PSL UNIVERSITÉ PARIS

HABILITATION À DIRIGER DES RECHERCHES

DE L'UNIVERSITÉ PSL

Présentée à l'Université Paris-Dauphine

Optimal Transport and applications to the study of some geometrical partial differential equations

Présentation des travaux par

Thomas Gallouët

Le 24 Janvier 2024

Discipline

Mathématiques

Composition du jury :

Didier, Bresch Directeur de recherche CNRS, Université Savoie Mont blanc	*Rapporteur*
Daniel, Matthes Professeur des universités, Technische Universität München	*Rapporteur*
Giuseppe, Savaré Professeur des universités, Université de Bocconi	*Rapporteur*
Yann, Brenier Directeur de recherche CNRS, Université Paris Saclay	*Examinateur*
Gabriel, Peyré Directeur de recherche CNRS, École Normale Supérieur, PSL	*Examinateur*
Sylvia, Serfaty Professeur des universités, New York University,	*Examinatrice*
Guillaume, Carlier Professeur des universités Université Paris Dauphine, PSL	*Coordinateur*

Dauphine | PSL UNIVERSITÉ PARIS

UNIVERSITÉ PARIS-DAUPHINE
CEREMADE

MÉMOIRE D'HABILITATION À DIRIGER LES RECHERCHES

Optimal Transport and applications to the study of some geometrical partial differential equations

---

présenté par

Thomas GALLOUËT

spécialité: Mathématiques

coordinateur des travaux: Guillaume Carlier

soutenu le 24 Janvier 2024 après avis de

Didier	Bresch
Daniel	Matthes
Giuseppe	Savare

devant le jury composé de :

Yann	Brenier
Didier	Bresch
Guillaume	Carlier
Daniel	Matthes
Gabriel	Peyré
Giuseppe	Savaré
Sylvia	Serfaty

Université Paris-Dauphine

Abstract

MIDO
Ceremade

Habilitation à diriger les recherches

Optimal Transport and applications to the study of some geometrical partial differential equations

by Thomas GALLOUËT

This document is about Optimal Transport and its application to partial differential equations such as gradient flows or Euler flows in the Wasserstein spaces. We investigate theoretical as well as numerical questions. On the theoretical side of optimal transport, we address questions such as Wasserstein splines, Wasserstein extrapolation and some questions related to the smoothness of Unbalanced Optimal Transport (Unbalanced Brenier polar projection, Unbalanced Monge-Ampère equations, a special class of Cone convex functions). We then apply the Wasserstein Gradient/Euler flow structure to the study of some PDEs.

On the one hand, the flow structure is used to prove theoretical results, notably the existence of solutions to the system of incompressible immiscible multiphase flows in porous media, and the definition of the notion of relaxed solution for the Camassa-Holm equations, which happens to be the counter part for the Unbalanced Optimal Transport of what Incompressible Euler is for the classical Optimal Transport. One the other hand, the geometrical structure is also used to design, implement and prove convergence for different numerical schemes. For instance we introduce the notion of variational Finite Volume schemes for Wasserstein Gradient flows. These schemes are finite volume schemes defined as the Euler-Lagrange equations for a space discretization of a minimizing movement (JKO) scheme, a "first discretize then optimize" approach. We also defined Lagrangian numerical schemes for a class of Gradient and Euler flows. These schemes are ODEs preserving the underlying geometrical structure with an approximated energy defined through semi discrete Optimal Transport. Through a splitting procedure and using Unbalanced Optimal Transport, all the effort undertaken for Wasserstein Gradient Flows can by extended to encompass more general and non conservative reaction diffusion equations.

Acknowledgements

Remerciements

First of all I would like to thank Didier Bresch, Daniel Matthes and Giuseppe Savare for agreeing to review this manuscript (and for the nice reports they wrote). Many thanks also to Yann Brenier, Guillaume Carlier, Gabriel Peyré and Sylvia Serfaty for agreeing to be on the jury. I'm impressed and have learned a lot from the work of each of the members of the jury. It is a real honor to have them here. I would like to add a special note to Sylvia Serfaty who was present in both my PhD and HDR jury, and to Guillaume Carlier for agreeing to be my HDR-coordinator and to manage the constraints and stress that comes with it.

J'ai également eu la chance ces dix dernières années de travailler dans des environnements exceptionnels, que ce soit à Lille, Bruxelles, Liège, Paris, Palaiseau, ou Orsay, en côtoyant des collègues qui étaient à la fois sympathiques et inspirants. Ils faisaient de ces lieux non pas un simple espace de travail mais un espace de vie. J'ai appris de mes expériences dans chacun de ces endroits. Chaque départ était accompagné de douces pensées mélancoliques, imaginant avoir pu rester plus longtemps. Chaque arrivée était chaleureuse, enthousiaste et l'occasion de découvrir, apprendre, et vivre de nouvelles choses. Un merci particulier à ceux qui ont rendu ces événements possibles : Claire Chainais-Hillairet, Antoine Gloria, Yann Brenier (merci Filippo pour l'idée), Yvik Swan, Guillaume Carlier, Jean-David Benamou, Gabriel Peyré.

Merci au Ceremade et à l'exceptionnelle équipe Mokaplan pour m'avoir accueilli ces dernières années. La composition de l'équipe varie fortement au cours du temps et je ne vais pas prendre le risque de citer tout le monde, mais en plus de ceux déjà nommés dans cette page je remercie Vincent Duval, Derya Gök, Luca Nenna, Paul Pegon, Martine Verneuille, et Irène Waldspurger qui forment un noyau stable de ce groupe de recherche depuis mon arrivée et avec qui ça a toujours été un plaisir de passer du temps.

Et pour la nouvelle page qui s'ouvre, merci au Laboratoire de Mathématiques d'Orsay et à Inria Saclay d'avoir soutenu notre projet d'équipe. Merci à tous les membres de ParMA pour avoir accepté et contribué à faire cette équipe et à Bertrand Thirion et Jean-David Benamou pour leur pression légère mais constante afin d'obtenir une date pour la soutenance de l'HDR.

Merci à toutes les personnes rencontrées à ces occasions, personnels administratifs, étudiant.e.s, doctorant.e.s., post-doctorant.e.s, chercheur.euse.s, Usbyressois.e.s, qui font que ces dernières années ont été scientifiquement et humainement un bonheur à vivre. Je pense évidemment à toutes ces discussions mathématiques, scientifiques qui sont l'essence de notre métier, avec une mention spéciale à celles inutiles a priori au sens d'une productivité L^1, mais tellement passionnantes et enrichissantes ainsi qu'à celles passées au téléphone, en visio, dans des endroits et/ou moments disons peu standards.

Mais je pense tout autant aux pauses cafés sans café, discussions sur les vélos pliables, cargo, bateau, les enfants, la politique, l'histoire, la géographie, philosophie, nature humaine, ... les séances d'escalades (avec un zeste de Verdon), de ski (débutant ou pas), tennis, squash, trail, ... et les moments passés en vos compagnies dans les bars du 12eme, de Bruxelles, de Liège, de la place de Tilff et particulièrement

d'une terrasse à Lisbonne où je suis probablement encore un peu assis.

J'ai rencontré beaucoup trop de personnes passionnantes ces dernières années pour me lancer dans une liste exhaustive. Parmi elles merci encore à mes collaborateur.trice.s : Jean-David Benamou, Vincent Calvez, Clément Cancès, Claire Chainais-Hillairet, Alessio Figalli, Roberta Ghezzi, Maxime Laborde, Quentin Mérigot, Guillaume Mijoule, Léonard Monsaingeon, Andrea Natale, Ludovic Rifford, Erwan Stämpfli, Yvik Swan, Gabriele Todeschi, François-Xavier Vialard. C'était et c'est un vrai plaisir de travailler avec chacun.e.s d'entre vous et j'espère que malgré tout c'est un peu réciproque.

Merci à Nathalie, Cécile, Luca, Andrea et Léonard pour leur disponibilités et relectures, pas seulement de ce document, et par ailleurs infiniment plus efficaces que ChatGPT.

Merci à vous tous avec qui j'ai partagé du temps ces dernières années pour avoir contribué à en faire de très bons souvenirs scientifiquement et humainement : ma famille au sens large, mes ami.e.s de toujours, le monde des mathématiques et de la montagne (ou à défaut de quelques collines boisées, cailloux et autre viaduc).

Et bien sûr merci à Cécile, Gwenaël, Alizé et Tiphaine pour être la source de mes plus grands moments de bonheur.

Contents

Chapter 1

Research summary

1.1 Introduction

This manuscript starts with a summary of some of my research papers of the last ten years. Then after a synthetic reminder of their main contributions, I present the articles in their entirety. All these works are somehow related to Optimal Transport. One of the main focus of my research was the study, from a theoretical and numerical point of view, of PDEs which happen to have a geometric structure in Wasserstein-like spaces, like Gradient flows or Euler Flows. This requires an in-depth understanding of tools that are specific to optimal transport, such as Wasserstein splines, Wasserstein geodesics extrapolation or a better understanding of Unbalanced Optimal Transport. We can organize these works in two parts.The first one is built around Optimal Transport and the second one around Unbalanced Optimal Transport. Each part being composed of three similar research directions. The first direction deals with structural properties of balanced/unbalanced Optimal Transport. The second axis details some numerical methods for the approximation of balanced/unbalanced gradient flows or more general reaction diffusion equations. The last axis focuses on Euler flows and numerical methods designed to approximate them.

This work was carried out with several collaborators I met in my life as a researcher. At different times, I was their student, their colleague or their supervisor. First, I give a summary of these collaborators and the contribution of research papers detailing each articles and the links between them. Then, I join the papers in the structure presented above, adding at the beginning a quick reminder of the main contributions and some research perspectives.

Collaborators

Co-authors

J.D. Benamou, C. Cancès, C. Chainais-Hillairet (Post-doc supervisor), R. Ghezzi, M. Laborde, Q. Mérigot, G. Mijoule, L. Monsaingeon, A. Natale, Gabriele Todeschi, Y. Swan (Post-doc supervisor), F.X. Vialard.

Post-doc students

- Andrea Natale, 2017-2020, co-supervision with F.X. Vialard and then Q. Mérigot.
- Guillaume Mijoule, 2018-2020.

PhD students

- Gabriele Todeschi, 2018- 2021, co-supervision with C. Cancès.

- Erwan Stämpfli, 2021-. 2023, co-supervision with Y. Brenier.

Undergraduate students

- Médard Govoeyi, Master 2, Avril 2023- Sep. 2023, co-supervision with M. Laborde.
- Erwan Stämpfli, Master 2, Avril 2021- Sep. 2021, co-supervision with Y. Brenier.
- Jean Jacques Godeme, Master 2, Avril 2020- Sep. 2020, co-supervision, with Léonard Monsaingeon.
- Gabriele Todeschi, Master 2, Avril 2018- Sep. 2018, co-supervision with C. Cancès.
- Jean Paul Greveni, L3, 2017
- Cédric Oms, Master 1, 2016

Research papers

The manuscript is composed of the following research papers, listed in order of appearance in the manuscript:

1. **Second order models for optimal transport and cubic splines on the Wasserstein space.** *Foundations of Computational Mathematics, Springer Verlag* (2019) https://hal.science/hal-01682107v2 J.D. Benamou, Gallouët T.O. et Vialard F.X.
2. **From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows.** *Under minor revision Mathematics of Computations* (2022) Gallouët T.O., Natale A. et Todeschi. G https://hal.science/hal-03f790981v2
3. **The gradient flow structure for incompressible immiscible two-phase flows in porous media.** *C. R. Acad. Sci. Paris, Ser. I(353) :985– 989* (2015). https://hal.science/hal-01122770. Cancès C., Gallouët T.O., Monsaingeon L.
4. **Incompressible immiscible multiphase flows in porous media: a variational approach.** *Analysis and PDE Vol. 10* (2017), No. 8, 1845–1876 https://arxiv.org/abs/1607.04009. Cancès C., Gallouët T.O., Monsaingeon L.
5. **Simulation of multiphase porous media flows with minimizing movement and finite volume schemes.)** *European Journal of Applied Mathematics, Cambridge University Press (CUP), 30 (6), pp.1123-1152* (2019). https://arxiv.org/abs/arXiv:1802.01321. Cancès C., Gallouët T.O., Laborde M., Monsaingeon L.
6. **A Lagrangian scheme *à la Brenier* for the incompressible Euler equations.** *Found Comput Math 18: 835* (2018). https://doi.org/10.1007/s10208-017-9355-y. Gallouët T.O. and Mérigot Q.
7. **Convergence of a Lagrangian discretization for barotropic fluids and porous media flow.** *SIAM Journal on Mathematical Analysis* (2021) https://hal.science/hal-03234144. Gallouët T.O., Mérigot Q., Natale A.

8. **Regularity theory and geometry of unbalanced optimal transport.** *Submitted 2023* Gallouët T.O., Ghezzi R. et Vialard F.X. https://hal.science/hal-03498098v1.

9. **A JKO splitting scheme for Kantorovich-Fisher-Rao gradient flows.** *SIAM Journal on Mathematical Analysis, Vol. 49, Issue 2.* (2017) https://arxiv.org/abs/1602.04457. Gallouët T.O. et Monsaingeon L.

10. **An unbalanced optimal transport splitting scheme for general advection-reaction-diffusion problems.** *ESAIM: Control, Optimisation and Calculus of Variations* (2018) https://hal.science/hal-01508911. Gallouët T.O., Laborde M. and Monsaingeon L.

11. **The Camassa-Holm equation as an incompressible Euler equation: a geometric point of view.** *Journal of Differential Equations, Volume 264, Issue 7, Pages 4199-4234.* (2018) https://arxiv.org/abs/1609.04006. Gallouët T.O. and Vialard F.X.

12. **Generalized compressible flows and solutions of the H(div) geodesic problem.** *Archive for Rational Mechanics and Analysis, Springer Verlag* (2020) https://hal.science/hal-01815531v3. Gallouët T.O., Natale A. et Vialard F.X.

1.2 Optimal Transport

1.2.1 Optimal Transport, Wasserstein space

The following two papers deal with some notions in the Wasserstein space namely Wasserstein splines and Wasserstein extrapolation. The first paper was realized with two collaborators I got when I arrived at Inria Paris: J.D. Benamou (DR Inria Paris) and F.X. Vialard (MCF Dauphine now Professeur at Paris Est). The second paper was written with A. Natale (former post doc student, now CR at Inria Lille) and G. Todeschi (former PhD student, currently Post-doc.)

Articles:

1. **Second order models for optimal transport and cubic splines on the Wasserstein space.** *Foundations of Computational Mathematics, Springer Verlag* (2019) https://hal.science/hal-01682107v2 J.D. Benamou, Gallouët T.O. et Vialard F.X.

2. **From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows.** *Under minor revision Mathematics of Computations* (2022) Gallouët T.O., Natale A. et Todeschi. G https://hal.science/hal-03f790981v2

Cubic Splines.

In this work realized in collaboration with J.D. Benamou and F.X. Vialard, we extend the Wasserstein geodesics, defined on the space of probability densities, to the case of higher-order interpolation such as cubic spline interpolation. Our motivation is to answer the practical question of the extension of cubic splines to the Wasserstein space and their numerical computation. First we present the natural extension of cubic splines to the Wasserstein space when considered as a Riemannian manifold. We then propose a simpler approach based on the relaxation of the variational problem

on the path space. This relaxation is defined on the space of densities using multimarginal optimal transport and yields a convex minimization problem. In short, the proposed method consists in minimizing, on the space of measures on the path space, under marginal constraints, the squared norm of the acceleration. This relaxation is performed in the spirit of generalized geodesics for Euler equations introduced by Brenier. In this setting, we show that two numerical approaches, classical in optimal transportation can be applied. One is based on entropic regularization and the Sinkhorn Algorithm, the other relies on the Semi-Discrete formulation of Optimal Transportation and the computation of Laguerre cells, a classical problem in computational geometry. We showcase our methodology on 1D and 2D data. To the best of our knowledge, this question has not been yet addressed in the literature on optimal transport until very recently in two independent and simultaneous preprints : [4] and 1 (this paper). Both works share the same idea of relaxing the cubic spline formulation in the space of measures using multi-marginal optimal transport. Our paper however explores a larger hierarchy of models and several numerical methods.

In our implementation the numerical methods we proposed shared the same drawback, for a reasonable computational time, they are limited to low dimension $d \leq 3$. Moreover the semi-discrete method is limited to the quadratic cost and is not convex in general. A smart initialization or optimization strategy is needed to obtain the convergence towards global minima. Recent advances in Semi-Discrete Optimal Transport Solvers open the door to an implementation in higher order dimension whereas for the entropic regularization problems classical multi-scale approaches can be used, bearing in mind that for the interpolation the dependence in the time-step discretization is of a higher order than the one in the case of Sinkhorn algorithm for classical optimal transport. This prevents us to decrease the regularized parameter ϵ as efficiently as done for Sinkhorn algorithm for classical optimal transport. At the end of the paper (Remark 6) we notice that this notion of interpolation with relaxed multi-marginal optimal transport can be used to define a Wasserstein extrapolation. This question of defining an interpolation was then pursued in the work described below with different collaborators.

Wasserstein extrapolation.

The study of Wasserstein geodesic extrapolation is a part of 2. In this paper it is used as a tool to define a 2nd order in time numerical scheme for Wasserstein gradient flows. However it has is own interest. This operation is not uniquely defined in general since after time 1 shocks can occur in the trajectory of particles associated to the Wasserstein geodesic. With Andrea Natale (former post doc) and Gabriele Todeschi (former PhD student) we proposed different definitions of Wasserstein extrapolation in the case where the cost is given by the square of the euclidian distance. These definitions are given via different formulations of Optimal Transport and leads to the definition of Free-flow, metric, viscosity extrapolations. Each of these corresponds to a different way of handling shocks: either a shockless traverse, or different types of dissipative collisions. We proved the well posedness of these notions as well as some important properties that we define such as consistency or dissipation. The metric formulation for instance is given by a convex optimization problem. This convexity is not obvious and can be obtained thanks to a dual convex formulation in the spirit of Toland duality [3]. We also proposed a numerical scheme and an implementation to approximate the viscosity extrapolation. However the metric extrapolation seems to us the more natural and richest definition. In a follow up work we study more

deeply the metric extrapolation and its dual formulation for more general costs. In the quadratic case we aim at proposing a numerical scheme and an implementation for this metric extrapolation based on a non convex reformulation of the dual problem and semi-discrete techniques.

1.2.2 Wasserstein Gradient flows

A large part of my research was focused on PDEs that can be recast as gradient flows in the Wasserstein space i.e. equations or system of equations that can be recast under the form

$$\partial_t \rho - \operatorname{div}\left(\rho \nabla \frac{\delta \mathcal{E}}{\delta \rho}(\rho)\right) = 0,$$

with a zero flux boundary condition and for a given energy $\mathcal{E}$ defined on the set of probability measures. We used this interpretation either to prove the existence of solutions for a system of PDEs but also to build variational, energy-diminishing schemes. The first three papers presented below deal with a particular system of PDEs: incompressible immiscible multiphase flows. The next two papers aim at building variational finite volume numerical scheme in order to compute numerical approximations of general Wasserstein gradient flows.

Articles:

3. **The gradient flow structure for incompressible immiscible two-phase flows in porous media.** *C. R. Acad. Sci. Paris, Ser. I(353) :985– 989* (2015). https://hal.science/hal-01122770. Cancès C., Gallouët T.O., Monsaingeon L.

4. **Incompressible immiscible multiphase flows in porous media: a variational approach.** *Analysis and PDE Vol. 10* (2017), No. 8, 1845–1876 https://arxiv.org/abs/1607.04009. Cancès C., Gallouët T.O., Monsaingeon L.

