_2^2(\rho, \rho_e^{n-1})\,, \end{aligned}$$

which, dividing by t and taking the limit $t \to 0$, gives (5.13). Substituting ρ^{n-1}, ρ^{n-2} in (5.13) and using (5.16), we obtain respectively the two estimates (5.14)-(5.15). □

From the inequality (5.13) we understand that problem (5.10) is a lower bound for the original formulation (5.5). Let us assume that the geodesic $\frac{4}{3}$-extrapolation from ρ^{n-2} to ρ^{n-1} exists. It is given by $(\pi_e)_\# \gamma^{n-1}$, where γ^{n-1} is the optimal transport plan between ρ^{n-1} and ρ^{n-2} and $\pi_e = \frac{1}{\alpha-\beta}(\alpha\pi_2 - \beta\pi_1)$, the maps π_1 and π_2 standing for the canonical projections on the first and second component of the space $\Omega \times \Omega$. Recalling that along the geodesic extrapolation particles move straight with constant speed and the total distance is therefore proportional to the travel time (see Sections 1.1.2-1.1.2), it holds

$$\begin{aligned} &\alpha\mathcal{W}_2^2((\pi_e)_\#\gamma^{n-1}, \rho^{n-1}) - \beta\mathcal{W}_2^2((\pi_e)_\#\gamma^{n-1}, \rho^{n-2}) = \\ &= \alpha\Big(\frac{\beta}{\alpha-\beta}\Big)^2\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) - \beta\Big(\frac{\alpha}{\alpha-\beta}\Big)^2\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) = -\frac{\alpha\beta}{\alpha-\beta}\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2})\,, \end{aligned}$$

which is the minimum for problem (5.11) by the lower bound (5.16). The geodesic $\frac{4}{3}$-extrapolation solves problem (5.11). This latter may not always exist, as the particles may bump into each others or the domain. However, the solution ρ_e^n to problem (5.11) does and it is in this sense its metric generalization. The estimates (5.14)-(5.15), which express the stability of this extrapolation, will be fundamental in the sequel. If the geodesic extrapolation exists, the equality holds in (5.14)-(5.15).

Let us consider now problem (5.10). Again for compactness of notation, we define the objective functional of problem (5.10) as:

$$\mathcal{G}(\rho^{n-1}, \rho^{n-2}; \rho) := \frac{1}{\tau}(\alpha - \beta)\mathcal{W}_2^2(\rho, \rho_e^{n-1}) + \mathcal{E}(\rho)\,.$$

Thanks to the assumptions on the energy, $\mathcal{G}(\rho^{n-1}, \rho^{n-2}; \rho)$ is lower semi-continuous with respect to the weak-* topology, bounded from below and strictly convex in the generalized geodesic sense. Then the next theorem follows quite easily.

Theorem 5.3. *At each step n, there exists a unique solution ρ^n to problem (5.10).*

Proof. The proof of existence is again an application of the direct method of the calculus of variations. We recall that the Wasserstein distance is 2-convex along generalized geodesics [3, Section 9.2]. Then the functional $\mathcal{G}(\rho^{n-1}, \rho^{n-2}; \rho)$ is as well 2-convex along generalized geodesics, thanks to the assumption on $\mathcal{E}$. The uniqueness of the minimizer follows. □

Differently from the JKO scheme, but analogously to the VIM and the original BDF2 schemes, at each step the energy is not necessarily diminished.

Lemma 5.4. *At each step n, the solution ρ^n satisfies the following inequality*

$$\frac{\alpha-2\beta}{\tau}\mathcal{W}_2^2(\rho^n,\rho^{n-1})+\mathcal{E}(\rho^n)\leq\frac{\beta}{\tau}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2})+\mathcal{E}(\rho^{n-1})\,. \tag{5.17}$$

Proof. By optimality of ρ^n and using (5.14), we can write

$$\begin{aligned}\frac{1}{\tau}(\alpha-\beta)\mathcal{W}_2^2(\rho^n,\rho_e^{n-1})+\mathcal{E}(\rho^n)&\leq\frac{1}{\tau}(\alpha-\beta)\mathcal{W}_2^2(\rho^{n-1},\rho_e^{n-1})+\mathcal{E}(\rho^{n-1})\\&\leq\frac{1}{\tau}\frac{\beta^2}{\alpha-\beta}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2})+\mathcal{E}(\rho^{n-1})\,.\end{aligned}$$

Using the inequality

$$\mathcal{W}_2^2(\rho^n,\rho^{n-1})\leq(1+\frac{1}{c})\mathcal{W}_2^2(\rho^n,\rho_e^{n-1})+(1+c)\mathcal{W}_2^2(\rho^{n-1},\rho_e^{n-1})\,,$$

for $c=\frac{\alpha-2\beta}{\beta}$ and again (5.14), we can estimate the left-hand side from below as

$$\begin{aligned}\frac{\alpha-\beta}{\tau}\mathcal{W}_2^2(\rho^n,\rho_e^{n-1})+\mathcal{E}(\rho^n)&\geq\frac{\alpha-\beta}{\tau}\left(\frac{c}{c+1}\mathcal{W}_2^2(\rho^n,\rho^{n-1})-c\mathcal{W}_2^2(\rho^{n-1},\rho_e^{n-1})\right)+\mathcal{E}(\rho^n)\\&\geq\frac{\alpha-2\beta}{\tau}\mathcal{W}_2^2(\rho^n,\rho^{n-1})-\frac{\beta(\alpha-2\beta)}{\tau(\alpha-\beta)}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2})+\mathcal{E}(\rho^n)\,.\end{aligned}$$

Rearranging, we obtain (5.17). □

Although the energy does not strictly decrease at each step, its possible growth is controlled. The energy is in any case bounded by a decreasing function as the inequality (5.19) shows in the following theorem, which also provides the usual classical estimate for gradient flows.

Theorem 5.5. *Given a fixed time horizon $T>0$, assuming there exists a constant $C_1>0$ such that $\mathcal{W}_2^2(\rho^1,\rho^0)\leq C_1\tau$ and assuming further $\mathcal{E}(\rho^1)<\infty$, it holds:*

$$\frac{1}{\tau}\sum_{n=0}^{N_\tau}\mathcal{W}_2^2(\rho^n,\rho^{n-1})\leq C_2 \tag{5.18}$$

$\forall\tau>0$, $N_\tau=\frac{T}{\tau}$, with the constant C_2 independent of τ.

Proof. Summing over n the inequality (5.17) we obtain

$$\frac{\alpha-3\beta}{\tau}\sum_{n=0}^{N}\mathcal{W}_2^2(\rho^n,\rho^{n-1})\leq\mathcal{E}(\rho^1)-\mathcal{E}(\rho^N)+\frac{\beta}{\tau}\mathcal{W}_2^2(\rho^1,\rho^0)\,, \tag{5.19}$$

Then, since $\alpha-3\beta>0$ for $\alpha=1,\beta=\frac{1}{4}$, and thanks to the lower bound on the energy, we have

$$\frac{1}{\tau}\sum_{n=0}^{N}\mathcal{W}_2^2(\rho^n,\rho^{n-1})\leq\frac{1}{\alpha-3\beta}\Big(\mathcal{E}(\rho^1)-\mathcal{E}(\rho^N)+\beta C_1\Big)\leq C_2\,.$$

□

Remark 5.6. *The condition $\mathcal{W}_2^2(\rho^1,\rho^0)\leq C_1\tau$ is satisfied for example considering the first step to be performed with a JKO step and ρ^0 such that $\mathcal{E}(\rho^0)<\infty$.*

5.2.2 Convergence towards the Fokker-Planck equation

The objective of the present section is to validate our modified BDF2 approach by showing the convergence of the discrete flow generated via the scheme (5.10)-(5.11) to the linear Fokker-Planck equation. We recall the form of the model problem and the energy functional that generates it: given a Lipschitz continuous exterior potential $V \in W^{1,\infty}(\Omega)$, the equation writes

$$\partial_t \varrho = \Delta \varrho + \nabla \cdot (\varrho \nabla V) \quad \text{in } [0,T) \times \Omega \,, \tag{5.20}$$

complemented with no-flux boundary conditions $\nabla \varrho + \varrho \nabla V \cdot \boldsymbol{n} = 0$ on $\partial\Omega$ and an initial condition $\varrho(0,\cdot) = \rho_0 \in \mathcal{P}(\Omega)$. Equation (5.20) represents a Wasserstein gradient flow with respect to the energy functional

$$\mathcal{E}(\rho) = \mathcal{U}(\rho) + \int_\Omega \rho V \,, \tag{5.21}$$

where the internal energy $\mathcal{U}$, the entropy, is defined as

$$\mathcal{U}(\rho) = \begin{cases} \int_\Omega \rho \log \rho & \text{if } \rho \text{ absolutely continuous} \,, \\ +\infty & \text{else} \,. \end{cases} \tag{5.22}$$

The functional (5.21) is lower semi-continuous with respect to the weak-* topology [115, Proposition 7.7], is bounded from below and convex along generalized geodesics [3, Proposition 9.3.9]. Therefore the measure ρ^n solution to problem (5.10) exists and is unique at each step n, and it is furthermore absolutely continuous. We consider a proper initial condition ρ^0, $\mathcal{E}(\rho^0) < \infty$, and the first measure ρ^1 to be generated via a JKO step with time parameter τ:

$$\rho^1 = \operatorname*{argmin}_{\rho \in \mathcal{P}(\Omega)} \frac{1}{2\tau} \mathcal{W}_2^2(\rho, \rho^0) + \mathcal{E}(\rho) \,.$$

Although the discrete flow does not move by strictly minimizing the energy at each step, we want to show that it converges to the maximal slope curve of $\mathcal{E}$. We will resort to the same ideas developed in the original work [73].

Relying on the estimate (5.18), the compactness arguments for obtaining a limit curve are rather standard. We introduce the two density curves

$$\rho_\tau(t) = \sum_{n=1}^{N} \rho^{n-1} \mathbb{1}_{(t^{n-1},t^n]} \,, \quad \rho_\tau(0) = \rho^0 \,,$$
$$\tilde{\rho}_\tau(t) = \sum_{n=1}^{N} \tilde{\rho}^{n} \mathbb{1}_{(t^{n-1},t^n]} \,, \quad \tilde{\rho}_\tau(0) = \rho^0 \,,$$

with $\tilde{\rho}^n$ geodesic between ρ^{n-1} and ρ^n on the time interval $[t^{n-1}, t^n]$ (that is the time interval in problem (1.15) is $[t^{n-1}, t^n]$ and not $[0,1]$). This means that there exist a vector field $\tilde{\boldsymbol{v}}_\tau$ which solves the continuity equation

$$\partial_t \tilde{\rho}_\tau + \nabla \cdot (\tilde{\rho}_\tau \tilde{\boldsymbol{v}}_\tau) = 0 \quad \text{in } [0,1] \times \Omega \,.$$

It is defined as the interpolation of the vector fields $\tilde{\boldsymbol{v}}^n(t,\boldsymbol{x})$ defined on each interval $[t^{n-1},t^n]$ as

$$\tilde{\boldsymbol{v}}^n(t,\boldsymbol{x}) = \left(\frac{\mathrm{T}^n - \mathrm{Id}}{\tau}\right) \circ \left(\frac{(t^n - t)}{\tau}\mathrm{Id}(\boldsymbol{x}) + \frac{(t - t^{n-1})}{\tau}\mathrm{T}^n(\boldsymbol{x})\right)^{-1}.$$

The vector field $\frac{\mathrm{T}^n - \mathrm{Id}}{\tau}$ is the constant velocity of the particles of ρ^{n-1} going to ρ^n. On each interval $[t^{n-1},t^n]$ it holds:

$$\mathcal{W}_2^2(\rho^n,\rho^{n-1}) = \tau \int_{t^{n-1}}^{t^n}\int_\Omega \tilde{\rho}_\tau |\tilde{\boldsymbol{v}}_\tau|^2 .$$

The curve ρ_τ is a piecewise constant measure curve whereas $\tilde{\rho}_\tau$ is a (absolutely) continuous one, interpolation of the discrete densities.

Proposition 5.7. *The sequence $(\rho_\tau)_\tau$ converges uniformly in the $\mathcal{W}_2$ distance to an absolutely continuous measure curve ϱ.*

Proof. The sequence of curves $(\tilde{\rho}_\tau)_{\tau\in\mathbb{R}_+}$, defined from $[0,T]$ to the compact Wasserstein space, is uniformly Hölder continuous. Indeed, for any $r,s\in[0,T], s>r$, denote N_r, N_s the two integers such that $r\in[t^{N_r},t^{N_r+1}], s\in[t^{N_s-1},t^{N_s}]$. Let us call $(\rho_*,\boldsymbol{v}_*)$ the optimal density displacement and velocity field between $\tilde{\rho}_\tau(r)$ and $\tilde{\rho}_\tau(s)$. We recall that along this curve the kinetic energy is constant in time. Since on the contrary $(\tilde{\rho}_\tau,\tilde{\boldsymbol{v}}_\tau)$ is not optimal between these two measures, it holds

$$\begin{aligned}
\mathcal{W}_2(\tilde{\rho}_\tau(s),\tilde{\rho}_\tau(r)) &= \int_r^s \left(\int_\Omega \rho_*|\boldsymbol{v}_*|^2\right)^{\frac12} \le (s-r)^{\frac12}\left(\int_r^s\int_\Omega \rho_*|\boldsymbol{v}_*|^2\right)^{\frac12} \\
&\le (s-r)^{\frac12}\left(\int_r^s\int_\Omega \tilde{\rho}_\tau|\tilde{\boldsymbol{v}}_\tau|^2\right)^{\frac12} \le (s-r)^{\frac12}\left(\sum_{n=N_r+1}^{N_s}\int_{t^{n-1}}^{t^n}\int_\Omega \tilde{\rho}_\tau|\tilde{\boldsymbol{v}}_\tau|^2\right)^{\frac12} \\
&= (s-r)^{\frac12}\left(\sum_{n=N_r+1}^{N_s}\frac{1}{\tau}\mathcal{W}_2^2(\rho^n,\rho^{n-1})\right)^{\frac12} \le C(s-r)^{\frac12}
\end{aligned} \tag{5.23}$$

where in the last inequality we used the estimate (5.18). By the (generalized) Ascoli-Arzelà theorem, the sequence converges uniformly in $\mathcal{W}_2$, up to a subsequence, to a limit curve ϱ. As the inequality (5.23) passes to the limit, ϱ is as well an absolutely continuous curve in the Wasserstein space. Finally, for any $r\in[0,T]$,

$$\mathcal{W}_2(\rho_\tau(r),\tilde{\rho}_\tau(r)) \le \int_{t^{N_r}}^{t^{N_r+1}}\int_\Omega \tilde{\rho}_\tau|\tilde{\boldsymbol{v}}_\tau|^2 \le C\sqrt{\tau},$$

by the same computations. The piecewise continuous curve ρ_τ converges uniformly with order $\sqrt{\tau}$ to the same limit curve ϱ.

□

To characterize the limit curve ϱ we have to rely on the optimality conditions of the objective functionals $\mathcal{G}$ and $\mathcal{F}$. Consider an absolutely continuous measure ρ and a smooth

vector field $\boldsymbol{\xi}$ tangent to the boundary of Ω. We define ω as the absolutely continuous curve solution to

$$\partial_s \omega + \nabla \cdot (\omega \boldsymbol{\xi}) = 0\,, \quad \text{in } (-\delta, \delta) \times \Omega\,, \quad \omega(0) = \rho\,, \tag{5.24}$$

for $\delta \in \mathbb{R}, \delta > 0$. We take variations along curves defined in this way. The hypothesis of absolute continuity of ρ is essential in order to define $\omega(s)$ at each time $s \in (-\delta, \delta)$ as pushforward of ρ.

Lemma 5.8. *Consider two measures $\rho, \nu \in \mathcal{P}(\Omega)$. Assume ρ absolutely continuous and denote by γ the optimal transport plan between them. For any $\boldsymbol{\xi} \in C^\infty(\Omega, \mathbb{R}^d)$ such that $\boldsymbol{\xi} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$, consider curves defined as in (5.24) centered in ρ. It holds:*

$$\frac{\mathrm{d}\mathcal{W}_2^2(\omega(s), \nu)}{\mathrm{d}s}|_{s=0} = 2 \int_{\Omega\times\Omega} \langle \boldsymbol{x} - \boldsymbol{y}, \boldsymbol{\xi}(\boldsymbol{x}) \rangle \, \mathrm{d}\gamma(\boldsymbol{x}, \boldsymbol{y})\,, \tag{5.25}$$

and

$$\frac{\mathrm{d}\mathcal{E}(\omega(s))}{\mathrm{d}s}|_{s=0} = -\int_\Omega (\nabla \cdot \boldsymbol{\xi}(\boldsymbol{x}))\rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \langle \nabla V(\boldsymbol{x}), \boldsymbol{\xi}(\boldsymbol{x}) \rangle \rho(\boldsymbol{x})\mathrm{d}\boldsymbol{x}\,. \tag{5.26}$$

Proof. See [73, Theorem 5.1]. □

We want to write the optimality conditions of the two problems (5.11)-(5.10). As the measure ρ^n is absolutely continuous, we can apply Lemma 5.8 for (5.10). Nevertheless, the measure ρ_e^{n-1} solution to (5.11) is not necessarily absolutely continuous, even though ρ^{n-1} and ρ^{n-2} are. Constructing a counter-example is simple, thanks to the equivalence between problem (5.11) and the geodesic extrapolation, when this latter exists. We cannot therefore apply directly formula (5.25). We can however write the optimality conditions of problem (5.11) by a regularization argument.

Lemma 5.9. *At each step n, let us denote by $\gamma_e^{n-1,1}$ and $\gamma_e^{n-1,2}$ respectively the optimal transport plans from ρ^{n-1} to ρ_e^{n-1} and from ρ^{n-2} to ρ_e^{n-1}. Then it necessarily holds*

$$\alpha \int_{\Omega\times\Omega} \langle \boldsymbol{x}_e - \boldsymbol{y}_1, \boldsymbol{\xi}(\boldsymbol{x}_e) \rangle \mathrm{d}\gamma_e^{n-1,1}(\boldsymbol{x}_e, \boldsymbol{y}_1) - \beta \int_{\Omega\times\Omega} \langle \boldsymbol{x}_e - \boldsymbol{y}_2, \boldsymbol{\xi}(\boldsymbol{x}_e) \rangle \mathrm{d}\gamma_e^{n-1,2}(\boldsymbol{x}_e, \boldsymbol{y}_2) = 0\,, \tag{5.27}$$

for any $\boldsymbol{\xi} \in C^\infty(\Omega, \mathbb{R}^d)$ such that $\boldsymbol{\xi} \cdot \boldsymbol{n} = 0$ on $\partial\Omega$.

Proof. In order to prove the result we construct a sequence of approximated smooth variational problems and pass to the limit in the optimality conditions. Let us define

$$\mathcal{F}_\varepsilon(\rho^{n-1}, \rho^{n-2}; \rho) := \mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho) + \varepsilon \mathcal{U}(\rho)\,, \tag{5.28}$$

and the regularized problem

$$\inf_{\rho \in \mathcal{P}(\Omega)} \mathcal{F}_\varepsilon(\rho^{n-1}, \rho^{n-2}; \rho)\,, \tag{5.29}$$

for $\varepsilon \in \mathbb{R}, \varepsilon > 0$. Problem (5.29) admits a solution $\rho_{e,\varepsilon}^{n-1}$ which is now absolutely continuous. The proof is the same as in Theorem 5.2. By applying Lemma 5.8, we can write down the necessary optimality conditions of problem (5.29):

$$\begin{aligned}\frac{\mathrm{d}\mathcal{F}_\varepsilon(\rho^{n-1}, \rho^{n-2}; \omega(s))}{\mathrm{d}s}|_{s=0} &= 2\alpha \int_{\Omega\times\Omega} \langle \boldsymbol{x}_e - \boldsymbol{y}_1, \boldsymbol{\xi}(\boldsymbol{x}_e) \rangle \mathrm{d}\gamma_{e,\varepsilon}^{n-1,1}(\boldsymbol{x}_e, \boldsymbol{y}_1) \\ &- 2\beta \int_{\Omega\times\Omega} \langle \boldsymbol{x}_e - \boldsymbol{y}_2, \boldsymbol{\xi}(\boldsymbol{x}_e) \rangle \mathrm{d}\gamma_{e,\varepsilon}^{n-1,2}(\boldsymbol{x}_e, \boldsymbol{y}_2) - \varepsilon \int_\Omega (\nabla \cdot \boldsymbol{\xi}(\boldsymbol{x})) \rho_{e,\varepsilon}^{n-1}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 0\,,\end{aligned} \tag{5.30}$$

for any $\boldsymbol{\xi} \in C^\infty(\Omega;\mathbb{R}^d)$ tangent to the boundary, where we now denote by $\gamma_{e,\varepsilon}^{n-1,1}$ and $\gamma_{e,\varepsilon}^{n-1,2}$ respectively the optimal transport plans from ρ^{n-1} to $\rho_{e,\varepsilon}^{n-1}$ and from ρ^{n-2} to $\rho_{e,\varepsilon}^{n-1}$.

We want to show that problem (5.29) Γ-converges towards problem (5.11) in order to pass to the limit in the optimality conditions. By the lower semi-continuity of $\mathcal{F}$ and the fact that $\mathcal{U}$ is non-negative, the Γ-lim inf is obvious,

$$\mathcal{F}(\rho^{n-1},\rho^{n-2};\rho) \leq \liminf_\varepsilon \mathcal{F}(\rho^{n-1},\rho^{n-2};\rho_\varepsilon) \leq \liminf_\varepsilon \mathcal{F}_\varepsilon(\rho^{n-1},\rho^{n-2};\rho_\varepsilon)\,,$$

for any $\rho_\varepsilon \to \rho$ in the Wasserstein sense. Concerning the Γ-lim sup, if $\mathcal{U}(\rho_e^{n-1}) < +\infty$ we can take $\rho_\varepsilon = \rho_e^{n-1}$ as recovering sequence. Otherwise, we take a sequence of absolutely continuous measure ρ_ε converging to ρ_e^{n-1} in the Wasserstein sense. The set of absolutely continuous measures is dense in $\mathcal{P}(\Omega)$ for the weak-* topology, justifying the existence of such sequence. Since $\mathcal{U}(\rho_e^{n-1}) = \infty$, up to a reparametrization we can assume that the entropy is increasing and that

$$\mathcal{U}(\rho_\varepsilon) \leq \frac{C}{\sqrt{\varepsilon}}\,,$$

for a constant C independent of ε. Then it holds:

$$\limsup_\varepsilon \mathcal{F}_\varepsilon(\rho^{n-1},\rho^{n-2};\rho_\varepsilon) = \lim_\varepsilon \mathcal{F}_\varepsilon(\rho^{n-1},\rho^{n-2};\rho_\varepsilon) = \mathcal{F}(\rho^{n-1},\rho^{n-2};\rho_e^{n-1})\,.$$

Therefore problem (5.29) Γ-converges to problem (5.11), which implies that $\rho_{e,\varepsilon}^{n-1} \to \rho_e^{n-1}$ in the Wasserstein sense. By the stability of optimal transport plans [120, Exercise 2.17]

$$\gamma_{e,\varepsilon}^{n-1,1} \longrightarrow \gamma_e^{n-1,1}\,, \quad \gamma_{e,\varepsilon}^{n-1,2} \longrightarrow \gamma_e^{n-1,2}\,,$$

for the weak-* convergence of measures. As the vector field $\boldsymbol{\xi}$ is smooth, passing to the limit in (5.30) we obtain (5.27).

□

We can now prove convergence of the sequence of curves $(\rho_\tau)_\tau$ towards a distributional solution of equation (5.20).

Theorem 5.10. *For all $\varphi \in C_c^\infty([0,T)\times\Omega)$, the limit curve ϱ satisfies:*

$$-\int_0^T\int_\Omega \partial_t\varphi\varrho - \int_\Omega \varphi(0)\varrho(0)\mathrm{d}\boldsymbol{x} - \int_0^T\int_\Omega \Delta\varphi\varrho + \int_0^T\int_\Omega \langle\nabla V,\nabla\varphi\rangle\varrho = 0\,. \tag{5.31}$$

Proof. Consider a smooth function $\varphi \in C_c^\infty([0,T]\times\Omega)$ such that $\nabla\varphi\cdot\boldsymbol{n} = 0$ on $\partial\Omega$. We define the sequence $(\varphi^n)_n \subset C_c^\infty(\Omega)$ as $\varphi^n = \varphi(t^n,\cdot)$. At each step $n > 2$, the derivatives (5.25)-(5.26) hold with $\xi = \nabla\varphi^{n-2}$. Consider then a curve ω defined as in (5.24) centered in ρ^n. We denote $\bar\gamma^n$ the optimal transport plan between ρ_e^{n-1} and ρ^n. Owing to (5.25)-(5.26), the necessary optimality condition of problem (5.10) for the measure ρ^n is:

$$\begin{aligned}\frac{\mathrm{d}\mathcal{G}(\rho^{n-1},\rho^{n-2};\omega(s))}{\mathrm{d}s}|_{s=0} &= \frac{2}{\tau}(\alpha-\beta)\int_{\Omega\times\Omega}\langle\boldsymbol{x}-\boldsymbol{x}_e,\nabla\varphi^{n-2}(\boldsymbol{x})\rangle\mathrm{d}\bar\gamma^n(\boldsymbol{x},\boldsymbol{x}_e)\\ &\quad -\int_\Omega(\Delta\varphi^{n-2}(\boldsymbol{x}))\rho^n(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega\langle\nabla V(\boldsymbol{x}),\nabla\varphi^{n-2}(\boldsymbol{x})\rangle\rho^n(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 0\,.\end{aligned} \tag{5.32}$$

Thanks to Proposition 5.7 and the regularity of φ, we immediately have

$$\left| \sum_{n=2}^{N} \tau \left(- \int_\Omega (\Delta \varphi^{n-2}) \rho^n + \int_\Omega \langle \nabla V, \nabla \varphi^{n-2} \rangle \rho^n \right) \right. \\ \left. - \left(- \int_0^T \int_\Omega \Delta \varphi \varrho + \int_0^T \int_\Omega \langle \nabla V, \nabla \varphi \rangle \varrho \right) \right| \longrightarrow 0 \,,$$

for $\tau \to 0$. In order to prove that the measure ϱ is a distributional solution of equation (5.20) we need to show that

$$I_1 = \left| \sum_{n=2}^{N} 2(\alpha - \beta) \int_{\Omega \times \Omega} \langle \boldsymbol{x} - \boldsymbol{x}_e, \nabla \varphi^{n-2}(\boldsymbol{x}) \rangle \mathrm{d}\gamma^n(\boldsymbol{x}, \boldsymbol{x}_e) \right. \\ \left. - \left(- \int_0^T \int_\Omega \partial_t \varphi \varrho - \int_\Omega \varphi(0) \varrho(0) \mathrm{d}\boldsymbol{x} \right) \right| \longrightarrow 0 \,,$$

as well. We can bound the latter quantity as $I_1 \leq I_2 + I_3$, where

$$I_2 = \left| \sum_{n=2}^{N} 2(\alpha - \beta) \int_{\Omega \times \Omega} \langle \boldsymbol{x} - \boldsymbol{x}_e, \nabla \varphi^{n-2}(\boldsymbol{x}) \rangle \mathrm{d}\gamma^n(\boldsymbol{x}, \boldsymbol{x}_e) \right. \\ \left. - 2 \int_\Omega ((\alpha - \beta)\rho^n - \alpha \rho^{n-1} + \beta \rho^{n-2}) \varphi^{n-2} \right| ,$$

and

$$I_3 = \left| \sum_{n=2}^{N} 2 \int_\Omega ((\alpha - \beta)\rho^n - \alpha \rho^{n-1} + \beta \rho^{n-2}) \varphi^{n-2} - \left(- \int_0^T \int_\Omega \partial_t \varphi \varrho - \int_\Omega \varphi(0) \varrho(0) \right) \right| .$$

Integrating by part the discrete derivative

$$\sum_{n=2}^{N} 2 \int_\Omega ((\alpha - \beta)\rho^n - \alpha \rho^{n-1} + \beta \rho^{n-2}) \varphi^{n-2} = \\ = \sum_{n=2}^{N} 2 \int_\Omega ((\alpha - \beta)\varphi^{n-2} - \alpha \varphi^{n-1} + \beta \varphi^n) \rho^n + \beta \rho^0 \varphi^0 + (\beta \varphi^1 - \alpha \varphi^0) \rho^1 \\ = \sum_{n=2}^{N} \int_\Omega (\frac{3}{2} \varphi^{n-2} - 2\varphi^{n-1} + \frac{1}{2} \varphi^n) \rho^n + (\varphi^1 - \varphi^0) \rho^1 + \frac{1}{2} \rho^0 \varphi^0 - \rho^1 \varphi^0 - \frac{1}{2} \varphi^1 \rho^1 \,,$$

we can see that, thanks to the smoothness of the function φ and Proposition 5.7, $I_3 \leq C\tau$ for some constant C independent of τ. Let us focus then on the term I_2.