5. **Simulation of multiphase porous media flows with minimizing movement and finite volume schemes.)** *European Journal of Applied Mathematics, Cambridge University Press (CUP), 30 (6), pp.1123-1152* (2019). https://arxiv.org/abs/arXiv:1802.01321. Cancès C., Gallouët T.O., Laborde M., Monsaingeon L.

6. **A variational finite volume scheme for Wasserstein gradient flows.** *Numerische Mathematik, Springer Verlag, 146 (3), pp 437 - 480* (2020). https://hal.science/hal-02189050. C.Cancès, Gallouët T.O., Todeschi. G

7. **From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows.** Under minor revision for *Mathematics of Computations* (2023) https://hal.science/hal-03790981 Gallouët T.O., Natale A. et Todeschi. G

Incompressible immiscible multiphase flows in porous media

This research was carried out in collaboration with C. Cancès and L. Monsaingeon. We were joined by M. Laborde for the numerical paper 5. The models for multiphase porous media flows have been widely studied in the last decades since they are of great interest in several fields of applications, like e.g. oil-engineering, carbon dioxide sequestration, or nuclear waste repository management. However in the case of more than three phases there were no existence results. The difficulty is that

the system of PDEs is not completely parabolic, making it difficult to obtain a priori estimates. Moreover the possible presence of vacuum for some phases leads to a series of technical difficulties. In 3 we highlight the Wasserstein gradient flow structure. Then in 4 we fully leverage this interpretation in order to prove the existence of solutions to the incompressible immiscible mutliphase-phase flow in a possibly heterogeneous porous medium. The proof is based on the convergence of a JKO scheme. It uses, among other things, flow interchange and duality methods in order to collect enough estimates for an Aubin-Lions convergence strategy to work. Finally in 5 we propose, implement and compare two numerical methods which are both designed to decrease the natural energy. One is based on a classical upstream mobility finite volume scheme, which is a reference for such equations. The other, ALG2-JKO, is a discretization of the JKO scheme. Both methods are well adapted for gradient flows equations, and more precisely they verify the following key properties for the numerical solutions; namely:

- preservation of positivity,
- conservation of mass and saturation constraints,
- energy dissipation along solutions.

We found that the ALG2-JKO scheme produces very similar results: same qualitative behaviour, conservation of the mass of each phase and preservation of the positivity while being more robust and adaptative. But the finite volume approach is under some conditions computationally more efficient. A natural question then arises: can we build a numerical scheme that would share the best of the two approaches? This is the object of the next section. Another direction of research is to understand what happens when the internal energy of the multiphase flow vanishes. All that remains are potential energies and constraints. In this vanishing internal-energy limit the system becomes hyperbolic instead of almost parabolic. This is the object of the ongoing thesis by Erwan Stampli's PhD that I co-supervise with Y. Brenier. We have two works in progress on this subject proving for instance the convergence of the parabolic system towards the hyperbolic counterpart on a torus in dimension 1.

Variational finite volume scheme

As seen above, a natural question arises from the numerical comparison between the ALG2-JKO scheme and the upstream mobility Finite Volume scheme presented in 5. Is there a way to combine the best of both methods? In other words can we build a Variational finite volume scheme that would exactly be the Euler-Lagrange equation of a fully discretized JKO step? A *first discretize then optimize* approach that would allow us to use a Newton method while keeping the variational structure. This was the starting point of Gabriele Todeschi's Phd done under the supervision of C. Cancès and myself. The first paper 6 answers this question positively, while in the second one 7 we propose to modify the variational structure in order to make the scheme 2nd order in time. During his PhD G. Todeschi developed together with A. Natale some methods to reach higher space orders within this class of schemes.

In 6, we then propose a variational finite volume scheme to approximate the solutions to Wasserstein gradient flows. The time discretization is based on a JKO formula and an implicit linearization of the Wasserstein distance expressed thanks to the Benamou-Brenier formula, whereas the space discretization relies on an upstream mobility two-point flux approximation finite volume scheme. The scheme is based on a *first discretize then optimize* approach in order to preserve the variational

structure at the discrete level. It can be applied to a wide range of energies and guarantees non-negativity of the discrete solutions as well as decay of the energy. We show that the scheme admits a unique solution whatever the convex energy involved in the continuous problem is, and we prove its convergence in the case of the linear Fokker-Planck equation with positive initial density. Numerical illustrations show that it is first order accurate in both time and space, and robust with respect to both the energy and the initial profile.

Then G. Todeschi was able to build in is PhD thesis higher order space approximations while keeping the variational structure. Later on in 7 we proposed a second order in time variational finite volume scheme. To do this we introduce a time discretization for Wasserstein gradient flows based on the classical Backward Differentiation Formula of order two. The main building block of the scheme is the notion of geodesic extrapolation in the Wasserstein space described in Section 1.2.1. We prove the convergence of the resulting scheme to the solution of the limit PDE in the case of the Fokker-Planck equation, and for a specific choice of extrapolation we also prove a more general result, that is convergence towards EVI flows. Finally, we propose a full discretization which numerically achieves second order accuracy in both space and time. This paper is inspired from previous works that were done in this direction but not completely satisfying to us from a numerical point of view see [16, 15, 11] for instance. The key difference between these papers and our work is a different interpretation of the BDF2 scheme. The method of proofs for the convergence of the scheme are then largely inspired from [16, 15].

1.2.3 Euler flows

Another class of PDEs relates to the Wasserstein space: the Euler flows where, instead of the speed, the acceleration is given by the Wasserstein gradient of an energy:

$$\partial_t \rho + \operatorname{div}(\rho v) = 0,\ \partial_t v + v \cdot \nabla v = -\nabla \frac{\delta \mathcal{E}}{\delta \rho}(\rho).$$

The incompressible Euler's equations fall into this category as well as some compressible Euler equations. Building Lagrangian numerical schemes for these equations was the object of the following works. As a by product of the second paper we also build Lagrangian numerical schemes for Wasserstein gradient flows such as the porous medium equation. One can interpret this flow as some high friction limit of Euler flows.

Articles:

8. **A Lagrangian scheme *à la Brenier* for the incompressible Euler equations.** *Found Comput Math 18: 835* (2018). https://doi.org/10.1007/s10208-017-9355-y. Gallouët T.O. and Mérigot Q.

9. **Convergence of a Lagrangian discretization for barotropic fluids and porous media flow.** *SIAM Journal on Mathematical Analysis* (2021) https://hal.science/hal-03234144. Gallouët T.O., Mérigot Q., Natale A.

The first paper is a collaboration with Q. Mérigot. It is based on the reinterpretation of Y. Brenier's old ideas and Q. Mérigot's new method that allows to deal numerically with semi-discrete Optimal Transport: a transport between sums of Dirac masses and a smooth measure. It was done when I was a post-doc of Y. Brenier. The

second paper is a collaboration with Q. Mérigot and A. Natale. At the time A. Natale was a post-doc under our supervision.

Incompressible Euler

In 8 we approximate the regular solutions of the incompressible Euler equations by the solution of ODEs on finite-dimensional spaces. This approach combines Arnold's interpretation of the solution of the Euler equations for incompressible and inviscid fluids as geodesics in the space of measure-preserving diffeomorphisms, and an extrinsic approximation of the equations of geodesics due to Brenier. Indeed the empirical measure of a system of particles cannot be uniform. In our scheme, the incompressibility constraint is relaxed by imposing that the Wasserstein distance between the uniform measure and the empirical measure should be small relative to a parameter ϵ. This is enforced in an Hamiltonian fashion, the Wasserstein distance acting as a spring attached to the manifold of measure preserving maps. Using recently developed semi-discrete optimal transport solvers, this approach yields a numerical scheme which is able to handle problems of realistic size in 2D at the time of the paper and by now much larger 3D systems composed of millions of particles. We prove the convergence of this scheme towards regular solutions of the incompressible Euler equations thanks to a relative entropy method. The key arguments allowing to apply a (double) Grönwall argument are the use of optimality, orthogonality properties and the degree of freedom in the pressure term: its mean. We also provide numerical experiments on a few simple test cases in 2D. Many extension of this work are possible. Two ongoing projects are the fluid-structure interactions and incompressible Navier-Stokes equations as our Lagrangian scheme is particularly adapted with the finite volume discretization of the Laplacian.

Barotropic fluids

When expressed in Lagrangian variables, the equations of motion for compressible (barotropic) fluids have the structure of a classical Hamiltonian system in which the potential energy is given by the internal energy of the fluid. The dissipative counterpart of such a system coincides with the porous medium equation, which can be cast in the form of a Wasserstein gradient flow for the same internal energy. Motivated by these related variational structures, we propose a particle method for both problems in which the internal energy is replaced by its Moreau-Yosida regularization in the L^2 sense, which can be efficiently computed as a semi-discrete optimal transport problem in the spirit of what we have done for the incompressible Euler equation. This last equation corresponds to the case where energy is the characteristic function of measure preserving maps. Again using a modulated energy argument which exploits the convexity of the problem in Eulerian variables, we prove quantitative convergence estimates towards smooth solutions. We verify such estimates by means of several numerical tests.

The main strength of these Lagrangian methods is that they are based on the physical energy and a nice geometrical structure for the PDE: either Gradient flows, Euler/Hamiltonian Flows, or Conservative flows where the velocity is given by the rotation of the Wasserstein gradient of the energy ($v = -J\nabla\frac{\delta\mathcal{E}}{\delta\rho}(\rho)$), with J an antisymmetric matrix). In particular we have two extensions in mind: the Keller-Segel model and semi-geostrophic equations, using some recent technics developed by D. Bresch and co-authors and S. Serfaty and co-authors [7, 2] in order to deal with the additional interaction term into the Grönwall arguments.

1.3 Unbalanced Optimal transport, geometry and PDE

Optimal Transport is a powerful tool to compare probability distributions and interpret PDEs with some geometrical structure. However in some applications or PDEs it is natural to consider change of total mass or different mass between the measures. This mass constraint can easily be alleviated with global renormalization but the obtained model will not be able to account for possible local change of mass. Considering this shortcoming [8, 1], it was natural to enrich the model using local change of mass as proposed by three research groups independently in [6, 5, 10, 12]. This definition shares a lot with classical optimal transport with primal, dual, static formulation and importantly a Riemannian submersion. Once again, the work I have contributed to on this subject can be divided in three categories: properties of Unbalanced Optimal Transport, applications to gradient flows for this metric and finally Euler flows and more specifically the counterpart of the Incompressible equation in this framework which is the Camassa-Holm equation.

1.3.1 Unbalanced Optimal transport, geometry and PDE

Articles:

10. **Regularity theory and geometry of unbalanced optimal transport.** Gallouët T.O., Ghezzi R. et Vialard F.X. https://hal.science/hal-03498098v1.

This work is done in collaboration with R.Ghezzi and F.X.Vialard. It is a preprint that will be shortly submitted for publication. Using the dual formulation only, we show that regularity of unbalanced optimal transport also called entropy-transport inherits from the regularity of standard optimal transport. We then provide detailed examples of Riemannian manifolds and costs for which unbalanced optimal transport is regular. Among all entropy-transport formulations the Wasserstein-Fisher-Rao metric, also called Hellinger-Kantorovich, stands out since it admits a dynamic formulation, which extends the Benamou-Brenier formulation of optimal transport. After demonstrating the equivalence between dynamic and static formulations on a closed Riemannian manifold, we prove a polar factorization theorem, similar to the one due to the Brenier-McCann one. As a byproduct, we formulate the Monge-Ampère equation associated with Wasserstein-Fisher-Rao (WFR) metric, which also holds for more general costs. This allows to give a sense to *Brenier's weak variational solutions* for this large class of PDEs composed of a "classical" Monge-Ampère operator combined with lower order non linear terms. This includes for instance the JKO scheme, moment maps, and is a key ingredient for the regularity of Unbalanced Optimal Transport maps. Last, we give explicit links between c-convex functions/c-segment for the cost induced by the WFR metric and c-convex functions/c-segment for the associated cost on the cone space. One of the main corollaries is that weak Ma-Trudinger-Wang condition on the cone implies it for the cost induced by WFR.

1.3.2 Unbalanced gradient flows and general reaction diffusion PDEs

Articles:

11. **A JKO splitting scheme for Kantorovich-Fisher-Rao gradient flows.** *SIAM Journal on Mathematical Analysis, Vol. 49, Issue 2.* (2017) https://arxiv.org/abs/1602.04457. Gallouët T.O. et Monsaingeon L.

12. **An unbalanced optimal transport splitting scheme for general advection-reaction-diffusion problems.** *Journal of Differential Equations ESAIM: Control, Optimisation and Calculus of Variations* (2018) https://hal.science/hal-01508911. Gallouët T.O., Laborde M. and Monsaingeon L.

The first article is written in collaboration with L. Monsaingeon. In this work we set up a splitting variant of the Jordan-Kinderlehrer-Otto scheme in order to handle gradient flows with respect to the Wasserstein-Fisher-Rao metric, defined on the space of positive Radon measure with varying masses. We perform successively a JKO time step for the quadratic Wasserstein/Monge-Kantorovich distance, and then for the Hellinger/Fisher-Rao distance. Exploiting the inf-convolution structure of the metric we show convergence of the whole process for the standard class of energy functionals under suitable compactness assumptions, and investigate in details the case of internal energies. The interest is twofolds: on the one hand, we prove existence of weak solutions for a certain class of reaction-advection-diffusion equations, and on the other hand this process is constructive and well adapted to available numerical solvers. From a technical point of view, this approach has the advantage of avoiding too detailed an examination of the geometry of the WFR space, which is now well known. [13].

Later and with M. Laborde in addition, we extended this work and showed that unbalanced optimal transport provides a convenient framework to handle more general reaction and diffusion processes in a unified metric setting. Using the same strategy of alternating minimizing movement schemes for the Wasserstein distance and for the Fisher-Rao distance, but with a different energy for each step, we prove existence of weak solutions for general scalar reaction-diffusion-advection equations or systems of multiple interacting species like prey-predator systems. We also consider an application to a very degenerate Hele-Shaw diffusion problem involving a Gamma-limit. Moreover we provide some numerical simulations using an ALG2-JKO strategy for the Wasserstein JKO step. This splitting strategy allows to transfer all recent developments on the JKO scheme to the case of reaction-diffusion equations such as Unbalanced gradient flows.

1.3.3 Camassa-Holm

Optimal Transport and Unbalanced Optimal transport costs share the same structure, in particular the existence of a right invariant action leading to a formal Riemannian submersion. In the optimal transport case the geodesic on the isotropy group of this action and for the induced metric are exactly the Incompressible Euler's equations. Y. Brenier used this remark to propose, among other things, the notion of *generalized solutions* for the Incompressible Euler's equation where the initial and final positions are given. The natural question we asked ourselves was: "what is the counterpart to the incompressible Euler's equations in the Unbalanced Optimal Transport framework". This question led to the following two works together with F.X. Vialard and then with F.X. Vialard and our shared postdoc student A. Natale. The counterpart of the Incompressible Euler's equations is identified to the Camassa-Holm equation when $d = 1$, and one of its possible multi-dimensional generalizations when $d > 1$: the geodesic on the group of diffeomorphisms for the $H(\mathrm{div})$ metric.

Articles:

13. **The Camassa-Holm equation as an incompressible Euler equation: a geometric point of view.** *Journal of Differential Equations, Volume 264, Issue 7, Pages 4199-4234.* (2018) https://arxiv.org/abs/1609.04006. Gallouët T.O. and Vialard F.X.

14. **Generalized compressible flows and solutions of the H(div) geodesic problem.** *Archive for Rational Mechanics and Analysis, Springer Verlag* (2020) https://hal.science/hal-01815531v3. Gallouët T.O., Natale A. et Vialard F.X.

The group of diffeomorphisms of a compact manifold endowed with the L^2 metric acting on the space of probability densities gives a unifying framework for the incompressible Euler equation and the theory of optimal mass transport. In 13, we show a similar relation between this unbalanced optimal transport problem and the $H(\mathrm{div})$ right-invariant metric on the group of diffeomorphisms, which corresponds to the Camassa-Holm equation in one dimension. It leads us to study this geodesic problem on the group of diffeomorphisms, equipped with the $H(\mathrm{div})$ metric. Geometrically, we present an isometric embedding of the group of diffeomorphisms endowed with this right-invariant metric in the automorphisms group of the fiber bundle of half densities endowed with an L^2 type of cone metric. This point of view has three applications: (1) We interpret solutions to the Camassa-Holm equation and one of its generalization in higher dimension as particular solutions of the incompressible Euler equation on the plane for a radial density which has a singularity within the origin. This correspondence can be introduced via a sort of Madelung transform. More precisely on S^1 it gives that solutions to the standard Camassa-Holm thus give radially 1-homogeneous solutions of the incompressible Euler equation on $\mathbb{R}^2$ which preserves a radial density that has a singularity at 0. (2) We generalize a result of Khesin et al. in [9] by computing the curvature of the group as a Riemannian submanifold. (3) Generalizing a result of Brenier to the case of Riemannian manifolds, which states that solutions of the incompressible Euler equations are length minimizing geodesics for sufficiently short times. We prove a similar result for the Camassa-Holm equation: smooth solutions of the Euler-Arnold equation for the $H(\mathrm{div})$ right-invariant metric are length minimizing geodesics for sufficiently short times.

We then pursue the analogy with Brenier's work for the Incompressible Euler's equations in 14. In particular we propose a relaxation *à la Brenier* of this problem, in which solutions are represented as probability measures on the space of continuous paths on the cone over the domain. We call the minimizers of such a relaxation *generalized solutions*. This approach allows us to obtain several results on the $H(\mathrm{div})$ geodesic problem. In particular, we show that: if the base space is convex, smooth $H(\mathrm{div})$ geodesics are globally length-minimizing for short times and in any dimension. This result generalizes the one in 13, which was only valid on the unit circle and was local otherwise. On the torus $S^1 \times S^1$, we show that there exists $h \in \mathrm{Diff}(S^1 \times S^1)$ such that the infimum of the action problem, that defined the generalized geodesics of Camassa-Holm equation, cannot be attained by any smooth flow. This result is within the spirit of Shnirelman's work on Incompressible Euler's equations [17]. On the contrary, for the same h there exists a generalized solution that arises as the limit of a minimizing sequence of smooth flows.There exists a unique pressure field in the sense of distribution associated with generalized solutions. To the best of the authors' knowledge, the pressure field we consider is a variable that has not been studied before in the literature on the Camassa-Holm equation or the $H(\mathrm{div})$ geodesic problem. It appears however as a natural variable in the generalized setting and deserves a closer look from a more conventional PDE perspective in order to obtain

a priori estimates. Finally, we propose a numerical scheme to construct generalized solutions on the cone and present some numerical results illustrating the relation between the generalized Camassa-Holm and incompressible Euler solutions.

Bibliography

[1] Benamou, Jean-David. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: M2AN*, 37(5):851–868, 2003.

[2] Didier Bresch, Pierre-Emmanuel Jabin, and Zhenfu Wang. Modulated free energy and mean field limit, 2019.

[3] Guillaume Carlier. Remarks on toland's duality, convexity constraint and optimal transport. 2008.

[4] Yongxin Chen, Giovanni Conforti, and Tryphon T. Georgiou. Measure-valued spline curves: an optimal transport viewpoint. 01 2018.

[5] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced Optimal Transport: Geometry and Kantorovich Formulation. *ArXiv e-prints*, August 2015.

[6] L. Chizat, B. Schmitzer, G. Peyré, and F.-X. Vialard. An Interpolating Distance between Optimal Transport and Fisher-Rao. *Found. Comp. Math.*, 2016.

[7] Antonin Chodron de Courcelle, Sylvia Serfaty, and Matthew Rosenzweig. Announced resultsin seminar. *In preparation*.

[8] Tryphon T. Georgiou, Johan Karlsson, and Mir Shahrouz Takyar. Metrics for power spectra: An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, 2009.

[9] B. Khesin, J. Lenells, G. Misiolek, and S. C. Preston. Curvatures of Sobolev metrics on diffeomorphism groups. *Pure and Applied Mathematics Quarterly*, 9(2):291 – 332, 2013.

[10] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite Radon measures. *Advances in Differential Equations*, 21(11/12):1117 – 1164, 2016.

[11] 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.

[12] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.

[13] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Fine properties of geodesics and geodesic λ-convexity for the hellinger-kantorovich distance-convexity for the hellinger-kantorovich distance. 08 2022.

[14] Bruno Lévy. Partial optimal transport for a constant-volume lagrangian mesh with free boundaries. *Journal of Computational Physics*, 451:110838, 2022.

[15] 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.

[16] Simon Plazotta. A bdf2-approach for the non-linear fokker-planck equation, 2018.

[17] Alexander I Shnirelman. Generalized fluid flows, their approximation and applications. *Geometric & Functional Analysis GAFA*, 4(5):586–620, 1994.

Chapter 2

Optimal transport geometry and PDE

2.1 Optimal Transport

Articles:

- **Second order models for optimal transport and cubic splines on the Wasserstein space.** *Foundations of Computational Mathematics, Springer Verlag* (2019) https://hal.science/hal-01682107v2 J.D. Benamou, Gallouët T.O. et Vialard F.X.
- **From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows.** *Under minor revision Mathematics of Computations* (2022) Gallouët T.O., Natale A. et Todeschi. G https://hal.science/hal-03f790981v2

Collaborators: The first paper has been done with two collaborators I got when I arrived at Inria Paris: J.D. Benamou (DR Inria Paris) and F.X. Vialard (MCF Dauphine then Professeur Paris Est). The second paper is done with A. Natale (former post doc student of mine now CR at Inria Lille) and G. Todeschi (former PhD student of mine now Post-doc.)

Main contributions:

Cubic Splines:

- We propose a notion of relaxed cubic splines in the Wasserstein Space.
- We propose and implement three different numerical methods in order to compute these cubic splines. Based on the one hand on entropic regularization and on the other hand on Semi discrete Optimal Transport.

Wasserstein extrapolation:

- We propose different notions of Wasserstein extrapolation for the quadratic cost. We prove the well-posedness of these notions.
- We propose different numerical scheme in order to approximate these definitions. We prove the convergence for a large class of scheme but not the one used in the numerical section.
- We implement one of these scheme for which we give numerical evidence of convergence.

Research directions: With A. Natale and G. Todeschi we are continuing our work on the extrapolation of Wasserstein geodesics especially for the metric extrapolation which seems to us to have the richest structure. We explore the different equivalent definitions (primal, dual,..) and aim to fill the gap left in the previous paper on the numerical implementation for this definition of Wasserstein extrapolation. On e application would be to build another variational order two in time numerical scheme for Wasserstein gradient flows see Section 2.2.2 for more details.

SECOND ORDER MODELS FOR OPTIMAL TRANSPORT AND CUBIC SPLINES ON THE WASSERSTEIN SPACE

JEAN-DAVID BENAMOU, THOMAS O. GALLOUËT, AND FRANÇOIS-XAVIER VIALARD

Abstract. On the space of probability densities, we extend the Wasserstein geodesics to the case of higher-order interpolation such as cubic spline interpolation. After presenting the natural extension of cubic splines to the Wasserstein space, we propose a simpler approach based on the relaxation of the variational problem on the path space. We explore two different numerical approaches, one based on multi-marginal optimal transport and entropic regularization and the other based on semi-discrete optimal transport.

1. Introduction

We propose a variational method to generalize cubic splines on the space of densities using multimarginal optimal transport. In short, the proposed method consists in minimizing, on the space of measures on the path space, under marginal constraints, the norm squared of the acceleration. In this setting, we show that two numerical approaches, classical in optimal transportation can be applied. One is based on entropic regularization and the Sinkhorn Algorithm, the other relies on the Semi-Discrete formulation of Optimal Transportation and the computation of Laguerre cells, a classical problem in computationnal geometry. We showcase our methodology on 1D and 2D data.

In the past few years, higher-order interpolations methods have been investigated for applications in computer vision or medical imaging, for time-sequence interpolation or regression. The most usual setting is when data are modeled as shapes, which can be understood as objects embedded in the Euclidean space with no preferred parametrization: space of unparametrized curves or surfaces, or images are some of the most important examples. These examples are infinite dimensional but the finite dimensional case of a Riemannian manifold was interesting for camera motion interpolation as first introduced in [22] and further developed in [6, 8]. Motivated by different applications, the problem of interpolation between two shapes is usually treated via the use of a Riemannian metric on the space of shapes and computing a geodesic between the two shapes. From a mathematical point of view, shape spaces are often infinite dimensional and thus, non-trivial analytical questions arise such as existence of minimizing geodesics or global well-posedness of the initial value problem associated with geodesics. A finite dimensional approximation is still possible such as in [29], in which spline interpolation is proposed for a diffeomorphic group action on a finite dimensional manifold. It has been extended for invariant higher-order lagrangians in [11, 12] on a group, still finite dimensional. A numerical implementation of the variational and shooting splines has been developed in [26] with applications to medical imaging. The question of existence of an extremum is not addressed in these publications. An attempt is given in [28] where the exact relaxation of the problem is computed in the case of the group of diffeomorphisms of the unit interval. In a similar direction, in [13], the authors discuss the convergence of the discretization of cubic splines in some particular infinite dimensional Riemannian context on the space of shapes.

As a shape space, we are interested in this article in probability measures endowed with the Wasserstein metric. Since the Wasserstein metric shares some similarities with a Riemannian metric on this space of probability densities, it is natural to study further higher-order models in this context. Our motivation is to answer the following practical question of the extension of cubic splines to the Wasserstein space and their numerical computation.

We present in Section 2 the notion of cubic splines on a Riemannian manifold and detail its variational formulation in Hamiltonian coordinates. We then discuss independently in Section 3 a

geometric approach to the Wasserstein space that will be useful for the introduction of our proposed method detailed in Section 4. Finally in Sections 5 we present the numerical entropic relaxation method and an alternative numerical method based on semi-discrete optimal transport. The reader not interested in geometric interpretation can skip directly to Section 4.

To the best of our knowledge, this question has not been yet addressed in the literature on optimal transport until very recently in two independant and simultaneous preprints : [31] and [14] (this paper). Both work share the same idea of relaxing the cubic spline formulation in the space of measure using multi-marginal optimal transport. Our paper however explores a larger hierarchy of models and several numerical methods.

2. Cubic splines on Riemannian manifolds

In this section, we present Riemannian cubics, which are the extension of variational splines to a Riemannian manifold (M, g) where g is the Riemannian metric. Variational cubic splines on a Riemannian manifold are the minimizers of the acceleration; that is, denoting $\frac{D}{Dt}$ the covariant derivative, minimization on the set of curves $x : [0, T] \to M$ of the functional

$$\mathcal{E}(x) = \int_0^1 g(x)\left(\frac{D}{Dt}\dot{x}, \frac{D}{Dt}\dot{x}\right) \mathrm{d}t\,, \tag{2.1}$$

subject to constraints on the path such as constraints on the tangent space, $(x(t_i), \dot{x}(t_i))$ are prescribed for a collection of times $t_i \in [0, 1]$, or constraints on the positions such as $x(t_i) = x_i$.

Under mild conditions on the constraints, if M is complete, minimizers exist, for instance in the case of constraints on the tangent space mentioned above. A pathological case where minimizers might not exist is when the initial speed is not prescribed. Consider for instance the two dimensional torus, where lines of irrational slopes are dense, it is possible to show that for any collection of points which do not lie on a line, the infimum of $\mathcal{E}$ is 0 while it is never reached, see [13]. The Euler-Lagrange equation associated to the functional $\mathcal{E}$ is

$$\frac{D^3}{Dt^3}\dot{x} - R\left(\dot{x}, \frac{D}{Dt}\dot{x}\right)\dot{x} = 0\,, \tag{2.2}$$

where R is the curvature tensor of the Riemannian manifold M. Note that this equation is similar to a Jacobi field equation.

We now formulate the variational problem in coordinates. In a coordinate chart around a point $x(t) \in M$, the geodesic equations are given by

$$\frac{D}{Dt}\dot{x} = \ddot{x} + \Gamma(x)(\dot{x}, \dot{x}) = 0\,, \tag{2.3}$$

where Γ is a short notation for the Christoffel symbols associated with the Levi-Civita connection. It is a second-order differential equation which is conveniently written as a first-order differential equation, via the Hamiltonian formulation. Again in local coordinates on T^*M the cotangent bundle of M, the geodesic equation can be written as

$$\begin{cases} \dot{p} + \partial_x H = 0 \\ \dot{x} - \partial_p H = 0\,, \end{cases} \tag{2.4}$$

where $H(x, p) = \frac{1}{2}g(x)^{-1}(p, p)$. Note that, the ODE (2.3) can be obtained from the Hamiltonian system using $\dot{x} = g(x)^{-1}p$. From these two equivalent formulations (2.3) and (2.4), it can be shown that $g^{-1}(x)(\dot{p} + \partial_x H) = \frac{D}{Dt}\dot{x}$. Therefore, it proves that the variational spline problem can be rewritten in Hamiltonian coordinates as follows

$$\inf_u \int_0^1 g(x)^{-1}(a, a)\,\mathrm{d}t\,,$$

under the constraint

$$\begin{cases} \dot{x} - g(x)^{-1}p = 0 \\ \dot{p} + \partial_x H(x, p) = a\,, \end{cases}$$

with initial conditions $x(0) = x_0$ and $p(0) = p_0$. It is natural to ask whether such variational problems carry over in infinite dimensional situations such as the Wasserstein space, which will be discussed in the rest of the paper.

3. A formal application of spline interpolation to the Wasserstein space

It is well known that the Hamiltonian formulation of geodesics on the Wasserstein space, define over a riemannian manifold M, are

$$\begin{cases} \dot\rho + \nabla\cdot(\rho\nabla\phi) = 0 \\ \dot\phi + \frac{1}{2}|\nabla\phi|^2 = 0\,, \end{cases} \tag{3.1}$$

where $\rho : M \mapsto \mathbb{R}_{\geq 0}$ and $\phi : M \mapsto \mathbb{R}$ *implicitly time dependant* are respectively a probability density and a function. Note that these equations are valid when working with smooth densities. The Hamiltonian is the following,

$$H(\rho,\phi) = \frac{1}{2}\int_M |\nabla\phi|^2\rho\, d\mu_0\,, \tag{3.2}$$

where μ_0 is a reference measure on M.

Remark 1. Taking the gradient of the equation governing ϕ, and denoting $v = \nabla\phi$, we get Burger's equation:

$$\dot v + (v,\nabla)v = 0\,, \tag{3.3}$$

where in coordinates, the operator (v,∇) is defined as $(v,\nabla)w \doteq \sum_{i=1}^n v_i\nabla w_i$ where v, w are vector fields and n is the dimension of the M. In Lagrangian coordinates, this equation implies that

$$\ddot\varphi = 0\,, \tag{3.4}$$

where $\varphi(t) : M \mapsto M$ is the Lagrangian flow associated with v ($\dot\varphi = v\circ\varphi$), which is well-defined under sufficient regularity conditions.

Remark 2. For the Wasserstein case, the operator is given by $g(\rho)^{-1}\phi = -\nabla\cdot[\rho\nabla\phi]$ so that the (formal) computation of the covariant derivative $\frac{D}{Dt}\dot\rho$ on the Wasserstein space is:

$$\frac{D}{Dt}\dot\rho = -\nabla\cdot\left[\rho\,(v + (v,\nabla)v)\right], \tag{3.5}$$

where $v = \nabla\phi$ is the horizontal lift associated with $\dot\rho$, that is $\dot\rho + \nabla\cdot(\rho\nabla\phi) = 0$. This result is proven rigorously in [18].

From a control viewpoint, we aim at minimizing $\frac{1}{2}\int_0^1 H(\rho,a)\,dt$ for the control system:

$$\begin{cases} \dot\rho + \nabla\cdot(\rho\nabla\phi) = 0 \\ \dot\phi + \frac{1}{2}|\nabla\phi|^2 = a\,, \end{cases} \tag{3.6}$$

where a is a time dependent function defined on M. Alternatively, in terms of the variables (ρ,ϕ), this amounts to minimize

$$\int_0^1\int_M |\nabla[\dot\phi + \frac{1}{2}|\nabla\phi|^2]|^2\rho\ \mathrm{d}\mu_0\ \mathrm{d}t\,, \tag{3.7}$$

under the continuity equation constraint $\dot\rho + \nabla\cdot(\rho\nabla\phi) = 0$. It is a nonconvex optimization problem in the couple (ρ,ϕ). The key issue here is that the variational problem itself is a priori not well-posed since our formulation is valid in a smooth setting and to make it rigorous on the space of measures, the tight relaxation of this problem is needed. However, we do not address this issue in our work and in the next section we turn our attention to a simple relaxation of the problem which is probably not tight.

4. A hierarchy of relaxed models

4.1. **Context.** We recall the classical optimal transport setting. We have the following well known equivalence [23, 30]

$$\begin{aligned} W_2^2(\rho_0,\rho_1) &= \inf_{\varphi} \int_0^1 \int_M |\dot\varphi|^2 \,\mathrm{d}\mu_0 \,\mathrm{d}t = \inf_{\rho,v} \int_0^1 \int_M |v|^2 \,\mathrm{d}\rho \,\mathrm{d}t \\ &= \inf_{\rho} \int_0^1 \inf_{v} \int_M |v|^2 \,\mathrm{d}\rho \,\mathrm{d}t = \inf_{\rho,\nabla\phi} \int_0^1 \int_M |\nabla\phi|^2 \,\mathrm{d}\rho \,\mathrm{d}t \end{aligned} \tag{4.1}$$

Under constraints that

$$[\varphi(t)]_*\mu_0 = \rho(t) \text{ for } t = 0, 1$$

($[\varphi(t)]_*\mu_0$ is the image measure of μ_0 : $\int_M f(y)\,\mathrm{d}[\varphi(t)]_*\mu_0(y) = \int f(T(x))\,\mathrm{d}\mu(x)$ for every measurable function $f : M \to \mathbb{R}$)
and the continuity equation

$$\dot\rho + \nabla\cdot(\rho v) = \dot\rho + \nabla\cdot(\rho\nabla\phi) = 0$$

with fixed initial and final conditions

$$\rho(0) = \rho_0 \text{ and } \rho(1) = \rho_1.$$

Moreover, geodesics in the space of densities for the Wasserstein metric are given by $[\varphi(t)]_*\mu_0 = \rho(t)$ and the associated displacement maps satisfy $v\circ\varphi = \dot\varphi$.

The last equality in (4.1) exactly says that the infimum $\inf_{v(t)} \int_M |v(t)|^2 \,\mathrm{d}\rho(t)$ among all $v(t)$ satisfying the continuity equation at each time t is achieved when $v(t)$ is a gradient. This property is a consequence of a Riemannian submersion and $\nabla\phi$ is called the horizontal lift of $\dot\rho$. It is this last formulation that formally gives a Riemannian structure on the space of probability measures. See the remark 1 below for more details on the geometrical structure.

For higher-order variational problems, e.g. the minimization of the acceleration, the reduction in the last inequality does not holds true in general, even if the Riemannian submersion structure is present as shown in [12]. It means in the case of acceleration that, a priori, with the same constraint as for (4.1) :

$$\begin{aligned} \inf_{\varphi} \int_0^1 \int_M |\ddot\varphi|^2 \,\mathrm{d}\mu_0 \,\mathrm{d}t &= \inf_{\rho,v} \int_0^1 \int_M |\dot v + (v,\nabla)v|^2 \,\mathrm{d}\rho \,\mathrm{d}t \\ &\neq \inf_{\rho,\nabla\phi} \int_0^1 \int_M |\dot\phi + (\nabla\phi,\nabla)\nabla\phi|^2 \,\mathrm{d}\rho \,\mathrm{d}t, \end{aligned} \tag{4.2}$$

where we have used that $\ddot\varphi = \dot v\circ\varphi + (v\circ\varphi,\nabla)v\circ\varphi$.

Remark 1. *From a geometrical point of view,* (4.1) *says the Wasserstein space can be seen, at least formally, as a homogeneous space as described in* [15, Appendix 5] *and originally in* [23]. *Consider the group of (smooth) diffeomorphisms of M a closed manifold,* $\mathrm{Diff}(M)$*, and the space of (smooth) probability densities* $\mathrm{Dens}(M)$*. The space of densities is endowed with a* $\mathrm{Diff}(M)$ *action defined by the pushforward, that is to a given $\varphi \in \mathrm{Diff}(M)$ and $\rho \in \mathrm{Dens}(M)$, the pushforward of ρ by φ is* $\mathrm{Jac}(\varphi^{-1})\rho\circ\varphi^{-1}$*. By Moser's lemma, this action is transitive, thus making the space of densities as a homogeneous space. More importantly, there exists a compatible Riemannian structure between* $\mathrm{Diff}(M)$ *and* $\mathrm{Dens}(M)$*. Once having chosen a reference density μ_0, the $L^2(M,\mu_0)$ metric on the diffeomorphism group descends to the Wasserstein L^2 metric on the space of densities, or in other words, the pushforward action $\varphi \mapsto \varphi_*\mu_0$ is a Riemannian submersion. An important property of Riemannian submersion is that geodesics on* $\mathrm{Dens}(M)$ *are in correspondence with geodesics on the group, given by horizontal lift. This property is actually contained in Brenier's polar factorization theorem, which shows that the horizontal lift is the gradient of a convex function.*

4.2. **The Monge formulation.** In Section 3 we used the formal Riemannian structure on the set of probability measure to define an intrinsic notion of splines, (3.7) is indeed the RHS of inequality (4.2). In this section we propose a simpler alternative definition of Wasserstein splines based on the LHS of inequality (4.2).

Definition 1 (Monge formulation)**.** Let $0 = t_0 < \ldots < t_n = 1$, $n \geq 2$ and $\rho_1, \ldots, \rho_n$ be n probability measures on M.

Minimize, among time dependent maps $\varphi(t) : M \mapsto M$,

$$\int_0^1 \int_M |\ddot{\varphi}|^2 \, \mathrm{d}\mu_0 \, \mathrm{d}t \,, \tag{4.3}$$

under the marginal constraints $\varphi(t_i)_* \mu_0 = \rho_i$. This minimizing problem is denoted by (MS).

It is a Monge formulation of the variational problem, similar to standard optimal transport. On a Riemannian manifold M, the notation $\ddot{\varphi}$ stands for $\frac{D}{Dt}\dot{\varphi}$. By the change of variable with the map φ, the problem can be written in Eulerian coordinates, that is using the vector field associated with the Lagrangian map φ, $\partial_t \varphi = v \circ \varphi$, one aims at minimizing for (ρ, u)

$$\int_0^1 \int_M |u|^2 \rho \, \mathrm{d}\mu_0 \, \mathrm{d}t \tag{4.4}$$

under the constraints

$$\begin{cases} \dot{\rho} + \mathrm{div}(\rho v) = 0 \\ \dot{v} + (v, \nabla) v = u \,, \end{cases} \tag{4.5}$$

with the marginals constraints $\rho(t_i) = \rho_i$.