At each step n, adding and subtracting $2(\alpha-\beta)\int_\Omega(\rho^n-\rho_e^{n-1})\varphi^{n-2}$, we can write:

$$\begin{aligned}
&\left|2(\alpha-\beta)\int_{\Omega\times\Omega}\langle \boldsymbol{x}-\boldsymbol{x}_e,\nabla\varphi^{n-2}(\boldsymbol{x})\rangle \mathrm{d}\gamma^n(\boldsymbol{x},\boldsymbol{x}_e)-2\int_\Omega((\alpha-\beta)\rho^n-\alpha\rho^{n-1}+\beta\rho^{n-2})\varphi^{n-2}\right|\\
&\leq 2(\alpha-\beta)\left|\int_{\Omega\times\Omega}\langle \boldsymbol{x}-\boldsymbol{x}_e,\nabla\varphi^{n-2}(\boldsymbol{x})\rangle \mathrm{d}\gamma^n(\boldsymbol{x},\boldsymbol{x}_e)-\int_\Omega(\rho^n-\rho_e^{n-1})\varphi^{n-2}\right|\\
&\quad+2\left|\int_\Omega(\alpha\rho^{n-1}-\beta\rho^{n-2}-(\alpha-\beta)\rho_e^{n-1})\varphi^{n-2}\right|\\
&=2(\alpha-\beta)I_4+2I_5\,.
\end{aligned}$$

Rewriting

$$\int_\Omega(\rho^n-\rho_e^{n-1})\varphi^{n-2}=\int_{\Omega\times\Omega}(\varphi^{n-2}(\boldsymbol{x})-\varphi^{n-2}(\boldsymbol{x}_e))\mathrm{d}\bar\gamma^n(\boldsymbol{x},\boldsymbol{x}_e)\,,$$

we can bound I_4 as

$$\begin{aligned}
I_4&=\left|\int_{\Omega\times\Omega}\varphi^{n-2}(\boldsymbol{x})-\varphi^{n-2}(\boldsymbol{x}_e)-\langle \boldsymbol{x}-\boldsymbol{x}_e,\nabla\varphi^{n-2}(\boldsymbol{x})\rangle\mathrm{d}\bar\gamma^n(\boldsymbol{x},\boldsymbol{x}_e)\right|\\
&\leq\frac{1}{2}||\mathrm{Hess}(\varphi^{n-2})||_\infty\left(\int_{\Omega\times\Omega}|\boldsymbol{x}-\boldsymbol{x}_e|^2\mathrm{d}\bar\gamma^n(\boldsymbol{x},\boldsymbol{x}_e)\right)\\
&=\frac{1}{2}||\mathrm{Hess}(\varphi^{n-2})||_\infty\mathcal{W}_2^2(\rho^n,\rho_e^{n-1})\\
&\leq||\mathrm{Hess}(\varphi^{n-2})||_\infty\left(\mathcal{W}_2^2(\rho^n,\rho^{n-1})+\mathcal{W}_2^2(\rho^{n-1},\rho_e^{n-1})\right)\\
&\leq||\mathrm{Hess}(\varphi^{n-2})||_\infty\Big(\mathcal{W}_2^2(\rho^n,\rho^{n-1})+\frac{\beta^2}{(\alpha-\beta)^2}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2})\Big)\,,
\end{aligned}$$

where we used the estimate (5.14). In order to bound the term I_5, we use the optimality condition of problem (5.11), namely equation (5.27). Rewriting

$$\begin{aligned}
\int_\Omega(\alpha\rho^{n-1}-\beta\rho^{n-2}-(\alpha-\beta)\rho_e^{n-1})\varphi^{n-2}=&\alpha\int_{\Omega\times\Omega}(\varphi^{n-2}(\boldsymbol{y}_1)-\varphi^{n-2}(\boldsymbol{x}_e))\mathrm{d}\gamma_e^{n-1,1}(\boldsymbol{x}_e,\boldsymbol{y}_1)\\
&-\beta\int_{\Omega\times\Omega}(\varphi^{n-2}(\boldsymbol{y}_2)-\varphi^{n-2}(\boldsymbol{x}_e))\mathrm{d}\gamma_e^{n-1,2}(\boldsymbol{x}_e,\boldsymbol{y}_2)\,,
\end{aligned}$$

we can bound I_5 as

$$\begin{aligned}
&I_5=\\
&\left|\alpha\int_{\Omega\times\Omega}(\varphi^{n-2}(\boldsymbol{y}_1)-\varphi^{n-2}(\boldsymbol{x}_e))\mathrm{d}\gamma_e^{n-1,1}(\boldsymbol{x}_e,\boldsymbol{y}_1)-\beta\int_{\Omega\times\Omega}(\varphi^{n-2}(\boldsymbol{y}_2)-\varphi^{n-2}(\boldsymbol{x}_e))\mathrm{d}\gamma_e^{n-1,2}(\boldsymbol{x}_e,\boldsymbol{y}_2)\right|\\
&\leq\left|\alpha\int_{\Omega\times\Omega}\langle \boldsymbol{x}_1-\boldsymbol{x}_e,\nabla\varphi^{n-2}(\boldsymbol{x}_e)\rangle\mathrm{d}\gamma_e^{n-1,1}(\boldsymbol{x}_e,\boldsymbol{y}_1)-\beta\int_{\Omega\times\Omega}\langle \boldsymbol{x}_2-\boldsymbol{x}_e,\nabla\varphi^{n-2}(\boldsymbol{x}_e)\rangle\mathrm{d}\gamma_e^{n-1,2}(\boldsymbol{x}_e,\boldsymbol{y}_2)\right|\\
&\quad+\frac{1}{2}||\mathrm{Hess}(\varphi^{n-2})||_\infty\left(\alpha\int_{\Omega\times\Omega}|\boldsymbol{y}_1-\boldsymbol{x}_e|^2\mathrm{d}\gamma_e^{n-1,1}(\boldsymbol{x}_e,\boldsymbol{y}_1)+\beta\int_{\Omega\times\Omega}|\boldsymbol{y}_2-\boldsymbol{x}_e|^2\mathrm{d}\gamma_e^{n-1,2}(\boldsymbol{x}_e,\boldsymbol{y}_2)\right)\\
&=\frac{1}{2}||\mathrm{Hess}(\varphi^{n-2})||_\infty\left(\alpha\mathcal{W}_2^2(\rho_e^{n-1},\rho^{n-1})+\beta\mathcal{W}_2^2(\rho_e^{n-1},\rho^{n-2})\right)\\
&\leq\frac{1}{2}||\mathrm{Hess}(\varphi^{n-2})||_\infty\left(\frac{\alpha\beta^2+\beta\alpha^2}{(\alpha-\beta)^2}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2})\right)\,,
\end{aligned}$$

where we used again the smoothness of φ, the optimality conditions (5.27), and the estimates (5.14)-(5.15). Finally, we can bound I_2 as

$$\begin{aligned} I_2 &\leq \tau ||\mathrm{Hess}(\varphi)||_\infty \sum_{n=2}^{N} \left(2(\alpha-\beta)\mathcal{W}_2^2(\rho^n,\rho^{n-1}) + (\alpha+2\beta)\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2})\right) \\ &\leq C\tau\,, \end{aligned}$$

using the estimate (5.18). The whole term I_1 is therefore converging to zero and ϱ satisfies equation (5.31).

□

Remark 5.11. *Note that the term I_5 could be controlled in other ways. First of all, it would be exactly zero if we considered the density to be extrapolated in the euclidean sense, i.e. taking $\rho_e^{n-1} = \frac{\alpha}{\alpha-\beta}\rho^{n-1} - \frac{\beta}{\alpha-\beta}\rho^{n-2}$, implying that this simple strategy would still lead to a convergent scheme although the accuracy would not be second order, as one can numerically verify. Considering instead the two optimal transport maps T^{n-1} and ${\mathrm{T}_e}^{n-1}$, mapping respectively ρ^{n-2} to ρ^{n-1} and ρ^{n-2} to ρ_e^{n-1}, we could rewrite*

$$\begin{aligned} I_2 &= \left| \int_\Omega (\alpha\varphi^{n-2}(\mathrm{T}^{n-1}(\boldsymbol{y}_2)) - \beta\varphi^{n-2}(\boldsymbol{y}_2) - (\alpha-\beta)\varphi^{n-2}({\mathrm{T}_e}^{n-1}(\boldsymbol{y}_2)))\mathrm{d}\rho^{n-2} \right| \\ &\leq \left| \int_\Omega \langle \alpha\mathrm{T}^{n-1}(\boldsymbol{y}_2)) - \beta\boldsymbol{y}_2 - (\alpha-\beta){\mathrm{T}_e}^{n-1}(\boldsymbol{y}_2), \nabla\varphi^{n-2}(\boldsymbol{y}_2)\rangle \mathrm{d}\rho^{n-2} \right| \\ &\quad + \frac{1}{2}||\mathrm{Hess}(\varphi^{n-2})||_\infty \left(\int_{\Omega\times\Omega} |\mathrm{T}^{n-1}(\boldsymbol{y}_2) - \boldsymbol{y}_2|^2 \mathrm{d}\rho^{n-2} + \int_{\Omega\times\Omega} |{\mathrm{T}_e}^{n-1}(\boldsymbol{y}_2) - \boldsymbol{y}_2|^2 \mathrm{d}\rho^{n-2} \right). \end{aligned}$$

The scheme would be consistent taking ρ_e^{n-1} equal to the pushforward of the map ${\mathrm{T}_e}^{n-1} = \frac{\alpha}{\alpha-\beta}\mathrm{T}^{n-1} - \frac{\beta}{\alpha-\beta}\mathrm{Id}$. If ${\mathrm{T}_e}^{n-1}$ is not the gradient of a convex function, the pushforward via this map is not necessarily the solution of problem (5.11). However, ${\mathrm{T}_e}^{n-1}$ may not be well-defined in case the mass leaves the domain, i.e. $\mathrm{Im}({\mathrm{T}_e}^{n-1}) \not\subset \Omega$.

5.2.3 Convergence in the EVI sense

Let us now make the further assumption that the energy functional $\mathcal{E}$ is λ-convex in the generalized geodesic sense, for $\lambda \in \mathbb{R}_+$. We will limit ourselves to the case $\lambda \geq 0$ for simplicity. We recall that a curve $\varrho : [0,T] \to \mathcal{P}(\Omega)$, $\varrho(0) = \rho^0$, is a Wasserstein gradient flow in the EVI sense if for any $\nu \in \mathcal{P}(\Omega)$ it holds

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\mathcal{W}_2^2(\varrho(t),\nu) \leq \mathcal{E}(\nu) - \mathcal{E}(\varrho(t)) - \frac{\lambda}{2}\mathcal{W}_2^2(\varrho(t),\nu), \quad \forall t \in (0,T)\,. \tag{5.33}$$

Equivalently, we can write it in integral form. The inequality (5.33) holds if and only if for all $r, s \in (0,T)$ with $r \leq s$ it holds

$$\frac{1}{2}\mathcal{W}_2^2(\varrho(s),\nu) - \frac{1}{2}\mathcal{W}_2^2(\varrho(r),\nu) \leq \mathcal{E}(\nu)(s-r) - \int_r^s (\mathcal{E}(\varrho(t)) + \frac{\lambda}{2}\mathcal{W}_2^2(\varrho(t),\nu))\mathrm{d}t\,. \tag{5.34}$$

The original BDF2 scheme [93] has been proven to converge to gradient flows in the EVI sense, i.e. the limit curve extracted from the time discretization (5.5) satisfies the inequality (5.34).

The convergence has been proven for general metric spaces. We want to show here that our modified BDF2 scheme (5.10)-(5.11) recovers the same convergence in the Wasserstein space.

We first show that for scheme (5.10)-(5.11) a discrete equivalent form of the inequality (5.34) holds. As the Wasserstein distance $\mathcal{W}_2^2(\cdot, \rho_e^{n-1})$ is 2-convex along any generalized geodesic centered in ρ_e^{n-1}, the overall functional

$$\mathcal{G}(\rho^{n-1}, \rho^{n-2}; \rho) = \frac{1}{\tau}(\alpha - \beta)\mathcal{W}_2^2(\rho, \rho_e^{n-1}) + \mathcal{E}(\rho)\,, \tag{5.35}$$

is $\frac{2}{\tau}(\alpha - \beta) + \lambda > 0$ convex along any generalized geodesic centered in ρ_e^{n-1} as well. We consider the case $\lambda \geq 0$ in order to avoid dealing with the conditions on the time step τ in order to have $\frac{2}{\tau}(\alpha - \beta) + \lambda > 0$, and simplify the presentation.

Lemma 5.12. *At each step n, $\forall \nu \in \mathcal{P}(\Omega)$, the following inequality holds:*

$$\begin{aligned}\Big(\frac{1}{\tau}(\alpha - \beta) + \frac{\lambda}{2}\Big)\mathcal{W}_2^2(\rho^n, \nu) - \frac{\alpha}{\tau}\mathcal{W}_2^2(\nu, \rho^{n-1}) + \frac{\beta}{\tau}\mathcal{W}_2^2(\nu, \rho^{n-2}) \leq \\ \leq \mathcal{E}(\nu) - \mathcal{E}(\rho^n) + \frac{\alpha\beta}{\tau(\alpha - \beta)}\mathcal{W}_2^2(\rho^{n-1}, \rho^{n-2}) - \frac{1}{\tau}(\alpha - \beta)\mathcal{W}_2^2(\rho^n, \rho_e^{n-1})\,.\end{aligned} \tag{5.36}$$

Proof. The functional $\mathcal{G}$ is $(2(\alpha-\beta)+\lambda)$-convex along generalized geodesics, therefore considering the generalized geodesic ω between ν and ρ^n with center ρ_e^{n-1}, and using the optimality of ρ^n, it holds

$$\begin{aligned}0 &\leq \mathcal{G}(\rho^{n-1}, \rho^{n-2}; \omega(t)) - \mathcal{G}(\rho^{n-1}, \rho^{n-2}; \rho^n) \\ &\leq t(\mathcal{G}(\rho^{n-1}, \rho^{n-2}; \nu) - \mathcal{G}(\rho^{n-1}, \rho^{n-2}; \rho^n)) - \frac{1}{2}\Big(\frac{2}{\tau}(\alpha - \beta) + \lambda\Big)t(1-t)\mathcal{W}(\rho^n, \nu).\end{aligned}$$

Dividing by t and taking the limit $t \to 0$ we obtain

$$\Big(\frac{1}{\tau}(\alpha - \beta) + \frac{\lambda}{2}\Big)\mathcal{W}_2^2(\rho^n, \nu) - \frac{1}{\tau}(\alpha - \beta)\mathcal{W}_2^2(\nu, \rho_e^{n-1}) \leq \mathcal{E}(\nu) - \mathcal{E}(\rho^n) - \frac{1}{\tau}(\alpha - \beta)\mathcal{W}_2^2(\rho^n, \rho_e^{n-1})\,.$$

Adding on both side the term $-\mathcal{F}(\rho^{n-1}, \rho^{n-2}; \rho_e^{n-1})$, using (5.13) on the left-hand side and (5.16) on the right-hand side, we conclude. □

We recall that thanks to the classical estimate (5.18) (Theorem 5.5), the piecewise constant curve

$$\rho_\tau(t) = \sum_{n=1}^{N} \rho^{n-1}\mathbb{1}_{(t^{n-1}, t^n]}\,, \quad \rho_\tau(0) = \rho^0\,,$$

converges uniformly in the $\mathcal{W}_2$ distance to an absolutely continuous limit curve ϱ (see Proposition 5.7). In order to prove convergence of the scheme in the EVI sense, we show that the this curve satisfies inequality (5.34). Thanks to the uniform convergence in time, the procedure is the same as in [93, Theorem 5.1].

Theorem 5.13. *The curve ϱ satisfies* (5.34).

Proof. For simplicity, assume that given $r, s \in (0,T), r \le s$, there exist $N_\tau, M_\tau \in \mathbb{N}, N_\tau \le M_\tau$, such that $r = N_\tau \tau, s = M_\tau \tau$, $\forall \tau$. We multiply by τ inequality (5.36) and sum over n from N_τ to M_τ to obtain the discrete integral form of the EVI:

$$\sum_{n=N_\tau}^{M_\tau} \left((\alpha-\beta)\mathcal{W}_2^2(\rho^n,\nu) - \alpha\mathcal{W}_2^2(\nu,\rho^{n-1}) + \beta\mathcal{W}_2^2(\nu,\rho^{n-2})\right) \le$$

$$\le \mathcal{E}(\nu)(t-s) - \sum_{n=N_\tau}^{M_\tau} \tau\Big(\mathcal{E}(\rho^n) + \frac{\lambda}{2}\mathcal{W}_2^2(\rho^n,\nu)\Big)$$

$$+ \sum_{n=N_\tau}^{M_\tau} \left(\frac{\alpha\beta}{\alpha-\beta}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2}) - (\alpha-\beta)\mathcal{W}_2^2(\rho^n,\rho_e^{n-1})\right). \quad (5.37)$$

By canceling out terms, the left-hand side is equal to

$$-\alpha\mathcal{W}_2^2(\nu,\rho^{N_\tau-1}) + \beta\mathcal{W}_2^2(\nu,\rho^{N_\tau}) + \beta\mathcal{W}_2^2(\nu,\rho^{N_\tau-1})$$
$$+ (\alpha-\beta)\mathcal{W}_2^2(\rho^{M_\tau-1},\nu) + (\alpha-\beta)\mathcal{W}_2^2(\rho^{M_\tau},\nu) - \alpha\mathcal{W}_2^2(\nu,\rho^{M_\tau-1}), \quad (5.38)$$

and thanks to the uniform convergence in the Wasserstein distance, (5.38) converges to

$$\frac{1}{2}\mathcal{W}_2^2(\varrho(s),\nu) - \frac{1}{2}\mathcal{W}_2^2(\varrho(r),\nu),$$

for $\tau \to 0$, where we recall $\alpha = 1, \beta = \frac{1}{4}$. Concerning the right-hand side, thanks again to the uniform convergence in the Wasserstein distance, the lower semi-continuity of $\mathcal{E}$ and Fatou's lemma, we have

$$\limsup_{n\to\infty} -\sum_{n=N_\tau}^{M_\tau} \tau\Big(\mathcal{E}(\rho^n) + \frac{\lambda}{2}\mathcal{W}_2^2(\rho^n,\nu)\Big) \le -\int_r^s \Big(\mathcal{E}(\varrho(t)) + \frac{\lambda}{2}\mathcal{W}_2^2(\varrho(t),\nu)\Big)\mathrm{d}t\,.$$

Finally, owing to bound (5.18), we estimate the last contribution of (5.37) as

$$\sum_n \frac{\alpha\beta}{\alpha-\beta}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2}) - (\alpha-\beta)\mathcal{W}_2^2(\rho^n,\rho_e^{n-1}) \le \sum_n \frac{\alpha\beta}{\alpha-\beta}\mathcal{W}_2^2(\rho^{n-1},\rho^{n-2}) \le C\tau,$$

which converges to zero. In the end, we recover the continuous inequality (5.34).

□

5.3 Finite volume discretization

Based on the modified BDF2 we proposed, we want to devise a second order accurate finite volume scheme in both space and time. The discretization of problem (5.10) will be done employing the techniques we presented in the previous Chapter 4. More involved is instead the discretization of problem (5.11), as it is not a convex optimization problem in the classical sense. We propose for this reason a non-variational approach.

5.3.1 Extrapolation in the viscosity sense

Let us consider two measures $\mu, \nu \in \mathcal{P}(\Omega)$ and assume that the $\frac{\alpha}{\alpha-\beta}$-extrapolation, in the geodesic sense, from μ to ν exists. We have seen that in this case, given the solution $(\tilde{\phi}, \tilde{\rho})$ to problem (1.26) which defines the geodesic interpolation, we can compute the extrapolated measure as the final value of the measure curve defined by

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \nabla \phi) = 0\,, \\ \partial_t \phi + \dfrac{1}{2}|\nabla \phi|^2 = 0\,, \end{cases} \quad \text{in } \left[1, \frac{\alpha}{\alpha-\beta}\right] \times \Omega\,, \tag{5.39}$$

complemented with the boundary conditions $\rho \nabla \phi \cdot \boldsymbol{n} = 0$ on $\partial\Omega$ and the initial conditions $\rho(1,\cdot) = \tilde{\rho}(1,\cdot), \phi(1,\cdot) = \tilde{\phi}(1,\cdot)$. System (5.39) is a system of two forward in time equations that can be solved separately, starting from the Hamilton-Jacobi equation and then recovering the measure evolution from the continuity equation. In alternative, once we have computed the solution ϕ to the Hamilton-Jacobi equation, the latter step can be written as finding

$$\rho_e \in \operatorname*{arginf}_{\rho \in \mathcal{P}(\Omega)} \frac{\alpha-\beta}{2\beta} \mathcal{W}_2^2(\rho, \nu) - \int_\Omega \phi_e \rho\,, \tag{5.40}$$

where $\phi_e = \phi(\frac{\alpha}{\alpha-\beta}, \cdot)$. Problem (5.40) is a JKO step with energy $\mathcal{E}(\rho) = -\int_\Omega \phi_e \rho$ and time step $\frac{\beta}{\alpha-\beta} = \frac{1}{3}$. If the geodesic extrapolation exists, the solution to problem (5.40) coincides with the one of (5.39). Indeed, in this case we know that the potential ϕ_e is given by

$$\phi_e = \frac{\alpha}{\alpha-\beta}\tilde{\phi}(0,\cdot) = \frac{4}{3}\tilde{\phi}(0,\cdot)$$

(see Section 1.1.2) and $\frac{|\boldsymbol{x}|^2}{2} - \phi_e$ is convex. The optimality condition of (5.40) then guarantees that ϕ_e is the optimal potential in the transport from the minimizer to ν (see [115, Proposition 7.20] and the Example 7.21 that follows). By uniqueness of the geodesic, the solution provided coincides with the solution of (5.39).

As we already said, the extrapolation in the geodesic sense may not always exist, the problem being that the particles moving in a straight trajectory may collide at some point after the final measure or encounter the boundary of the domain. If this happens, the solution to the Hamilton-Jacobi equation cannot be a classical one. We need to consider solutions that dissipate the possible shock, namely viscosity solutions. The weak solution (ϕ, ρ) provided in this way by system (5.39) is well defined thanks to the semi-concavity of the initial condition $\tilde{\phi}(1,\cdot)$, see [12]. We can always extrapolate in this sense and move past a potential shock. If ϕ is not a classical solution to the Hamilton-Jacobi equation, that is in the case the geodesic extrapolation does not exist, problems (5.39) and (5.40) do not provide the same measure. In this case, we can therefore define the extrapolation ρ_e in two different ways, either integrating the continuity equation in (5.39) or solving the variational problem (5.40). We will use both strategies in our numerical approach. If the geodesic extrapolation exists, the two approaches are consistent, which motivates their use for numerical computations. As the numerical experiments will show in Section 5.4, this choice not only provides a consistent scheme but it is also second order accurate.

5.3.2 A second order finite volume scheme

Following the same ideas we exposed in Chapters 3-4, we want to approximate the scheme (5.10)-(5.39) with the least computational effort possible, preserving the accuracy of the overall approach. Problem (5.10) can be simplified using again the weighted H^{-1} norm and resorting to the same space discretization we introduced in Chapter 4. This time, the choice of the arithmetic average of the densities as weight for the H^{-1} norm will be fundamental in order to achieve second order accuracy in time. We will focus here on the strategy to compute ρ_e^{n-1}, which can be divided in three consecutive steps. First, we need to solve the optimal transport problem between the measures ρ^{n-2} and ρ^{n-1} in order to evaluate the optimal potential $\tilde{\phi}^{n-1}$ pushing one into the other. Then, we evolve $\tilde{\phi}^{n-1}$ forward in time with the Hamilton-Jacobi equation. Finally, we recover the density from the first equation in (5.39), or solving problem (5.40). We will reduce these problems to one step discretizations.

Discrete setting

We consider TPFA finite volumes which requires sufficiently regular partitioning of the domain, according to Definition 1.1. We recall that the spaces of discrete variables defined on cells and diamond cells, $\mathbb{R}^{\mathcal{T}}$ and $\mathbb{R}^{\Sigma}$, are endowed with the two scalar products $\langle\cdot,\cdot\rangle_{\mathcal{T}}$ and $\langle\cdot,\cdot\rangle_{\Sigma}$:

$$\langle\cdot,\cdot\rangle_{\mathcal{T}} : (\boldsymbol{a},\boldsymbol{b}) \in [\mathbb{R}^{\mathcal{T}}]^2 \mapsto \sum_{K\in\mathcal{T}} a_K b_K m_K\,, \quad \langle\cdot,\cdot\rangle_{\Sigma} : (\boldsymbol{u},\boldsymbol{v}) \in [\mathbb{R}^{\Sigma}]^2 \mapsto \sum_{\sigma\in\Sigma} u_\sigma v_\sigma m_\sigma d_\sigma\,.$$

The space of discrete probability measures $\mathbb{P}_{\mathcal{T}} \subset \mathbb{R}^{\mathcal{T}}$ is

$$\mathbb{P}_{\mathcal{T}} = \{\boldsymbol{\rho} \in \mathbb{R}_+^{\mathcal{T}} : \langle\boldsymbol{\rho},\mathbf{1}\rangle_{\mathcal{T}} = \langle\boldsymbol{\rho}^0,\mathbf{1}\rangle_{\mathcal{T}}\}\,.$$

Finally, the space of consevative fluxes is defined as:

$$\mathbb{F}_{\mathcal{T}} = \{\boldsymbol{F} = (F_{K,\sigma}, F_{L,\sigma})_{\sigma\in\Sigma} \in \mathbb{R}^{2\Sigma} : F_{K,\sigma} + F_{L,\sigma} = 0\}\,.$$

We recall that we denote $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$ and that by convention $|\boldsymbol{F}| = (F_\sigma)_{\sigma\in\Sigma} \in \mathbb{R}^{\Sigma}$ and $(\boldsymbol{F})^2 = (F_\sigma^2)_{\sigma\in\Sigma} \in \mathbb{R}^{\Sigma}$, for $\boldsymbol{F} \in \mathbb{F}_{\mathcal{T}}$. The discrete divergence $\mathrm{div}_{\mathcal{T}} : \mathbb{F}_{\mathcal{T}} \to \mathbb{R}^{\mathcal{T}}$ and the discrete gradient $\nabla_\Sigma : \mathbb{R}^{\mathcal{T}} \to \mathbb{F}_{\mathcal{T}}$ are defined by:

$$(\mathrm{div}_{\mathcal{T}}\boldsymbol{F})_K = \mathrm{div}_K\boldsymbol{F} = \frac{1}{m_K}\sum_{\sigma\in\Sigma_K} F_{K,\sigma} m_\sigma\,,$$
$$(\nabla_\Sigma \boldsymbol{a})_{K,\sigma} = \nabla_{K,\sigma}\boldsymbol{a} := \frac{a_L - a_K}{d_\sigma}\,.$$

We will use a centered reconstruction for the mobility in order to attain the second order accuracy in space. For this purpose, we will use again the weighted arithmetic average operator $\mathcal{L}_\Sigma : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}^{\Sigma}$ and its adjoint $\mathcal{L}_\Sigma^* : \mathbb{R}^{\Sigma} \to \mathbb{R}^{\mathcal{T}}$ (with respect to the two scalar products):

$$(\mathcal{L}_\Sigma \boldsymbol{a})_\sigma = \frac{d_{K,\sigma}}{d_\sigma} a_K + \frac{d_{L,\sigma}}{d_\sigma} a_L\,, \quad (\mathcal{L}_\Sigma^* \boldsymbol{u})_K = \sum_{\sigma\in\Sigma_K} u_\sigma \frac{m_\sigma d_{K,\sigma}}{m_K}\,,$$

for $\boldsymbol{a} \in \mathbb{R}^{\mathcal{T}}$ and $\boldsymbol{u} \in \mathbb{R}^{\Sigma}$.

Discrete extrapolation

Let us consider two discrete densities $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{P}_{\mathcal{T}}$. We are going to show a possible implementation to extrapolate from $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$, following what we proposed in Section 5.3.1. As we did in Chapters 3-4, we will approximate the Wasserstein distance with a discrete weighted H^{-1} norm. We stress that we will consider now the arithmetic average of the densities as weight.