Remark 2. *Remark that formally when $v = \nabla \phi$, this new model reduces to the formulation* (3.7)*. Therefore, it justifies the fact that Problem* (4.3) *is a relaxation of* (3.7)*. However, as already mentioned, this relaxation is probably not tight.*

Another formal geometric argument in the direction of proving that the two formulations are different is that the Wasserstein space has nonnegative curvature if the underlying space M has nonnegative curvature, but the space of maps in the Euclidean space is flat. Therefore, the two Euler-Lagrange equations (2.2) *lead to a different evolution equations: for instance, if M is the Euclidean space then the Euler-Lagrange equation for the second model is simply $\dddot{\varphi} = 0$, which is a priori different from the splines Euler-Lagrange equation in the Wasserstein case.*

4.3. **The Kantorovich relaxation.** Since, as is well-known in standard optimal transport, the Monge formulation is not well-posed for general given margins $\rho_1, \ldots, \rho_n$, we propose instead to solve yet another relaxation of the problem on the space of curves which takes the form:

Definition 2 (Kantorovich relaxation)**.** Let $0 = t_1 < \ldots < t_n = 1$, $n \geq 3$ and $\rho_1, \ldots, \rho_n$ be n probability measures on M.

Minimize on the space of probability measures on the path space $H^2([0,1], M)$ denoted by $\mathcal{H}$ in short,

$$\min_{\mu} \int_{\mathcal{H}} |\ddot{x}|^2 \, \mathrm{d}\mu(x) \,, \tag{4.6}$$

which is a linear functional of $\mathrm{d}\mu$. The curves of densities is given by its marginals in time

$$t \mapsto \rho(t)\mu_0 := [e_t]_*(\mu) \,, \tag{4.7}$$

e_t is the evaluation function at time t : if $\gamma \in H^2([0,1], M) \subset C^0([0,1], M)$ then $e_t(\gamma) = \gamma(t, .) \in M$. The notation $[e_t]_* \mu$ is the image measure by the map e_t defined by duality :
$\int_M f(y) \, \mathrm{d}[e_t]_* \mu(y) = \int_{\mathcal{H}} f(e_t(x)) \, \mathrm{d}\mu(x)$ for every measurable function $f : M \to \mathbb{R}$. Note that x is a path on $[0,1] \times M$ while y is a point on M.

With these notations, the marginal constraint at given time t_i are

$$[e_t]_*(\mu) = \rho_i \, \mu_0 \,. \tag{4.8}$$

By standard arguments, the Kantorovich relaxation admits minimizers under general hypothesis on the manifold M, which we do not detail here. It is straightforward to check that existence of minimizers holds when $M = \mathbb{R}^d$.

As expected, the Kantorovich formulation is the relaxation of the Monge formulation in Definition 1.

Theorem 1. *Let $M = \mathbb{R}^d$, $0 = t_1 < \ldots < t_n = 1$, $n \geq 3$ and $\rho_1, \ldots, \rho_n \in$ be n probability measures on $\mathbb{R}^d$ with compact support and ρ_1 being atomless. Then, under the constraints* (4.8)*, the infimums of the variational problem* (4.3) *and* (4.6) *coincide, moreover, the infimum is attained for the latter.*

Proof. See the proof of a more general result in Appendix A. □

First we remark that we can reformulate both the Monge and Kantorovich problems on the set of cubic splines. It is the purpose of the following lemmas and corollaries, whose proofs are straightforward.

Definition 3 (Cubic interpolant)**.** Let $(x_1, \ldots, x_n) \in \mathbb{R}^d$ be n given points and $(t_1 < \ldots < t_n)$ be n timepoints. There exists a unique cubic spline minimizing the acceleration of the curve $x(t)$ such that $x(t_i) = x_i$. This unique curve is called cubic interpolant and is denoted by $c_{x_1,\ldots,x_n}$, depending implicitly on the timepoints.

Lemma 2. *When the supports of the measures ρ_i are compact on $\mathbb{R}^d$, the support of every minimizing μ in Definition 2 is included in the set the cubic interpolants $c_{x_1,\ldots,x_n}$ for $(x_1, \ldots, x_n) \in \mathrm{Supp}(\rho_1) \times \ldots \times \mathrm{Supp}(\rho_n)$.*

Proof. The constraints are the marginal constraints $[e_{t_i}]_*(\mu) = \rho_i$ for $i \geq 3$ which implies that set of paths charged by an optimal measures satisfies $x(t_i) \in \mathrm{Supp}(\rho_i)$. In particular, any path in this set can be replaced by its minimal spline energy, the cubic interpolant $c_{x_1,\ldots,x_n}$. □

Corollary 3. *As a consequence, the set of paths charged by an optimal plan are uniformly C^2 and for every smooth function $\eta : \mathbb{R}^d \mapsto \mathbb{R}$ with compact support, the map $t \mapsto \langle \mu(t), \eta \rangle$ is C^2.*

Proof. The set of cubic interpolants is compact since the map $(x_1, \ldots, x_n) \mapsto c_{x_1,\ldots,x_n}$ is continuous from $\mathbb{R}^{dn}$ to the space of C^2 fonctions (solution of an invertible linear system) and $\mathrm{Supp}(\rho_i)$ are compact. Therefore, the set of maps are uniformly C^1. The last point follows directly. □

Corollary 4. *The Kantorovich problem in Definition 2 on $\mathbb{R}^d$ reduces to a multimarginal optimal transport problem, as follows, let $c(x_1, \ldots, x_n)$ be the continuous cost of the cubic interpolant at times $t_1, \ldots, t_n$, the minimization of* (4.6) *reduces to the minimization of*

$$\int_{M^n} c(x_1, \ldots, x_n)\, \mathrm{d}\pi(x_1, \ldots, x_n) \qquad (K) \tag{4.9}$$

on the space of probability measures $\pi \in \mathcal{P}(M^n)$ and under the marginal constraints $(p_i)_*(\pi) = \rho_i$ where p_i is the projection of the i^{th} factor.

Proof. Direct consequence of Lemma 2. □

Similarly

Corollary 5. *The Monge problem in Definition 1 on $\mathbb{R}^d$ reduces to a Monge multimarginal optimal transport problem, as follows, let $c(x_1, \ldots, x_n)$ be the continuous cost of the cubic interpolant at times $t_1, \ldots, t_n$, the minimization of* (4.3) *reduces to the minimization of*

$$\int_M c\left(x, \varphi(t_1, x), \ldots, \varphi(t_n, x)\right)\, \mathrm{d}\mu_0(x), \tag{4.10}$$

on the space of path $\varphi \in C^2([0,1], M)$ (or even cubic splines) and under the marginal constraints $(\varphi(t_i))_*\mu_0 = \rho_i$.

The dual formulation of the minimization problem (K) is also well known [16, Theorem 2.1]

Definition 4 (Kantorovich dual problem (KP))**.** Let $\mathcal{Q} = \left\{\phi_i \in L^1(\rho_i\,\mu_0)\,, i = 1..n\right\}$ be the space of integrable n-uplet. Maximize on $\mathcal{Q}$

$$\sum_{i=1}^{n} \int_M \phi_i \rho_i\,\mu_0, \text{ under the constraint } \sum_{i=1}^{n} \phi_i(x_i) \leq c(x_1, ..., x_n). \tag{4.11}$$

And the following duality results holds true:

Proposition 6. *There exists a n-uplet $(\phi_i)_{i=1..n} \in \mathcal{Q}$ optimal for (KP). Moreover (K)=(KP) and for any π optimal in (4.9) there holds $\sum_1^n \phi_i(x_i) = c(x_1, ..., x_n)$, π almost everywhere.*

A natural question is whether the solution of the Kantorovich problem (K) is admissible in the Monge formulation (MS) (Definition 1). With the formulation reduced above the spline, given by (4.9) and (4.10), one can try to apply existing theory to answer to this question, see [16, 24] and references therein for precise criterion. However our cost does not satisfy any of those known criterion. In fact, we have the following result which proves that the relaxation to plans are necessary even in the context of Theorem 1.

Proposition 7. *(Counter Example) Given the three-marginals problems of minimizing the acceleration, there exist data (ρ_0, ρ_1, ρ_2) such that ρ_0 is atomless and such that the solution of (K) is not a (measurable) Monge map.*

Proof. Consider $\rho_0(x) = \mathbf{1}_{[-1,1]}$ and the Dirac masses $a = \delta_1$ and $b = \delta_{-1}$ and the maps T_a, T_b that respectively pushforward ρ_0 onto a and b. These maps are uniquely determined and affine. Consider now $\rho_2 = \frac{1}{2}(T_a)_*\rho_0 + \frac{1}{2}(T_b)_*\rho_0 = \frac{a}{2} + \frac{b}{2}$. Then, introducing $(T^{1/2}) = \frac{1}{2}(\mathrm{Id} + T)$, we consider $\rho_1 = \frac{1}{2}(T_a^{1/2})_*\rho_0 + \frac{1}{2}(T_b^{1/2})_*\rho_0$, note that it is equal to ρ_0 since the maps $T_{a,b}^{1/2}$ are affine.

By construction, the minimization of the acceleration for (ρ_0, ρ_1, ρ_2) is null since it is a mixture of plans supported by straight lines. If there existed an optimal Monge solution it is necessarily supported by only one map denoted by T and since the cost is null, the map at time $1/2$ is necessarily $T^{1/2}$ defined above. The preimage of 1 (resp. -1) by T is a measurable set A (resp. B). Then, necessarily, $\rho_1 = (T^{1/2})_*\chi_A + (T^{1/2})_*\chi_B$, and in fact, $T_{|A} = T_a$ and $T_{|B} = T_b$ (since the image of the map is known). Therefore, we have $\rho_1 = 2\chi_A \circ (T_a^{1/2})^{-1} + 2\chi_B \circ (T_b^{1/2})^{-1}$ which is not equal to the uniform Lebesgue measure on $[-1, 1]$. □

Remark 3. *It is an open question to prove or disprove a similar result when the final density ρ_2 is atomless. The counterexample explained above strongly uses the fact that the final density is a sum of Dirac masses and it might not be robust when replacing the final density by a uniform density on a small interval.*

4.4. **The corresponding interpolation problem on the tangent space.** The relaxed problem on the space of curves can be used to define variational interpolation problem on the phase space, or more precisely on the tangent space TM. Since the space $H^2([0,T], M)$ is contained in $C^1([0,T], M)$, one can formulate the optimal transport problem on phase space (identified with the tangent space) for the acceleration cost.

Definition 5 (Optimal transport on phase space)**.** Let $\bar{\rho_0}, \bar{\rho_1}$ be two probability measures on TM. Minimize on the space of probability measures on $\mathcal{H}$,

$$\min_{\mu} \int_{\mathcal{H}} |\ddot{x}|^2 \,\mathrm{d}\mu(x)\,, \tag{4.12}$$

which is a linear functional of μ under the marginal constraints

$$[j_0]_*(\mu) = \bar{\rho_0}\,, \text{ and } [j_1]_*(\mu) = \bar{\rho_1}\,, \tag{4.13}$$

where $j_t : H^2([0,T], M) \to TM$ is defined by $j_t(x) = (x(t), \dot{x}(t))$.

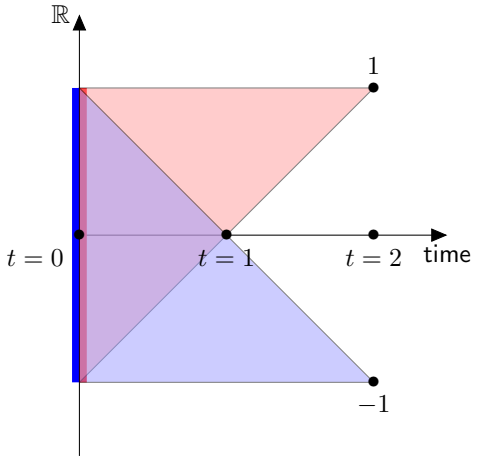


FIGURE 1. The inital density at time 0 is described with a mixture of two densities colored in red and blue which are evolving indepently along straight lines in time. The blue density is mapped onto -1 and the red density is mapped onto 1. The acceleration cost is null and the proof of Proposition 7 shows that it is not possible to reproduce the density at time $1/2$ by a map.

Proposition 8 (Optimal interpolation on phase space)**.** *The support of every optimal solution is contained in the set of cubic splines interpolating between* $(x,v) \in \mathrm{Supp}(\bar{\rho_0})$ *and* $(y,w) \in \mathrm{Supp}(\bar{\rho_1})$*. Moreover if* $M = \mathbb{R}^d$ *and if* $\bar{\rho_0}$ *has density with respect to the Lebesgue measure, then the unique solution to Problem* (4.12) *is characterized by a map* $\varphi : TM \mapsto TM$*.*

Remark that the optimal solution in the last part of Proposition 8 provides an interpolation on the phase space using $[j_t]_*(\mu)$.

Proof. The proof of the first part is similar to Lemma 2 and the second part follows by application of Brenier's theorem since the total cost of the cubic splines between (x,v) and (y,w) can be explicitly computed as

$$c_{ph}((x,v),(y,w)) = 12|x-y|^2 + 4(|v|^2 + |w|^2 + \langle v,w\rangle + 3\langle v+w, x-y\rangle) \tag{4.14}$$

and satisfies the twisted condition, so [30, Theorem 10.28] applies. □

Note that this problem is very different from using the Wasserstein distance on $\mathcal{P}(TM)$ where the tangent space TM is endowed with the direct product metric. Indeed, the cost c_{ph} does not vanish on the diagonal $(x,v) = (y,v)$ contrarily to the quadratic cost on TM.

Interestingly, let us remark that the multimarginal problem can be recast as the minimization problem on $\Pi \in \mathcal{P}(\underbrace{TM \times \ldots \times TM}_{\text{n times}})$, denoting $\Pi_{t_i,t_{i+1}}$ the pushforward on $TM \times TM$ at times (t_i, t_{i+1}),

$$\min_{\pi} \sum_{i=1}^{n-1} \langle \Pi_{t_i,t_{i+1}}, c_{ph}((x_i,v_i),(x_{i+1},v_{i+1}))\rangle \tag{4.15}$$

under the constraints that $[e_{t_i}]_*(\Pi_{i,i+1}) = \rho_i$. From the numerical point of view, this rewriting might be useful since the cost used on the multimarginal problem is now separable in time. This relaxation to the tangent space is used in the semidiscrete algorithm in Section 5.3.1. Obviously, up to the minimization on the variables v_i, we retrieve the minimization problem (K) since one has a cost c which is defined on M^n

$$c(x_0,\ldots,x_n) = \min_{v_0,\ldots,v_n} \sum_{i=1}^{n-1} c_{ph}((x_i,v_i),(x_{i+1},v_{i+1})) \tag{4.16}$$

where the index i runs over the marginals.

5. Numerical Study

We have discussed several variational relaxation of the classical definition of splines, applied to the Wasserstein space of densities. At least two different numerical techniques from Optimal Transportation can be used in this setting. We apply the Entropic regularisation and Sinkhorn (briefly recalled in appendix B first to a simple Hermite interpolation problem (section 5.1) and then in section to the multimarginal problem (4.9). In section 5.3, we use the semi-discrete Optimal Transportation approach in the spirit of [21] directly to problem (4.6) without the time discretisation in (4.9).

5.1. **Hermite interpolation.** In this section, we are interested in the problem of interpolation on the phase space described in the previous. The marginals $[e_t]_*(\mu)$ are densities defined on the tangent space TM. If we only specify the marginals at time 0 and 1 as empirical measures: $[e_0]_*(\mu) = \sum_{i=1}^k \alpha_i \, \delta_{x_i} \delta_{v_i}$ and $[e_1]_*(\mu) = \sum_{j=1}^k \beta_j \, \delta_{y_j} \delta_{w_j}$, as explained in Section 4.4, we can simplify the Kantorovich using the exact L^2 norm of the acceleration of the spline between (x_v) and (y, w), whose cost is given in Formula (4.14). Again, let us underline that this cost is *not* a Riemannian cost on the tangent space of $\mathbb{R}^d$ since if $v = w$ and x, y are close, the cost is dominated by the term $4(|v|^2 + |w|^2 + \langle v, w\rangle)$ which need not be zero. Then, the Kantorovich problem reduces to the minimization of

$$\sum_{i,j=1}^{k,l} \pi_{i,j} c((x_i, v_i), (y_j, w_j)) \,, \tag{5.1}$$

under the constraints

$$\begin{cases} \sum_{i=1}^k \pi_{i,j} = \beta_j \\ \sum_{j=1}^l \pi_{i,j} = \alpha_i \,. \end{cases} \tag{5.2}$$

It is straightforward to apply entropic regularization/Sinkhorn in this case which amounts to add, for a positive parameter ε, $\varepsilon \sum_{i,j} \pi_{i,j} \log(\pi_{i,j})$ to the previous linear functional and to numerically solve the corresponding variational problem with the Sinkhorn algorithm [27, 9] (See also appendix B where Sinkhorn algorithm is detailed in the more general multimarginal case). It is interesting to note that the choice of ε is more delicate than in the standard case of a quadratic distance cost.

In Figure 2, we present the convergence rate of this method with respect to two different values of ε and the most likely deterministic plan given the optimal plan π^ε. Note that this entropic regularization method scales with the number of points as N^2 and is valid in every dimension.

5.2. **MultiMarginal formulation.** This is the direct discretization of (4.6) which avoids working in phase space with the cost (4.16) thus enabling fast computations in 2D. In what follows, the time cylinder $[0, 1] \times M$ is discretized in time as $\bigotimes_{i=0,N} M_i$, the product space of $N + 1$ copies of M at each of the $N + 1$ time steps. We will use a regular time step discretization $\tau_i = i \, d\tau$ where $d\tau = \frac{1}{N}$. Using a classic finite difference approach, the time discretization of (4.6) is

$$\min_{\mu_{d\tau}} \int_{\bigotimes_{i=0,N} M_i} c_{d\tau}(x_1, ..., x_N) \, \mathrm{d}\mu_{d\tau}(x_1, .., x_N) \,, \tag{5.3}$$

where $\mu_{d\tau}$ now spans the space of probability measures on $\bigotimes_{i=0,N} M_i$ representing the space of piecewise linear curves passing through $x_0, x_1, ..., x_N$ at times $\tau_0, ..., \tau_N$.

A straighforward computation gives

$$c_{d\tau}(x_1, ..., x_N) := \sum_{i=1,N-1} \frac{\|x_{i+1} + x_{i-1} - 2\, x_i\|^2}{d\tau^3} \tag{5.4}$$

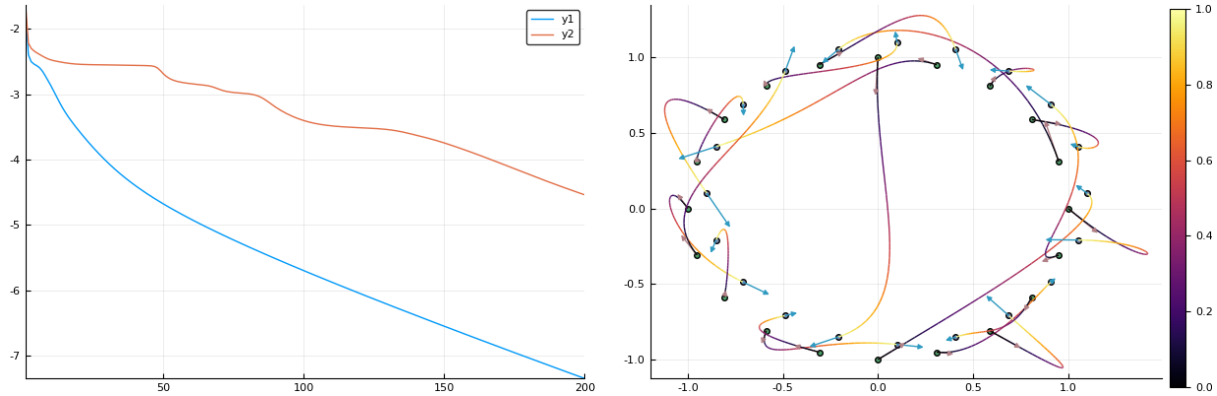


FIGURE 2. Convergence (left) and Hermite interpolation problem between Two empirical measure in phase space (right). We represent the most likely splines in the position space.

For all times, marginals (4.7) are computed as :

$$\tau_j \mapsto \int_{\bigotimes_{i\neq j} M_i} \mathrm{d}\mu_{d\tau}(x_1, ..x_N) \tag{5.5}$$

In order to simplify the presentation we will assume that the marginal constraints (4.8) are set at times $t_1, ..t_n$ which coincide with times steps of the discretization (of course $n < N$, meaning the number of constraint is not the same as the number of time steps).

In short, there exist $(j_1, ..j_n) \in [0, N]$ such that

$$(t_1, .., t_n) = (\tau_{j_1}, ..., \tau_{j_n}).$$

The constraint (4.8) becomes for all $k = 1, ..n$

$$\int_{\bigotimes_{i\neq j_k} M_i} \mathrm{d}\mu_{d\tau}(x_1, ..x_N) = \rho_{j_k}(x_{j_k}) \tag{5.6}$$

where ρ_{j_k} is the prescribed density to interpolate at time $\tau_{j_k} = t_k$.

The time discretized problem is the multimarginal problem (5.3 -5.6).

The simplest space discretization strategy is to use a regular cartesian grid. In dimension 2 and for $M = [0, 1]^2$ and at time t_i, the grid will be denoted $x_{\alpha_i,\beta_i} = (\alpha_i\, h, \beta_i h)$ for $(\alpha_i, \beta_i) \in [0, N_x]$ and $h = \frac{1}{N_x}$, $a = \{\alpha_i\}$ and $b = \{\beta_i\}$ will be the vectors of indices.

The time and space discretization of the problem then becomes

$$\min_T \sum_{a,b} C_{a,b}\, T_{a,b} \tag{5.7}$$

Where T is the $N \times N_x \times N_x$ tensor of grid values $\mu_{d\tau}(x_{\alpha_1,\beta_1}, .., x_{\alpha_N,\beta_N})$ and

$$C_{a,b} = c_{d\tau}(x_{\alpha_1,\beta_1}, .., x_{\alpha_N,\beta_N}) \tag{5.8}$$

The marginals (5.5) at all times τ_j are given by

$$\sum_{a\setminus\{\alpha_j\},\, b\setminus\{\beta_j\}} T_{a,b} \tag{5.9}$$

The constraints (5.6) therfore becomes for all k

$$\sum_{a\setminus\{\alpha_{j_k}\},\, b\setminus\{\beta_{j_k}\}} T_{a,b} = \rho_{j_k}(x_{\alpha_{j_k}, \beta_{j_k}}) \tag{5.10}$$

$a \setminus \{\alpha_{j_k}\}$ denotes the set of indices a minus α_{j_k}.

The Entropic regularized problem is

$$\min_{T^\epsilon} \sum_{a,b} \{C_{a,b}\, T^\epsilon_{a,b} + \epsilon\, T^\epsilon_{a,b}\, \log(T^\epsilon_{a,b})\} \tag{5.11}$$

and easier to solve. See Appendix B for a description of Sinkhorn algorithm.

Numerical Simulations.
1D case: We present, figures 3 and 4, a 1D test case to highlight some of the qualitative properties of the cubic splines interpolation on the space of densities.

We consider four interpolation time points and the corresponding data are mixture of Gaussians of different standard deviations. We use a discretization of 140 points on the interval $[0, 1]$ with 16 time steps. The doted line represent the reconstructed density curve in time. This experiment shows that the mass can concentrate or diffuse in some situation.

Another important point here is that the entropic regularization parameter has an important impact on this concentration/diffusion effects: we show the simulations for $\varepsilon = 0.002$ and $\varepsilon = 8.10^{-5}$. In the simulation with a large ε, the concentration effect is not present and it is due to the diffusion on the path space.
2D case: We present a 2D test case which computes a Wasserstein spline in the sense of (5.7) interpolating four Gaussian identical densities at time 1, 5, 13, and 17, see figure 5. We use a time step $d\tau = 1$ and 17 $N = 17$ time steps. The space discretization is $Nx = 50$. The entropic regularization parameter is $\epsilon = 0.002$, note that the stability of the method depends on this parameter. It also generates artificial diffusion as it becomes more costly top concentrate the available mass on fewer Euclidean splines between the points of the support of the four Gaussians. We can compute the interpolating densities at intermediate times using (5.9) but is more interesting to represent in figure 6 the contour line of the third quartile, i.e. the highest values of the densities representing 1/4 of the total mass. Comparing with figure 7, it seems clear that the Entropy diffusion spreading pollutes the solution of the original problem (without entropic regularization).

We compare this solution with the classical Quadratic cost Optimal Transport interpolation, i.e. with the speed instead of the acceleration in the cost. More precisely taking :

$$c_{d\tau}(x_1, ..., x_N) := \sum_{i=0,N-1} \frac{\|x_{i+1} - x_i\|^2}{d\tau} \tag{5.12}$$

As expected the mass follows respectively the linear interpolation or the Euclidean spline interpolation of the center of the Gaussians which are represented as thick red lines in figure 5.

Finally we show the convergence of the Sinkhorn iterate for both simulations in figure 6. The convergence is much slower for the speed case but we did not optimize the implementation which does not need tensors and instead just used a degraded version of the acceleration code. This may be the reason for this strange difference.

5.3. **Semi-Discrete approach.** We propose another numerical scheme based on the semi-discrete approach introduced by Mérigot in [19] in dimension 2 and developed by Levy [17] in dimension 3. Here we approximate the optimal plan π in the formulation (4.9) by a sum of N tensor product of diracs masses. That is $\pi_N = \sum_{j=1}^N \left(\bigotimes_{i=1}^n \frac{1}{N}\delta_{X_j^i}\right) = \sum_{j=1}^N \frac{1}{N}\delta_{(X_j^1,...,X_j^n)}$.

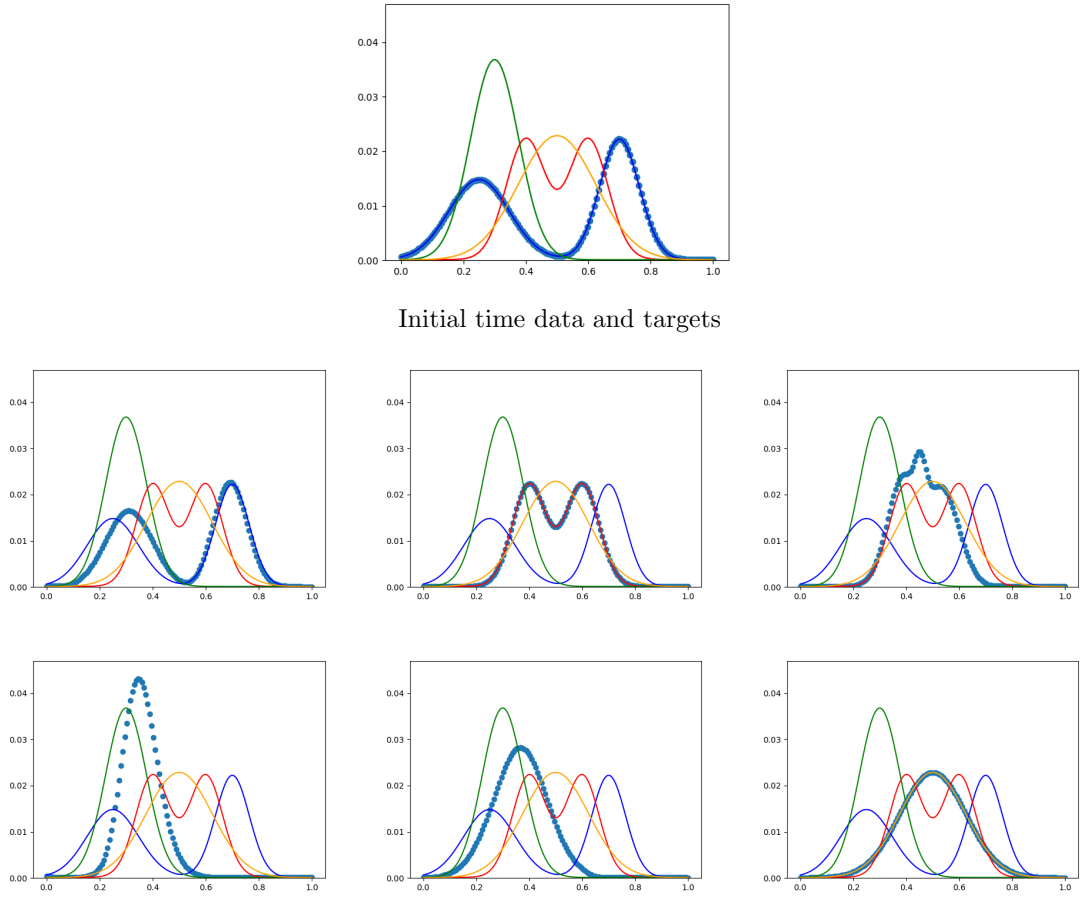


FIGURE 3. Four interpolation timepoints, $1, 6, 11, 16$ and representation of the four density configurations, as well as 6 intermediate times. The doted line represent the reconstructed density curve in time. This experiment underlines that the spline curve has more smoothness in time and can present some concentration or diffusion effects depending on the data which would not be present for the usual Wasserstein geodesic. The entropic regularization parameter is $\varepsilon = 8.10^{-5}$.

Remark 4. *Since there is a unique corresponds between n points $\left(X_j^1, \ldots, X_j^n\right)$ and the spline $c_{X_j^1,\ldots,X_j^n}$ passing through these points at time $(t_1, \ldots t_n)$ the measure π_N can also be seen as N direct masses defined over the set of splines: $\pi_N = \sum_{j=1}^N \frac{1}{N} \delta_{c_{X_j^1,\ldots,X_j^n}}$.*

We then have to relax the constraint $(p_i)_*(\pi) = \rho_i$ since $(p_i)_*(\pi_N) = \sum_{j=1}^N \frac{1}{N} \delta_{X_j^i}$ cannot be absolutely continuous. It leads to the following variational problem.

Definition 6 (Semi-discrete variational problem)**.** Let $\epsilon > 0$, $0 = t_1 < \ldots < t_n = 1$, $n \geq 3$ and $(\rho_i)_{i=1\ldots n}$ be n absolutely continuous measures. Recall that $c(Y_1, \ldots, Y_n)$ is the cost of the cubic spline passing through the points $(Y_1, \ldots, Y_n)$ at time $(t_1, \ldots t_n)$. Let

$$\mathcal{Q}^N = \left\{ \sum_{j=1}^N \frac{1}{N} \delta_{(X_j^1,\ldots,X_j^n)} \middle| (X_j)_{j=1,\ldots,N} \in M^n \right\}.$$

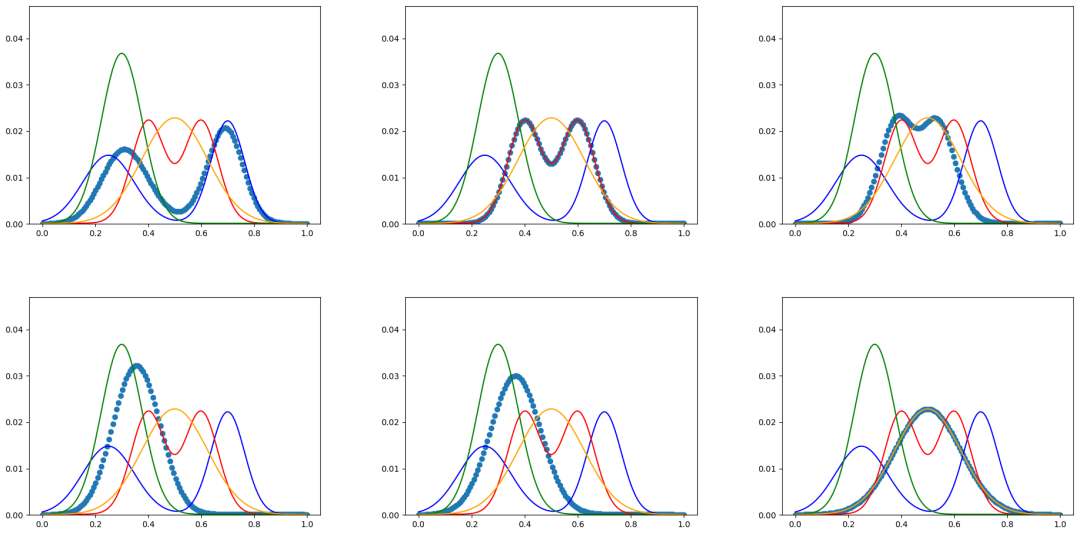


FIGURE 4. The same experiment with a larger entropic regularization parameter $\varepsilon = 0.002$. As expected, we observe less concentration of mass.

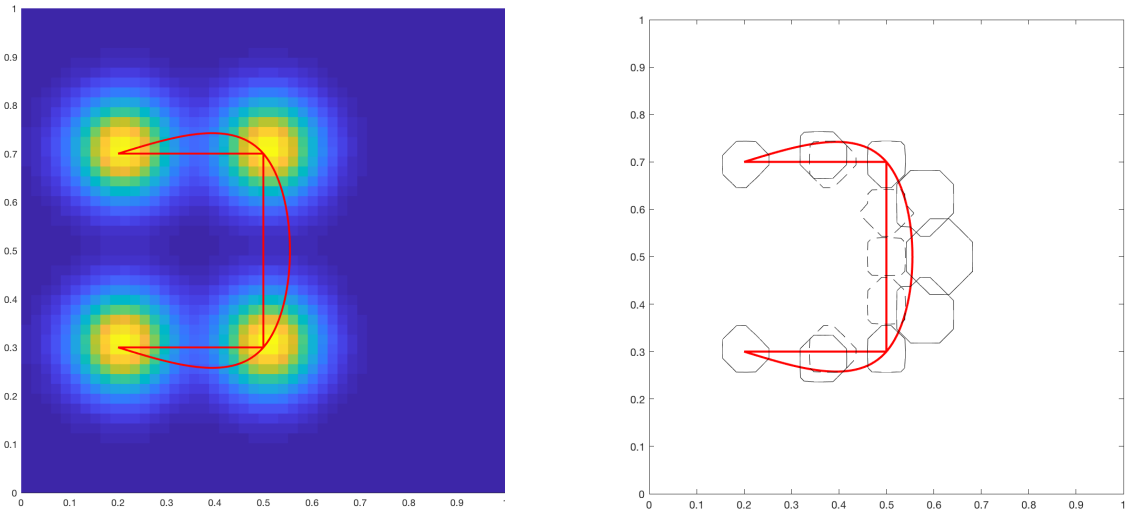


FIGURE 5. Spline interpolation of Four Gaussians with 17 times steps. Left : the data and the linear and classic cubic spline interpolation of the of Gaussian center point. Right : the level curve of the third quartile of the density every 2 time steps, in solid line for our Spline Wasserstein interpolation and in dashed line for the classic quadratic cost (speed) interpolation.

Then the semi-discrete variational problem, (SDV), is given by

$$(SDV) = \min_{\mathcal{Q}^N} \frac{1}{N}\sum_{j=1}^{N} c(X_j^1, \dots, X_j^n) + \sum_{i=1}^{n} \frac{1}{2\epsilon^2} W_2^2 \left(\sum_{j=1}^{N} \frac{1}{N} \delta_{X_j^i}, \rho_i \right), \tag{5.13}$$

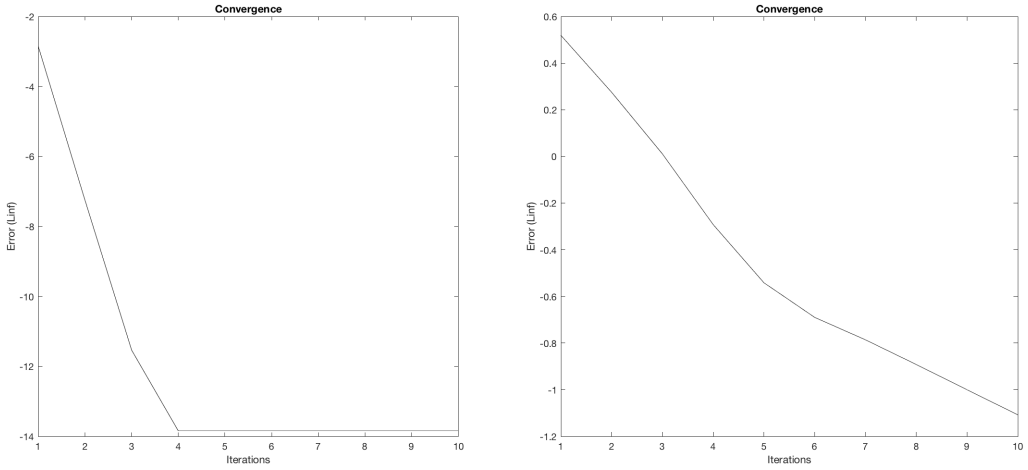


FIGURE 6. Convergence, i.e. Infinity norm of the difference of the Dual unknown between to Sinkhorn iteration. This is computed every 10 iterations. Left :for the acceleration cost, right : for the speed cost .

where W_2 is the classical Wasserstein distance given by the quadratic cost.

The main drawback of this method is that, as illustrated in the numerical simulations below, the problem (SDV) is not convex.

5.3.1. *Implementation.* In order to solve numerically the minimization problem (SDV) we use the reformulation of the spline cost in the phase space, that is in $\mathbb{R}^d$, with $t_{i+1} - t_i = \delta_i$:

$$c\left(Y_1, \ldots, Y_n\right) = \min_{(V_1, .. V_n) \in (\mathbb{R}^d)^n} \sum_{i=1}^{n-1} \frac{1}{\delta_i^3} c_{ph}\left[\left(Y_i, \delta_i V_i\right), \left(Y_{i+1}, \delta_i V_{i+1}\right)\right] \tag{5.14}$$

where

$$c_{ph}[(x,v),(y,w)] = 12|x-y|^2 + 4(|v|^2 + |w|^2 + \langle v, w\rangle + 3\langle v+w, x-y\rangle). \tag{5.15}$$

The advantage of the formulation (5.14) is that the cost is separable in the phase space and the gradient with respect to speeds and positions is easy to compute.

We thus implement a gradient descent in the phase space using the lbfgs function in python. We compute the gradient by automatic differentiation. The Wasserstein terms in the minimization problem (5.13) depends only on the positions and are computed thanks to Mérigot Library [1] in dimension 2. To do simulations in dimension 3 one has to use Lévy Library [2]. The density constraints ρ_i are given trough linear functions on a triangulation.

Remark 5. *Other problems can be addressed using similar optimization problem as in Definition 6. For instance the quadratic cost in* (5.13) *leads to Wasserstein interpolation. We can also interpolate with curves as smooth as we want, using for instance the L^2 norm of the derivative of order m of the curve or even other classical interpolating curves.*

5.3.2. *Numerical simulations.* We propose three numerical simulations, one to compare the qualitative results with respect to the multi marginal approach and especially Figure 5. A second one in order to illustrate the non-convexity issue and a third one for applications in images.