The first step is to compute the interpolation between the two discrete densities. By approximating the Wasserstein distance with the weighted H^{-1} norm between $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$[3],

$$\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}; \boldsymbol{\mu}-\boldsymbol{\nu}\Big) = \sup_{\boldsymbol{\phi}\in\mathbb{R}^{\mathcal{T}}} \langle \boldsymbol{\mu}-\boldsymbol{\nu}, \boldsymbol{\phi}\rangle_{\mathcal{T}} - \langle \mathcal{L}\Big(\frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}\Big), (\nabla_{\Sigma}\boldsymbol{\phi})^2\rangle_{\Sigma}\,, \tag{5.41}$$

we simply have to compute the optimal potential $\widetilde{\boldsymbol{\phi}}$ in (5.41). See Section 4.2 for details on formula (5.41). This potential can be considered as an approximation of the continuous one at the midpoint of the time interval $[0,1]$. The optimality conditions of the problem provide the one step discretization of the continuity equation:

$$\boldsymbol{\mu}-\boldsymbol{\nu} + \mathrm{div}_{\mathcal{T}}(\mathcal{L}\Big(\frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}\Big) \odot \nabla\widetilde{\boldsymbol{\phi}}) = 0\,. \tag{5.42}$$

Notice that a solution to problem (5.41) is not necessarily unique as the Dirichlet energy is weighted by the time-space reconstruction of the density, which may vanish. In this case, in order to uniquely defined $\widetilde{\boldsymbol{\phi}}$ and consequently the extrapolation, we consider the solution with minimal Dirichlet energy.

Once we have computed the potential $\widetilde{\boldsymbol{\phi}}$ which transports $\boldsymbol{\mu}$ to $\boldsymbol{\nu}$, the second step can be realized integrating forwardly in time the Hamilton-Jacobi equation explicitly. We integrate till the time $1+\frac{\beta}{2(\alpha-\beta)} = \frac{7}{6}$ in order to approximate the potential in the midpoint of the interval $[1, \frac{\alpha}{(\alpha-\beta)}] = [1, \frac{4}{3}]$. In the following step we will compute the extrapolated density $\boldsymbol{\rho}_e$ as solution to the continuity equation discretized with the midpoint rule, whence this choice. The length of the temporal step is therefore $\frac{\alpha}{2(\alpha-\beta)} = \frac{2}{3}$ and the potential $\boldsymbol{\phi}_e \in \mathbb{R}^{\mathcal{T}}$ is explicitly given by:

$$\boldsymbol{\phi}_e = \widetilde{\boldsymbol{\phi}} - \frac{2(\alpha-\beta)}{\alpha}\frac{1}{2}\mathcal{L}^*_{\mathcal{T}}(\nabla_{\Sigma}\widetilde{\boldsymbol{\phi}})^2\,, \tag{5.43}$$

To discretize the Hamilton-Jacobi operator and reconstruct the norm squared of the gradient on each cell $K \in \mathcal{T}$, we used the adjoint of the linear reconstruction $\mathcal{L}$. As this step is not variational, this choice is not mandatory and we could use in principle any second order discretization. Notice in particular that at the discrete level, since we approximate the equation in one step, we don't need to recover a viscous solution. We can therefore be more flexible in the discretization. Finally, the density $\boldsymbol{\rho}_e$ is computed as solution to the discrete continuity equation with given velocity field $\nabla_{\Sigma}\boldsymbol{\phi}_e$:

$$\rho_e - \boldsymbol{\nu} + \frac{(\alpha-\beta)}{\beta}\mathrm{div}_{\mathcal{T}}(\mathcal{L}\Big(\frac{\rho_e+\boldsymbol{\nu}}{2}\Big) \odot \nabla_{\Sigma}\boldsymbol{\phi}_e) = 0\,. \tag{5.44}$$

Equation (5.44) discretizes the continuuity equation with the midpoint rule on the time interval $[1, \frac{\alpha}{(\alpha-\beta)}] = [1, \frac{4}{3}]$. We used again the linear reconstruction operator, although this choice is not mandatory, and the arithmetic average of the densities to reconstruct the mobility.

[3]Notice that with respect to the definitions we gave in Chapters 3-4 we are implicitly changing the sign of the potential.

The first step of this procedure to compute the discrete extrapolation is variational and can be solved in an efficient and robust way. Although for vanishing mobility the non-uniqueness could be an issue to carefully deal with, in practice the density will be always strictly greater than zero. At each step in fact the density $\boldsymbol{\rho}^n$ will be computed thanks to the same interior point method we presented in Chapter 4. Therefore, we can compute the solution $\widetilde{\boldsymbol{\phi}}$ solving directly the linear system (5.42). The second step is explicit and does not pose any problem. The last one is not variational and a solution is not guaranteed to exist or to be unique in general, depending on the velocity field $\nabla_\Sigma \boldsymbol{\phi}_e$. Furthermore, this discretization does not guarantee the positivity of the density $\boldsymbol{\rho}_e$. Designing a second order space discretization in order to preserve it is not immediate due to the unavoidable arithmetic average in time for the mobility, which is necessary to preserve the order two accuracy of the approach.

In order to overcome this last issue, we present an alternative method for computing $\boldsymbol{\rho}_e$ based on problem (5.40). After computing the potential $\widetilde{\boldsymbol{\phi}}$ as solution to problem (5.41), we can evolve it until the final time $\frac{\alpha}{\alpha-\beta} = \frac{4}{3}$, that is considering a temporal step of length $\frac{1}{2} + \frac{\beta}{\alpha-\beta} = \frac{\alpha+\beta}{2(\alpha-\beta)} = \frac{5}{6}$:

$$\boldsymbol{\phi}_e = \tilde{\boldsymbol{\phi}} - \frac{2(\alpha-\beta)}{\alpha+\beta}\frac{1}{2}\mathcal{L}^*_{\mathcal{T}}(\nabla_\Sigma \tilde{\boldsymbol{\phi}})^2 \,. \tag{5.45}$$

Then, we approximate problem (5.40) using again the discrete H^{-1} norm and we compute the density $\boldsymbol{\rho}_e$ as[4]

$$\boldsymbol{\rho}_e \in \operatorname*{arginf}_{\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}} \frac{\beta}{(\alpha-\beta)}\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-1}}{2}; \boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}\Big) - \langle \boldsymbol{\phi}_e, \boldsymbol{\rho}\rangle_{\mathcal{T}} \,, \tag{5.46}$$

which is an LJKO step with step length $\frac{\beta}{\alpha-\beta}$ and discrete energy $-\langle \boldsymbol{\phi}_e, \boldsymbol{\rho}\rangle_{\mathcal{T}}$.

Modified BDF2 discrete scheme

We can finally formulate our second order finite volume scheme. Consider a strictly convex discrete energy function $\mathcal{E}_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$. Given two initial densities $\boldsymbol{\rho}^0, \boldsymbol{\rho}^1 \in \mathbb{R}^{\mathcal{T}}$, with the same total discrete mass $\langle \boldsymbol{\rho}^0, \mathbf{1}\rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^1, \mathbf{1}\rangle_{\mathcal{T}}$, and a time step $\tau > 0$, we compute the sequence of densities $(\boldsymbol{\rho}^n)_{n\geq 2} \subset \mathbb{P}_{\mathcal{T}}$ defined by the following recursive scheme:

$$\begin{aligned}
&\text{1) compute the } \frac{4}{3}\text{-extrapolation } \boldsymbol{\rho}_e^{n-1} \text{ from } \boldsymbol{\rho}^{n-2} \text{ to } \boldsymbol{\rho}^{n-1};\\
&\text{2) compute } \boldsymbol{\rho}^n \text{ solution to the LJKO step:}\\
&\qquad \inf_{\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}} \frac{2}{\tau}(\alpha-\beta)\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}_e^{n-1}}{2}; \boldsymbol{\rho}_e^{n-1}-\boldsymbol{\rho}\Big) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) \,.
\end{aligned} \tag{5.47}$$

The first step can be realized either via (5.41)-(5.43)-(5.44) or (5.41)-(5.45)-(5.46). In order to find solutions to the LJKO steps, we use the same interior point strategy presented in Section 4.4. Since extrapolating using the continuity equation (5.44) does not guarantee the positivity of the density, we need to carefully handle it. If the density $\boldsymbol{\rho}_e^{n-1}$ becomes negative,

[4] We recall that the factor $\frac{1}{2}$ in front of the Wasserstein distance squared is absorbed in the definition of $\mathcal{A}_{\mathcal{T}}$, see Section 4.2.

the functional in the LJKO step in (5.47) is not anymore convex on the whole $\mathbb{R}_+^{\mathcal{T}}$ and it may also be unbounded from below. To avoid it, we can modify the step as

$$\inf_{\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}} \frac{2}{\tau}(\alpha-\beta)\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+(\boldsymbol{\rho}_e^{n-1})^+}{2};\boldsymbol{\rho}_e^{n-1}-\boldsymbol{\rho}\Big)+\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})$$

where $(\boldsymbol{\rho}_e^{n-1})^+ = \big(\max((\boldsymbol{\rho}_e^{n-1})_K,0)\big)_{K\in\mathcal{T}}$.

5.3.3 Other implementations

As we said in Section 5.1.1, the VIM scheme and the original variational BDF2 are not particularly convenient approaches for designing a discrete scheme. We can nevertheless briefly show that their implementation is possible and leads to second order accurate schemes (see Section 5.4.2). We will compare their solutions to the solutions provided by scheme (5.47).

As we said in Section 5.1.1, the VIM scheme can be implemented by first solving a JKO step with time step $\frac{\tau}{2}$ and then computing the 2-extrapolation. We can use the same idea for extrapolating that we presented in the previous section. We can then propose a discrete VIM scheme as: given the initial density $\boldsymbol{\rho}^0\in\mathbb{P}_{\mathcal{T}}$ and a time step $\tau>0$, at each step n,

1) compute $\tilde{\boldsymbol{\rho}}^n$ solution to the LJKO step:

$$\inf_{\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}} \frac{2}{\tau}\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-1}}{2};\boldsymbol{\rho}^{n-1}-\boldsymbol{\rho}\Big)+\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})\,; \tag{5.48}$$

2) compute $\boldsymbol{\rho}^n$ as the 2-extrapolation from $\boldsymbol{\rho}^{n-1}$ to $\tilde{\boldsymbol{\rho}}^n$.

The time parameter in the LJKO step is $\frac{\tau}{2}$. The extrapolation step can again be computed either via (5.43)-(5.44) or (5.45)-(5.46). In order to realize the 2-extrapolation, we need to consider the values $\alpha=1,\beta=\frac{1}{2}$. In this case, the value for the potential $\widetilde{\boldsymbol{\phi}}^{n-1}$ is already known from the LJKO step and does not need to be computed. As before, the discrete LJKO steps can be computed thanks to the interior point strategy presented in Section 4.4.

We can also propose a naive discretization of scheme (5.5). Consider two initial conditions $\boldsymbol{\rho}^0,\boldsymbol{\rho}^1\in\mathbb{P}_{\mathcal{T}}$ and the time parameter $\tau>0$. At each step n, for $\boldsymbol{\rho}^{n-1},\boldsymbol{\rho}^{n-2}\in\mathbb{P}_{\mathcal{T}}$, compute $\boldsymbol{\rho}^n$ as solution to

$$\inf_{\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}} \frac{2}{\tau}\Big(\alpha\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-1}}{2};\boldsymbol{\rho}-\boldsymbol{\rho}^{n-1}\Big)-\beta\mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-2}}{2};\boldsymbol{\rho}-\boldsymbol{\rho}^{n-2}\Big)\Big)+\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})\,, \tag{5.49}$$

with again $\alpha=1,\beta=\frac{1}{4}$. We approximate each Wasserstein distance in (5.5) with a weighted discrete H^{-1} norm. Problem (5.49) is not a convex optimization problem. Notice that it is not even bounded from below in general. The function $\mathcal{A}_{\mathcal{T}}(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-2}}{2};\boldsymbol{\rho}^{n-2}-\boldsymbol{\rho})$ is not indeed bounded from above for all $\boldsymbol{\rho}^{n-2}\in\mathbb{P}_{\mathcal{T}}$[5]. We can nevertheless try to compute stationary points of the objective function in (5.49). A stationary point can be easily recognized to satisfy the

[5] Just consider that, if the density $\boldsymbol{\rho}^{n-2}$ is not supported everywhere, one can chose a density $\boldsymbol{\rho}\in\mathbb{P}_{\mathcal{T}}$ in order to realize a finite displacement with vanishing mobility $\frac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-2}}{2}$, i.e. infinite kinetic energy.

following system of equations:

$$\begin{cases} \dfrac{\boldsymbol{\rho}-\boldsymbol{\rho}^{n-1}}{\tau} - \mathrm{div}_{\mathcal{T}}\Big(\mathcal{L}\Big(\dfrac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-1}}{2}\Big)\odot\nabla\boldsymbol{\phi}_1\Big) = 0, \\ \dfrac{\boldsymbol{\rho}-\boldsymbol{\rho}^{n-2}}{\tau} - \mathrm{div}_{\mathcal{T}}\Big(\mathcal{L}\Big(\dfrac{\boldsymbol{\rho}+\boldsymbol{\rho}^{n-2}}{2}\Big)\odot\nabla\boldsymbol{\phi}_2\Big) = 0, \\ \dfrac{2\alpha}{\tau}\big(\boldsymbol{\phi}_1 + \dfrac{1}{2}\mathcal{L}^*_{\mathcal{T}}(\nabla_\Sigma\boldsymbol{\phi}_1)^2\big) - \dfrac{2\beta}{\tau}\big(\boldsymbol{\phi}_2 + \dfrac{1}{2}\mathcal{L}^*_{\mathcal{T}}(\nabla_\Sigma\boldsymbol{\phi}_2)^2\big) - \nabla_{\boldsymbol{\rho}}\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) \le 0. \end{cases} \tag{5.50}$$

where $\nabla_{\boldsymbol{\rho}}\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) = (\frac{\partial\mathcal{E}_{\mathcal{T}}}{\partial\rho_K}(\boldsymbol{\rho}))_{K\in\mathcal{T}} \in \mathbb{R}^{\mathcal{T}}$. The first two continuity equations are the optimality conditions of the two discrete H^{-1} norm, whereas the third one is the stationarity with respect to $\boldsymbol{\rho}$. We can try to compute solutions to system (5.50) using again the interior point strategy.

5.4 Numerical validation of the modified BDF2 approach

We want to show now that the scheme (5.47), based on the modified BDF2 temporal discretization we proposed, is second order accurate. We will consider both the implementation we proposed. We will also show qualitatively the behavior of the scheme with simple one dimensional tests. We will compare in this case the solution provided by scheme (5.47) with the solutions provided by schemes (5.48) and (5.49). We will further show that also these schemes are second order accurate. When two initial conditions $\boldsymbol{\rho}^0, \boldsymbol{\rho}^1$ are needed, we compute first $\boldsymbol{\rho}^1$ from $\boldsymbol{\rho}^0$ via the LJKO scheme (4.11).

We consider for these purposes two specific problems that exhibit a gradient flow structure in the Wasserstein space: the Fokker-Planck equation we presented in Section 5.2.2 and the porous medium equation. We recall that this latter equation writes

$$\partial_t\varrho = \Delta\varrho^\delta + \nabla\cdot(\varrho\nabla V)\,, \tag{5.51}$$

and it is a Wasserstein gradient flow with respect to the energy

$$\mathcal{E}(\rho) = \int_\Omega \frac{1}{\delta-1}\rho^\delta + \rho V\,, \tag{5.52}$$

for a given δ strictly greater than one and with $V \in W^{1,\infty}(\Omega)$ a Lipschitz continuous exterior potential [108]. The energy functionals (5.21) and (5.52) are both of the form $\mathcal{E}(\rho) = \int_\Omega E(\rho)\mathrm{d}\boldsymbol{x}$ for a real valued scalar function E. They can be straightforwardly discretized as $\mathcal{E}_{\mathcal{T}} = \sum_{K\in\mathcal{T}} E(\rho_K)m_K$.

5.4.1 Comparison between the three approaches

We perform one dimensional tests in order to show how the scheme (5.47) works and compare it with the other two approaches (5.48) and (5.49). We consider a discretization of the domain $\Omega = [0,1]$ in subintervals of length $m_K = 0.02$. We will compute the extrapolation thanks to (5.45)-(5.46).

We first address the diffusion equation, which is problem (5.20) with zero external potential V. We take as initial condition

$$\rho^0 = \exp\Big(-50\Big(x-\frac{1}{2}\Big)^2\Big)\,,$$

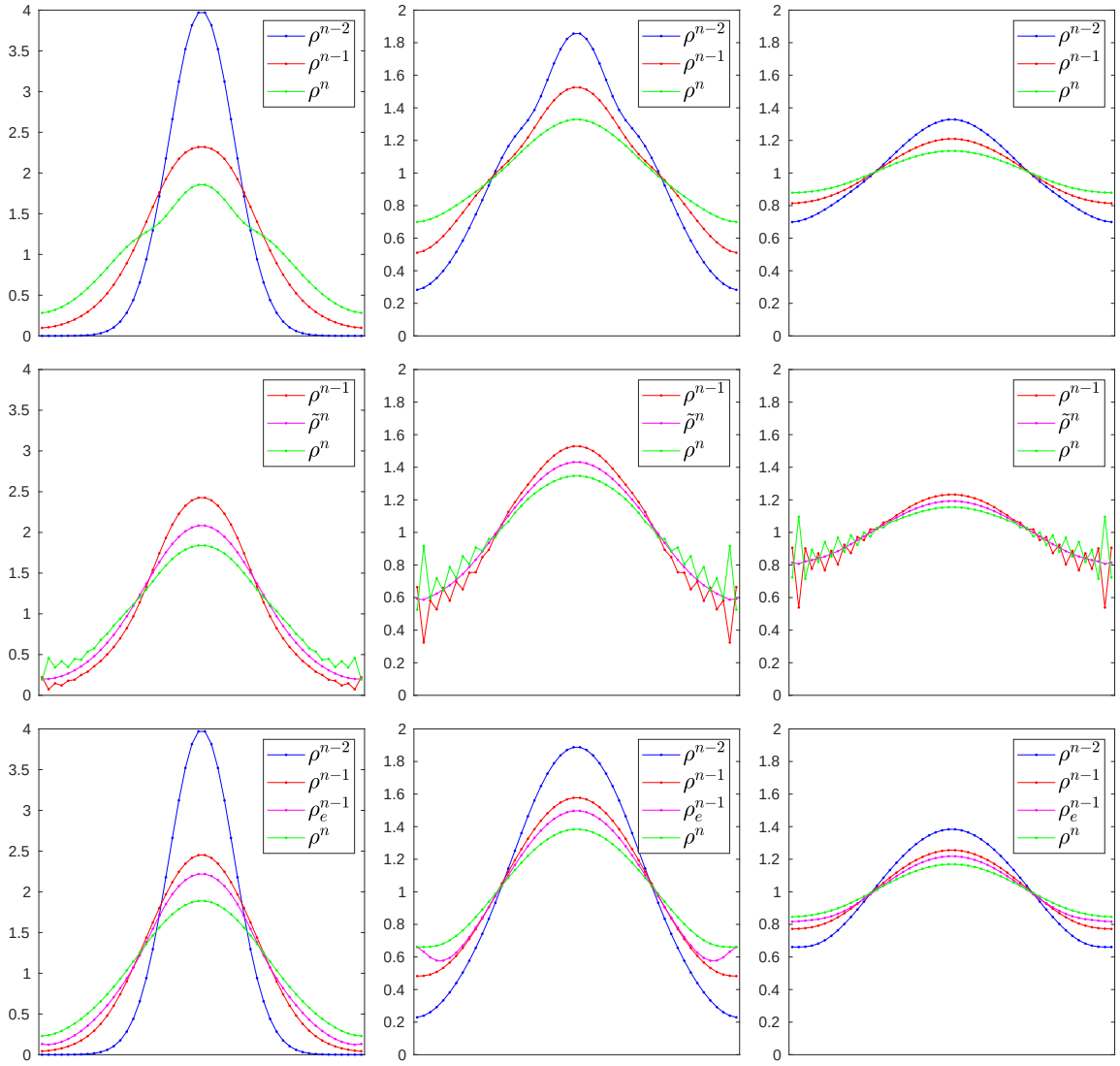


Figure 5.1: Comparison between the three schemes for the diffusion equation. From top to bottom, the BDF2 scheme (5.49), the VIM scheme (5.48) and the modified BDF2 scheme (5.47). From left to right, three different time steps: $t = 0.02, 0.04, 0.06$.

which we discretize as $\boldsymbol{\rho}^0 = \rho^0(\boldsymbol{x}_K)$, and the time step $\tau = 0.01$. We solve the problem with the three schemes, the BDF2 scheme (5.49), the VIM scheme (5.48) and the modified BDF2 scheme (5.47). The results are shown in Figure 5.1. The mass diffuses from the center of the domain, where it is initially concentrated, towards the boundary. Extrapolating along this dynamics means that the particles have to move further towards the boundary. It happens therefore that they reach it, concentrate and start diffusing again, which in the end may generate oscillations. In the VIM scheme (5.48), when we first solve a JKO step and then extrapolate, there appear oscillations that are higly unstable and persist along the integration in time. In our approach instead, the modified BDF2 scheme (5.47), extrapolating before and then realizing the JKO step smooths and stabilizes the flow. The BDF2 scheme (5.49) does

not suffer at all from this problem, yet again notice that the dynamics may slightly differ from the pure diffusion (for the time step chosen). The oscillations attenuates and tend to disappear for smaller and smaller time step τ.

Remark 5.14. *Our modification of the BDF2 scheme and the VIM scheme perform the same operations, extrapolation and JKO step, but in a different order. Up to a temporal shift, the two operations can be considered to be synchronized. Hence, it is the different length of the steps, rather than the order of the operations, that explains the difference in the observed regularity. In scheme (5.47) the length of the extrapolation step is $\frac{1}{3}$, whereas it is $\frac{1}{2}$ for scheme (5.48). The length of the JKO step is $\frac{2}{3}$ for the former and $\frac{1}{2}$ for the latter. The modified BDF2 performs a smaller extrapolation and a longer JKO step.*

Consider now the porous medium equation (5.51) with $\delta = 2$ and the external potential $V(x) = -x$, which drifts the mass towards the positive direction. We take as initial condition

$$\rho^0(x) = \mathbb{1}_{x \leq \frac{3}{10}} ,$$

discretized again as $\boldsymbol{\rho}^0 = \rho^0(\boldsymbol{x}_K)$, and the time step $\tau = 0.002$. In this case, the naive implementation we proposed for the BDF2 scheme (5.49) does not converge. Notice that the objective function in (5.49) is unbounded from below. We compute the discrete flow with the VIM scheme (5.48) and our modified BDF2 scheme (5.47). The results are shown in Figure 5.2. Again, the VIM sheme is unstable whereas the modified BDF2 controls and smooths the oscillations generated by the extrapolation step, and the solution is reliable. In this case the oscillations are not due to the boundary, as the mass is flowing away from it. The oscillations are due to the compact support of the measure and the explicit integration in time of the Hamilton-Jacobi equation: in the extrapolation step the mass cannot flow outside the support, which acts then like a boundary.

5.4.2 Convergence tests

We assess now the second order accuracy of the three schemes. We will consider a simple one dimensional test to compare them. For the approach we proposed (5.47), we will also perform two dimensional cases. We use two explicit solutions ϱ, for the Fokker-Planck equation (5.20) and the porous medium equation (5.51). We consider then a sequence of meshes $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m}\right)$ with decreasing meshsize $h_m = h_{\mathcal{T}_m}$ and a sequence of decreasing time steps τ_m such that $\frac{h_{\mathcal{T}_{m+1}}}{h_{\mathcal{T}_m}} = \frac{\tau_{m+1}}{\tau_m}$. We solve the discrete problem for each couple $(h_{\mathcal{T}_m}, \tau_m)$ and evaluate the convergence with respect to the discrete $L^1((0,T); L^1(\Omega))$ error:

$$\epsilon_m = \sum_n \tau \sum_{K\in\mathcal{T}_m} |\rho_K^n - \varrho(\boldsymbol{x}_K, n\tau)| m_K .$$

We compute the rate of convergence as:

$$\frac{\log(\epsilon_{m-1}) - \log(\epsilon_m)}{\log(\tau_{m-1}) - \log(\tau_m)} .$$

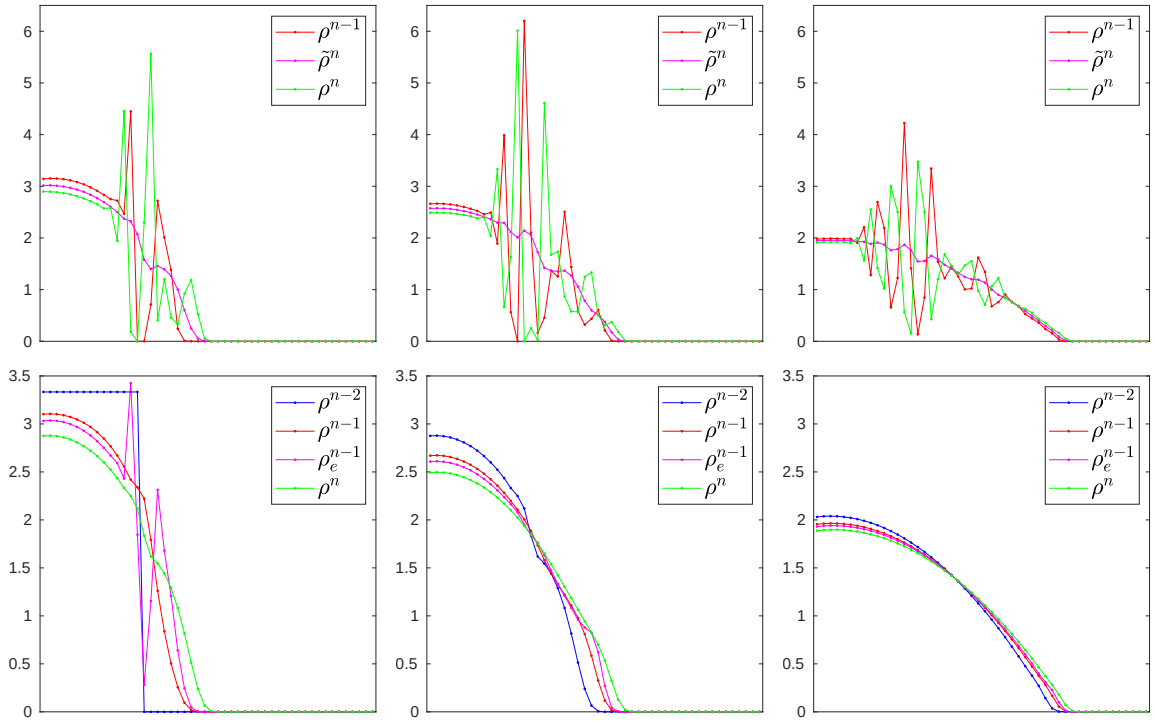


Figure 5.2: Comparison between the VIM scheme (5.48) (top row) and the modified BDF2 scheme (5.47) (bottom row) for the porous medium equation. The BDF2 scheme (5.49) does not converge in this case. From left to right, three different time steps: $t = 0.004, 0.008, 0.020$.

One dimensional case

On the domain $\Omega = [0,1]$ and for the external potential $V(x) = -gx$, we consider the density

$$\varrho(t,x) = \exp\left(-\left(\pi^2 + \frac{g^2}{4}\right)t + \frac{g}{2}x\right)\left(\pi\cos(\pi x) + \frac{g}{2}\sin(\pi x)\right) + \pi\exp\left(g\left(x - \frac{1}{2}\right)\right), \quad (5.53)$$

analytical solution to the Fokker-Planck equation (5.20). We consider the value $g = 1$. For each mesh $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\boldsymbol{x}_K)_{K\in\mathcal{T}_m}\right)$ and time step τ_m, we compute then the discrete solution using the three schemes, starting from the initial condition $(\boldsymbol{\rho}_K^0)_{K\in\mathcal{T}} = (\varrho(0,\boldsymbol{x}_K))_{K\in\mathcal{T}}$. We will use (5.45)-(5.46) to compute the extrapolation. The results are presented in Table 5.1. The BDF2 and our modified approach are second order accurate. The VIM scheme suffers the problem of the oscillations on the time interval $[0, 0.25]$. Repeating the test on the interval $[0.05, 0.25]$, the convergence significantly improves and attains the second order accuracy as well.

Two dimensional case

We want to test convergence of the scheme we proposed (5.47) on two dimensional cases. For this purpose, we use the same sequence of grids we used in Chapters 3-4 (see Figure 3.1).

We repeat first the test on the Fokker-Planck equation in two dimensions. We consider the same solution (5.53) on the domain $\Omega = [0,1]^2$. We test both the two different ways of

Table 5.1: Time-space convergence for the three schemes. Integration time $[0, 0.25]$ for the first three cases, $[0.05, 0.25]$ for the last one.