The rotation case: Figure 7. In this case we compute Wasserstein splines passing through four gaussians with variance 15 and center of masses respectively $(0,2),(10,0),(10,6),(0,4)$ with constraint parameter $\epsilon = 10^{-3}$. The number of points is 2000. In this case the result is a global minimizer and is not sensible to initialization. The lack of convexity is not an issue. Compare to

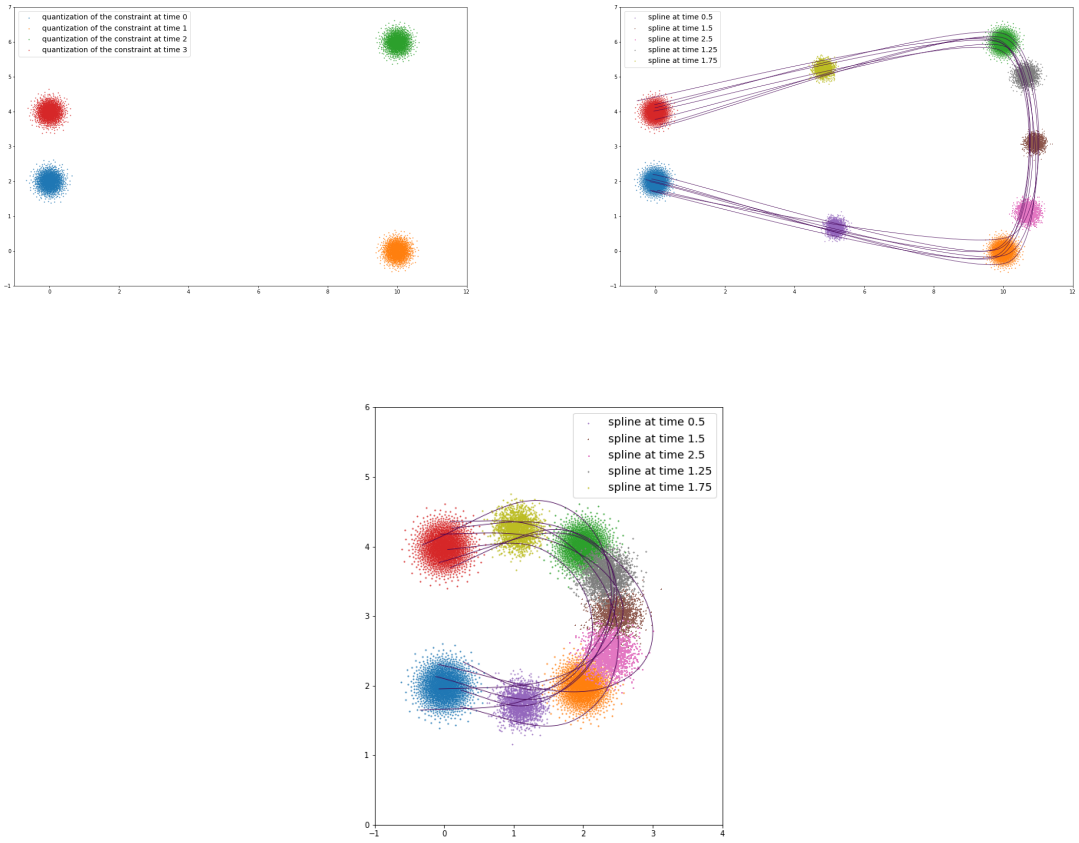


FIGURE 7. Spline interpolation for gaussians with 2000 Dirac masses for each measure, $\epsilon = 1O^{-3}$. Left: sample of each density constraints ρ_i, i=1,2,3,4. Right: Some trajectory of diracs masses randomly chosen, marginals at the constrained time $0, 1, 2, 3$ and marginals at time $0.5, 1.2, 1.5, 1.7, 2.5$. Second Line : the same configuration as in figure 5.

Figure 5, this approach gives a better a approximation of the intermediate densities especially with less diffusion.

The crossing case: Figure 8, 9. Here we compute Wasserstein splines starting from a mixture of two gaussians with centrer $(0, -1), (0, 1)$ and variance 15 then passing through a gaussian with center $(0, 0)$ and variance 15 and finishing at a translation of the initial mixture. The number of points is 2000, ϵ will value 1 or 1000.

We expect the global minimizer to be straight lines crossing around the middle constraint and with a low cost. Numerically depending on the initial conditions, we can recover different local minimizers, the local minimum which is reached is extremely correlated with the initial coupling. In Figure 8 we observe that changing ϵ but keeping a similar initial coupling, all points are given by a quantization of the middle density with a random enumeration and 0 initial speed, yields to a similar local minimum.

Finding a good initial coupling is the hard part in order reach the global maximum. One solution is to initialize with points close to each other and a very large ϵ. Then one as to add some noise in the gradient and decreases slowly ϵ. Unfortunately we didn't find a systematic approach for this

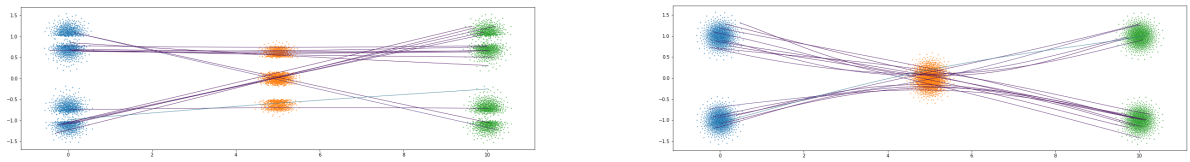


FIGURE 8. Spline interpolation for a mixture of gaussians with 2000 Dirac masses. Same initial coupling for both figure. Left: $\epsilon = 1$. Right: $\epsilon = 1000$.

random multi-scale method and one as to fit the parameters case by case. In Figure 9 the global minimizer is achieved by first computing the spline with a relaxed constraint, i.e. large ϵ, only for the final time (in pratice $\epsilon = [1000, 1000, 1]$. Then we use this result, which has the good initial coupling, as and initial condition and set $\epsilon = 1000$ for all the constraints. We also compare this results with the interpolation with a different initial condition and the Wasserstein geodesics. In all these simulations we clearly observe that particles can cross along the dynamic appart from the optimal transport inthis situation.

Note that this spline approach is related to the problem of finding minimal geodesics along volume preserving maps done by Mérigot and Mirebeau [20] : in their work the constraints ρ_i are the Lebesgue measure, the cost is changed by the quadratic cost between two points and they have a coupling constraint. Therefore their minimization problem is also non convex but the coupling is given as a constraint so the non convexity issue didn't rise as clearly as in this spline problem.
Image interpolation: pour l'instant c'est pas presentable, ca passe vraiment au milieu. Je vais relancer dans la semaine mais je propose de faire une version sans.

Remark 6 (Extrapolation)**.** *The minimization of the acceleration can be used to provide time extrapolation of Wasserstein geodesic in a natural way: particles follow straight lines. This can be implemented in a* 3*-marginal problem with the acceleration cost* $c(x_1, x_2, x_3) = \frac{1}{\lambda^2}|x_3 - 2x_2 + x_1|^2 + \frac{1}{\lambda}|x_2 - x_1|^2$ *under marginal constraints at time* 1 *and* 2*. Note that, in the spline model, the formulation we proposed does not prevent particles from crossing each other. They are completely independent. Therefore, the particles following simply geodesic lines and after a shock, the evolution is not geodesic in the Wasserstein sense (since shocks do not occur but at initial and final times). The implementation of time extrapolation using entropic regularization is straightforward. Figures 10 and 11 show some experiments on* $[0, 1]$ *discretized with* 100 *points and* $\varepsilon = 0.015$*. The translation experiment recovers what is expected however the effect of the diffusion can be seen with a twice larger* ε*. We also show two other simulations, one is a splitting simulation and the last one is a merging of two "bumps" into a single one. The extrapolation shows an other bimodal distribution which is explained by particle crossings. Note that this extrapolation scheme may proven useful in the development of higher-order schemes for the JKO algorithm.*

6. PERSPECTIVES

In this paper, we presented natural approaches to define cubic splines on the space of probability measures. We have presented a Monge formulation and its Kantorovich relaxation on the path space as well as their corresponding reduction on minimal cubic spline interpolation. We leave for future work theoretical questions such as the study of conditions under which the existence of a Monge map as a minimizer occurs, as well as the relaxation of cubic spline in the Wasserstein metric. Our main contributions focus on the numerical feasibility of the minimization of the acceleration on the path space with marginal constraints. We have developed the entropic regularization scheme for the

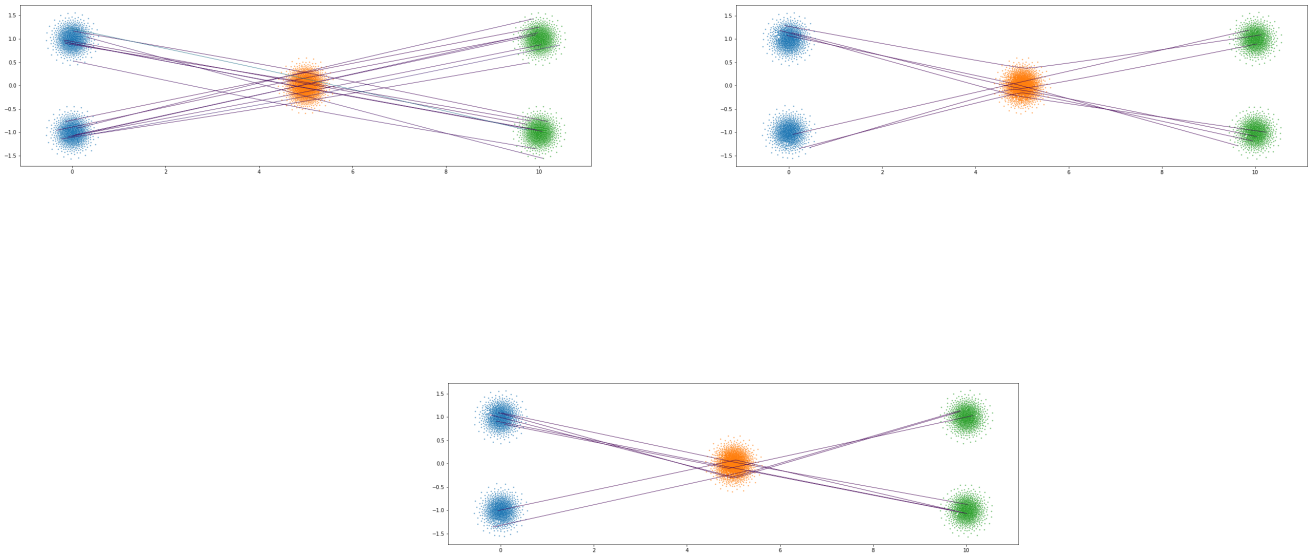


FIGURE 9. Spline interpolation for a mixture of gaussians with 2000 Dirac masses for each measure. $\epsilon = 1000$. Top Left: Initialization with a good coupling, total cost = 302. Top Right: Initialization with a quantization of the middle density and no speed, total cost = 804 (local minima). Bottom: Interpolation with the Wasserstein geodesic. $\epsilon = 1000$, cost = 930.

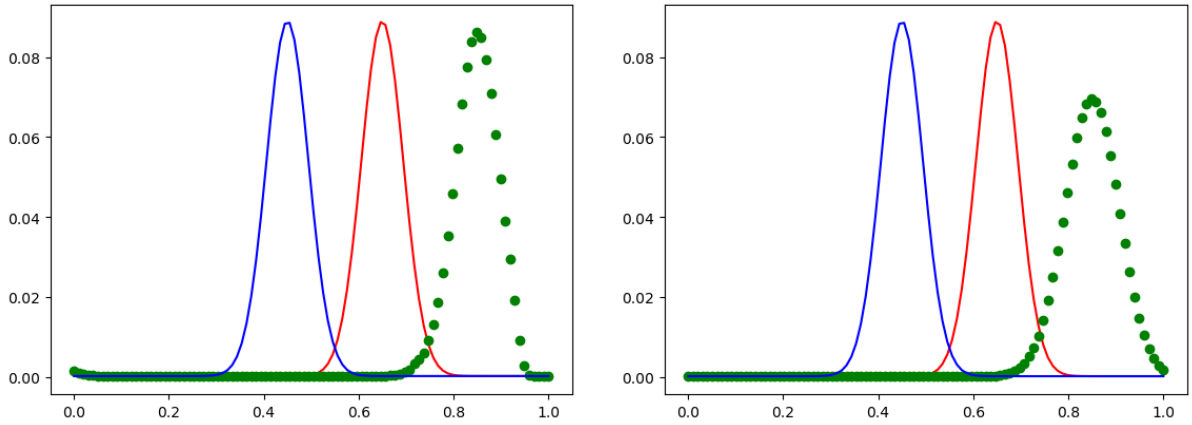


FIGURE 10. Extrapolation of a translation with two different $\varepsilon = 0.015$ and $\varepsilon = 0.03$

acceleration and shown simulations in 1D and 2D. Future work will address the 3D case which is out of reach with the methods presented in the first sections of this paper but possibly tackled with the semi-discrete method presented en Section 5.3. In a similar direction, the application of this

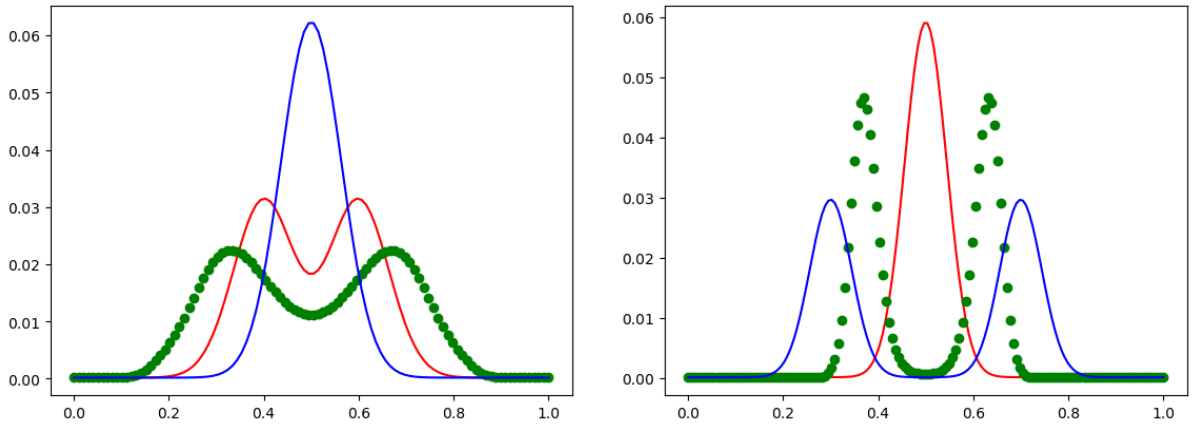


FIGURE 11. On the left, a splitting experiment and on the right, a merging experiment.

approach to the unbalanced case in the spirit of [7] seems challenging due to the this dimensionality constraint and could be achieved within the semi-discrete setting.

In the Lagrangian setting, i.e. semi-discrete method, the extrapolation of a Wasserstein geodesic between ρ_0 and ρ_1 is obtained using three positions with the following formulation : let

$$\mathcal{Q}^N = \left\{ \sum_{j=1}^{N} \frac{1}{N} \delta_{(X_j^1, X_j^2, X_j^3)} \middle| (X_j)_{j=1,\dots,N} \in M^n \right\},$$

then

$$(SDextra) = \min_{\mathcal{Q}^N} \frac{1}{N} \sum_{j=1}^{N} \frac{d^2}{2}(X_j^1, X_j^2) + \frac{1}{N} \sum_{j=1}^{N} c(X_j^1, X_j^2, X_j^3) + \sum_{i=1}^{2} \frac{1}{2\epsilon^2} W_2^2 \left(\sum_{j=1}^{N} \frac{1}{N} \delta_{X_j^i}, \rho_i \right), \tag{6.1}$$

where d is the distance on M and $c(X_j^1, X_j^2, X_j^3)$ the cost of the cubic spline. In particular this formulation forces the curve to be a Wasserstein geodesic between ρ_1 and ρ_2, using the quadratic cost, and let free the final marginal. The implementation is completely similar as in Section 5.3 and the trajectory of each dirac masses is a straight line.

APPENDIX A. PROOF OF THEOREM 1

The proof is a rewriting of the proof of [25, Theorem 1.33] when the initial and final spaces do not have the same dimension. In particular we prove that transport plans concentrated on a graph of a map $T : \mathbb{R}^d \to \mathbb{R}^p$ are dense into transport plans in $\mathbb{R}^d \times \mathbb{R}^p$ and deduce, taking $p = (n-1)d$, that for any continuous cost the multimarginal Kantorovich problem is the relaxation of the multimarginal Monge problem.

Theorem 9. *Let $M = \mathbb{R}^d$ and $c : M^n \to \mathbb{R}$ be a continuous cost fonction. Let $(\rho_i)_{i\in 1,\dots,n}$ be n probability measures on M. We define the Monge Problem (M_c) as*

$$(M_c) = \inf \int_M c\left(x, T_2(x), \dots, T_n(x)\right) \rho_1 \,,$$

over the set of map $\Pi_T = \left\{ T : M \to M^{n-1},\ x \mapsto (T_i(x))_{i=2,\dots,n} \middle| (T_i)_* (\rho_1) = \rho_i,\ , i = 2, \dots, n \right\}$. The Kantorovich problem (K_c) is defined by

$$(K_c) = \inf \int_{M^n} c\left(x_1, \dots, x_n\right) \pi\left(x_1, \dots, x_n\right) \,,$$

over the set of plan $\Pi = \{\pi \in \mathcal{P}(M^n) | (p_i)_*(\pi) = \rho_i,\ i = 1, \dots, n\}$, *where* p_i *is the projection of the* i^{th} *factor. Then, if all* $(\rho_i)_{i\in 1,\dots,n}$ *have compact support and* ρ_1 *is atomless there holds* $(M_c) = (K_c)$.

In order to prove Theorem 9 we first remark that [25, Corrollary 1.29 and Theorem 1.32] have their multimarginal counterpart.

Lemma 10. *Let* $\mu \in \mathcal{P}(\mathbb{R}^d)$ *be atomless measure and* $\nu \in \mathcal{P}(\mathbb{R}^p)$, *then there exists a transport map* $T : \mathbb{R}^d \to \mathbb{R}^p$ *such that* $T_*\mu = \nu$.

Proof of Lemma 10. Let $\sigma_d : \mathbb{R}^d \to \mathbb{R}$ (resp $\sigma_p : \mathbb{R}^p \to \mathbb{R}$) be an injective Borel map with Borel inverse (see [25, Lemma 1.28] for instance for a very simple proof of existence in this case). Since μ is atomless $(\sigma_d)_*\mu$ is also atomless. Let $t : \mathbb{R} \to \mathbb{R}$ be the optimal transport map from $(\sigma_d)_*\mu$ to $(\sigma_p)_*\nu$ for the quadratic cost. $t_* \left((\sigma_d)_*\mu\right) = (\sigma_p)_* \nu$. Thus $T = \sigma_p^{-1} \circ t \circ \sigma_d$ is a map pushing forward μ to ν. □

Theorem 11. *With the notation of Theorem 9, if the support of all* ρ_i *are included in a compact domain then the set of plans* Π_T *induced by a transport is dense, for the weak topology, in the set of plans* Π *whenever* ρ_1 *is atomless.*

Remark 7. *Theorem 11 is in fact very general, one can consider M N be only Polish spaces for instance. Then there exists invertible Borel maps from M (resp N) to* $[0, 1]$. *This is enough to obtain Lemma 10. Then one just need to consider a uniformly small partition of* Ω *to prove the density Theorem 11.*

Proof of Theorem 11. Again the proof is based on [25, Theorem 1.32]. In particular the strategy of the proof is to approach a transport plan by transport maps defined on small sets on which the measure is preserved.

We consider a compact domain $\Omega = \Omega_d \times \Omega_p \in (\mathbb{R}^d \times \mathbb{R}^p)$ and $\pi \in \mathcal{P}(\Omega_d \times \Omega_p)$ such that $(p_{\mathbb{R}^d})_*(\pi) = \mu$ is atomless. For any m set a partition of Ω_p (resp Ω_q) into (disjoint) sets $K_{i,m}$ (resp $L_{j,m}$) with diameter smaller than $1/2m$. Then $C_{i,j,m} = K_{i,m} \times L_{j,m}$ is a partition of Ω into sets with diameter smaller than $1/m$. Let $\pi_{i,m}$ be the restriction of π on $K_{i,m} \times \Omega_p$ and $\mu_{i,m} = (p_{\mathbb{R}^d})_*(\pi_{i,m})$ and $\nu_{i,m} = (p_{\mathbb{R}^d})_*(\pi_{i,m})$. Since μ is atomless $\mu_{i,m} = \mu_{|K_{i,m}}$ is also atomless and thanks to Lemma 10 there exists $t_{i,m}$ such that $(t_{i,m})_*\mu_{i,m} = \nu_{i,m}$. By definition

(A.1)

$$\pi[C_{i,j,m}] = \pi_{i,m}[C_{i,j,m}] = \mu_{i,m}[K_{i,j}]\nu_{i,m}[L_{j,m}] = (\mathrm{Id}, \mathrm{t_{i,m}})_*(\mu_{\mathrm{i,m}})([\mathrm{C_{i,j,m}}]) = (\mathrm{Id}, \mathrm{t_m})_*(\mu)[\mathrm{C_{i,j,m}}],$$

where t_m is define on Ω by $t_{|K_{i,m}} = t_{i,m}$. In particular $(t_m)_*(\mu) = \nu$. Equation (A.1) and the definition of the partition sets $C_{i,j,m}$ implies that $(\mathrm{Id}, \mathrm{t_m})_*(\mu)$ weakly converges toward π as $m + \infty$ (they give same masses to any set of the partition). See [Theorem 1.31]santambrogio2015optimal for instance. To finish the proof let us remark that we can set $p = d(n-1)$ then $\mu = \rho_1$ is atomless and $t_m : \mathbb{R}^d \to \mathbb{R}^{d(n-1)}$ defines $(t_{2,n}, ..., t_{n,m})$. □

Proof of Theorem 9 . The continuity of the cost c and the density Theorem 11 implies that $(K_c) \leq (M_c)$. Since the converse is always true we have $(M_c) = (K_c)$. □

Remark 8. *Theorem 1 is a consequence of Theorem A since both the Monge and the Kantorovich (Definition 1 and 2) problems reduces on* M^n *with the spline cost which is continuous (see Corollary 4 and 5.*

Appendix B. Entropic Regularisation and Sinkhorn

B.1. **Entropic regularization and Sinkhorn algorithm.** The linear programming problems (5.7-5.10) is extremely costly to solve numerically and a natural strategy, which has received a lot of attention recently following the pionneering works of [10] and [9] is to approximate these problems by strictly convex ones by adding an entropic penalization. It has been used with good results on a number of multi-marginal optimal transport problems [3] [4] [5]. Here is a rapid and simplified description, see the references above for more details.

The regularized problem is

$$\min_{T^\epsilon} \sum_{a,b} \{C_{a,b}\, T^\epsilon_{a,b} + \epsilon\, T^\epsilon_{a,b}\, \log(T^\epsilon_{a,b})\} \tag{B.1}$$

It is strictly convex. Denoting $u^k_{\alpha_{j_k},\beta_{j_k}}$ the Lagrange multipliers of the k constraints (5.10), we obtain the optimality conditions:

$$T^\epsilon_{a,b} = K_{a,b}\, \Pi_{k=1}^N U^k_{j_k} \tag{B.2}$$

where

$$U^k_{j_k} = e^{\frac{1}{\epsilon} u^k_{\alpha_{j_k},\beta_{j_k}}} \qquad K_{a,b} = e^{-\frac{1}{\epsilon} C_{a,b}}$$

Equation (B.2) caracterize the optimal tensor as a scaling of the Kernel K depending on the dual unknown U^k. Inserting this factorization into the constrains (5.10) the dual problem takes the form of the set of equations ($\forall k \in [1, n]$)

$$U^k_{j_k} = \rho_{j_k}(x_{\alpha_{j_k},\beta_{j_k}})\Big(\sum_{a\setminus\{\alpha_{j_k}\},\, b\setminus\{\beta_{j_k}\}} K_{a,b}\, \Pi_{k'\in\{1,..n\}\setminus k}\, U^{k'}_{j_{k'}}\Big)^{-1} \tag{B.3}$$

Sinkhorn algorithm simply amounts to perform a Gauss-Seidel type iterative resolution of the system (B.3) and therefore consists in computing the sums on the right-hand side and then perform the (grid) point wise division.

B.2. **Implementation.** In dimension 2, each unknown U_k has dimension N_x^2, the cost of one full Gauss Seidel cycle, i.e. on Sinkhorn iteration on all unknowns, will therefore be $n \times N_x^2 \times$ the cost to compute the tensor matrix products in the denominator of (B.3). Remember that n is the number of time steps with constraints and N the total number of time steps. The given tensor Kernel $K_{a,b}$ is a priori a large $N \times N_x \times N_x$ tensor with indices $a, b = \alpha_1, ..\alpha_N, \beta_1, .., \beta_N$. It can however advantageously be tensorized both along dimensions and also margins. First, using (5.4-5.8) we see that the Kernel is the product of smaller tensors

$$K_{a,b} = \Pi_{i=1,N-1} K^0_{i-1,i,i+1}, \text{ with } K^0_{i-1,i,i+1} := e^{-\frac{1}{\epsilon\, d\tau^3}\|x_{\alpha_{i+1},\beta_{i+1}} + x_{\alpha_{i-1},\beta_{i-1}} - 2\, x_{\alpha_i,\beta_i}\|^2}.$$

Moreover as we chose to work on a cartesian grid at all time steps, K^0 tensorize again into

$$K^0_{i-1,i,i+1} = K^\alpha_{i-1,i,i+1}\, K^\beta_{i-1,i,i+1} \text{ with } K^\alpha_{i-1,i,i+1} := e^{-\frac{h^2}{\epsilon\, d\tau^3}\|\alpha_{i+1} + \alpha_{i-1} - 2\alpha_i\|^2}$$

Finally our large kernel $K_{a,b}$ can be represented a the product of $2\,(N-2)$ identical tensors of size $N_x \times N_x \times N_x$. Assuming a cubic cost n^3 for the multiplication of two $(n \times n)$ matrix, we see oru algorithm is of order $O(N\, N_x^4)$ in dimension 2.

References

[1] `https://github.com/mrgt/PyMongeAmpere`.

[2] `https://members.loria.fr/Bruno.Levy/GEOGRAM/vorpaview.html`.

[3] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[4] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. A Numerical Method to solve Optimal Transport Problems with Coulomb Cost. working paper or preprint, May 2015.

[5] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. working paper or preprint, October 2017.

[6] M. Camarinha, F. Silva Leite, and P.Crouch. Splines of class $\mathcal{C}^k$ on non-euclidean spaces. *IMA Journal of Mathematical Control & Information*, 12:399–410, 1995.

[7] L. Chizat, B. Schmitzer, G. Peyré, and F.-X. Vialard. An Interpolating Distance between Optimal Transport and Fisher-Rao. *Found. Comp. Math.*, 2016.

[8] P. Crouch and F. Silva Leite. The dynamic interpolation problem: On Riemannian manifold, Lie groups and symmetric spaces. *Journal of dynamical & Control Systems*, 1:177–202, 1995.

[9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[10] Alfred Galichon and Bernard Salanié. Matching with Trade-offs: Revealed Preferences over Competiting Characteristics. working paper or preprint, April 2010.
[11] F. Gay-Balmaz, D. D. Holm, D. M. Meier, T. S. Ratiu, and F.-X. Vialard. Invariant Higher-Order Variational Problems. *Communications in Mathematical Physics*, 309:413–458, January 2012.
[12] F. Gay-Balmaz, D. D. Holm, D. M. Meier, T. S. Ratiu, and F.-X. Vialard. Invariant Higher-Order Variational Problems II. *Journal of NonLinear Science*, 22:553–597, August 2012.
[13] B. Heeren, M. Rumpf, and B. Wirth. Variational time discretization of Riemannian splines. *ArXiv e-prints*, November 2017.
[14] François-Xavier Vialard Jean-David Benamou, Thomas Gallouët. Second order models for optimal transport and cubic splines on the wasserstein space. *Preprint arXiv:1801.04144*, 2018.
[15] B. Khesin and R. Wendt. *The geometry of infinite-dimensional groups*, volume 51. Springer Science & Business Media, 2008.
[16] Young-Heon Kim and Brendan Pass. A general condition for monge solutions in the multi-marginal optimal transport problem. *SIAM Journal on Mathematical Analysis*, 46(2):1538–1550, 2014.
[17] Lévy, Bruno. A numerical algorithm for l2 semi-discrete optimal transport in 3d. *ESAIM: M2AN*, 49(6):1693–1715, 2015.
[18] J. Lott. Some geometric calculations on Wasserstein space. *Communications in Mathematical Physics*, 277(2):423–437, 2008.
[19] Quentin Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30 (5):1583–1592, 2011.
[20] Quentin Mérigot and Jean-Marie Mirebeau. Minimal geodesics along volume preserving maps, through semi-discrete optimal transport. *arXiv preprint arXiv:1505.03306*, 2015.
[21] Quentin Mérigot and Jean-Marie Mirebeau. Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport. *SIAM J. Numer. Anal.*, 54(6):3465–3492, 2016.
[22] L. Noakes, G. Heinzinger, and B. Paden. Cubic splines on curved spaces. *IMA Journal of Mathematical Control & Information*, 6:465–473, 1989.
[23] F. Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[24] Pass, Brendan. Multi-marginal optimal transport: Theory and applicationsâ^—. *ESAIM: M2AN*, 49(6):1771–1790, 2015.
[25] F. Santambrogio. Optimal transport for applied mathematicians. *Progress in Nonlinear Differential Equations and their applications*, 87, 2015.
[26] Nikhil Singh, François-Xavier Vialard, and Marc Niethammer. Splines for diffeomorphisms. *Medical Image Analysis*, 25(1):56–71, 2015.
[27] R. Sinkhorn. Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly*, 74:402–405, 1967.
[28] R. Tahraoui and F.-X. Vialard. Riemannian cubics on the group of diffeomorphisms and the Fisher-Rao metric. *ArXiv e-prints*, June 2016.
[29] F.-X. Vialard and A. Trouvé. Shape Splines and Stochastic Shape Evolutions: A Second Order Point of View. *Quart. Appl. Math.*, 2012.
[30] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[31] Tryphon T Georgiou Yongxin Chen, Giovanni Conforti. Measure-valued spline curves: An optimal transport viewpoint. *Preprint arXiv:1801.03186*, 2018.

INRIA, Project team Mokaplan, Université Paris-Dauphine, PSL Research University, Ceremade
Email address: jean-david.benamou@inria.fr

INRIA, Project team Mokaplan, Université Paris-Dauphine, PSL Research University, Ceremade
Email address: thomas.gallouet@inria.fr

Université Paris-Dauphine, PSL Research University, Ceremade, INRIA, Project team Mokaplan
Email address: fxvialard@normalesup.org

FROM GEODESIC EXTRAPOLATION TO A VARIATIONAL BDF2 SCHEME FOR WASSERSTEIN GRADIENT FLOWS

THOMAS O. GALLOUËT, ANDREA NATALE, AND GABRIELE TODESCHI

ABSTRACT. We introduce a time discretization for Wasserstein gradient flows based on the classical Backward Differentiation Formula of order two. The main building block of the scheme is the notion of geodesic extrapolation in the Wasserstein space, which in general is not uniquely defined. We propose several possible definitions for such an operation, and we prove convergence of the resulting scheme to the limit PDE, in the case of the Fokker-Planck equation. For a specific choice of extrapolation we also prove a more general result, that is convergence towards EVI flows. Finally, we propose a variational finite volume discretization of the scheme which numerically achieves second order accuracy in both space and time.

Keywords: Optimal transport, Wasserstein extrapolation, Wasserstein gradient flows, BDF2

MSC(2020)**:** 49Q22, 35A15, 65M08

1. INTRODUCTION

In this paper we are concerned with the construction of second-order in time discretizations for the following system of PDEs, describing the time evolution of a density $\varrho : [0,T]\times\Omega \to \mathbb{R}_+$ on a convex compact domain Ω and over the time interval $[0,T]$:

$$\partial_t \varrho - \operatorname{div}\left(\varrho \nabla \frac{\delta \mathcal{E}}{\delta \rho}(\varrho)\right) = 0 \quad \text{on } (0,T)\times\Omega\,, \tag{1.1}$$

with initial and boundary conditions:

$$\varrho(0,\cdot) = \rho_0\,, \quad \varrho \nabla \frac{\delta \mathcal{E}}{\delta \rho}(\varrho)\cdot n_{\partial\Omega} = 0 \quad \text{on } (0,T)\times\partial\Omega\,, \tag{1.2}$$

for a given initial density ρ_0, and where $n_{\partial\Omega}$ denotes the outward pointing normal to $\partial\Omega$. In equation (1.1), $\mathcal{E} : L^1(\Omega;\mathbb{R}_+) \to \mathbb{R}$ is a functional of the density and describes the energy of the system. Different choices for $\mathcal{E}$ yield different equations modeling a wide range of phenomena. Typical examples are the Fokker-Planck equation [22], the porous medium equation [32] or the Keller-Segel equation [6], but also more complex cases such as multiphase flows [10, 24, 11] or crowd motion models [36] can be considered.

Since the density satisfies the continuity equation with zero boundary flux, its total mass is conserved. Moreover, the energy decreases along the evolution:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\varrho(t,\cdot)) \le 0\,.$$

This behaviour is a consequence of the fact that system (1.1), under suitable assumptions on the energy, can be interpreted as a gradient flow in the space of probability measures $\mathcal{P}(\Omega)$ equipped with the Wasserstein distance W_2. This interpretation is well-known since the pioneering work of Jordan, Kinderlehrer and Otto [22], who showed that one recovers the

Date: October 18, 2023.

Fokker-Planck equation when following the steepest descent curve of an entropy functional with respect to the Wasserstein metric. Such result is best explained in the time-discrete setting: given a uniform decomposition $0 = t_0 < t_1 < \ldots < t_N = T$ of the interval $[0, T]$ with time step $\tau := t_{n+1} - t_n$, consider the sequence $(\rho_n)_n$ defined for $1 \leq n \leq N$ by

$$\rho_n = \operatorname*{argmin}_{\rho \in \mathcal{P}(\Omega)} \frac{W_2^2(\rho, \rho_{n-1})}{2\tau} + \mathcal{E}(\rho)\,, \tag{1.3}$$

where the energy is given by

$$\mathcal{E}(\rho) = \int_\Omega V\rho + \rho \log \rho\,, \tag{1.4}$$

with $V : \Omega \to \mathbb{R}$ being a Lipschitz function, if ρ is absolutely continuous with respect to the Lebesgue measure and $+\infty$ otherwise. Then, one can show that the discrete curve $t \mapsto \tilde{\varrho}(t)$, defined by $\tilde{\varrho}(t, \cdot) = \rho_{n-1}$ for $t \in (t_{n-1}, t_n]$ and $1 \leq n \leq N$, converges uniformly in the W_2 distance to the unique solution of the Fokker-Planck equation

$$\partial_t \varrho - \operatorname{div}(\varrho \nabla V) - \Delta \varrho = 0 \quad \text{on } (0, T) \times \Omega\,, \tag{1.5}$$

satisfying (1.2).

The numerical scheme defined in equation (1.3) is known as JKO scheme and it allows one to interpret many different models as Wasserstein gradient flows. It also provides a convenient framework both for the analysis of such models (e.g., to prove existence of solutions or exponential convergence towards steady states) [2, 35], and for the design of numerical discretizations [5, 15, 12, 26, 14]. In fact, reproducing the JKO scheme at the discrete level generally implies energy stability even in very degenerate settings. Moreover in the case of convex energies one can use robust convex optimization tools that, e.g., can easily take into account the positivity constraint on the density or even other type of strong constraints (as in the case of incompressible immiscible multiphase flows in porous media, see Section 7.3).

Since the JKO scheme is a variational version of the implicit Euler scheme, it is an order one method. Recently, several higher-order alternatives to the JKO scheme have been proposed, but it is not trivial to translate them into a fully-discrete setting (see [29, 27], and Section 1.2 below for a detailed description of such approaches). In fact, to the best of our knowledge, there exists no viable fully-discrete approach able to compute with second order accuracy general Wasserstein gradient flows while preserving (to some extent) the underlying variational structure.

In this paper we contribute to this quest by reformulating the classical multi-step scheme based on the Backward Differentiation Formula of order two (BDF2) as the composition of two inner steps: a geodesic extrapolation step, and a standard JKO step. We refer to the resulting scheme as Extrapolated Variational BDF2 (EVBDF2) scheme. As the extrapolation step is not uniquely defined (since Wasserstein geodesics may not be globally defined in time), we provide several natural notions of extrapolation and for some of these we provide convergence guarantees for the resulting scheme. For a particular choice of extrapolation, which unfortunately is not covered by our theory, we also propose a simple and efficient (space-time) discretization. Importantly, we find numerically that this does indeed produce second-order accurate solutions both in space and time.

1.1. **Description of the BDF2 approach and main results.**

In the Euclidean setting, the gradient flow associated to a smooth real-valued convex function $F : \mathbb{R}^d \to \mathbb{R}$ and a

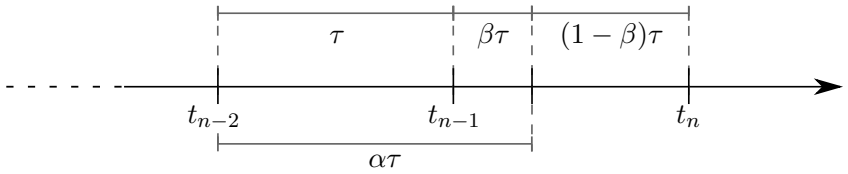


FIGURE 1. A graphical representation of the time intervals involved in the definition of the EVBDF2 scheme.

starting point $x_0 \in \mathbb{R}^d$, is the unique solution to the Cauchy problem

$$\begin{cases} x'(t) = -\nabla F(x(t))\,, & \forall\, t > 0\,, \\ x(0) = x_0\,. \end{cases} \tag{1.6}$$

The BDF2 scheme applied to such a system, with time step $\tau > 0$, can be written as follows: given $x_0, x_1 \in \mathbb{R}^d$, for $n \geq 2$ find $x_n \in \mathbb{R}^d$ satisfying

$$\frac{3}{2\tau}\Big(x_n - \frac{4}{3}x_{n-1} + \frac{1}{3}x_{n-2}\Big) = -\nabla F(x_n)\,. \tag{1.7}$$

This can be interpreted as an implicit Euler step, with starting point

$$x^\alpha_{n-1} \coloneqq x_{n-2} + \alpha(x_{n-1} - x_{n-2}) = x_{n-1} + \beta(x_{n-1} - x_{n-2})\,,$$

where $\alpha = 4/3$ and $\beta = \alpha - 1 = 1/3$, and with time step $(1-\beta)\tau = 2\tau/3$. In turn, x^α_{n-1} coincides with the Euclidean extrapolation at time α, from x_{n-2} (at time 0) to x_{n-1} (at time 1), with respect to a fictitious time variable (see Figure 1 for a graphical representation of the time intervals involved in the scheme).