		BDF2 (5.49)		Mod. BDF2 (5.47)		VIM (5.48)		VIM (5.48)	
h_m	τ_m	ϵ_m	rate	ϵ_m	rate	ϵ_m	rate	ϵ_m	rate
0.100	0.050	2.091e-02	/	2.217e-02	/	5.895e-02	/	4.667e-03	/
0.050	0.025	6.376e-03	1.713	7.016e-03	1.660	3.615e-02	0.706	1.024e-03	2.188
0.025	0.013	1.791e-03	1.832	2.044e-03	1.779	2.294e-02	0.656	2.517e-04	2.025
0.013	0.006	4.849e-04	1.885	5.653e-04	1.854	1.468e-02	0.644	6.264e-05	2.007
0.006	0.003	1.280e-04	1.922	1.508e-04	1.906	1.234e-02	0.251	1.562e-05	2.003
0.003	0.002	3.324e-05	1.945	3.933e-05	1.939	9.983e-03	0.306	3.901e-06	2.002

Table 5.2: Time-space convergence for the modified BDF2 scheme (5.47), with two different type of extrapolations, for the Fokker-Planck equation.

		(5.43)-(5.44)		(5.45)-(5.46)	
h_m	τ_m	ϵ_m	rate	ϵ_m	rate
0.2986	0.0500	2.122e-02	/	2.111e-02	/
0.1493	0.0250	6.802e-03	1.641	6.800e-03	1.634
0.0747	0.0125	2.002e-03	1.765	2.017e-03	1.754
0.0373	0.0063	5.585e-04	1.842	5.669e-04	1.831
0.0187	0.0031	1.501e-04	1.896	1.535e-04	1.884

computing the extrapolation, namely (5.43)-(5.44) and (5.45)-(5.46). The results are listed in Table 5.2. They can be compared with the results in Tables 3.1-4.1, for the schemes we proposed respectively in Chapters 3 and 4. The scheme is second order accurate, using both extrapolation approaches.

We consider now an explicit solution of the porous medium equation (5.51) with zero exterior potential V. This equation admits a solution called Barenblatt profile [108]:

$$\varrho(t, \boldsymbol{x}) = \frac{1}{t^{d\lambda}} \Big(\frac{\delta - 1}{\delta} \Big)^{\frac{1}{\delta - 1}} \max \Big(M - \frac{\lambda}{2} \Big| \frac{\boldsymbol{x} - \boldsymbol{x}_0}{t^{\lambda}} \Big|^2, 0 \Big)^{\frac{1}{\delta - 1}}, \tag{5.54}$$

where $\lambda = \frac{1}{d(\delta - 1) + 2}$, d standing for the space dimension, and $\boldsymbol{x}_0$ is the point where the mass is centered. The parameter M can be chosen to fix the total mass. The value

$$M = \Big(\frac{\delta}{\delta - 1} \Big)^{-\frac{1}{\delta}} \Big(\frac{\lambda \delta}{2\pi(\delta - 1)} \Big)^{\frac{\delta - 1}{\delta}}$$

sets it equal to one. The function (5.54) solves (5.51) on the domain $\Omega = [0, 1]^d$, with $\boldsymbol{x}_0$ in the interior of the Ω, starting from $t_0 > 0$ and for a sufficiently small time horizon T, that is as long as the mass does not reach the boundary of the domain. We consider the two-dimensional case and $\boldsymbol{x}_0 = (0.5, 0.5)$. We solve the problem for $\delta = 2, 3, 4$, with initial condition $(\boldsymbol{\rho}_K^0)_{K \in \mathcal{T}} = (\varrho(t^0, \boldsymbol{x}_K))_{K \in \mathcal{T}}$, starting respectively from $t^0 = 10^{-4}, 10^{-5}, 10^{-6}$. We consider an integration time of 10^{-3}. The extrapolation is in this case computed via (5.45)-(5.46). The results are presented in Table 5.3. The convergence profile is not clean, probably

Table 5.3: Time-space convergence for the modified BDF2 scheme (5.47) for the porous medium equation.

		$\delta = 2$		$\delta = 3$		$\delta = 4$	
h_m	τ_m	ϵ_m	rate	ϵ_m	rate	ϵ_m	rate
0.2986	2.000e-04	5.139e-04	/	7.515e-04	/	9.537e-04	/
0.1493	1.000e-04	1.999e-04	1.363	2.780e-04	1.435	3.085e-04	1.628
0.0747	5.000e-05	6.429e-05	1.636	4.630e-05	2.586	1.103e-04	1.485
0.0373	2.500e-05	1.471e-05	2.127	2.903e-05	0.674	3.847e-05	1.519
0.0187	1.250e-05	4.129e-06	1.833	7.521e-06	1.949	1.340e-05	1.522

due to the low precision of the discretization in space. We can nevertheless notice that in the case $\delta = 2$ the rate of convergence is slowly assessing order two. In the cases $\delta = 3, 4$, where the solutions are less regular, we can notice that the order tends to 1.5.

Appendix A

A monotone discretization for the computation of geodesics

We have shown in Section 1.1.2 that solutions (ϕ, ρ) to the optimal transport problem (1.19)-(1.20) can be characterized as solutions to the system of equations

$$\begin{cases} \partial_t \rho + \nabla \cdot (\rho \nabla \phi) = 0\,, \\ \partial_t \phi + \frac{1}{2}|\nabla \phi|^2 = 0\,, \end{cases} \qquad \text{in } [0,1] \times \Omega\,, \tag{A.1}$$

complemented with the no-flux boundary condition and the initial and final conditions. The Hamilton-Jacobi equation in (A.1) has been saturated (i.e. the equality holds everywhere) thanks to its monotonicity (see Section 1.2.2). We want to present here a strategy for discretizing problem (1.19)-(1.20) in order to obtain a discrete version of system (A.1). That is, we want to preserve the monotonicity of the Hamilton-Jacobi equation at the discrete level. We present this approach for completeness and to be able to show that preserving the monotonicity does not solve the stability issues related to the TPFA discretization, see Section 2.7. For the latter reason, we will avoid to resort to the two nested meshes discretization introduced in Chapter 2, which would just complicate the presentation and in the end will not be exploited in this particular case.

The first key ingredient, as we explained in Section 2.3, is to discretize the kinetic energy (1.18) with a left/right endpoint approximation, in order to preserve the monotonicity of the Hamilton-Jacobi equation in either one time direction or the other. Let us consider a left endpoint approximation, which will lead to an implicit scheme for the Hamilton-Jacobi equation in the positive direction of time. The second key ingredient is to use a monotone discretization in space: we will rely on the upwind reconstruction for the mobility, as we did in Chapter 3.

A.1 Discrete setting

Let us recall first of all the TPFA finite volume discrete setting. The unknown density and potential are discretized on a regular partitioning of the domain, according to Definition 1.1. The discrete spaces are $\mathbb{R}^{\mathcal{T}}$ and $\mathbb{R}^{\Sigma}$, the spaces of discrete variables defined on cells and

diamond cells, and are endowed with the scalar products:

$$\langle \cdot, \cdot \rangle_{\mathcal{T}} : (\boldsymbol{a}, \boldsymbol{b}) \in [\mathbb{R}^{\mathcal{T}}]^2 \mapsto \sum_{K \in \mathcal{T}} a_K b_K m_K \,,$$
$$\langle \cdot, \cdot \rangle_{\Sigma} : (\boldsymbol{u}, \boldsymbol{v}) \in [\mathbb{R}^{\Sigma}]^2 \mapsto \sum_{\sigma \in \Sigma} u_\sigma v_\sigma m_\sigma d_\sigma \,.$$

Vector fields are discretized instead in the space $\mathbb{F}_{\mathcal{T}}$, the space of conservative fluxes:

$$\mathbb{F}_{\mathcal{T}} = \{ \boldsymbol{F} = (F_{K,\sigma}, F_{L,\sigma})_{\sigma \in \Sigma} \in \mathbb{R}^{2\Sigma} : F_{K,\sigma} + F_{L,\sigma} = 0 \} \,. \tag{A.2}$$

We recall the following notation: $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$, $|\boldsymbol{F}| = (F_\sigma)_{\sigma \in \Sigma} \in \mathbb{R}^{\Sigma}$ and $(\boldsymbol{F})^2 = (F_\sigma^2)_{\sigma \in \Sigma} \in \mathbb{R}^{\Sigma}$, for $\boldsymbol{F} \in \mathbb{F}_{\mathcal{T}}$.

The discrete divergence and gradient operators are defined on $\mathbb{F}_{\mathcal{T}}$ and $\mathbb{R}^{\Sigma}$. Precisely, $\mathrm{div}_{\mathcal{T}} : \mathbb{F}_{\mathcal{T}} \to \mathbb{P}_{\mathcal{T}}$ and $\nabla_\Sigma : \mathbb{R}^{\mathcal{T}} \to \mathbb{F}_{\mathcal{T}}$ are defined by:

$$(\mathrm{div}_{\mathcal{T}} \boldsymbol{F})_K = \mathrm{div}_K \boldsymbol{F} = \frac{1}{m_K} \sum_{\sigma \in \Sigma_K} F_{K,\sigma} m_\sigma \,,$$
$$(\nabla_\Sigma \boldsymbol{a})_{K,\sigma} = \nabla_{K,\sigma} \boldsymbol{a} \coloneqq \frac{a_L - a_K}{d_\sigma} \,.$$

The duality relation holds at the discrete level: $\langle \nabla_\Sigma \boldsymbol{a}, \boldsymbol{F} \rangle_{\mathbb{F}_{\mathcal{T}}} = -\langle \boldsymbol{a}, \mathrm{div}_{\mathcal{T}} \boldsymbol{F} \rangle_{\mathcal{T}}$, $\forall \boldsymbol{a} \in \mathbb{R}^{\mathcal{T}}$. In order to reconstruct the density on the diamond cells, we use the upwind reconstruction presented in Chapter 3. Given a discrete vector field $\boldsymbol{v} \in \mathbb{F}_{\mathcal{T}}$, we can define the upwind reconstruction operator $\mathcal{U}_\Sigma[\boldsymbol{v}] : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}^{\Sigma}$ as

$$(\mathcal{U}_\Sigma[\boldsymbol{v}]\boldsymbol{a})_\sigma = a_K \, \mathbb{1}_{v_{K,\sigma} \leq 0} + a_L \, \mathbb{1}_{v_{L,\sigma} < 0} \,, \quad \forall \boldsymbol{a} \in \mathbb{R}^{\mathcal{T}} \,. \tag{A.3}$$

A.2 Monotone discretization

Consider an integer $N > 0$ and a discretization of the time interval $[0, 1]$ in $N+1$ subintervals of constant length $\Delta t = \frac{1}{N+1}$. We denote by $\boldsymbol{\rho} = (\boldsymbol{\rho}^k)_{k=0}^{N+1}$ and $\boldsymbol{F} = (\boldsymbol{F}^k)_{k=1}^{N+1}$ the time evolutions of discrete density and momentum, where $\boldsymbol{\rho}^k \in \mathbb{R}^{\mathcal{T}}$ and $\boldsymbol{F}^k \in \mathbb{F}_{\mathcal{T}}$. Using the reconstruction (A.3) and a left endpoint approximation for the discretization of the kinetic energy, the discrete kinetic energy functional $\mathcal{B}_{N,\mathcal{T}} : [\mathbb{R}^{\mathcal{T}}]^{N+2} \times [\mathbb{F}_{\mathcal{T}}]^{N+1} \to [0, +\infty]$ is given by

$$\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) \coloneqq \begin{cases} \sum_{k=1}^{N+1} \Delta t \sum_{\sigma \in \Sigma} B\big((\mathcal{U}_\Sigma[\boldsymbol{F^k}] \boldsymbol{\rho}^{k-1})_\sigma, F_\sigma^k \big) m_\sigma d_\sigma & \text{if } \rho_K^k \geq 0, \\ +\infty & \text{else,} \end{cases} \tag{A.4}$$

where we recall that $B : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty]$ is defined by

$$B(p, \boldsymbol{Q}) \coloneqq \begin{cases} \frac{|\boldsymbol{Q}|^2}{2p} & \text{if } p > 0 \,, \\ 0 & \text{if } p = 0, \, \boldsymbol{Q} = 0 \,, \\ +\infty & \text{else} \,. \end{cases} \tag{A.5}$$

Denote $x^- = \min(x, 0)$ and $x^+ = \max(x, 0)$. Thanks to the definition of the space of conservative fluxes (A.2) and the definition of the upwind reconstruction (A.3), the function $\mathcal{B}_{N,\mathcal{T}}$ can be rewritten as[1]

$$\mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) = \sum_{k=1}^{N+1} \Delta t \sum_{\sigma\in\Sigma} B\big(\boldsymbol{\rho}_L^{k-1}, (F_{K,\sigma}^k)^+\big) m_\sigma d_\sigma + B\big(\boldsymbol{\rho}_K^{k-1}, (F_{L,\sigma}^k)^+\big) m_\sigma d_\sigma \,, \tag{A.6}$$

and it is therefore convex and lower semi-continuous, as composition of convex and lower semi-continuous functions.

Given two discrete densities $\boldsymbol{\rho}^{in}, \boldsymbol{\rho}^f \in \mathbb{R}_+^{\mathcal{T}}$ with the same total mass, $\langle \boldsymbol{\rho}^{in}, \mathbf{1}\rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}^f, \mathbf{1}\rangle_{\mathcal{T}}$, the discrete transport problem is formulated as

$$\inf_{(\boldsymbol{\rho},\boldsymbol{F})\in\mathcal{C}_{N,\mathcal{T}}} \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) \tag{A.7}$$

where $\mathcal{C}_{N,\mathcal{T}} \subset [\mathbb{R}^{\mathcal{T}}]^{N+2} \times [\mathbb{F}_{\mathcal{T}}]^{N+1}$ is the convex subset whose elements $(\boldsymbol{\rho}, \boldsymbol{F})$ satisfy both the discrete continuity equation

$$\frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} + \mathrm{div}_{\mathcal{T}} \boldsymbol{F}^k = 0 \,, \quad \forall k \in \{1, .., N+1\}, \tag{A.8}$$

and the initial and final conditions

$$\boldsymbol{\rho}^0 = \boldsymbol{\rho}^{in}, \quad \boldsymbol{\rho}^{N+1} = \boldsymbol{\rho}^f \,. \tag{A.9}$$

We will enforce explicitly this last constraint. As for the discrete problem (2.11) we presented in Chapter 2, we can state existence of a discrete solution, the proof being identical (Theorem 2.3). (A.7) can be written as

$$\inf_{(\boldsymbol{\rho},\boldsymbol{F})} \sup_{\boldsymbol{\phi}} \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{F}) \,, \tag{A.10}$$

where the Lagrangian $\mathcal{L}_{N,\mathcal{T}}$ is defined as in (2.15):

$$\mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{F}) = \mathcal{B}_{N,\mathcal{T}}(\boldsymbol{\rho}, \boldsymbol{F}) + \sum_{k=1}^{N+1} \Delta t \langle \boldsymbol{\phi}^k, \frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} + \mathrm{div}_{\mathcal{T}} \boldsymbol{F}^k \rangle_{\mathcal{T}} \,.$$

Again, strong duality holds and a primal-dual solution exists, which we can characterize as the saddle point of the Lagrangian function. The dual problem is given by

$$\sup_{\boldsymbol{\phi}} \inf_{(\boldsymbol{\rho},\boldsymbol{F})} \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}, \boldsymbol{F}) \,. \tag{A.11}$$

Taking in (A.11) the optimality conditions in $\boldsymbol{F}$, thanks to (A.6) and (A.5), provides

$$F_{K,\sigma}^k = \begin{cases} \rho_K^{k-1} \nabla_{K,\sigma} \boldsymbol{\phi}^k & \text{if } \phi_K^k \geq \phi_L^k \,, \\ \rho_L^{k-1} \nabla_{K,\sigma} \boldsymbol{\phi}^k & \text{otherwise} \,, \end{cases} \tag{A.12}$$

[1] The upwind notation here and in the following is switched with respect to the notation we set in Chapter 3, due to the fact that the flux $\boldsymbol{F}$ we are considering here has the opposite sign with respect to the flux considered there.

$\forall K \in \mathcal{T}, \forall \sigma \in \Sigma_K, \forall k \in \{1, .., N+1\}$. In order to simplify the notation, let us denote the reconstructed density by $\boldsymbol{\rho}_\Sigma \in [\mathbb{R}^\Sigma]^{N+1}$,

$$(\boldsymbol{\rho}_\Sigma^k)_\sigma = \rho_\sigma^k = \begin{cases} \rho_K^{k-1} & \text{if } \phi_K^k \geq \phi_L^k \,, \\ \rho_L^{k-1} & \text{otherwise} \,, \end{cases} \tag{A.13}$$

$\forall k \in \{1, .., N\}$. We can write then the optimal flux as

$$\boldsymbol{F}^k = \boldsymbol{\rho}_\Sigma^k \odot \nabla_\Sigma \boldsymbol{\phi}^k, \quad \forall k \in \{1, .., N+1\} \,. \tag{A.14}$$

Replacing (A.14) inside the Lagrangian we reduce it to:

$$\mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}) = -\frac{\Delta t}{2} \sum_{k=1}^{N+1} \langle \boldsymbol{\rho}_\Sigma^k, (\nabla_\Sigma \boldsymbol{\phi}^k)^2 \rangle_\Sigma + \sum_{k=1}^{N+1} \Delta t \langle \boldsymbol{\phi}^k, \frac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} \rangle_\mathcal{T} \,. \tag{A.15}$$

In order to take the optimality conditions with respect to $\boldsymbol{\rho}$, notice that we can rewrite the kinetic energy term, for each time step $k \in \{1, .., N+1\}$, as:

$$\begin{aligned} \langle \boldsymbol{\rho}_\Sigma^k, (\nabla_\Sigma \boldsymbol{\phi}^k)^2 \rangle_\Sigma &= \sum_{\sigma \in \Sigma} \rho_\sigma^k (\nabla_{K,\sigma} \boldsymbol{\phi}^k)^2 m_\sigma d_\sigma \\ &= \sum_{\sigma \in \Sigma} \rho_K^{k-1} \Big((\nabla_{K,\sigma} \boldsymbol{\phi}^k)^- \Big)^2 + \rho_L^{k-1} \Big((\nabla_{L,\sigma} \boldsymbol{\phi}^k)^- \Big)^2 m_\sigma \mathrm{d}_\sigma \\ &= \sum_{K \in \mathcal{T}} \rho_K^{k-1} \Big(\sum_{\sigma \in \Sigma_K} \Big((\nabla_{K,\sigma} \boldsymbol{\phi}^k)^- \Big)^2 \frac{m_\sigma d_\sigma}{m_K} \Big) m_K \,. \end{aligned}$$

The optimality condition in $\boldsymbol{\rho}$, the discrete Hamilton-Jacobi equation, is then

$$\frac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t} + \frac{1}{2} \mathcal{U}_\mathcal{T} \Big((\nabla_\Sigma \boldsymbol{\phi}^{k+1})^- \Big)^2 \leq 0 \,, \quad \forall k \in 1, .., N \,, \tag{A.16}$$

with the vectorial notation $(\boldsymbol{v}^-)^2 = ((v_{K,\sigma}^-)^2)_{K,\sigma} \in \mathbb{R}^{2\Sigma}$ for $\boldsymbol{v} \in \mathbb{F}_\mathcal{T}$, and defining $\mathcal{U}_\mathcal{T} : \mathbb{R}^{2\Sigma} \to \mathbb{R}^\mathcal{T}$ the reconstruction operator given by:

$$(\mathcal{U}_\mathcal{T} \, \boldsymbol{u})_K = \sum_{\sigma \in \Sigma_K} u_{K,\sigma} \frac{m_\sigma d_\sigma}{m_K} \,.$$

In order to reconstruct the squared norm of the gradient of the discrete potential, on each cell, only the negative gradients are retained. Note that $(\boldsymbol{v}^+)^2$ is not a conservative flux, as the negative part on each edge is discarded. As the equality holds in (A.16) in the cells K where $\rho_K^k > 0$, we can finally write the discrete dual problem as

$$\sup_{\boldsymbol{\phi} \in \mathcal{K}_{N,\mathcal{T}}} \langle \boldsymbol{\phi}^{N+1}, \boldsymbol{\rho}^f \rangle_\mathcal{T} - \langle \boldsymbol{\phi}^1 + \frac{\Delta t}{2} \mathcal{U}_\mathcal{T} \left((\nabla_\Sigma \boldsymbol{\phi}^1)^- \right)^2, \boldsymbol{\rho}^{in} \rangle_\mathcal{T} \tag{A.17}$$

where $\mathcal{K}_{N,\mathcal{T}} \subset [\mathbb{R}^\mathcal{T}]^{N+1}$ is the convex subset of potentials $\boldsymbol{\phi}$ verifying (A.16).

A saddle point $(\boldsymbol{\phi}, \boldsymbol{\rho})$ of the Lagrangian (A.15) satisfies the following system of necessary and sufficient optimality conditions:

$$\begin{cases} \dfrac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} + \mathrm{div}_{\mathcal{T}}(\boldsymbol{\rho}_{\Sigma}^k \odot \nabla_{\Sigma}\boldsymbol{\phi}^k) = 0\,, \quad \forall k \in 1, .., N+1, \\ \dfrac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t} + \dfrac{1}{2}\mathcal{U}_{\mathcal{T}}\Big((\nabla_{\Sigma}\boldsymbol{\phi}^{k+1})^-\Big)^2 \leq 0\,, \quad \forall k \in 1, .., N, \end{cases} \tag{A.18}$$

We want to show that the discretization we obtained for the Hamilton-Jacobi equation preserves its monotonicity in the positive direction of time and, consequently, that the equation can be saturated. For this, we will need again Lemma 3.6 that we proved in Chapter 3. We introduce again the function $\boldsymbol{\mathcal{H}} \in C^1(\mathbb{R}^{\mathcal{T}}; \mathbb{R}^{\mathcal{T}})$ defined as[2]

$$\mathcal{H}(\boldsymbol{a}) := \boldsymbol{a} + \frac{\Delta t}{2}\mathcal{U}_{\mathcal{T}}\Big((\nabla_{\Sigma}\boldsymbol{a})^-\Big)^2\,, \quad \forall \boldsymbol{a} \in \mathbb{R}^{\mathcal{T}}\,.$$

Theorem A.1. *A solution $(\boldsymbol{\phi}, \boldsymbol{\rho})$ of the following system of equations,*

$$\begin{cases} \dfrac{\boldsymbol{\rho}^k - \boldsymbol{\rho}^{k-1}}{\Delta t} + \mathrm{div}_{\mathcal{T}}(\boldsymbol{\rho}_{\Sigma}^{k-1} \odot \nabla_{\Sigma}\boldsymbol{\phi}^k) = 0\,, \quad \forall k \in 1, .., N+1, \\ \dfrac{\boldsymbol{\phi}^{k+1} - \boldsymbol{\phi}^k}{\Delta t} + \dfrac{1}{2}\mathcal{U}_{\mathcal{T}}\left((\nabla_{\Sigma}\boldsymbol{\phi}^{k+1})^-\right)^2 = 0\,, \quad \forall k \in 1, .., N, \end{cases} \tag{A.19}$$

is a saddle-point of the Lagrangian (A.15).

Proof. We know that there exists a saddle point $(\widehat{\boldsymbol{\phi}}, \boldsymbol{\rho})$ for the Lagrangian (A.15) and it satisfies the system of necessary and sufficient optimality conditions (A.18). We introduce a new potential $\boldsymbol{\phi} \in [\mathbb{R}^{\mathcal{T}}]^{N+1}$ recursively defined as follows: $\boldsymbol{\phi}^1 = \widehat{\boldsymbol{\phi}}^1$ and $\forall k \in \{2, .., N+1\}$, $\boldsymbol{\phi}^k$ is defined as the solution to the equation $\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^k) = \boldsymbol{\phi}^{k-1}$. The potential $\boldsymbol{\phi}$ defined in this way is solution to the second equation of (A.19). By Lemma 3.6, given the intial condition $\widehat{\boldsymbol{\phi}}^1$, $\boldsymbol{\phi}$ exists and is unique. We want to prove by recurrence that $\boldsymbol{\phi}^k \geq \widehat{\boldsymbol{\phi}}^k$, $\forall k \in \{1, .., N+1\}$. For $k = 2$, following the proof of Theorem 3.5, by setting as in Lemma 3.6 $\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^2) = \boldsymbol{f} = \widehat{\boldsymbol{\phi}}^1$ and $\boldsymbol{\mathcal{H}}(\widehat{\boldsymbol{\phi}}^2) = \widehat{\boldsymbol{f}} \leq \widehat{\boldsymbol{\phi}}^1$, we have

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^2) \geq \boldsymbol{\mathcal{H}}(\widehat{\boldsymbol{\phi}}^2) \quad \Longrightarrow \quad \boldsymbol{\phi}^2 \geq \widehat{\boldsymbol{\phi}}^2\,.$$

For $k > 2$, $\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^k) = \boldsymbol{f} = \boldsymbol{\phi}^{k-1}$, $\boldsymbol{\mathcal{H}}(\widehat{\boldsymbol{\phi}}^k) = \widehat{\boldsymbol{f}} \leq \widehat{\boldsymbol{\phi}}^{k-1} \leq \boldsymbol{\phi}^{k-1}$ and we have

$$\boldsymbol{\mathcal{H}}(\boldsymbol{\phi}^k) \geq \boldsymbol{\mathcal{H}}(\widehat{\boldsymbol{\phi}}^k) \quad \Longrightarrow \quad \boldsymbol{\phi}^k \geq \widehat{\boldsymbol{\phi}}^k\,.$$

By recurrence, $\boldsymbol{\phi}^k \geq \widehat{\boldsymbol{\phi}}^k, \forall k \in \{1, .., N+1\}$.

For $k = N+1$, $\boldsymbol{\phi}^{N+1} \geq \widehat{\boldsymbol{\phi}}^{N+1}$, therefore

$$\langle \boldsymbol{\phi}^{N+1}, \boldsymbol{\rho}^f \rangle_{\mathcal{T}} \geq \langle \widehat{\boldsymbol{\phi}}^{N+1}, \boldsymbol{\rho}^f \rangle_{\mathcal{T}}\,,$$

[2] The function $\boldsymbol{\mathcal{H}}$ is the same as the one introduced in Chapter 3 due to the change in sign of the gradient. The Lemma 3.6 is therefore again valid.

and consequently $\boldsymbol{\phi}$ is optimal as well for the dual problem (A.17). We are left to show that the couple $(\boldsymbol{\phi}, \boldsymbol{\rho})$ is still a saddle point in order to prove that it verifies the necessary first order optimality conditions (A.18), where the Hamilton-Jacobi equation can be satured, that is it satisfies equations (A.19). For any $(\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\rho}}) \in [\mathbb{R}^{\mathcal{T}}]^{N+1} \times [\mathbb{R}_+^{\mathcal{T}}]^N$, it holds

$$\mathcal{L}_{N,\mathcal{T}}(\widetilde{\boldsymbol{\phi}}, \boldsymbol{\rho}) \leq \mathcal{L}_{N,\mathcal{T}}(\widehat{\boldsymbol{\phi}}, \boldsymbol{\rho}) = \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \boldsymbol{\rho}) \leq \mathcal{L}_{N,\mathcal{T}}(\boldsymbol{\phi}, \widetilde{\boldsymbol{\rho}}) \,,$$

which is indeed the condition on $(\boldsymbol{\phi}, \boldsymbol{\rho})$ to be a saddle point. The first inequality derives from the fact that $(\widehat{\boldsymbol{\phi}}, \boldsymbol{\rho})$ is a saddle point for $\mathcal{L}_{N,\mathcal{T}}$, the equality and the second inequality (which is actually again an equality) are due to the fact that the Hamilton-Jacobi equation is satisfied everywhere by $\boldsymbol{\phi}$. □

Thanks to theorem A.1, one can find a solution to the discrete optimal transport problem (A.7) by solving directly the system of equations (A.19). Comparing this system of equations to the system (2.19), this simply means that thanks to the specific discretization chosen, among the different possible choices for the Lagrange multiplier for the positivity constraint on the density, zero is admissible. We remark that the solution $\boldsymbol{\phi}$ which saturates the Hamilton-Jacobi equation is not necessarily unique, as it depends on the possibly non-unique initial condition (see Remark 2.5, which holds true also in this case).

The approach we presented in this section is not useful in practice for the computation of Wasserstein geodesics for several reasons. We already explained in Section 2.3 that a left/right endpoint approximation for the kinetic energy gives restrictions on the initial and final densities. The discretization is also presumably of order one in time and space (the upwind reconstruction is of order one). Although one could try to solve directly the system of equations (A.19) with a Newton scheme, the problem is too hard and it requires in any case a continuation method like the one we introduced in Section 2.6. Finally, we have shown in Section 2.7 that the preservation of the monotonicity does not prevent the stability issues.

Appendix B

A mixed finite element discretization of dynamical optimal transport

We present here another possible strategy for computing solutions to the dynamical optimal transport problem. We present a discretization based on locally conservative finite elements, which are closely related to finite volumes. Finite elements have been already employed to discretize this problem [16]. However, not in a locally conservative form. Furthermore, in [16] the authors did not discretize the variational problem rather its optimality conditions. This partially motivated this work. In this sense, what we present is a generalization of the variational and locally conservative finite difference approach proposed in [109]. In [109] the authors considered proximal splitting algorithms, an instance of primal dual methods, to solve the discrete optimization problem. We use here the same strategy. The original motivation of this work has been indeed to inspect the use of this type of techniques, more common in the optimal transport community, which could have proved useful for our finite volume discretizations. Nevertheless, we realized that they are not easy to apply to arbitrary discretizations of the problem (especially on unstructured grids). More importantly, they are efficient only as far as high accuracy is not mandatory and uniform grids are used. These reasons led us to introduce our interior point strategy (see Section 2.6).

This work is issued from:

A. Natale, G. Todeschi. A mixed finite element discretization of dynamical optimal transport. HAL: hal-02501634, 2020.

B.1 Introduction

Optimal transport provides a convenient framework for density interpolation as a convex optimization problem. Its most remarkable feature is its sensitivity to horizontal displacement, which generally allows one to retrieve translations when interpolating between two densities. This property has motivated the application of optimal transport to many imaging problems, especially in the context of physical sciences and fluid dynamics. A typical example comes from satellite image interpolation in oceanography. In this case, one is interested in recon-

structing the evolution of a quantity of interest such as Sea Surface Temperature (SST) or Sea Surface Height (SSH) between two given observations. As highlighted in [69], for this type of applications one needs to include appropriate regularization terms to avoid the appearance of unphysical phenomena such as mass concentration in the reconstructed density evolution.

In this work we propose a finite element approach to solve the dynamical formulation of optimal transport with quadratic cost on unstructured meshes (and therefore can be easily implemented on complex domains) and that can be easily modified to include different type of regularizations which are relevant for the dynamic reconstruction and interpolation of physical quantities. For some choices of finite element spaces, using the framework introduced in [79], we can prove convergence of our discrete solutions to the ones of the continuous problem.

The dynamical formulation of optimal transport inspired some of the first numerical methods for this problem. This reads as follows: given two probability measures $\rho_0, \rho_1 \in \mathcal{P}(D)$ on a compact domain $D \subset \mathbb{R}^d$, find the curve $t \in [0,1] \mapsto \rho(t,\cdot) \in \mathcal{P}(D)$ which solves

$$\inf_{\rho,v} \left\{ \int_0^1 \int_D \frac{|v(t,\cdot)|^2}{2} \mathrm{d}\rho(t,\cdot)\mathrm{d}t \,;\, \partial_t \rho + \mathrm{div}_x(\rho v) = 0, \rho(0) = \rho_0,\, \rho(1) = \rho_1 \right\} \tag{B.1}$$

where $v : [0,1] \times D \to \mathbb{R}^d$ is a time-dependent velocity field on D tangent to the boundary ∂D, and $|\cdot|$ denotes the Euclidean norm. In other words, problem (B.1) selects the curve of minimal kinetic energy with fixed endpoints ρ_0 and ρ_1.

Benamou and Brenier [14] realized that introducing the momentum $m \coloneqq \rho v$, problem (B.1) can be recast into a convex optimization problem in the variables (ρ, m), with a linear constraint, since the continuity equation becomes

$$\partial_t \rho + \mathrm{div}_x m = 0\,. \tag{B.2}$$

If we define $\sigma \coloneqq (\rho, m)$, regarded as a measure on $[0,1] \times D$, this constraint is equivalent to $\mathrm{div}\,\sigma = 0$, where now div denotes the divergence operator on the space-time domain $[0,1] \times D$. Introducing the dual variable $q = (a,b)$ where $a \in C([0,1] \times D)$ and $b \in C([0,1] \times D; \mathbb{R}^d)$, the kinetic energy minimized in (B.1) can be written in the form

$$\sup_q \left\{ \int_0^1 \int_D q \cdot \mathrm{d}\sigma \,;\, a + \frac{|b|^2}{2} \le 0 \right\}.$$

Combining this expression with (B.1) we obtain a saddle point problem in the variables (q, σ) with a nonlinear constraint on q and a linear one on σ.

The numerical method proposed in [14] involves discretizing q and σ by their values on a regular grid, and expressing the constraint on σ via a Lagrange multiplier; then the dual problem can be solved by an Augmented Lagrangian ADMM approach, optimizing separately in q and the Lagrange multiplier and then performing a gradient descent step on σ. Disregarding the discretization in space-time, the convergence of the method has been studied in [66, 70]. The same approach was used to discretize different problems related to optimal transport (e.g., gradient flows [17], mean field games [16], unbalanced optimal transport [61]) using a finite element discretization in space-time. Importantly, in these cases the numerical method is obtained by discretizing the several steps of the augmented Lagrangian approach rather than as a discrete optimization algorithm. This implies that in general it is difficult to establish the convergence of the discrete algorithms. Moreover, for these type of methods,

convergence results towards the continuous solutions with mesh refinement are only available for specific settings (e.g., the L^1-type optimal transport problems studied in [71]), but they are not available for the optimal transport problem (B.1).

Papadakis, Peyré, and Oudet proposed in [109] a staggered finite difference discretization on regular grids of the optimal transport problem (B.1), and they considered different proximal splitting algorithms to solve it. The computational bottleneck for these methods as well as for the original augmented Lagrangian approach is the projection onto the space of divergence-free vector fields σ, which amounts to solving a Poisson equation at each iteration. This however can be avoided by exploiting the Helmholtz decomposition of vector fields, as recently showed in [68], or adding regularization terms as in [85]. Recently, Carrillo and collaborators [41] proposed a finite difference scheme similar to that in [109] (in the context of the discretization of Wasserstein gradient flows), for which they could also prove its convergence with mesh refinement, but only upon strong regularity assumptions on the solutions of the continuous problem.

In [80] a numerical scheme was proposed using tools from finite element and finite volume methods, where one explicitly constructs a duality structure for the discrete variables. Later Lavenant [79] proved convergence of this scheme, unconditionally with respect to the time/space step size, to the solutions of the optimal transport problem, proposing a general framework for convergence of discretizations of problem (B.1) between two arbitrary probability measures. This filled a critical gap for the analysis of discrete dynamical transport models, since previously convergence results were only known in case of sufficiently smooth solutions (as in [41]) or conditional to the relative time/space step sizes (e.g., in the context of finite volume methods, combining the results in [52] and [64]).

B.1.1 Contributions and structure of the chapter

We propose here a mixed finite element discretization of (B.1) which generalizes to the finite element setting the finite difference scheme proposed by Papadakis et al. [109]. We derive our method by discretizing a saddle point formulation of the dynamic optimal transport problem on Hilbert spaces, where one looks for a solution $(q, \sigma) \in L^2([0,1] \times D; \mathbb{R}^{d+1})^2$. Nonetheless, we stress that the method we obtain is still well-defined when the initial and final data are arbitrary probability measures. By using $H(\mathrm{div})$-conforming spaces for the variable σ, we are able to construct discrete solutions that satisfy exactly the weak form of the continuity equation (B.2).

Using the framework of [79], we also show that our discrete solutions, for specific choices of finite element spaces, converge towards the solutions of the optimal transport problem between two arbitrary measures, and therefore even when the solution σ is only a measure (see Theorem B.7). Such a result carries over also to a slight modification of the finite difference scheme proposed in [109], which can be viewed as a particular instance of our discretization on a uniform quadrilateral grid (see Remark B.9).

Finally, as in [109], we solve the discrete problem using a proximal splitting algorithm [112]. Importantly, this is not only a discretization of the same algorithm applied to the continuous saddle point formulation as in previous works, but also a genuine optimization scheme applied to the finite dimensional problem. Furthermore, we observe numerically that the proposed modification of the finite difference scheme in [109] (which we derived to prove convergence with mesh refinement) also yields a remarkable speedup for the convergence of

the proximal splitting algorithm itself, keeping approximately the same computational cost per iteration.

The chapter is structured as follows. We establish the notation in Section B.2. In Section B.3 we give the precise formulation of problem (B.1) and describe the proximal splitting algorithm applied to the continuous problem in the Hilbert space setting. In Section B.4 we introduce and discuss the main finite element tools we use for our method. In Section B.5 we define our finite element discretization of problem (B.1) and state the convergence result. In Section B.6 we detail the steps required for solving our discrete optimal transport problem with a proximal splitting algorithm. In Section B.7 we describe how to introduce regularization terms in the formulation. Finally in Section B.8 we present some numerical results.

B.2 Notation

Throughout this work we will denote by $D \subset \mathbb{R}^d$ a convex polytope, with $d \in \{2, 3\}$, and by $\Omega \coloneqq [0, 1] \times D$ the space-time domain. For differential operators such as ∇ or div, we use the subscript x to emphasize that these are defined on D rather than Ω, but we will drop this subscript when this is clear from the context.

We use the standard notation for Sobolev spaces on D or Ω. In particular, $L^p(D; \mathbb{R}^d)$ denotes the space of functions $f : D \to \mathbb{R}^d$ whose Euclidean norm $|f|$ is in $L^p(D)$. We use a similar notation for functions taking values on a subset $K \subset \mathbb{R}^d$, or defined on Ω. We denote by $H(\mathrm{div}; D)$ the space of vector fields $f : D \to \mathbb{R}^d$ in $L^2(D; \mathbb{R}^d)$ whose divergence is in $L^2(D)$. Similarly, $H(\mathrm{div}; \Omega)$ the space of vector fields $f : \Omega \to \mathbb{R}^{d+1}$ in $L^2(\Omega; \mathbb{R}^{d+1})$ whose divergence is in $L^2(\Omega)$.

Finally, we denote by $\mathcal{M}(D)$ the set of finite signed measures on D, by $\mathcal{M}_+(D) \subset \mathcal{M}(D)$ the convex subset of positive measures; by $\mathcal{P}(D) \subset \mathcal{M}_+(D)$ the set of positive measures of total mass equal to one; and by $C(D)$ the space of continuous functions on D. We use a similar notation for the spaces of measures and continuous functions on Ω. We use $\langle \cdot, \cdot \rangle$ to denote either the duality pairing between measures and continuous functions or the L^2 inner product, on either D or Ω, according to the context.

B.3 Dynamical formulation of optimal transport

The dynamical optimal transport problem (B.1) can be formulated as a saddle point problem on the space of measures $\sigma \coloneqq (\rho, m) \in \mathcal{M}(\Omega) \times \mathcal{M}(\Omega)^d$. This can be written as follows

$$\inf_{\sigma \in \mathcal{C}} \mathcal{A}(\sigma), \quad \mathcal{A}(\sigma) \coloneqq \sup_{q \in C(\Omega; K)} \langle q, \sigma \rangle, \tag{B.3}$$

where $\mathcal{C}$ is the set of measures $\sigma \in \mathcal{M}(\Omega)^{d+1}$ satisfying $\mathrm{div}\, \sigma = 0$ in distributional sense with boundary conditions

$$\sigma \cdot n_{\partial\Omega} = \mathcal{X}\,, \quad \mathcal{X} \coloneqq \begin{cases} \rho_0 & \text{on } \{0\} \times D, \\ \rho_1 & \text{on } \{1\} \times D, \\ 0 & \text{otherwise,} \end{cases} \tag{B.4}$$

with $\rho_0, \rho_1 \in \mathcal{P}(D)$, and where $C(\Omega; K)$ is the space of continuous functions on Ω taking value in the convex set

$$K \coloneqq \left\{ (a, b) \in \mathbb{R} \times \mathbb{R}^d \,;\, a + \frac{|b|^2}{2} \leq 0 \right\}. \tag{B.5}$$

It will be convenient to treat time and space as separate variables. In particular we will also use the action defined by

$$A(\rho, m) \coloneqq \sup_{(a,b) \in C(D;K)} \langle \rho, a \rangle + \langle m, b \rangle \,,$$

for any $(\rho, m) \in \mathcal{M}(D)^{d+1}$. Then, $A(\rho, m)$ is finite if and only if m has a density with respect to ρ and in that case $A(\rho, m) = \int_D B(\rho, m)$, where $B : \mathbb{R} \times \mathbb{R}^d \to [0, +\infty]$ is the function given by

$$B(a, b) \coloneqq \begin{cases} \frac{|b|^2}{2a} & \text{if } a > 0, \\ 0 & \text{if } a = 0, b = 0, \\ +\infty & \text{if } a = 0, b \neq 0 \text{ or } a < 0 \,. \end{cases}$$

Due to the definition of the function B, any saddle point of problem (B.3) must satisfy $\rho \geq 0$.

The value of the infimum of problem (B.3) coincides with $W_2^2(\rho_0, \rho_1)/2$, where $W_2(\cdot, \cdot)$ denotes the Wasserstein distance associated with the L^2 cost (see Theorem 5.28 in [115]). Moreover the infimum itself is attained by a measure $\sigma = (\rho, m)$, where ρ is known as the Wasserstein geodesic between ρ_0 and ρ_1 (see proposition 5.32 in [115]). We refer the reader to [115] for more details on the links between the dynamical formulation and the Wasserstein distance.

B.3.1 Hilbert space setting and proximal splitting

Before discussing the discretization of problem (B.3), we review its reformulation on Hilbert spaces, and discuss the convergence of the proximal splitting algorithm.

Proposition B.1 (Guittet [66]; Hug et al. [70])**.** *Suppose $\rho_0, \rho_1 \in L^2(D)$. Then problem (B.3) is equivalent to*

$$\inf_{\sigma \in \mathcal{C}} \sup_{q \in L^2(\Omega; K)} \langle q, \sigma \rangle \,, \tag{B.6}$$

where $\mathcal{C}$ is the set of functions $\sigma \in H(\operatorname{div}; \Omega)$ satisfying $\operatorname{div} \sigma = 0$ in weak sense with boundary conditions given by (B.4). Moreover, assuming that $\operatorname{supp}(\rho_0) \cup \operatorname{supp}(\rho_1) \subset \mathring{D}$, there exists a saddle point $(\sigma^*, q^*) \in \mathcal{C} \times L^2(\Omega; K)$ solving problem (B.6).

The equivalence of problem (B.6) to (B.3) can be easily deduced by a regularization argument on σ and then applying Lusin's theorem as in Proposition 5.18 in [115]. The proof for the existence of a saddle point problem is more delicate and can be found in [70].

In order to apply a proximal splitting algorithm to solve problem (B.6), we first write it in the form

$$\inf_{\sigma \in L^2(\Omega; \mathbb{R}^{d+1})} \sup_{q \in L^2(\Omega; \mathbb{R}^{d+1})} \langle q, \sigma \rangle + \iota_{\mathcal{C}}(\sigma) - \iota_{\mathcal{K}}(q) \,, \tag{B.7}$$

where ι denotes the convex indicator function and

$$\mathcal{K} \coloneqq L^2(\Omega; K) = \{ q \in L^2(\Omega; \mathbb{R}^{d+1}) \,;\, q \in K \text{ a.e.} \}.$$

Note in particular that $\mathcal{C}$ and $\mathcal{K}$ are closed convex sets of L^2.

We apply to (B.7) the primal-dual projection algorithm proposed in [112]. In particular, given $\tau_1, \tau_2 > 0$ and an admissible $(\sigma^0, q^0) \in \mathcal{C} \times \mathcal{K}$, we define the sequence $\{(\sigma^k, q^k)\}_k$ by the two-step algorithm:

$$\textbf{Step 1}: \quad \sigma^{k+1} = P_{\mathcal{C}}(\sigma^k - \tau_1 q^k)\,. \tag{B.8a}$$

$$\textbf{Step 2}: \quad q^{k+1} = P_{\mathcal{K}}(q^k + \tau_2(2\sigma^{k+1} - \sigma^k))\,. \tag{B.8b}$$

where $P_{\mathcal{C}}$ and $P_{\mathcal{K}}$ are the L^2 projections on the closed convex sets $\mathcal{C}$ and $\mathcal{K}$, respectively. The projection onto $\mathcal{C}$ amounts to computing the Helmholtz decomposition of $\sigma^k - \tau_1 q^k$ and selecting the divergence-free part, whereas the projection onto $\mathcal{K}$ is a pointwise projection applied to a representative of $q^k + \tau_2(2\sigma^{k+1} - \sigma^k)$.

The proof of convergence in [112] holds also in our setting. More precisely, the following convergence theorem holds.

Theorem B.2 (Pock et al. [112])**.** *If $\tau_1\tau_2 < 1$ then $(\sigma^k, q^k) \to (\sigma^*, q^*) \in \mathcal{C} \times \mathcal{K}$ which solves (B.6).*

Discretizing problem (B.7), and consequently the proximal splitting algorithm (B.8), with finite elements requires choosing finite-dimensional spaces for σ and q so that the steps in (B.8) are well-posed and computationally feasible. However, satisfying these requirements is not enough to guarantee convergence of the discrete solutions to the ones of the infinite dimensional problem. Hereafter we will identify a class of finite element spaces for which the theory developed in [79] applies, which allows us to deduce convergence to the solutions of problem (B.3), i.e. even when ρ_0 and ρ_1 are arbitrary probability measures and the Hilbert space setting presented in this section is not well-defined.

B.4 Mixed finite element setting

B.4.1 Finite element spaces on D

We recall that D is a convex polytope in $\mathbb{R}^d$, with $d \in \{2, 3\}$. We consider a triangulation of D which we denote $\mathcal{T}_h$, i.e. a decomposition of D in either simplicial or quadrilateral (disjoint) elements, where h is the maximum diameter of the elements in $\mathcal{T}_h$. We assume that there exists a constant C_{mesh} such that

$$|h|^d \leq C_{mesh}|T|\,, \quad \forall\, T \in \mathcal{T}_h\,. \tag{B.9}$$

This implies that the mesh is quasiuniform, meaning that the ratio of any two element diameters is uniformly bounded by a constant depending only on C_{mesh}, and shape-regular, that is, for each element $T \in \mathcal{T}_h$, the ratio of its diameter and the diameter of the largest inscribed ball is uniformly bounded by a constant depending only on C_{mesh} (see, e.g., [8]).

For any $T \in \mathcal{T}_h$, we denote by $\mathcal{P}_k(T)$ the space of polynomials of degree up to k on T. If T is a quadrilateral element, i.e., in general, if T is obtained by an affine transformation $\phi : I^d \to T$ where I is the unit interval, then we define $\mathcal{P}_{k_1,\dots k_d}(I^d) \coloneqq \mathcal{P}_{k_1}(I) \otimes \dots \otimes \mathcal{P}_{k_d}(I)$ and $\mathcal{P}_{k_1,\dots k_d}(T) \coloneqq \mathcal{P}_{k_1,\dots k_d}(I^d) \circ \phi^{-1}$.

We now define the finite element spaces Q_h and V_h which will serve to construct approximations of the density ρ and the momentum m, respectively. We set

$$Q_h \coloneqq \{\varphi \in L^2(D)\,;\, \varphi|_T \in \mathcal{P}_0(T),\ \forall T \in \mathcal{T}_h\},$$
$$V_h \coloneqq \{v \in H(\mathrm{div}; D)\,;\, v|_T \in V_h(T),\ \forall T \in \mathcal{T}_h\}.$$

where $V_h(T)$ is the so-called shape function space. We distinguish two cases:

1. for simplicial elements (triangles or tetrahedrons), we take $V_h(T)$ to be either

$$\mathcal{RT}_0(T) \coloneqq \{v = v_0 + v_1\hat{x}\,;\, v_0 \in (\mathcal{P}_0(T))^d\,, v_1 \in \mathcal{P}_0(T)\} \subset (\mathcal{P}_1(T))^d,$$

where $\hat{x} = (x_1, \dots, x_d) \in (\mathcal{P}_1(T))^d$, which generates the lowest order Raviart-Thomas space; or $\mathcal{BDM}_1(T) = (\mathcal{P}_1(T))^d$, which generates the lowest order Brezzi-Douglas-Marini $H(\mathrm{div})$-conforming space;

2. for quadrilateral elements, we set $T = \phi(I^d)$, where I is an interval and ϕ an affine transformation, and we take $V_h(T)$ to be the tensor product space which generates the lowest order Raviart-Thomas space on quadrilateral elements. This is defined as follows:

$$\mathcal{RT}_{[0]}(T) \coloneqq \begin{cases} \mathcal{P}_{1,0}(T)e_1 + \mathcal{P}_{0,1}(T)e_2 & \text{if } d = 2\,, \\ \mathcal{P}_{1,0,0}(T)e_1 + \mathcal{P}_{0,1,0}(T)e_2 + \mathcal{P}_{0,0,1}(T)e_3 & \text{if } d = 3\,, \end{cases}$$

where $\{e_i\}_i$ is the basis for $\mathbb{R}^d$ aligned with the edges of T.

In other words, the space V_h is chosen as one of the standard lowest order $H(\mathrm{div})$-conforming spaces. In fact, the property of being piece-wise linear will be crucial in the following, namely to prove the convergence result in Theorem B.7 (see, in particular, Proposition B.12 in the appendix). A graphical representation of the degrees of freedom associated with these spaces is shown in figure B.1.

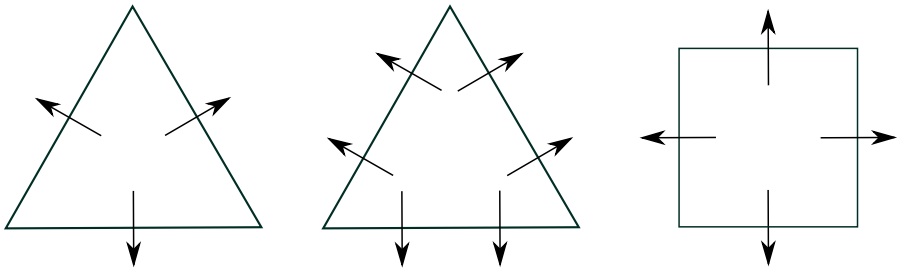
Importantly, with the choices mentioned above, one can define projection operators $\Pi_{Q_h} : L^2(D) \to Q_h$ and $\Pi_{V_h} : \mathcal{V}_D \subset H(\mathrm{div}; D) \to V_h$ that commute with the divergence operator [8, 24], where $\mathcal{V}_D$ is a dense subset of sufficiently smooth vector fields. By an appropriate regularization procedure of such operators (see, e.g., Section 5.4 in [8]), one can construct bounded projections $\tilde{\Pi}_{Q_h} : L^2(D) \to Q_h$ and $\tilde{\Pi}_{V_h} : H(\mathrm{div}; D) \to V_h$ satisfying a similar property. In other words, the following diagram commutes

$$\begin{array}{ccc} H(\mathrm{div}; D) & \xrightarrow{\mathrm{div}} & L^2(D) \\ \downarrow{\scriptstyle \tilde{\Pi}_{V_h}} & & \downarrow{\scriptstyle \tilde{\Pi}_{Q_h}} \\ V_h & \xrightarrow{\mathrm{div}} & Q_h \end{array}$$

As a consequence, the divergence operator is surjective onto Q_h when restricted on V_h, i.e. $\mathrm{div}\, V_h = Q_h$. Finally, we let $Q_h^+ \subset Q_h$ the convex subset of non-negative piecewise constant functions.

Remark B.3. *For the proof of Theorem B.7 in the appendix, we will consider as commuting projections Π_{V_h} and Π_{Q_h} the canonical projections defined in Section 5.2 of [8]. Here, we will only need the explicit definition of Π_{Q_h}, which is given by*

$$\Pi_{Q_h}\rho|_T = \frac{1}{|T|}\int_T \rho\,, \tag{B.10}$$



(a) $\mathcal{RT}_0$ (b) $\mathcal{BDM}_1$ (c) $\mathcal{RT}_{[0]}$

Figure B.1: Degrees of freedom for different choices of shape function space $V_h(T)$

for any $T \in \mathcal{T}_h$. Note, in particular, that Π_{Q_h} is well-defined on $\mathcal{M}(D)$ and its restriction on $\mathcal{M}_+(D)$ is surjective onto Q_h^+.

B.4.2 Finite element spaces on Ω

We now introduce finite element spaces on the space-time domain $[0,1] \times D$. We first define a decomposition $\mathcal{T}_{h,\tau}$, obtained by a tensor product construction. In other words, we assume that $\mathcal{T}_{h,\tau}$ is obtained by tensor product of a triangulation $\mathcal{T}_h$ of D and a decocomposition of $[0,1]$ of maximum size τ, so that any element $S \in \mathcal{T}_{h,\tau}$ is of the form $S = [t_0, t_1] \otimes T$ where $T \in \mathcal{T}_h$.

We now define the finite element spaces $F_{h,\tau}$ and $Z_{h,\tau}$ on the space-time domain. The space $Z_{h,\tau}$ will be constructed using the standard tensor product construction based on the spaces Q_h and V_h defined on D, and continuous $\mathcal{P}_1$ and discontinuous $\mathcal{P}_0$ spaces on $[0,1]$. In our discretization, the space-time vector field (ρ, m) will be an element of $Z_{h,\tau}$ whereas $F_{h,\tau}$ will be the space of discrete Lagrange multipliers associated with the continuity equation, which is equivalent to the constraint that the space-time divergence of (ρ, m) is zero.

More precisely, we define

$$F_{h,\tau} \coloneqq \{\phi \in L^2(\Omega)\,;\, \phi|_S \in \mathcal{P}_0(S),\ \forall\, S \in \mathcal{T}_{h,\tau}\},$$

$$Z_{h,\tau} \coloneqq \{v \in H(\mathrm{div};\Omega)\,;\, v|_S \in Z_{h,\tau}(S),\ \forall\, S \in \mathcal{T}_{h,\tau}\}.$$

For $S = [t_0, t_1] \otimes T$, the shape function space $Z_{h,\tau}(S)$ is built by defining a shape function space for the density, in the space-time domain, which is given by

$$Q_{h,\tau}(S) \coloneqq \mathcal{P}_1([t_0, t_1]) \otimes Q_h(T)$$

(i.e. the density is piecewise linear in time), and a shape function space for the momentum, in the space-time domain, which is given by

$$V_{h,\tau}(S) \coloneqq \mathcal{P}_0([t_0, t_1]) \otimes V_h(T)$$

(i.e. the momentum is piecewise constant in time). Then, we set

$$Z_{h,\tau}(S) \coloneqq (Q_{h,\tau}(S)\,\hat{t}) \oplus V_{h,\tau}(S)\,,$$

where $\hat{t}$ is the unit vector oriented in the time direction. The spaces $F_{h,\tau}$ and $Z_{h,\tau}$ inherit from Q_h and V_h the commuting diagram property mentioned above. In particular, there exist bounded projections $\tilde{\Pi}_{F_{h,\tau}} : L^2(\Omega) \to F_{h,\tau}$ and $\tilde{\Pi}_{Z_{h,\tau}} : H(\mathrm{div};\Omega) \to Z_{h,\tau}$ for which the following diagram commutes

$$\begin{array}{ccc} H(\mathrm{div};\Omega) & \xrightarrow{\mathrm{div}} & L^2(\Omega) \\ \downarrow{\scriptstyle \tilde{\Pi}_{Z_{h,\tau}}} & & \downarrow{\scriptstyle \tilde{\Pi}_{F_{h,\tau}}} \\ Z_{h,\tau} & \xrightarrow{\mathrm{div}} & F_{h,\tau} \end{array} \tag{B.11}$$

where the divergence is the one associated with the space-time domain Ω. Then, as before, the divergence operator is surjective onto $F_{h,\tau}$ when restricted on $Z_{h,\tau}$, i.e. $\mathrm{div}\, Z_{h,\tau} = F_{h,\tau}$. Note that the precise definition for the projection operators on tensor product meshes can be found in [7].

B.4.3 Discrete projection on the divergence-free subspace

Denote by $\mathcal{B}$ the kernel of the divergence operator on $H(\mathrm{div};\Omega)$. Given $\xi \in L^2(\Omega)$ we define the projection $P_{\mathcal{B}}(\xi)$ to be the divergence-free vector field σ minimizing the L^2 distance from ξ. This can be obtained solving the following problem for $(\sigma,\phi) \in H(\mathrm{div};\Omega) \times L^2(\Omega)$

$$\begin{cases} \langle \sigma, v\rangle + \langle \phi, \mathrm{div}\, v\rangle = \langle \xi, v\rangle & \forall v \in H(\mathrm{div};\Omega)\,, \\ \langle \mathrm{div}\,\sigma, \psi\rangle = 0 & \forall \psi \in L^2(\Omega)\,. \end{cases} \tag{B.12}$$

Let $\mathcal{B}_{h,\tau}$ be the kernel of the divergence operator restricted on $Z_{h,\tau}$. We define the projection $P_{\mathcal{B}_{h,\tau}}(\xi)$ to be the divergence-free vector field $\sigma_{h,\tau} \in Z_{h,\tau}$ minimizing the L^2 distance from ξ. This can be obtained solving the following problem for $(\sigma_{h,\tau},\phi_{h,\tau}) \in Z_{h,\tau} \times F_{h,\tau}$

$$\begin{cases} \langle \sigma_{h,\tau}, v\rangle + \langle \phi_{h,\tau}, \mathrm{div}\, v\rangle = \langle \xi, v\rangle & \forall v \in Z_{h,\tau}\,, \\ \langle \mathrm{div}\,\sigma_{h,\tau}, \psi\rangle = 0 & \forall \psi \in F_{h,\tau}\,. \end{cases} \tag{B.13}$$

The commuting diagram (B.11) implies well-posedness of the discrete system. In particular, it implies the following inf-sup condition: there exists a constant $\beta > 0$ independent of h and τ such that

$$\inf_{\phi\in F_{h,\tau}} \sup_{\sigma\in Z_{h,\tau}} \frac{\langle \phi, \mathrm{div}\,\sigma\rangle}{\|\sigma\|_{H(\mathrm{div})}\|\phi\|_{L^2}} \geq \beta\,,$$

see for example proposition 5.4.2 in [24]. Then, problem (B.13) is well-posed, i.e. it has a unique solution $(\sigma_{h,\tau},\phi_{h,\tau})$ which verifies $\sigma_{h,\tau} \in \mathcal{B}$ and

$$\|\sigma_{h,\tau}\|_{L^2} \leq C_1\|\xi\|_{L^2}\,,$$

$$\|\phi_{h,\tau}\|_{L^2} \leq C_2\|\xi\|_{L^2}\,,$$

$$\|\sigma_{h,\tau} - \sigma\|_{L^2} + \|\phi_{h,\tau} - \phi\|_{L^2} \leq C_3\|\xi_{h,\tau} - \xi\|_{L^2}\,,$$

where $C_1, C_2, C_3 > 0$ are constants independent of h and τ, $\xi_{h,\tau}$ is the L^2 projection of ξ onto $Z_{h,\tau}$ and (σ,ϕ) is the unique solution of problem (B.12) (e.g., these results can be derived as particular cases of Theorems 4.3.2, 5.2.1 and 5.2.5 in [24]).

In the following we will need to compute the discrete version of the L^2 projection onto $\mathcal{C}$. In particular we define

$$\mathcal{C}_{h,\tau} \coloneqq \{\sigma \in \mathcal{B}_{h,\tau}\,,\, \sigma \cdot n_{\partial\Omega} = \mathcal{X}_{h,\tau}\}\,, \tag{B.15}$$

where, since Π_{Q_h} can be defined on $\mathcal{M}(D)$ (see equation (B.10)), we set

$$\mathcal{X}_{h,\tau} \coloneqq \begin{cases} \Pi_{Q_h}\rho_0 & \text{on } \{0\}\times D, \\ \Pi_{Q_h}\rho_1 & \text{on } \{1\}\times D, \\ 0 & \text{otherwise.} \end{cases} \tag{B.16}$$

The well-posedness results described above for the L^2 projections onto $\mathcal{B}$ and $\mathcal{B}_{h,\tau}$ hold also for the L^2 projections onto $\mathcal{C}$ and $\mathcal{C}_{h,\tau}$ up to adding Neumann boundary conditions to the spaces $H(\mathrm{div};\Omega)$ and $Z_{h,\tau}$, and replacing $L^2(\Omega)$ and $F_{h,\tau}$ by $L^2(\Omega)/\mathbb{R}$ and $F_{h,\tau}/\mathbb{R}$, respectively.

B.5 Discrete dynamical formulation and convergence

In this section we formulate the discrete problem and state a convergence result obtained by applying the theory developed in [79]. For this, we need to introduce a space for the discrete dual variable q. We adopt the same notation as for the spaces defined in Section B.4. In particular, we set for $r \in \{0,1\}$,

$$X_h^r \coloneqq \{\phi \in L^2(D)\,;\, \phi|_T \in X_h^r(T),\ \forall\, T \in \mathcal{T}_h\}.$$

The superscript r denotes the polynomial order of the shape function space $X_h^r(T)$. We distinguish two cases:

1. for simplicial elements (triangles or tetrahedrons), we take $X_h^r(T) \coloneqq \mathcal{P}_r(T)$.
2. for quadrilateral elements, we set $T = \phi(I^d)$, where I is an interval and ϕ an affine transformation, and we take $X_h^r(T) \coloneqq \mathcal{P}_r(I)^d \circ \phi^{-1}$.

The associated space-time space is defined by

$$X_{h,\tau}^r \coloneqq \{\phi \in L^2(\Omega)\,;\, \phi|_S \in X_{h,\tau}^r(S),\ \forall\, S \in \mathcal{T}_{h,\tau}\},$$

with $X_{h,\tau}^r(S) = \mathcal{P}_0([t_0,t_1]) \otimes X_h^r(T)$. In order to simplify the notation, we will omit the superscript r when not relevant to the discussion.

Remark B.4. *The choice $r \in \{0,1\}$ is dictated by computational feasibility of the algorithm. In fact, for these cases, we can compute explicitly the projection on $\mathcal{K} \cap X_{h,\tau}^r$ (with respect to appropriate inner products) as it will be explained in the next section. On the other hand, we restrict ourselves to piecewise constant functions in time since this is crucial for the convergence of the algorithm, as shown in [79].*

The discrete action (at fixed time) is defined as follows:

$$A_h(\rho,m) \coloneqq \sup_{(a,b)\in(X_h)^{d+1}} \{\langle \rho, a\rangle + \langle m, b\rangle\,;\, (a,b) \in K\ a.e.\}$$

for any $(\rho,m) \in Q_h \times V_h$. By construction, $A_h : Q_h \times V_h \to [0,+\infty]$ is a proper convex function -1-positively homogeneous in its first variable and 2-positively homogeneous in its second

variable. Moreover, it is non-increasing in its first argument, i.e. $A_h(\rho_1+\rho_2,m)\le A_h(\rho_1,m)$ for any $\rho_1,\rho_2\in Q_h^+$ and $m\in V_h$. In fact, suppose that $A_h(\rho_1+\rho_2,m)<+\infty$. Then there exists $(a^*,b^*)\in(X_h)^{d+1}\cap\mathcal{K}$ such that $\langle\rho_1+\rho_2,a^*\rangle+\langle m,b^*\rangle=A_h(\rho_1+\rho_2,m)$; in particular $a^*\le 0$. Then

$$A_h(\rho_1,m)\ge A_h(\rho_1+\rho_2,m)-\langle\rho_2,a^*\rangle\ge A_h(\rho_1+\rho_2,m)\,,$$

and by a similar reasoning we obtain that if $A(\rho_1+\rho_2,m)=+\infty$ then we also have $A(\rho_1,m)=+\infty$.

The space-time discretization of problem (B.3) is given by

$$\inf_{\substack{\sigma\in\mathcal{C}_{h,\tau},\\ \rho\ge 0}}\mathcal{A}_{h,\tau}(\sigma),\quad \mathcal{A}_{h,\tau}(\sigma):=\sup_{q\in(X_{h,\tau}^r)^{d+1}\cap\mathcal{K}}\langle q,\sigma\rangle\,. \tag{B.17}$$

Note that, by definition, $\mathcal{A}_{h,\tau}$ is convex and non-negative. Therefore, problem (B.17) always admits minimizers.

Suppose that the time discretization is given by a decomposition of the interval $[0,1]$ in N elements, i.e. fixing the points $0=t_0<t_1<\ldots<t_{N+1}=1$. Given $\sigma=(\rho,m)\in Z_{h,\tau}$, we can identify the density ρ with the collection $\{\rho_i\}_{i=0}^{N+1}$ with $\rho_i\in Q_h$, and the momentum m with the collection $\{m_i\}_{i=1}^{N+1}$ with $m_i\in V_h$. Since q is piecewise constant in time, we have the following equivalent formulation

$$\mathcal{A}_{h,\tau}(\sigma)=\sum_{i=1}^{N+1}A_h\left(\frac{\rho_i+\rho_{i-1}}{2},m_i\right)|t_i-t_{i-1}|\,. \tag{B.18}$$

Note that in order to obtain (B.18) from (B.17), we relied on the particular choice of finite element spaces for density (piecewise linear in time), momentum (piecewise constant in time) and the corresponding dual variables (piecewise constant in time).

Remark B.5 (Continuity constraint)**.** *The choice of a $H(\mathrm{div})$-conforming finite element space for σ implies that the weak form of the continuity equation $\partial_t\rho+\mathrm{div}_x\, m=0$ is satisfied exactly by any solution of the discrete saddle point problem (B.17) (this is also directly implied by the definition of the constraint set $\mathcal{C}_{h,\tau}$ in (B.15)) .*

Remark B.6 (Positivity constraint)**.** *Note that removing the positivity constraint in the formulation (B.17), we obtain a different scheme. In that case, since the action is evaluated on the mean density (in time), the positivity constraint $\rho\ge 0$ is then only enforced on $\frac{\rho_i+\rho_{i-1}}{2}$, rather than on each ρ_i separately.*

The objects introduced until now define a finite dimensional model of optimal transport in the sense of Definition 2.5 in [79]. The framework developed therein can be used to deduce a convergence result for our scheme.

Theorem B.7. *Let $\rho_0,\rho_1\in\mathcal{P}(D)$ be given and $\{\mathcal{T}_{h,\tau}\}_{h,\tau>0}$ a family of tensor-product decomposition of Ω such that the time discretization is uniform, i.e. $t_i-t_{i-1}=\tau$ for all $i=1,\ldots,N+1$, and the space discretization $\mathcal{T}_h$ satisfies equation (B.9). Let $\sigma_{h,\tau}$ be a minimizer of problem (B.17) associated with $\mathcal{T}_{h,\tau}$ and for $r=1$. Then, as $h,\tau\to 0$, up to extraction of a subsequence, $\sigma_{h,\tau}$ converges weakly to $\sigma\in\mathcal{M}(\Omega)^{d+1}$ a minimizer of problem (B.3).*

The proof is essentially an extension of the one presented in [79] and is postponed to the appendix.

Remark B.8 (Stability)**.** *The existence of bounded projections sastifying the commuting diagram* (B.11) *ensures stability of the projection onto* $\mathcal{C}_{h,\tau}$ *(see* (B.14)*). Such commuting projections are also crucial to estabilish the convergence result in Theorem B.7: in [79], they are used to sample the continuous solution into a discrete one satisfying the continuity equation, therefore providing an admissible candidate for the discrete problem. Nonetheless, due to the nonlinear constraint* $q \in \mathcal{K}$*, one cannot apply the standard linear theory in [24], for example, so the commuting diagram condition does not imply directly a stability result analogous to* (B.14) *for the saddle point problem* (B.17) *(even if we see it as a discretization of the Hilbert space formulation in Proposition B.1). Numerically (see Section B.8) the finite element pairs considered here* $(Z_{h,\tau}, X^r_{h,\tau})$ *appear to be stable when* $r = 1$*, but strong oscillations may occur for* $r = 0$ *and* $V_h = \mathcal{BDM}_1$*, providing empirical evidence of the instablity of the discretization for this case.*

Remark B.9. *Suppose that* $D = [0,1]^d$ *and that* $\mathcal{T}_{h,\tau}$ *is a uniform quadrilateral discretization of* $\Omega = [0,1]^{d+1}$*. Then for* $r = 0$ *and removing the constraint* $\rho \geq 0$ *(see Remark B.6), the discrete problem* (B.17) *coincides with the discretization proposed in [109]. Theorem B.7 shows that modifying this method with* $r = 1$ *and adding the positivity constraint at all times, one can prove convergence to the solution of the continuous problem* (B.3)*.*

B.6 The proximal splitting algorithm

We now describe in detail the discrete version of the proximal splitting algorithm introduced in Section B.3.1, in the simplest setting where we remove the additional positivity constraint on the density, i.e. we solve

$$\inf_{\sigma \in \mathcal{C}_{h,\tau}} \mathcal{A}_{h,\tau}(\sigma), \quad \mathcal{A}_{h,\tau}(\sigma) \coloneqq \sup_{q \in (X_{h,\tau})^{d+1} \cap \mathcal{K}} \langle q, \sigma \rangle \,.$$

As mentioned in Remark B.6, this amounts to enforcing positivity only on the mean density in time between consecutive time-steps. Using this formulation rather than (B.17) we can reproduce the structure of the continuous version of the scheme, described in Section B.3.1. Note, however, that one can actually solve problem (B.17) with a similar strategy, e.g., by first reformulating the problem intrudicing a Lagrange multiplier to enforce the continuity equation, and then applying the same proximal splitting algorithm considered here but with the new variables and with an appropriate choice of norms.

We start by defining

$$\mathcal{K}^r_{h,\tau} \coloneqq \mathcal{K} \cap (X^r_{h,\tau})^{d+1} \coloneqq \{ q \in (X^r_{h,\tau})^{d+1} \,;\, q \in K \, a.e. \} \,.$$

We write the discrete problem as follows:

$$\inf_{\sigma \in L^2(\Omega;\mathbb{R}^{d+1})} \sup_{q \in L^2(\Omega;\mathbb{R}^{d+1})} \langle q, \sigma \rangle + \iota_{\mathcal{C}_{h,\tau}}(\sigma) - \iota_{\mathcal{K}_{h,\tau}}(q) \,, \tag{B.19}$$

where $\mathcal{C}_{h,\tau}$ is defined in (B.15). Then, the proximal splitting algorithm of section B.3.1 applied to problem (B.19) can be formulated as follows: given $\tau_1, \tau_2 > 0$ and an admissible

$(\sigma^0, q^0) \in \mathcal{C}_{h,\tau} \times \mathcal{K}_{h,\tau}$, we define the sequence $\{(\sigma^k, q^k)\}_k$ by performing iteratively the following two steps:

$$\textbf{Step 1}: \qquad \sigma^{k+1} = P_{\mathcal{C}_{h,\tau}}(\sigma^k - \tau_1 q^k)\,. \tag{B.20a}$$

$$\textbf{Step 2}: \qquad q^{k+1} = P_{\mathcal{K}_{h,\tau}}(q^k + \tau_2(2\sigma^{k+1} - \sigma^k))\,. \tag{B.20b}$$

The convergence result in Theorem B.2 clearly holds also in the discrete setting and gives convergence of the algorithm to a discrete saddle point $(\sigma_{h,\tau}, q_{h,\tau})$, if the condition $\tau_1\tau_2 < 1$ is satisfied. The two steps in the algorithm can be computed as follows.

Step 1 As discussed in Section B.4.3, the projection $P_{\mathcal{C}_{h,\tau}}$ can be computed modifying the system given by (B.13) by adding the Neumann boundary conditions associated with the function (B.16).

Step 2 Since $P_{\mathcal{K}_{h,\tau}}$ is an L^2 projection, we have that $P_{\mathcal{K}_{h,\tau}} = P_{\mathcal{K}_{h,\tau}} \circ P_{(X^r_{h,\tau})^{d+1}}$, where $P_{(X^r_{h,\tau})^{d+1}}$ denotes the L^2 projection onto $(X^r_{h,\tau})^{d+1}$. This means that we only need to be able to compute $P_{\mathcal{K}_{h,\tau}}$ when applied to an element of $X^r_{h,\tau}$. In addition, since $X^r_{h,\tau}$ is discontinuous across elements, we can compute the projection element by element, and since functions in $X^r_{h,\tau}(S)$ are constant along the time direction, we can also eliminate the time variable in the projection. In other words, we only need to solve for each element $[t_0, t_1] \times S$ a problem in the form

$$\xi_{\mathcal{K}} \coloneqq \operatorname{argmin}\{\|\xi - q\|^2_{L^2(T)}\,;\, q \in (X^r_h(T))^{d+1}\,,\ q(x) \in K\ \forall x \in T\} \tag{B.21}$$

for a given $\xi \in (X^r_h(T))^{d+1}$. We distinguish two cases:

1. if $r = 0$, the projection (B.21) is just the projection of a vector $\xi \in \mathbb{R}^{d+1}$ onto the convex set K;
2. if $r = 1$, any $\xi \in (X^1_h)^{d+1}$ is fully determined by its value on the vertices $\{v_i\}_i$ of T, and the condition $\xi \in \mathcal{K}$, is equivalent to $\xi(v_i) \in K$, by convexity of the set K (see equation (B.5)). However the problem is coupled in these variables when computing the projection in the L^2 norm. Here, we use instead a different projection and we simply set
$$\xi_{\mathcal{K}}(v_i) = \operatorname{argmin}\{|\xi(v_i) - q|^2\,;\, q \in K\}\,.$$
Note that this is a variational crime, but it can be avoided by reformulating the algorithm using as inner product on X^1_h a weighted ℓ^2 inner product on the degrees of freedom.

In both cases we only need to compute for each degree of freedom the projection of a given vector $(\bar a, \bar b) \in \mathbb{R} \times \mathbb{R}^d$ onto K. If $(\bar a, \bar b) \notin K$, such a projection is given explicitly by the vector

$$\left(-\frac{\mu^2}{2}, \mu\,\frac{\bar b}{|\bar b|}\right)$$

where $\mu \geq 0$ is the largest real root of the third order polynomial

$$x \mapsto \frac{x^3}{2} + x(\bar a + 1) - |\bar b|\,.$$

Remark B.10. *As for the finite difference discretization studied in [109], different optimization techniques could be applied to solve problem (B.17). In particular, it should be noted that the ADMM approach orginally proposed by Benamou and Brenier [14] could also be applied. This would lead to a very similar algorithm to (B.20), but it would require the introduction of an additional variable which avoids coupling of the degrees of freedom in the optimization step with respect to q. In other words, this is needed in order to be able to perform the projection on K for each degree of freedom separately. More details on this issue can be found in [109] for the discretization studied therein, and they hold also in the finite element setting.*

B.7 Regularization

The optimal transport problem does not involve any regularizing effect on the interpolation between two measures. In fact, one can even expect a loss of regularity in some cases, namely if one is interpolating between two smooth densities on a smooth but non-convex domain. Such a loss of regularity (which is often unphysical when the density represents a physical quantity) can be avoided introducing additional regularization terms in the formulation. In this section we describe how to do so, and how these modifications translate at the algorithmic level.

We consider the Hilbert space setting discribed in Section B.3.1 and we study problems in the form

$$\inf_{\sigma\in\mathcal{C}} \mathcal{A}(\sigma) + \alpha\mathcal{R}(\sigma) \tag{B.22}$$

where $\mathcal{R} : L^2(\Omega) \to \mathbb{R}$ is a convex, proper and l.s.c. functional, and $\alpha > 0$. For this type of problem, we can still apply the proximal splitting algorithm (B.8) replacing the projection onto $\mathcal{C}$ by $\mathrm{prox}_{\tau_1\mathcal{F}}$, the proximal operator of $\mathcal{F} \coloneqq \iota_{\mathcal{C}} + \alpha\mathcal{R}$, defined by

$$\mathrm{prox}_{\tau_1\mathcal{F}}(\xi) = \operatorname*{argmin}_{\eta\in L^2(\Omega;\mathbb{R}^{d+1})} \frac{\|\xi-\eta\|^2}{2\tau_1} + \mathcal{F}(\eta)\,.$$

This leads to the so-called PDGH algorithm, which for $\tau_1\tau_2 < 1$ can be seen just as a proximal point method applied to a monotone operator [45], and therefore we still have convergence in the Hilbert space setting. As mentioned in [79] convergence of the discrete problem with mesh refinement is more delicate and will not be discussed here.

B.7.1 Mixed L^2-Wasserstein distance

Define for any $\sigma = (\rho, m) \in L^2(\Omega) \times L^2(\Omega;\mathbb{R}^d)$

$$\mathcal{R}(\sigma) \coloneqq \begin{cases} \frac{1}{2}\|\partial_t\rho\|^2_{L^2(\Omega)} & \text{if } \partial_t\rho \in L^2(\Omega)\,, \\ +\infty & \text{otherwise}\,. \end{cases}$$

With this functional, problem (B.22) yields an interpolation between the Wasserstein distance and the L^2 distance. It was originally considered in [15], where a conjugate gradient method was proposed to compute the minimizers. Let $V \coloneqq H^1([0,1];L^2(D)) \times L^2([0,1];H(\mathrm{div};D))$ and let $\mathring{V}$ be the same space with homogenous boundary conditions on the fluxes. For any $\xi \in$

$L^2(\Omega)^{d+1}$, $\sigma = \mathrm{prox}_{\tau_1 \mathcal{F}}(\xi)$ is obtained by solving the following system for $(\sigma, \phi) \in V \times L^2(\Omega)/\mathbb{R}$

$$\begin{cases} \langle \sigma, v \rangle + \alpha\tau_1 \langle \partial_t \rho, \partial_t v_t \rangle + \langle \phi, \operatorname{div} v \rangle = \langle \xi, v \rangle \,, & \forall v \in \mathring{V} \,, \\ \langle \operatorname{div} \sigma, \psi \rangle = 0 \,, & \forall \psi \in L^2(\Omega)/\mathbb{R} \,, \\ \sigma \cdot n_{\partial\Omega} = \mathcal{X} \,, \end{cases}$$

where $v_t = v \cdot \hat{t}$ is the component of v in the time direction. Well-posedness can be obtained by standard methods for saddle point problems [24] and it translates directly into well-posedness of the discrete system obtained by replacing V with $Z_{h,\tau}$, $L^2(\Omega)$ with $F_{h,\tau}$, and $\mathcal{X}$ with $\mathcal{X}_{h,\tau}$.

B.7.2 H^1 regularization

Define for any $\sigma = (\rho, m) \in L^2(\Omega) \times L^2(\Omega; \mathbb{R}^d)$

$$\mathcal{R}(\sigma) \coloneqq \begin{cases} \frac{1}{2} \|\nabla_x \rho\|^2_{L^2(\Omega)} & \text{if } \rho \in L^2([0,1]; H^1(D)) \,, \\ +\infty & \text{otherwise} \,. \end{cases} \tag{B.23}$$

In this case we set $V \coloneqq H(\operatorname{div}; \Omega)$, $W \coloneqq L^2([0,1]; H(\operatorname{div}_x; D))$ and let $\mathring{V}$ and $\mathring{W}$ be the same spaces with homogenous boundary conditions on the fluxes. Then, for any $\xi \in L^2(\Omega)^{d+1}$, $\sigma = \mathrm{prox}_{\tau_1 \mathcal{F}}(\xi)$ is obtained by solving the following system for $(\sigma, \eta, \phi) \in V \times \mathring{W} \times L^2(\Omega)/\mathbb{R}$

$$\begin{cases} \langle \sigma, v \rangle - \alpha\tau_1 \langle \operatorname{div}_x \eta, v_t \rangle + \langle \phi, \operatorname{div} v \rangle = \langle \xi, v \rangle \,, & \forall v \in \mathring{V} \,, \\ \langle \rho, \operatorname{div}_x w \rangle + \langle \eta, w \rangle = 0 \,, & \forall w \in \mathring{W} \,, \\ \langle \operatorname{div} \sigma, \psi \rangle = 0 \,, & \forall \psi \in L^2(\Omega)/\mathbb{R} \,, \\ \sigma \cdot n_{\partial\Omega} = \mathcal{X} \,, \end{cases}$$

where $v_t = v \cdot \hat{t}$ is the component of v in the time direction. As before, well-posedness can be obtained by standard methods for saddle point problems [24].

We introduce the space $W_{h,\tau} \subset L^2([0,1]; H(\operatorname{div}_x; D))$ whose shape functions on $S = [t_0, t_1] \otimes T$ are given by

$$W_{h,\tau}(S) \coloneqq \mathcal{P}_1([t_0, t_1]) \otimes V_h(T).$$

We denote by $\mathring{W}_{h,\tau}$ the same space with the boundary conditions $\eta \cdot n_{\partial\Omega} = 0$ on $[0,1] \times \partial D$. Denote by $\nabla_x^h : L^2(\Omega) \to \mathring{W}_{h,\tau}$ the adjoint of $-\operatorname{div}_x$ defined by

$$\langle \nabla_x^h \phi, \eta \rangle = -\langle \phi, \operatorname{div}_x \eta \rangle \,, \quad \forall\, (\phi, \eta) \in L^2(\Omega) \times \mathring{W}_{h,\tau} \,.$$

We define a discrete version of (B.23) as follows:

$$\mathcal{R}_{h,\tau}(\sigma) \coloneqq \frac{1}{2} \|\nabla_x^h \rho\|^2_{L^2(\Omega)}.$$

Let $\mathcal{F}_{h,\tau} \coloneqq \iota_{\mathcal{C}_{h,\tau}} + \alpha \mathcal{R}_{h,\tau}$. Then for any $\xi \in L^2(\Omega)^{d+1}$, $\sigma = \mathrm{prox}_{\tau_1 \mathcal{F}_{h,\tau}}(\xi)$ is obtained by solving the following system for $(\sigma, \eta, \phi) \in \mathring{V}_{h,\tau} \times \mathring{W}_{h,\tau} \times F_{h,\tau}/\mathbb{R}$:

$$\begin{cases} \langle \sigma, v \rangle - \alpha\tau_1 \langle \operatorname{div}_x \eta, v_t \rangle + \langle \phi, \operatorname{div} v \rangle = \langle \xi, v \rangle \,, & \forall v \in \mathring{V}_{h,\tau} \,, \\ \langle \rho, \operatorname{div}_x w \rangle + \langle \eta, w \rangle = 0 \,, & \forall w \in \mathring{W}_{h,\tau} \,, \\ \langle \operatorname{div} \sigma, \psi \rangle = 0 \,, & \forall \psi \in F_{h,\tau}/\mathbb{R} \,, \\ \sigma \cdot n_{\partial\Omega} = \mathcal{X}_{h,\tau} \,. \end{cases}$$

B.8 Numerical results

In this section we describe two numerical tests that demonstrate the behaviour of the proposed discretization both qualitatively and in terms of convergence of the algorithm. For both tests the time discretization is uniform, but we will use different meshes and finite element spaces for the discretization in space. For all tests, we set $\tau_1 = \tau_2 = 1$ as parameters of the proximal splitting algorithm (B.20). The results shown hereafter have been obtained using the finite element software Firedrake [114] (see [97, 21], for the tensor product constructions) and the linear solver for the mixed Poisson equation is based on PETSc [9, 10]. The code to perform the tests in this section can be found in the repository `https://github.com/andnatale/dynamic-ot.git`.

B.8.1 Qualitative behaviour and convergence of the proximal-splitting algorithm

We set $D = [0,1]^2$, and consider either a structured triangular mesh, an unstructured one, or a uniform Cartesian mesh (shown in figures B.2, B.3 and B.4), and $\tau := |t_{i+1} - t_i| = 1/20$. The initial and final densities are given by

$$\rho_0(x) \propto \frac{3}{2} + \cos(2\pi|x - x_0|), \quad \rho_1(x) \propto \frac{3}{2} - \cos(2\pi|x - x_0|)\,, \tag{B.24}$$

where $x_0 = (0.5, 0.5)$, and they are normalized so that the total mass is equal to one. In figures B.2, B.3 and B.4, the interpolation at time $t = 0.5$ is shown for different choices of spaces V_h and X_h and different meshes. The discretization corresponding to the couple $V_h = \mathcal{BDM}_1$ and X_h^0 appears to yield the wrong solution both on the structured and unstructured mesh, which is also very oscillatory. Oscillations appear also for $V_h = \mathcal{RT}_0$, although the correct solution is recovered. For this latter case, the oscillations seem to be very sensitive to the structure of the mesh and are attenuated when choosing X_h^1 instead of X_h^0. On the Cartesian mesh the scheme does not generate any oscillations, with the choice of the space X_h^1 leading to slightly more diffusive results. Note that the appearance of oscillations is not related to the positivity constraint since in the case considered here the interpolation is strictly positive. On the other hand, we remark that for tests leading to pure translation of compactly supported densities (not shown) the oscillations disappear almost entirely even for the couple $V_h = \mathcal{BDM}_1$, X_h^0.

In figure B.5, the different schemes are compared in terms of convergence of the proximal splitting algorithm. The cases corresponding to X_h^1 appear to converge significantly faster than those corresponding to X_h^0. Note that in the case $V_h = \mathcal{RT}_{[0]}$, X_h^0, the resulting discretization as well as the optimization algorithm coincide with the ones proposed in [109], since here we consider a uniform Cartesian grid. Also in this case, replacing X_h^0 with X_h^1 (besides providing a convergence guarantee, see Remark B.9) yields a considerable speedup of the algorithm.

B.8.2 Non-convex domain

We now consider a non-convex polygonal domain D, with the spatial mesh $\mathcal{T}_h$ represented in figure B.6 and $\tau := |t_{i+1} - t_i| = 1/30$. Note that even if the case of a non-convex domain is beyond the domain of applicability of the convergence results presented in this chapter, our

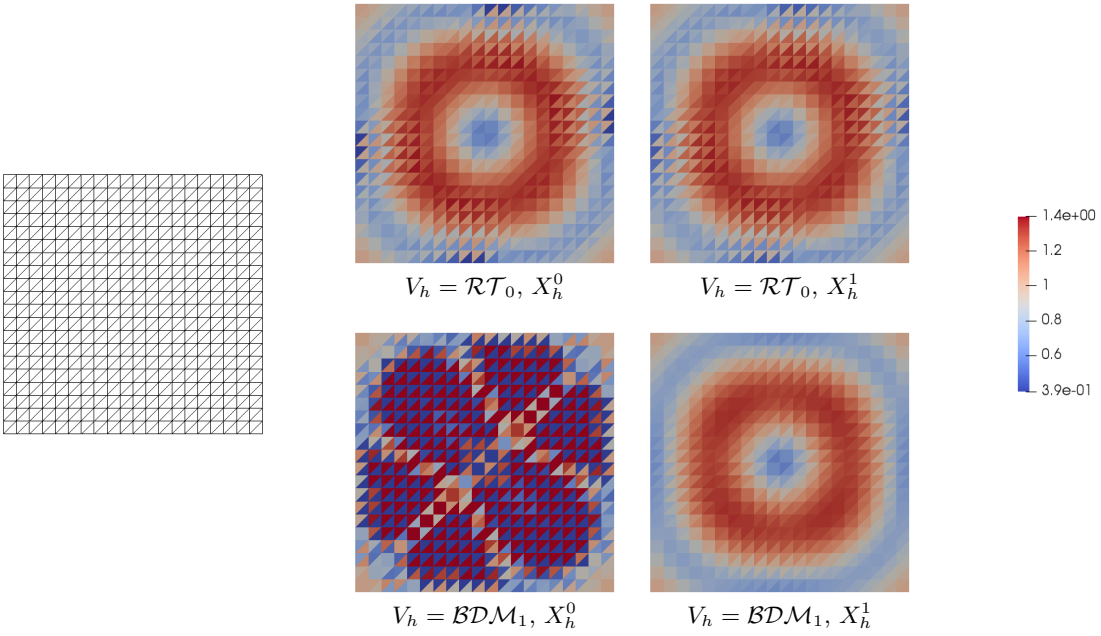


Figure B.2: Comparison between optimal transport interpolations of the densities in (B.24) for different spaces on a structured triangular mesh. Note that in the case $V_h = \mathcal{BDM}_1$, X_h^0, the data exceeds the color map range.

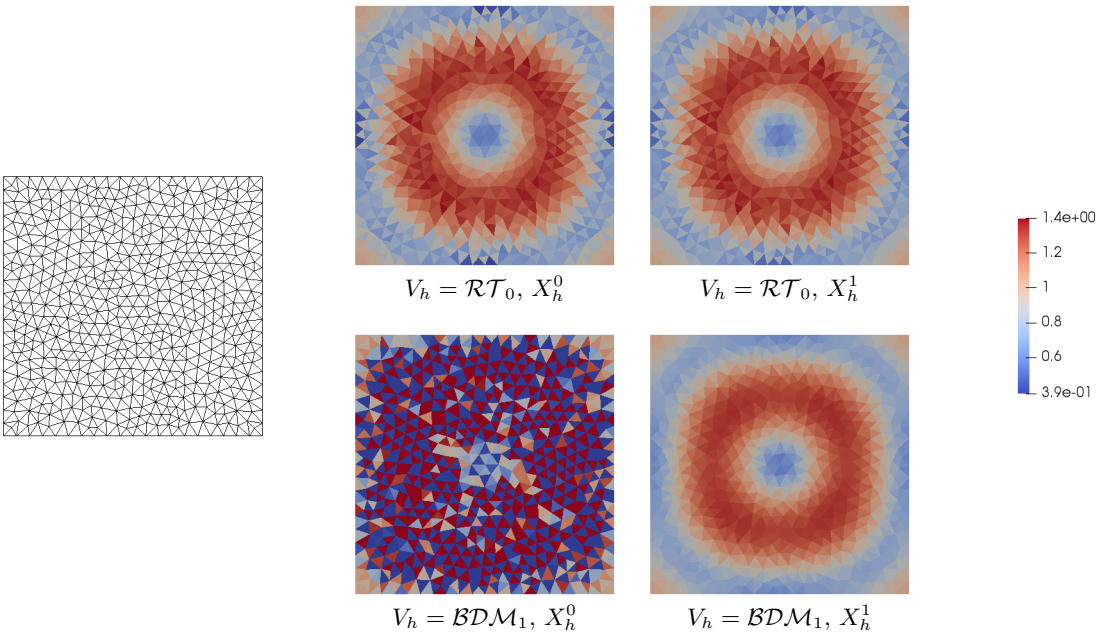
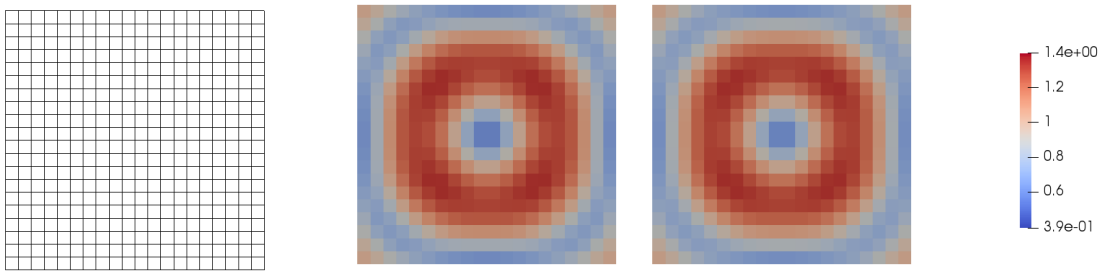


Figure B.3: Comparison between optimal transport interpolations of the densities in (B.24) for different spaces on an unstructured triangular mesh. Note that in the case $V_h = \mathcal{BDM}_1$, X_h^0, the data exceeds the color map range.



$V_h = \mathcal{RT}_{[0]},\ X_h^0$ $V_h = \mathcal{RT}_{[0]},\ X_h^1$

Figure B.4: Comparison between optimal transport interpolations of the densities in (B.24) for different spaces on an uniform Cartesian mesh.

scheme is still well-defined for this case. The boundary conditions are given by

$$\rho_0(x) = \exp\left(-\frac{|x-x_0|^2}{2s^2}\right), \quad \rho_1(x) = \exp\left(-\frac{|x-x_1|^2}{2s^2}\right),$$

with $s = 0.1$, $x_0 = (0.5, 0.1)$ and $x_1 = (0.5, 0.9)$. Such boundary conditions are illustrated in figure B.6. In this case, the exact density interpolation is not absolutely continuous, since mass concentrates on the segment connecting the two non-convex corners of the domain. Note that we have therefore refined the mesh along the diagonal where we expect the mass to concentrate.

In figure B.7, B.9 and B.8 we show the density evolution up to time $t = 0.5$ (the other half of the time evolution being symmetric in space given the boundary conditions and the domain shape) for the non-regularized case, the H^1 regularization and the L^2 regularization, respectively. For both regularizations the density profile appears to be smoothened, but only the H^1 regularization avoids concentration at the corners.

The proximal operator of the projection on the continuity equation is more expensive computationally for the H^1 regularization than for the other two cases, since we have to solve a larger mixed system at each iteration. However, for both regularizations, the proximal splitting algorithm itself converges much faster than the non-regularized case, as it can be seen in figure B.10.

B.9 Proof of theorem B.7

Applying Theorem 2.16 in [79], in order to prove Theorem B.7 it is sufficient to check that the conditions listed in definition 2.9 of [79] are verified. Such conditions translated to our finite element settings are listed in Proposition B.12 below.

From now on, we assume $r = 1$, and X_h stands for X_h^1. We also denote by $\mathcal{I}$ is the standard nodal interpolant onto X_h, defined element by element.

First of all, we introduce some notation and list some technical results [49]. Denote by P_{X_h} and P_{V_h} the L^2 projections onto X_h and V_h, respectively. Then,

$$\|P_{X_h}\varphi\|_{L^p} \leq C\|\varphi\|_{L^p}, \quad \forall \varphi \in L^p, 1 \leq p \leq \infty,$$

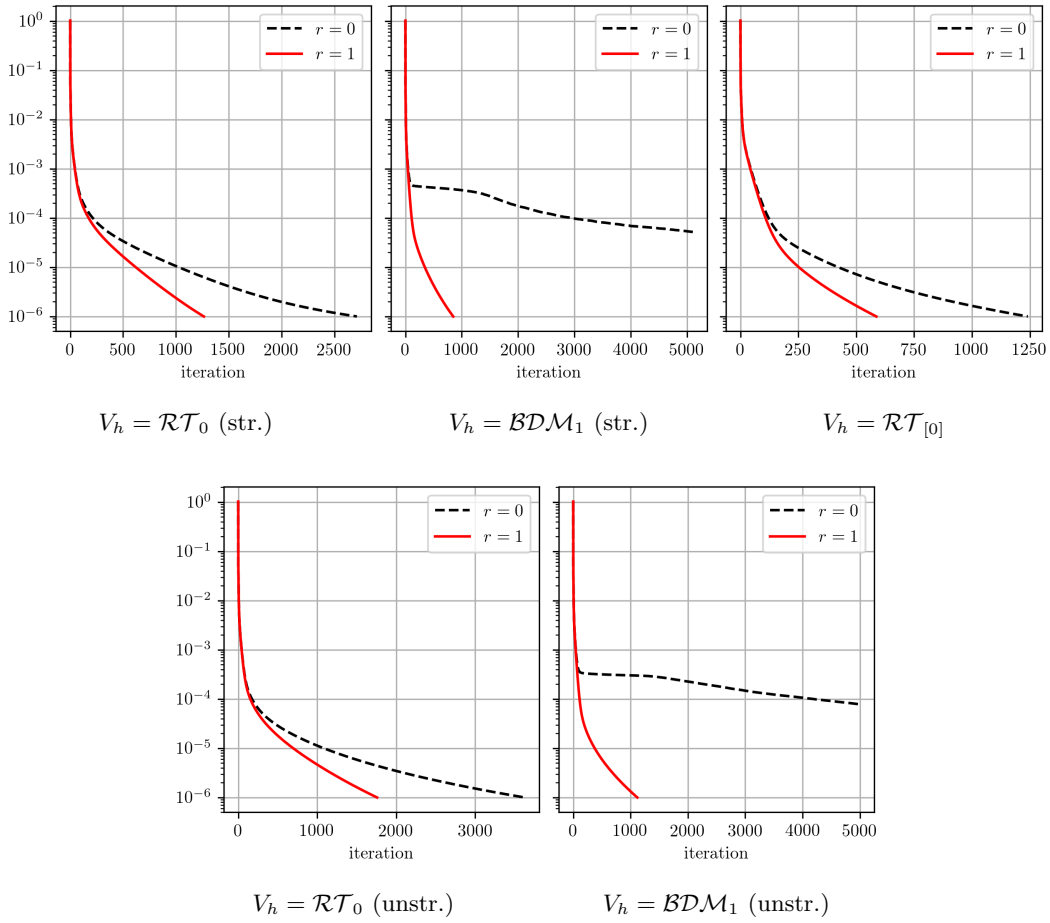


Figure B.5: Convergence of the proximal splitting algorithm measured by $\|\sigma_{n+1} - \sigma_n\|_{L^2(\Omega)}$ for different spaces V_h and X_h^r on the structured (str.) and unstructured (unstr.) triangular mesh, and on the Cartesian mesh.

and moreover $\forall\, T \in \mathcal{T}_h$

$$\|\varphi - P_{X_h}\varphi\|_{L^p(T)} \leq C h_T \|\nabla\varphi\|_{L^p(T)}\,, \quad \forall \varphi \in W^{1,p}(T)\,,\, 1 \leq p \leq \infty\,,$$

where, with an abuse of notation, we have used P_{X_h} to denote the L^2 projection onto $X_h(T)$. These imply the following lemma.

Lemma B.11. *Given the regularity assumption in* (B.9) *on* $\mathcal{T}_h$*, we have*

$$\|\mathcal{I}|P_{V_h} b|^2\|_{L^\infty} \leq C\|b\|^2_{L^\infty}\,,$$

for any $b \in L^\infty(D)$*, and*

$$\|\mathcal{I}|P_{V_h} b|^2 - |b|^2\|_{L^\infty} \leq Ch|b|_{W^{1,\infty}}\|b\|_{L^\infty}\,,$$

for any $b \in W^{1,\infty}(D)$*.*

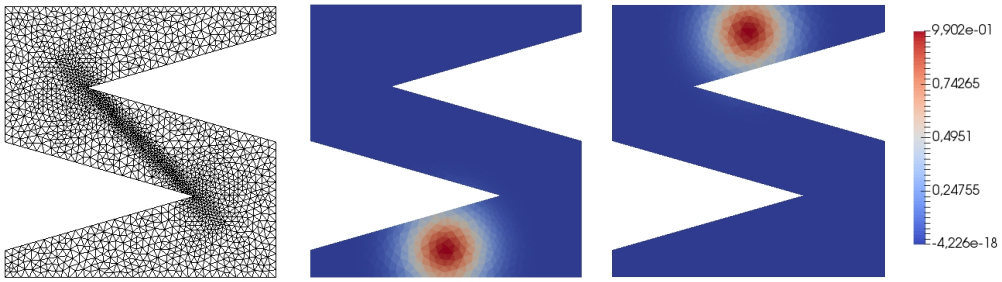


Figure B.6: Mesh, and initial and final density for the non-convex domain test.

Proof. For the first inequality, using standard inverse inequalities, we have

$$\begin{aligned}
\|\mathcal{I}|P_{V_h}b|^2\|_{L^\infty} &\leq \||P_{V_h}b|^2\|_{L^\infty} \\
&\leq Ch^{-d}\||P_{V_h}b|^2\|_{L^1} \\
&= Ch^{-d}\|P_{V_h}b\|_{L^2}^2 \\
&\leq Ch^{-d}\|P_{(X_h)^d}b\|_{L^2}^2 \\
&\leq C\|P_{(X_h)^d}b\|_{L^\infty}^2 \\
&\leq C\|b\|_{L^\infty}^2 .
\end{aligned}$$

For the second inequality , we observe that

$$\|\mathcal{I}|P_{V_h}b|^2 - |b|^2\|_{L^\infty} \leq \|\mathcal{I}|P_{V_h}b|^2 - \mathcal{I}|b|^2\|_{L^\infty} + \|\mathcal{I}|b|^2 - |b|^2\|_{L^\infty} .$$

The second term of the right-hand side is easy to control. For the first term, we have

$$\begin{aligned}
\|\mathcal{I}|P_{V_h}b|^2 - \mathcal{I}|b|^2\|_{L^\infty} &\leq \||P_{V_h}b|^2 - |b|^2\|_{L^\infty} \\
&\leq \||P_{V_h}b|^2 - |P_{(X_h)^d}b|^2\|_{L^\infty} + \||b|^2 - |P_{(X_h)^d}b|^2\|_{L^\infty} .
\end{aligned}$$

Again, the second term is easy to control. For the first tem, using the same reasoning as above,

$$\begin{aligned}
\||P_{V_h}b|^2 - |P_{(X_h)^d}b|^2\|_{L^\infty} &\leq Ch^{-d}\||P_{V_h}b|^2 - |P_{(X_h)^d}b|^2\|_{L^1} \\
&\leq Ch^{-d}\sum_{i=1}^{d}\|(P_{V_h}b)_i^2 - (P_{X_h}b_i)^2\|_{L^1} \\
&\leq Ch^{-d}\sum_{i=1}^{d}\|(P_{V_h}b)_i - P_{X_h}b_i\|_{L^1}\|b\|_{L^\infty} \\
&\leq Ch^{-\frac{d}{2}}\|P_{V_h}b - P_{X_h}b\|_{L^2}\|b\|_{L^\infty} \\
&\leq Ch\|\nabla P_{X_h}b\|_{L^\infty}\|b\|_{L^\infty} \leq Ch\|\nabla b\|_{L^\infty}\|b\|_{L^\infty} .
\end{aligned}$$

□

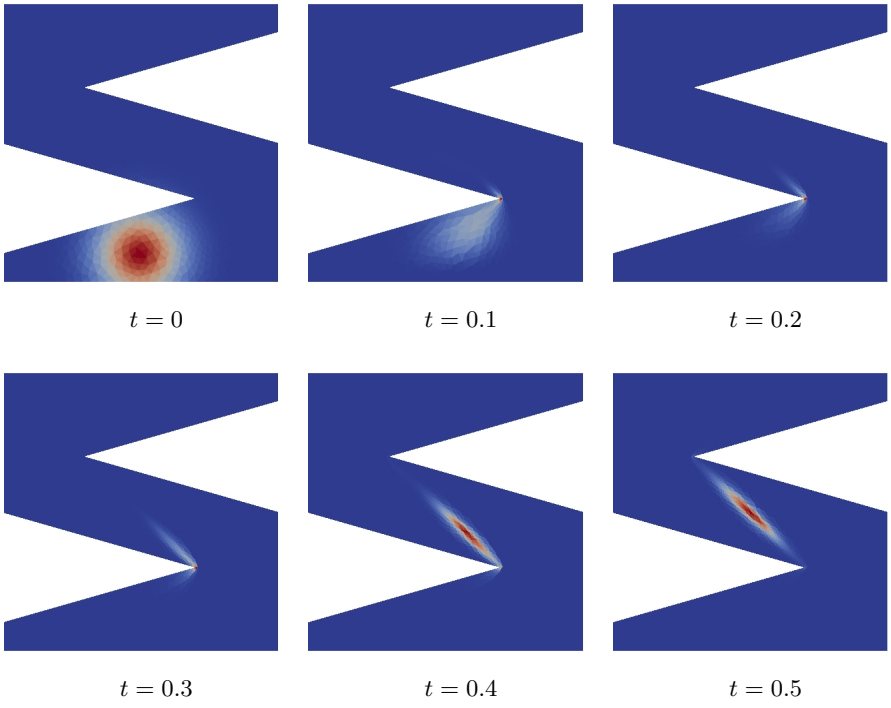


Figure B.7: Density evolution on the non-convex domain without regularization, $V_h = \mathcal{RT}_0$, X_h^1 (color scale is rescaled to fit data range).

As mentioned in Section B.4.1, there exist projection operators $\Pi_{Q_h} : L^2(D) \to V_h$ and $\Pi_{V_h} : \mathcal{V}_D \to Q_h$ commuting with the divergence operator, where $\mathcal{V}_D$ is a dense subset of $H(\mathrm{div}; D)$. We pick these to be the canonical projections introduced in Section 5.2 of [8], and in particular Π_{Q_h} as in equation (B.10). Such operators verify the following approximation properties (see Theorem 5.3 in [8]): for any $\varphi \in H^1(D)$ and $\eta \in H^1(D)^d$

$$\|\Pi_{Q_h}\varphi - \varphi\|_{L^2(D)} \leq Ch\|\varphi\|_{H^1(D)}\,, \quad \|\Pi_{V_h}\eta - \eta\|_{L^2(D)^d} \leq Ch\|\eta\|_{H^1(D)^d}\,. \tag{B.25}$$

Notice in particular that given the mesh regularity assumption (B.9), equation (B.25) is a standard property for Π_{Q_h} as defined in equation (B.10).

Proposition B.12 below contains the properties needed for convergence: it can be seen as a specific instance of Definition 2.9 of [79]. Note that a few of the properties listed therein are omitted here because they are either unnecessary or true by construction in our setting. Note also that the sampling operators used in [79] are replaced here with the canonical projections Π_{Q_h} and Π_{V_h}, where Π_{Q_h} can be naturally extended to $\mathcal{M}(D)$ (see equation (B.10)) and Π_{V_h} is considered to be defined on a dense subset of $\mathcal{M}(D)^d$. Moreover the reconstruction operators are simply the injection operators from Q_h and V_h to $\mathcal{M}(D)$ and $\mathcal{M}(D)^d$, respectively. Finally, we define for any $(\rho, b) \in \mathcal{M}(D) \times C(D; \mathbb{R}^d)$

$$A^*(\rho, b) \coloneqq \int_D \frac{|b|^2}{2}\rho\,,$$



$t = 0$ $t = 0.1$ $t = 0.2$



$t = 0.3$ $t = 0.4$ $t = 0.5$

Figure B.8: Density evolution on the non-convex domain with L^2 regularization, $\alpha = 0.002$, $V_h = \mathcal{RT}_0$, X_h^1 (color scale is rescaled to fit data range).

so that if $(\rho, m) \in \mathcal{M}_+(D) \times \mathcal{M}(D)^d$ then

$$A(\rho, m) = \sup_{b \in C(D;\mathbb{R}^d)} \langle m, b \rangle - A^*(\rho, b);$$

and for any $(\rho, b) \in Q_h \times V_h$,

$$A_h^*(\rho, b) \coloneqq \sup_{m \in V_h} \langle m, b \rangle - A_h(\rho, m)\,.$$

Proposition B.12. *The following properties hold:*

1. *For any $\rho \in \mathcal{M}_+(D)$, $\Pi_{Q_h}\rho \to \rho$ as $h \to 0$ weakly in $\mathcal{M}(D)$.*

2. *Let $B \subset (C^1(D))^d$ a bounded subset. Then there exists a constant ε_h tending to 0 as $h \to 0$ such that for any $b \in B$ and $\rho \in Q_h$*

$$A_h^*(\rho, P_{V_h} b) \leq A^*(\rho, b) + \epsilon_h \|\rho\|\,,$$

where P_{V_h} denotes the L^2 projection onto V_h. Moreover there exists a constant $C \geq 1$ such that for any $b \in C(D)^d$, there holds

$$A_h^*(\rho, P_{V_h} b) \leq \frac{C}{2} \|\rho\| \|b\|_{L^\infty}^2\,.$$

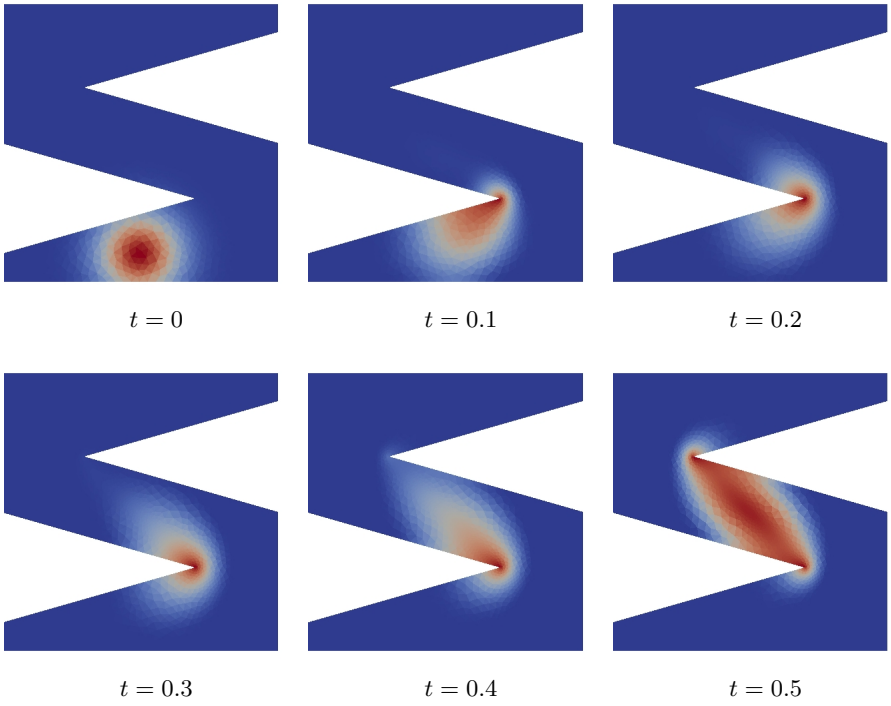


Figure B.9: Density evolution on the non-convex domain with H^1 regularization, $\alpha = 0.002$, $V_h = \mathcal{RT}_0$, X_h^1 (color scale is rescaled to fit data range).

3. *Let $B \subset C^0(D) \cap H^1(D)$ a bounded subset such that for all $\rho \in B$ there holds $\rho > C > 0$, and let $B' \subset (C^0(D) \cap H^1(D))^d$ a bounded subset. There exists a constant ε_h tending to 0 as $h \to 0$ such that, given $(\rho, m) \in \mathcal{M}(D)^{d+1}$ such that ρ has density in B and m in B', then*

$$A_h(\Pi_{Q_h}\rho, \Pi_{V_h}m) \leq A(\rho, m) + \varepsilon_h.$$

4. *There exists ε_h tending to 0 as $h \to 0$ and a continuous function ω satisfying $\omega(0) = 0$ such that: for any $x, y \in D$ and $h > 0$ there exists $\rho \in Q_h^+$ and $m_1, m_2 \in V_h$ satisfying*

$$\begin{cases} \operatorname{div} m_1 = \rho - \Pi_{Q_h}(\delta_x) \\ \operatorname{div} m_2 = \rho - \Pi_{Q_h}(\delta_y) \end{cases} \quad \text{and} \quad A_h(\rho, m_i) \leq \omega(|x - y|) + \varepsilon_h \,, \forall i \in \{1, 2\}. \tag{B.26}$$

Remark B.13. *In [79] point (3) of Proposition B.12 is stated with B and B' bounded subsets of $C^1(D)$ and $C^1(D)^d$, respectively. The condition we require here is stronger, but it is needed since we considered a convex polytope domain rather than a domain with a smooth boundary as in [79]. As a matter of fact, in [79] one applies the condition (3) on a regularized measure $(\tilde{\rho}, \tilde{m}) \in \mathcal{M}(D)^{d+1}$ obtained by convolution with the heat kernel and by solving an appropriate elliptic problem (see proposition 3.2 in [79]). For a convex polytope domain this procedure yields a couple $(\tilde{\rho}, \tilde{m})$ with densities which are not C^∞ given the singularities of the boundary. By classical elliptic regularity estimates on non-smooth domains (e.g., [118] and [86]), the regularity we require in condition (3) is however sufficient for the proof in [79] to apply without changes.*

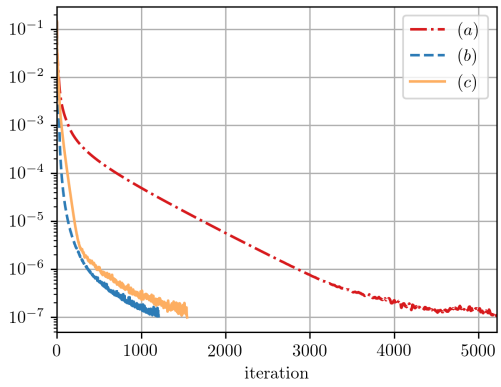


Figure B.10: Convergence of the proximal splitting algorithm measured by $\|\sigma_{n+1}-\sigma_n\|_{L^2(\Omega)}$ for non-convex domain test without regularization (a); with the H^1 regularization and $\alpha = 0.002$ (b); with the L^2 regularization and $\alpha = 0.002$ (c).

Proof. The first point is immediate from the definition of Π_{Q_h} in equation (B.10). For (2), we observe that

$$A_h(\rho, m) = \sup_{b \in X_h} \langle m, b\rangle - \frac{1}{2}\langle \rho, \mathcal{I}|b|^2\rangle\,,$$

where we recall that $\mathcal{I}$ is the standard element-wise nodal interpolant onto X_h. In fact, for any $b \in (X_h)^d$, we have $b^2 \le \mathcal{I}|b|^2$, and therefore when $\rho \ge 0$ we can "saturate" the constraint setting $a = -\mathcal{I}|b|^2/2$. On the other hand if $\rho < 0$ on some element both sides of the equality are $+\infty$. For $(\rho, b, m) \in Q_h \times V_h \times V_h$ define

$$A^*_{\mathcal{I},h}(\rho, b) \coloneqq \frac{1}{2}\langle \rho, \mathcal{I}|b|^2\rangle, \quad \bar{A}_{\mathcal{I},h}(\rho, m) \coloneqq \sup_{b \in V_h} \langle m, b\rangle - A^*_{\mathcal{I},h}(\rho, b).$$

Then, since when $\rho < 0$ on some element $A^*_h(\rho, b) = -\infty$,

$$A_h(\rho, m) \ge \bar{A}_{\mathcal{I},h}(\rho, m), \quad A^*_h(\rho, b) \le \bar{A}^*_{\mathcal{I},h}(\rho, b) \le A^*_{\mathcal{I},h}(\rho, b),$$

and we can prove (2) for $A^*_{\mathcal{I},h}$. In particular, we have

$$A^*_{\mathcal{I},h}(\rho, P_{V_h} b) \le A^*(\rho, b) + \frac{1}{2}\|\mathcal{I}|P_{V_h} b|^2 - |b|^2\|_{L^\infty}\|\rho\|,$$

and we obtain the result applying Lemma B.11. Using again Lemma B.11, we easily obtain the second bound as well.

For point (3), observe first that $A_h(\Pi_{Q_h}\rho, \Pi_{V_h} m) \le A(\Pi_{Q_h}\rho, \Pi_{V_h} m)$ by definition. Then given the assumption on ρ and m we can simply write

$$\begin{aligned} A_h(\Pi_{Q_h}\rho, \Pi_{V_h} m) - A(\rho, m) &\le \int_D \frac{|\Pi_{V_h} m|^2}{2\Pi_{Q_h}\rho} - \frac{|m|^2}{2\rho} \\ &\le \frac{1}{2}\int_D \Big|\frac{|\Pi_{V_h} m|^2 - |m|^2}{\Pi_{Q_h}\rho}\Big| + \Big|\frac{|m|^2}{\Pi_{Q_h}\rho} - \frac{|m|^2}{\rho}\Big| \\ &\le C(\|\Pi_{Q_h}\rho - \rho\|_{L^2} + \||\Pi_{V_h} m|^2 - |m|^2\|_{L^1})\,, \end{aligned}$$

where the constant C depends on the uniform lower bound on ρ and on the L^∞ norm of $|m|$. We conclude using Cauchy–Schwarz inequality on the second term and then equation (B.25).

For the last point, we will establish a connection between our scheme and the one proposed by Gladbach, Kopfer and Maas [64] and then use propoperty (B.26) for this scheme which was proved in [79]. We will consider only the case of a simplicial mesh and $V_h = \mathcal{RT}_0$ (which covers also the case of $V_h = \mathcal{BDM}_1$, since $\mathcal{RT}_0 \subset \mathcal{BDM}_1$). The quadrilateral case with $V_h = \mathcal{RT}_{[0]}$ can be dealt with in a completely analogous way.

First, we introduce some notation. For each $T \in \mathcal{T}_h$, let $\mathcal{T}_{h,T}$ be the set of neighbouring elements $L \in \mathcal{T}_h$ such that $f_{T,L} \coloneqq \overline{T} \cap \overline{L} \neq \emptyset$, which we assume to be oriented. Define by $\mathcal{F}_h$ the set of $(d-1)$-dimensional facets in the triangulation. Let $T, L \in \mathcal{T}_h$ be neighbouring elements, we denote by $\varphi_{T,L} \in \mathcal{RT}_0$ the canonical basis function associated with the oriented facet $f_{T,L}$. Then, any $m \in \mathcal{RT}_0$ can be written as

$$m = \sum_{f_{T,L} \in \mathcal{F}_h} m_{T,L} \varphi_{T,L} \,,$$

where $m_{T,L}$ is the flux of m on the oriented facet $f_{T,L}$. In other words we can identify functions in $(\rho, m) \in Q_h \times \mathcal{RT}_0$ with their finite volume representation $\{\rho_T, m_{T,L}\}_{T,L}$. Then, we can interpret the action for the finite volume scheme [64], which we denote by $A_h^{FV}(\rho, m)$, as a function on $Q_h \times \mathcal{RT}_0$. This is given by the following expression

$$A_h^{FV}(\rho, m) \coloneqq \sum_{f_{T,L} \in \mathcal{F}_h} \frac{m_{T,L}^2}{2\theta(\rho_T, \rho_L)} |f_{T,L}||x_T - x_L| \,,$$

where $\theta : \mathbb{R}^+ \times \mathbb{R}^+ \to \mathbb{R}^+$ is an appropriate function (see [64]) which we take to be the harmonic mean.

Now, in order to construct $\rho \in Q_h^+$ and $m_1, m_2 \in \mathcal{RT}_0$ satisfying (B.26), we use the same construction as in [79] for the finite volume scheme, and interpolate these to the spaces $\mathcal{RT}_0$ and Q_h^+ to obtain ρ, m_1 and m_2 satisfying

$$\begin{cases} \operatorname{div} m_1 = \rho - \Pi_{Q_h}(\delta_x) \,, \\ \operatorname{div} m_2 = \rho - \Pi_{Q_h}(\delta_y) \,. \end{cases}$$

In particular the support of ρ, m_1 and m_2 is a chain of neighbouring elements $T_1, \dots, T_N$. To prove the bound on the action, we observe that $A_h(\rho, m_i) \leq A(\rho, m_i)$. Then, we only need to bound $A(\rho, m_i)$ by the action of the finite-volume scheme $A_h^{FV}(\rho, m_i)$, since A_h^{FV} satisfies the desired inequality thanks to the regularity assumption (B.9) on the mesh [79]. By the regularity assumption on the triangulation, we can assume

$$\int_{T \cup L} |\varphi_{T,L}|^2 \, \mathrm{d}x \leq C |f_{T,L}||x_T - x_L|$$

uniformly. Then, by explicit calculations we obtain $A(\rho, m_i) \leq C A_h^{FV}(\rho, m_i)$ and we are done. □

Bibliography

[1] Yves Achdou, Fabio Camilli, and Italo Capuzzo-Dolcetta. Mean field games: Numerical methods for the planning problem. *SIAM J. Control Optim.*, 50(1):77–109, 2012.

[2] Ahmed Ait Hammou Oulhaj. Numerical analysis of a finite volume scheme for a seawater intrusion model with cross-diffusion in an unconfined aquifer. *Numer. Methods Partial Differential Equations*, 34(3):857–880, 2018.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[4] Luigi Ambrosio, Edoardo Mainini, and Sylvia Serfaty. Gradient flow of the Chapman-Rubinstein-Schatzman model for signed vortices. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 28(2):217–246, 2011.

[5] Luigi Ambrosio and Sylvia Serfaty. A gradient flow approach to an evolution problem arising in superconductivity. *Comm. Pure Appl. Math.*, 61(11):1495–1539, 2008.

[6] Boris Andreianov, Clément Cancès, and Ayman Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic–elliptic PDEs. *J. Funct. Anal.*, 273(12):3633–3670, 2017.

[7] Douglas N Arnold, Daniele Boffi, and Francesca Bonizzoni. Finite element differential forms on curvilinear cubic meshes and their approximation properties. *Numer. Math.*, 2014. arXiv:1204.2595.

[8] Douglas N Arnold, Richard S Falk, and Ragnar Winther. Finite element exterior calculus, homological techniques, and applications. *Acta numerica*, 15:1–155, 2006.

[9] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William D. Gropp, Dmitry Karpeyev, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, Hong Zhang, and Hong Zhang. PETSc users manual. Technical Report ANL-95/11 - Revision 3.11, Argonne National Laboratory, 2019.

[10] Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith. Efficient management of parallelism in object oriented numerical software libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.

[11] Martino Bardi and Lawrence C. Evans. On hopf's formulas for solutions of hamilton-jacobi equations. *Nonlinear Analysis: Theory, Methods & Applications*, 8(11):1373–1381, 1984.

[12] Bachir Ben Moussa and Georgios T. Kossioris. On the system of hamilton–jacobi and transport equations arising in geometrical optics. *Communications in Partial Differential Equations*, 28:1085 – 1111, 2003.

[13] Jean-David Benamou. Optimal transportation, modelling and numerical simulation. *Acta Numerica*, 30:249–325, 2021.

[14] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[15] Jean-David Benamou and Yann Brenier. Mixed l 2-wasserstein optimal mapping between prescribed density functions. *Journal of Optimization Theory and Applications*, 111(2):255–271, 2001.

[16] Jean-David Benamou and Guillaume Carlier. Augmented lagrangian methods for transport optimization, mean field games and degenerate elliptic equations. *Journal of Optimization Theory and Applications*, 167(1):1–26, 2015.

[17] Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented lagrangian approach to wasserstein gradient flows and applications. *ESAIM: Proceedings and Surveys*, 54:1–17, 2016.

[18] Jean-David Benamou, Guillaume Carlier, and Filippo Santambrogio. *Variational Mean Field Games*, pages 141–171. Springer International Publishing, Cham, 2017.

[19] Jean-David Benamou and Mélanie Martinet. Capacity Constrained Entropic Optimal Transport, Sinkhorn Saturated Domain Out-Summation and Vanishing Temperature. working paper or preprint, May 2020.

[20] Michele Benzi, Gene Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 05 2005.

[21] Gheorghe-Teodor Bercea, Andrew T. T. McRae, David A. Ham, Lawrence Mitchell, Florian Rathgeber, Luigi Nardi, Fabio Luporini, and Paul H. J. Kelly. A structure-exploiting numbering algorithm for finite elements on extruded meshes, and its performance evaluation in firedrake. *Geoscientific Model Development*, 9(10):3803–3815, 2016.

[22] Marianne Bessemoulin-Chatard. A finite volume scheme for convection-diffusion equations with nonlinear diffusion derived from the Scharfetter-Gummel scheme. *Numer. Math.*, 121(4):637–670, 2012.

[23] Adrien Blanchet. A gradient flow approach to the Keller-Segel systems. RIMS Kokyuroku's lecture notes, vol. 1837, pp. 52–73, June 2013.

[24] Daniele Boffi, Franco Brezzi, Michel Fortin, et al. *Mixed finite element methods and applications*, volume 44. Springer, 2013.

[25] François Bolley, Ivan Gentil, and Arnaud Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8):2430–2457, 2012.

[26] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[27] Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

[28] Haïm Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[29] Giuseppe Buttazzo, Chloé Jimenez, and Edouard Oudet. An optimization problem for mass transportation with congested dynamics. *SIAM Journal on Control and Optimization*, 48(3):1961–1976, 2009.

[30] Vincent Calvez and Thomas O. Gallouët. Particle approximation of the one dimensional Keller-Segel equation, stability and rigidity of the blow-up. *Discr. Cont. Dyn. Syst. A*, 36(3):1175–1208, 2016.

[31] Clément Cancès. Energy stable numerical methods for porous media flow type problems. *Oil & Gas Science and Technology-Rev. IFPEN*, 73:1–18, 2018.

[32] Clément Cancès, Thomas O. Galloëut, and Gabriele Todeschi. A variational finite volume scheme for wasserstein gradient flows. *Numerische Mathematik*, 146:437–480, 10 2020.

[33] Clément Cancès, Thomas O. Gallouët, Maxime Laborde, and Léonard Monsaingeon. Simulation of multiphase porous media flows with minimizing movement and finite volume schemes. HAL: hal-01700952, to appear in European J. Appl. Math., 2018.

[34] Clément Cancès, Thomas O. Gallouët, and Léonard Monsaingeon. Incompressible immiscible multiphase flows in porous media: a variational approach. *Anal. PDE*, 10(8):1845–1876, 2017.

[35] Clément Cancès and Cindy Guichard. Convergence of a nonlinear entropy diminishing Control Volume Finite Element scheme for solving anisotropic degenerate parabolic equations. *Math. Comp.*, 85(298):549–580, 2016.

[36] Clément Cancès and Cindy Guichard. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Found. Comput. Math.*, 17(6):1525–1584, 2017.

[37] Clément Cancès, Daniel Matthes, and Flore Nabet. A two-phase two-fluxes degenerate Cahn-Hilliard model as constrained Wasserstein gradient flow. *Arch. Ration. Mech. Anal.*, 233(2):837–866, 2019.

[38] Clément Cancès, Flore Nabet, and Martin Vohralík. Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations. HAL: hal-01894884, 2018.

[39] Pierre Cardaliaguet. Weak solutions for first order mean field games with local coupling, 2013.

[40] José A. Carrillo, Katy Craig, and Francesco S. Patacchini. A blob method for diffusion. *Calc. Var. Partial Differential Equations*, 58(2):53, 2019.

[41] José A. Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows. *Foundations of Computational Mathematics*, 2021.

[42] José A. Carrillo, Marco DiFrancesco, Alessio Figalli, Thomas Laurent, and Dejan Slepčev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.*, 156(2):229–271, 2011.

[43] José A. Carrillo, Bertram Düring, Daniel Matthes, and David S. McCormick. A Lagrangian scheme for the solution of nonlinear diffusion equations using moving simplex meshes. *J. Sci. Comput.*, 73(3):1463–1499, 2018.

[44] Claire Chainais-Hillairet, Jian-guo Liu, and Yue-Jun Peng. Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. *Mathematical Modelling and Numerical Analysis*, 37:319–338, 03 2003.

[45] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.

[46] Lénaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fisher–Rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[47] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[48] Ennio De Giorgi. New problems on minimizing movements. *Ennio de Giorgi: Selected Papers*, pages 699–713, 1993.

[49] Jim Douglas, Todd Dupont, and Lars Wahlbin. The stability in L^q of the L^2-projection into finite element function spaces. *Numerische Mathematik*, 23(3):193–197, 1974.

[50] Matthias Erbar and Jan Maas. Gradient flow structures for discrete porous medium equations. *Discrete Contin. Dyn. Syst.*, 34(4):1355–1374, 2014.

[51] Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 07 2017.

[52] Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon. Computation of optimal transport on discrete metric measure spaces. *Numerische Mathematik*, 144(1):157–200, 2020.

[53] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Discretization of heterogeneous and anisotropic diffusion problems on general nonconforming meshes sushi: a scheme using stabilization and hybrid interfaces. *IMA Journal of Numerical Analysis*, 30(4):1009–1043, 06 2009.

[54] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. In *Solution of Equation in* $\mathbb{R}^n$ *(Part 3), Techniques of Scientific Computing (Part 3)*, volume 7 of *Handbook of Numerical Analysis*, pages 713–1018. Elsevier, 2000.

[55] Robert Eymard and Gallouët Thierry. H-convergence and numerical schemes for elliptic problems. *SIAM J. Numerical Analysis*, 41:539–562, 04 2003.

[56] Enrico Facca, Franco Cardin, and Mario Putti. Towards a stationary monge–kantorovich dynamics: The physarum polycephalum experience. *SIAM Journal on Applied Mathematics*, 78, 10 2016.

[57] Enrico Facca, Sara Daneri, Franco Cardin, and Mario Putti. Numerical solution of monge–kantorovich equations via a dynamic formulation. *Journal of Scientific Computing*, 82, 09 2017.

[58] A. Forsgren, Philip E. Gill, and Margaret H. Wright. Interior methods for nonlinear optimization. *SIAM Review*, 44(4), 2002.

[59] Jürgen Fuhrmann. Existence and uniqueness of solutions of certain systems of algebraic equations with off-diagonal nonlinearity. *Appl. Numer. Math.*, 37:359–370, 2001.

[60] FVCAV. Benchmark. https://www.i2m.univ-amu.fr/fvca5/benchmark/Meshes/index.html.

[61] Thomas Gallouët, Maxime Laborde, and Leonard Monsaingeon. An unbalanced optimal transport splitting scheme for general advection-reaction-diffusion problems. *ESAIM: Control, Optimisation and Calculus of Variations*, 25:8, 2019.

[62] Klaus Gärtner and Lennard Kamenski. Why do we need voronoi cells and delaunay meshes? In Vladimir A. Garanzha, Lennard Kamenski, and Hang Si, editors, *Numerical Geometry, Grid Generation and Scientific Computing*, pages 45–60, Cham, 2019. Springer International Publishing.

[63] Nicola Gigli and Jan Maas. Gromov-Hausdorff convergence of discrete transportation metrics. *SIAM J. Math. Anal.*, 45(2):879–899, 2013.

[64] Peter Gladbach, Eva Kopfer, and Jan Maas. Scaling limits of discrete optimal transport. *arXiv preprint arXiv:1809.01092*, 2018.

[65] Jacek Gondzio. Interior point methods 25 years later. *European Journal of Operational Research*, 218(3):587–601, may 2012.

[66] Kevin Guittet. On the time-continuous mass transport problem and its approximation by augmented lagrangian techniques. *SIAM Journal on Numerical Analysis*, 41(1):382–399, 2003.

[67] Martin Heida. Convergences of the squareroot approximation scheme to the Fokker-Planck operator. *Math. Models Methods Appl. Sci.*, 28(13):2599–2635, 2018.

[68] Morgane Henry, Emmanuel Maitre, and Valérie Perrier. Primal-dual formulation of the dynamic optimal transport using helmholtz-hodge decomposition. 2019.

[69] Romain Hug, Emmanuel Maitre, and Nicolas Papadakis. Multi-physics optimal transportation and image interpolation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1671–1692, 2015.

[70] Romain Hug, Emmanuel Maitre, and Nicolas Papadakis. On the convergence of augmented lagrangian method for optimal transport between nonnegative densities. 2017.

[71] Noureddine Igbida and Van Thanh Nguyen. Augmented lagrangian method for optimal partial transportation. *IMA Journal of Numerical Analysis*, 38(1):156–183, 03 2017.

[72] Matt Jacobs, Inwon Kim, and Alpár R. Mészáros. Weak solutions to the Muskat problem with surface tension via optimal transport. arXiv:1905.05370, 2019.

[73] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[74] Oliver Junge, Daniel Matthes, and Horst Osberger. A fully discrete variational scheme for solving nonlinear Fokker–Planck equations in multiple space dimensions. *SIAM J. Numer. Anal.*, 55(1):419–443, 2017.

[75] Leonid V. Kantorovich. On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201, 1942.

[76] David Kinderlehrer, Léonard Monsaingeon, and Xiang Xu. A Wasserstein gradient flow approach to Poisson-Nernst-Planck equations. *ESAIM Control Optim. Calc. Var.*, 23(1):137–164, 2017.

[77] David Kinderlehrer and Noel J. Walkington. Approximation of parabolic equations using the Wasserstein metric. *M2AN Math. Model. Numer. Anal.*, 33(4):837–852, 1999.

[78] Philippe Laurençot and Bogdan-Vasile Matioc. A gradient flow approach to a thin film approximation of the Muskat problem. *Calc. Var. Partial Differential Equations*, 47((1-2)):319–341, 2013.

[79] Hugo Lavenant. Unconditional convergence for discretizations of dynamical optimal transport. *arXiv preprint arXiv:1909.08790*, 2020.

[80] Hugo Lavenant, Sebastian Claici, Edward Chien, and Justin Solomon. Dynamical optimal transport on discrete surfaces. *ACM Transactions on Graphics (TOG)*, 37(6):1–16, 2018.

[81] Hugo Leclerc, Quentin Mérigot, Filippo Santambrogio, and Federico Stra. Lagrangian discretization of crowd motion and linear diffusion. arXiv: 1905.08507, 2019.

[82] Guillaume Legendre and Gabriel Turinici. Second-order in time schemes for gradient flows in wasserstein and geodesic metric spaces. *Comptes Rendus Mathematique*, 355:345–353, 03 2017.

[83] Jean Leray and Jules Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup.*, 51((3)):45–78, 1934.

[84] Wuchen Li, Jianfeng Lu, and Li Wang. Fisher information regularization schemes for wasserstein gradient flows. *Journal of Computational Physics*, 416:109449, 2020.

[85] Wuchen Li, Penghang Yin, and Stanley Osher. Computations of optimal transport distance with fisher information regularization. *Journal of Scientific Computing*, 75(3):1581–1595, 2018.

[86] Gary M. Lieberman. Oblique derivative problems in Lipschitz domains. II. Discontinuous boundary data. *J. reine angew. Math*, 389:1–21, 1988.

[87] Christian Léonard. A survey of the schrödinger problem and some of its connections with optimal transport. *Discrete & Continuous Dynamical Systems*, 34(4):1533–1574, 2014.

[88] Jan Maas. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, 261(8):2250–2292, 2011.

[89] Jan Maas and Daniel Matthes. Long-time behavior of a finite volume discretization for a fourth order diffusion equation. *Nonlinearity*, 29(7):1992–2023, 2016.

[90] Daniel Matthes, Robert J. McCann, and Giuseppe Savaré. A family of nonlinear fourth order equations of gradient flow type. *Comm. Partial Differential Equations*, 34:1352–1397, 2009.

[91] Daniel Matthes and Horst Osberger. Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation. *ESAIM Math. Model. Numer. Anal.*, 48(3):697–726, 2014.

[92] Daniel Matthes and Horst Osberger. A convergent Lagrangian discretization for a nonlinear fourth-order equation. *Found. Comput. Math.*, 17(1):73–126, 2017.

[93] Daniel Matthes and Simon Plazotta. A variational formulation of the bdf2 method for metric gradient flows. *ESAIM: Mathematical Modelling and Numerical Analysis*, 53(1):145–172, 2019.

[94] Bertrand Maury, Aude Roudneff-Chupin, and Filippo Santambrogio. A macroscopic crowd motion model of gradient flow type. *Math. Models Methods Appl. Sci.*, 20(10):1787–1821, 2010.

[95] Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Mathematical Journal*, 80(2):309–323, 1995.

[96] Robert J. McCann. A convexity principle for interacting gases. *Advances in Mathematics*, 128(1):153–179, 1997.

[97] Andrew T. T. McRae, Gheorghe-Teodor Bercea, Lawrence Mitchell, David A. Ham, and Colin J. Cotter. Automated generation and symbolic manipulation of tensor product finite elements. *SIAM Journal on Scientific Computing*, 38(5):S25–S47, 2016.

[98] Quentin Mérigot and Boris Thibert. Chapter 2 - optimal transport: discretization and algorithms. In Andrea Bonito and Ricardo H. Nochetto, editors, *Geometric Partial Differential Equations - Part II*, volume 22 of *Handbook of Numerical Analysis*, pages 133–212. Elsevier, 2021.

[99] Alexander Mielke. A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.

[100] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[101] Ayman Moussa. Some variants of the classical Aubin-Lions Lemma. *J. Evol. Equ.*, 16(1):65–93, 2016.

[102] Thomas J. Murphy and Noel J. Walkington. Control volume approximation of degenerate two-phase porous media flows. *SIAM J. Numer. Anal.*, 57(2):527–546, 2019.

[103] Andrea Natale and Gabriele Todeschi. A mixed finite element discretization of dynamical optimal transport. working paper or preprint, May 2020.

[104] Andrea Natale and Gabriele Todeschi. TPFA Finite Volume Approximation of Wasserstein Gradient Flows. In *Finite Volumes for Complex Applications IX - Methods, Theoretical Aspects, Examples*, pages 193–201. Springer International Publishing, 2020.

[105] Andrea Natale and Gabriele Todeschi. Computation of optimal transport with finite volumes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(5):1847–1871, September 2021.

[106] Luìs Neves de Almeida, Federica Bubba, Benoît Perthame, and Camille Pouchol. Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations. arXiv:1803.10629, 2018.

[107] Felix Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory. *Arch. Rational Mech. Anal.*, 141(1):63–103, 1998.

[108] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.

[109] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[110] Rémi Peyre. Comparison between w_2 distance and h^{-1} norm, and localization of $w$asserstein distance. *ESAIM: COCV*, 24(4):1489–1501, 2018.

[111] Gabriel Peyré and Marco Cuturi. Computational optimal transport, 2020.

[112] Thomas Pock, Daniel Cremers, Horst Bischof, and Antonin Chambolle. An algorithm for minimizing the mumford-shah functional. In *2009 IEEE 12th International Conference on Computer Vision*, pages 1133–1140. IEEE, 2009.

[113] Imre Pólik and Tamás Terlaky. *Interior Point Methods for Nonlinear Optimization. In: Di Pillo G., Schoen F. (eds) Nonlinear Optimization. Lecture Notes in Mathematics*, volume 1989. Springer Berlin Heidelberg, 2010.

[114] Florian Rathgeber, David A. Ham, Lawrence Mitchell, Michael Lange, Fabio Luporini, Andrew T. T. McRae, Gheorghe-Teodor Bercea, Graham R. Markall, and Paul H. J. Kelly. Firedrake: automating the finite element method by composing abstractions. *ACM Trans. Math. Softw.*, 43(3):24:1–24:27, 2016.

[115] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.

[116] Filippo Santambrogio. Euclidean, Metric, and Wasserstein gradient flows: an overview, 2016.

[117] Filippo Santambrogio and Xu-Jia Wang. Convexity of the support of the displacement interpolation: Counterexamples. *Applied Mathematics Letters*, 58:152–158, 2016.

[118] Guido Stampacchia. Problemi al contorno ellittici, con dati discontinui, dotati di soluzioni hölderiane. *Annali di Matematica pura ed applicata*, 51(1):1–37, 1960.

[119] Zheng Sun, José A. Carrillo, and Chi-Wang Shu. A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials. *J. Comput. Phys.*, 352:76–104, 2018.

[120] Cédric Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

[121] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

[122] Augusto Visintin. *Models of phase transitions*, volume 28 of *Progress in nonlinear differential equations and their applications*. Birkhäuser Boston, 1996.

RÉSUMÉ

Cette thèse a pour objet la construction de schémas numériques localement conservatif et préservant la structure pour des flots de gradient Wasserstein, c'est à dire des courbes de descente maximale dans l'espace de Wasserstein. Les discrétisations en temps reposent sur des formulations variationnelles imitant au niveau discret ce comportement de courbes de descente maximale. Ces discrétisation font intervenir le calcul de la distance de Wasserstein, un exemple de problèmes de transport optimal. Les discrétisations en espaces sont basées sur des approximations volumes finis avec reconstructions à deux points des flux, également appelés schémas TPFA. Ces méthodes sont bien connues et particulièrement adaptées pour discrétiser des équations conservatives. Afin de conserver les structures variationnelles au niveau discret, notre approche est de d'abord discrétiser puis optimiser. Dans une première partie nous présentons des discrétisations TPFA pour la distance de Wasserstein, basées sur la formulation dynamique de Benamou-Brenier du transport optimal. Nous montrons des problèmes de stabilité liés à ces discrétisations et proposons une méthode permettant de les surmonter. Nous dérivons des estimations quantitatives de convergence pour ce model discret. Afin de résoudre le problème d'optimisation discret, nous introduisons une stratégie de point intérieur. Ensuite nous proposons des schémas d'ordre un puis deux pour des flots de gradients Wasserstein. Afin de réduire la complexité numérique des problèmes étudiés nous utilisons une linéarisation implicite de la distance de Wasserstein. En exploitant la monotonie de la reconstruction upwind, nous proposons un schéma d'ordre un que l'on peut résoudre efficacement avec une méthode de Newton et nous montrons sa convergence vers des solutions faibles de l'équation de Fokker-Planck. Pour augmenter l'ordre de convergence en espace, nous utilisons une reconstruction centrée qui nécessite une technique d'optimisation différente. Nous utilisons à nouveau la stratégie du point intérieur pour cela. Finalement, pour monter en ordre en temps, nous proposons une version modifiée de la discrétisation variationnelle BDF2 pour laquelle nous prouvons la convergence vers des flots de gradient Wasserstein. À l'aide de ces nouvelles discrétisations, nous construisons un schéma d'ordre deux en espace et en temps. Tous les schémas proposés sont accompagnés de nombreux résultats numériques.

MOTS CLÉS

Transport optimal, Flots de gradient Wasserstein, Volumes finis, Optimisation

ABSTRACT

This thesis is devoted to the design of locally conservative and structure preserving schemes for Wasserstein gradient flows, i.e. steepest descent curves in the Wasserstein space. The time discretization is based on variational approaches that mimic at the discrete in time level the behavior of steepest descent curves. These discretizations involve the computation of the Wasserstein distance, an instance of optimal transport problem. The space discretization is based on Two-Point Flux Approximation (TPFA) finite volumes, a well-known methodology particularly suited for the discretization of partial differential equations that present a conservative structure. In order to preserve the variational structure at the discrete level, we follow a first discretize then optimize approach. We start by presenting TPFA discretizations for the Wasserstein distance based on the Benamou-Brenier dynamical formulation. We expose some stability issues related to these discetizations, propose a possible solution to overcome them and derive quantitative estimate on the convergence of the discrete model. To solve the discrete optimization problem, we introduce an interior point strategy. Then, we propose first and second order accurate schemes for Wasserstein gradient flows. At this level, to reduce the computational complexity, we use an implicit linearization of the Wasserstein distance. By taking adavantage of the monotonicity of the upwind reconstruction, we propose a first order scheme which can be efficiently solved with a Newton method and show its convergence towards distributional solutions of the Fokker-Planck equation. In order to higher the accuracy in space, we use a centered reconstruction, which requires a different optimization technique. We use again the interior point strategy for this purpose. Finally, we propose a modified variational BDF2 time discretization and prove its convergence towards Wasserstein gradient flows. Thanks to these new discretizations, we design a second order accurate scheme in both time and space. All our approaches are validated with several numerical results.

KEYWORDS

Optimal transport, Wasserstein gradient flows, Finite volumes, Optimization