In order to define a counterpart to the BDF2 scheme (1.7) for Wasserstein gradient flows, one needs to replace the Euclidean extrapolation at time $\alpha > 1$ by an analogous operation in the space of probability measures equipped with the W_2 metric. In this paper, we will represent such an operation by a map $\mathsf{E}_\alpha : \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$ (where $\mathcal{P}_2(\mathbb{R}^d)$ is the set of probability measures on $\mathbb{R}^d$ with finite second moments), which we will refer to as an α-extrapolation operator. Given such a map, we define the EVBDF2 scheme as follows: given $\rho_0, \rho_1 \in \mathcal{P}(\Omega)$, for $n \geq 2$ find $\rho_n \in \mathcal{P}(\Omega)$ satisfying

$$\rho_n \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\Omega)} \frac{W_2^2(\rho, \rho^\alpha_{n-1})}{2(1-\beta)\tau} + \mathcal{E}(\rho)\,, \quad \rho^\alpha_{n-1} = \mathsf{E}_\alpha(\rho_{n-2}, \rho_{n-1})\,, \tag{1.8}$$

where here $\mathcal{E} : \mathcal{P}(\Omega) \to \mathbb{R}$ is defined on the whole space $\mathcal{P}(\Omega)$.

The extrapolation operator E_α plays a crucial role in the scheme, but it is not trivial to propose an appropriate definition for it due to the structure of W_2 geodesics on $\mathcal{P}_2(\mathbb{R}^d)$. To clarify this, recall that a (globally length-minimizing) geodesic with respect to the W_2 metric is a curve $\omega : [t_0, t_1] \to \mathcal{P}_2(\mathbb{R}^d)$ such that

$$W_2(\omega(s_0), \omega(s_1)) = \frac{|s_1 - s_0|}{|t_1 - t_0|} W_2(\omega(t_0), \omega(t_1))\,, \tag{1.9}$$

for all $s_0, s_1 \in (t_0, t_1)$. Given two measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ there always exists a geodesic connecting the two. Furthermore, due to Brenier's theorem, supposing that μ_0 is absolutely continuous with respect to the Lebesgue measure, there exists a unique geodesic $\omega : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ such that $\omega(0) = \mu_0$ and $\omega(1) = \mu_1$, and this has a very simple expression:

$$\omega(t) = ((1-t)\mathrm{Id} + t\nabla u)_\# \mu_0\,, \tag{1.10}$$

where Id is the identity map on $\mathbb{R}^d$ and $u : \mathbb{R}^d \to \mathbb{R}$ is a convex function. This means that particles travel on straight lines along the interpolation, without colliding into each other. However, for a given $\alpha > 1$, there may exist no geodesic defined on $[0, \alpha]$ that coincide on $[0, 1]$ with ω. This is because following their straight trajectories particles may collide immediately after time $t = 1$, even if both μ_0 and μ_1 have smooth and strictly positive densities. This means that one cannot use such geodesic extensions to define the extrapolation operator E_α in a unique way. Therefore, instead of focusing on a particular definition, we only require a uniform stability bound on the extrapolation which we will need to prove the convergence of the scheme. In particular, we will focus on extrapolation operators that are dissipative in the following sense:

Definition 1.1 (Dissipative extrapolations)**.** An extrapolation operator E_α is θ-dissipative if it satisfies

$$W_2(\mu_1, \mathsf{E}_\alpha(\mu_0, \mu_1)) \leq \theta W_2(\mu_0, \mu_1)\,, \tag{1.11}$$

for any $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and for a constant $\theta \geq 0$.

Note that by equation (1.9), if the extrapolation is consistent with the geodesic extension when this exists, then we must have $\theta \geq \alpha - 1 =: \beta$. Upon adding a further consistency assumption on the extrapolation given in equation (1.12) below (see Remark 3.6 for more comments on the role of our main assumptions), we can establish the following convergence result:

Theorem 1.2. *Let $\rho_0 \in \mathcal{P}(\Omega)$ and $\mathcal{E}$ given by (1.4). For any given $N \geq 1$, let $(\rho_n)_{n=0}^N$ be the discrete solution defined by the scheme (1.8) for given $\rho_1 \in \mathcal{P}(\Omega)$ (dependent on N), with time step $\tau = T/N$, and with E_α being a θ-dissipative extrapolation operator with $0 \leq \beta = \alpha - 1 < 1$ and $\theta < 1/2$, and such that for all $\mu_0, \mu_1 \in \mathcal{P}(\Omega)$ and $\varphi \in C_c^\infty(\mathbb{R}^d)$ verifying $\nabla\varphi \cdot n_{\partial\Omega} = 0$ on $\partial\Omega$,*

$$\left| \int_{\mathbb{R}^d} \varphi \left(\mathsf{E}_\alpha(\mu_0, \mu_1) - \alpha\mu_1 + \beta\mu_0 \right) \right| \leq C_\varphi W_2^2(\mu_0, \mu_1)\,, \tag{1.12}$$

where $C_\varphi > 0$ only depends on α, φ and Ω. Suppose that $W_2^2(\rho_0, \rho_1) \leq C\tau$, for a constant $C > 0$ independent of τ, and that $\mathcal{E}(\rho_1) \leq \mathcal{E}(\rho_0)$. Then, the curve $t \mapsto \tilde{\rho}_\tau(t)$ defined by $\tilde{\rho}_\tau(t) \coloneqq \rho_{n-1}$ for all $t \in (t_{n-1}, t_n]$ and $1 \leq n \leq N$, converges as $N \to \infty$, uniformly in the W_2 distance, to a distributional solution to the Fokker-Planck equation on $[0, T] \times \Omega$ and initial conditions given by ρ_0.

Of course, in order to achieve second order accuracy, we must set $\alpha = 4/3$ and require in addition that, if there exists a geodesic $\omega : [0, \alpha] \to \mathcal{P}(\Omega)$ such that $\omega|_{[0,1]}$ is a geodesic from μ_0 to μ_1, then $\mathsf{E}_\alpha(\mu_0, \mu_1)$ must coincide with $\omega(\alpha)$. Importantly, we will show that there exist several different ways to define such an operator, providing therefore different convergent approaches. We highlight that there is no inconsistency between the scheme (1.8), defined on $\mathcal{P}(\Omega)$, and an extrapolation operator E_α valued in $\mathcal{P}_2(\mathbb{R}^d)$. In fact, both for theoretical or numerical reasons, one may be led to define an extrapolation operator on the whole space to avoid issues with the boundary of Ω. Nevertheless, scheme (1.8) is well-defined and, as long as the consistency assumption (1.12) is satisfied, the convergence result of Theorem 1.2 holds.

One approach for producing an operator E_α, which enjoys a particularly rich structure, consists in reproducing the variational characterization of the linear extrapolation in the metric setting. Given two points $x_0, x_1 \in \mathbb{R}^d$, the Euclidean extrapolation at time α from x_0

to x_1 is the point $x_\alpha = \alpha x_1 - \beta x_0$ with $\beta = \alpha - 1$. This can be obtained as the unique solution to

$$x_\alpha = \operatorname*{argmin}_{x\in\mathbb{R}^d} \alpha|x - x_1|^2 - \beta|x - x_0|^2 . \tag{1.13}$$

Similarly, we define the metric extrapolation in the Wasserstein space as follows:

$$\mathsf{E}_\alpha(\mu_0, \mu_1) \coloneqq \operatorname*{argmin}_{\rho\in\mathcal{P}_2(\mathbb{R}^d)} \alpha W_2^2(\rho, \mu_1) - \beta W_2^2(\rho, \mu_0) . \tag{1.14}$$

Problem (1.14) is not a convex optimization problem in the classical sense. To see this, consider the following simple counterexample. In dimension $d = 1$, take

$$\mu_0 = (\delta_{-1} + \delta_1)/2, \quad \mu_1 = \delta_0, \quad \nu_0 = \delta_{-1}, \quad \nu_1 = \delta_1.$$

Along the interpolation $\nu(t) = (1 - t)\nu_0 + t\nu_1$, the first term of the functional in (1.14) is constant whereas the second one is concave. Nonetheless, we will show that problem (1.14) always admits a unique solution (see Proposition 4.10) and it also satisfies the assumptions in Theorem 1.2. Furthermore, exploiting the variational formulation of the metric extrapolation (1.14), we can prove a more general convergence result using the Evolution Variational Inequality (EVI) characterization of gradient flows in metric spaces. More precisely, we prove the following result:

Theorem 1.3. *Let $\rho_0 \in \mathcal{P}(\Omega)$ and $\mathcal{E} : \mathcal{P}(\Omega) \to \mathbb{R}$ being a λ-convex energy in the generalized geodesic sense, for $\lambda \in \mathbb{R}_+$. For any given $N \geq 1$, let $(\rho_n)_{n=0}^N$ be the discrete solution defined by the scheme (1.8) for given $\rho_1 \in \mathcal{P}(\Omega)$ (dependent on N), with time step $\tau = T/N$, and with E_α being the metric extrapolation (1.14) with $\beta = \alpha - 1$. Suppose that $W_2^2(\rho_0, \rho_1) \leq C\tau$, for a constant $C > 0$ independent of τ, and that $\mathcal{E}(\rho_1) \leq \mathcal{E}(\rho_0)$. Then, the curve $t \mapsto \tilde{\rho}_\tau(t)$ defined by $\tilde{\rho}_\tau(t) \coloneqq \rho_{n-1}$ for $t \in (t_{n-1}, t_n]$ and $1 \leq n \leq N$, converges as $N \to \infty$, uniformly in the W_2 distance, to the unique absolutely continuous curve $\varrho : [0, T] \to \mathcal{P}(\Omega)$ satisfying $\varrho(0) = \rho_0$ and such that for any $\nu \in \mathcal{P}(\Omega)$ it holds*

$$\frac{d}{dt}\frac{1}{2}W_2^2(\varrho(t), \nu) \leq \mathcal{E}(\nu) - \mathcal{E}(\varrho(t)) - \frac{\lambda}{2}W_2^2(\varrho(t), \nu), \quad \forall t \in (0, T) .$$

Remarkably, problem (1.14) admits a convex dual formulation, see Remark 4.14.

1.2. **Relation with previous works and numerical implementation issues.** Going back to the discretization of system (1.6), each step of the BDF2 scheme (1.7) can also be obtained as the optimality conditions of the following problem:

$$x_n = \operatorname*{argmin}_{x\in\mathbb{R}^d} \alpha\frac{|x - x_{n-1}|^2}{2(1 - \beta)\tau} - \beta\frac{|x - x_{n-2}|^2}{2(1 - \beta)\tau} + F(x) . \tag{1.15}$$

This suggests defining a similar formulation in Wasserstein space as follows

$$\rho_n \in \operatorname*{argmin}_{\rho\in\mathcal{P}(\Omega)} \alpha\frac{W_2^2(\rho, \rho_{n-1})}{2(1 - \beta)\tau} - \beta\frac{W_2^2(\rho, \rho_{n-2})}{2(1 - \beta)\tau} + \mathcal{E}(\rho) . \tag{1.16}$$

This approach has been proposed by Matthes and Plazotta [29, 33], who proved equivalent versions of Theorem 1.2 and 1.3. Even if in the Euclidean setting the analogue problems to (1.16) and (1.8) yield the same solutions, one can check that this is not the case in the Wasserstein space (see, e.g., the example in Figure 2). However, just as for the metric extrapolation problem (1.14), (1.16) is not a convex optimization problem in the classical sense. For this reason, it is not easy to provide a numerical implementation of (1.16) when $d \geq 2$. The same is true for the EVBDF2 scheme (1.8) when using the metric extrapolation. Nonetheless,

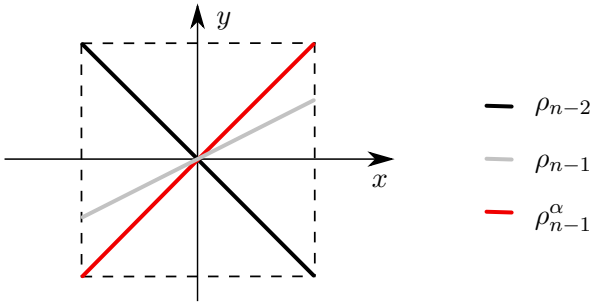


FIGURE 2. An example for which the schemes (1.8) and (1.16) provide different results, e.g., for the energy given by the convex indicator function of the set $\{\mu : \mu(\mathbb{R}^d \setminus \{x = 0\}) = 0\}$. In the figure ρ_{n-2}, ρ_{n-1} and ρ^{α}_{n-1} are uniformly distributed on the segments $(t, -t)$, $(t, (1-\beta)t/\alpha)$ and (t, t) for $t \in [-1, 1]$, respectively (in this case the geodesic from ρ_{n-2} to ρ_{n-1} on the time interval $[0, 1]$ can be extendend up to time α, yielding ρ^{α}_{n-1}). For the scheme (1.8) the measure ρ_n is uniformly distributed on the segment $(0, t)$ for $t \in [-1, 1]$, whereas for the scheme (1.16) the measure ρ_n can be obtained as the extrapolation of the projections of ρ_{n-2} and ρ_{n-1} on the axis y, and can be shown to have a strictly smaller support.

the advantage of using the EVBDF2 scheme is that one has some freedom in choosing the extrapolation operator, which makes it more amenable to computations.

Another second-order variation of the JKO scheme was proposed by Legendre and Turinici [27], and it is based on the implicit midpoint rule, which applied to system (1.6) leads to the scheme: for $n \geq 1$ find $x_n \in \mathbb{R}^d$ satisfying

$$\frac{1}{\tau}(x_n - x_{n-1}) = -\nabla F\Big(\frac{x_n + x_{n-1}}{2}\Big)\,,$$

which can be obtained as the optimality conditions of the problem

$$x_n = \operatorname*{argmin}_{x \in \mathbb{R}^d} \frac{|x - x_{n-1}|^2}{2\tau} + 2F\Big(\frac{x + x_{n-1}}{2}\Big)\,. \tag{1.17}$$

Translating such a scheme to the Wasserstein setting yields the Variational Implicit Midpoint (VIM) scheme proposed in [27]: for $n \geq 1$ find $\rho_n \in \mathcal{P}(\Omega)$ satisfying

$$\rho_n \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\Omega)} \frac{W_2^2(\rho, \rho_{n-1})}{2\tau} + 2\mathcal{E}(\rho_{n-1/2})\,, \tag{1.18}$$

where $\rho_{n-1/2}$ is the midpoint of the (not necessarily unique) geodesic between ρ and ρ_{n-1}. Also in this case, it is not evident how to implement such a scheme, as it requires an explicit formula for the midpoint given the initial and final measures. This may also lead to convexity issues. Notice however that in the same spirit of our formulation of the BDF2 scheme, the implicit midpoint scheme can be formulated in the following alternative way: for $n \geq 1$ find $\rho_n \in \mathcal{P}(\Omega)$ satisfying

$$\rho_n = \mathsf{E}_2(\rho_{n-1}, \rho_{n-1/2})\,, \quad \rho_{n-1/2} \in \operatorname*{argmin}_{\rho \in \mathcal{P}(\Omega)} \frac{W_2^2(\rho, \rho_{n-1})}{\tau} + \mathcal{E}(\rho)\,, \tag{1.19}$$

where $\mathsf{E}_2(\rho_{n-1}, \rho_{n-1/2})$ denotes the extrapolation at time $\alpha = 2$ of a geodesic from ρ_{n-1} (at time 0) to $\rho_{n-1/2}$ (at time 1). In general, this leads to a different discrete solution than the

one obtained with (1.18), although the two schemes coincide if there exists a unique geodesic extension from ρ_{n-1} to $\rho_{n-1/2}$ which stays globally length-minimizing up to time 2 for all n. Nevertheless, the behavior of scheme (1.19) is radically different from that of the EVBDF2 (1.8), due to the different way JKO steps and extrapolations are performed. Namely, the order of the operations as well as the length of the steps play a crucial role. We will investigate this phenomenon numerically by considering a fully-discrete version of the VIM scheme and show that in general this approach may lead to persistent oscillations in the solution (Section 7.1).

Providing a fully discrete version of problem (1.1), via the EVBDF2 scheme (1.8), comes with an additional challenge since the chosen space discretization should also be second-order accurate in space, in order to exploit the increased accuracy of the time discretization. We propose a discretization in the Eulerian framework of finite volumes. Specifically, we implement Two Point Flux Approximation (TPFA) finite volumes, which have been extensively analyzed lately for the discretization of optimal transport and Wasserstein gradient flows [21, 17, 31, 12, 30]. Following these last two works in particular, we propose a scheme in which the Wasserstein distance is locally linearized, at each step of the scheme, in order to decrease the computational complexity of the approach, without dropping the second-order accuracy in time. In addition, we propose one possible discrete version of the extrapolation in this setting, which can be implemented in a robust way, and we verify numerically the second-order accuracy of the resulting approach.

We stress that the space discretization of the EVBDF2 scheme that we propose, even if maintaining its variational structure, relies on substantial simplifications of the original problem. As a consequence, our theoretical results do not apply directly, and further work is required for a fully discrete convergence proof. Given this, the numerical results presented in Section 7 are only preliminary and they are mainly meant to demonstrate the feasibility of the approach.

2. Preliminaries and notation

Let $\mathcal{P}_2(\mathbb{R}^d)$ be the space of probability measures with finite second moments. Given $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, we denote by $W_2(\mu_0, \mu_1)$ the L^2-Wasserstein distance between μ_0 and μ_1 (see, e.g., Chapter 5 in [34]). This can be defined via the following minimization problem:

$$W_2^2(\mu_0, \mu_1) \coloneqq \min_{\gamma \in \Pi(\mu_0, \mu_1)} \int |x - y|^2 \, \mathrm{d}\gamma(x, y) \,, \tag{2.1}$$

where $\Pi(\mu_0, \mu_1)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ_0 and μ_1. This problem always admits a solution γ^*, although it is not necessarily unique, which we refer to as an optimal transport plan from μ_0 to μ_1. By linearity of the constraint and of the function minimized in (2.1), one can easily check that the function W_2^2 is jointly convex with respect to its arguments (with respect to the linear structure of $\mathcal{P}_2(\mathbb{R}^d)$). We will refer to the space of probability measures $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the metric W_2 as the Wasserstein space.

Problem (2.1) admits an alternative dynamical formulation, which was introduced by Benamou and Brenier in [4], and which reads as follows:

$$W_2^2(\mu_0, \mu_1) = (t_1 - t_0) \min_{(\omega, v) \in \mathcal{C}} \int_{t_0}^{t_1} \mathrm{d}t \int \omega(t) |v(t, \cdot)|^2 \tag{2.2}$$

where $\mathcal{C}$ is the set of curves (ω, v) with finite total kinetic energy, with $\omega : [t_0, t_1] \to \mathcal{P}_2(\mathbb{R}^d)$ and $v : [t_0, t_1] \to L^2(\omega(t); \mathbb{R}^d)$, satisfying weakly the continuity equation

$$\partial_t \omega + \operatorname{div}(\omega v) = 0 \tag{2.3}$$

with zero flux boundary conditions (i.e. $\omega\, v \cdot n_{\partial\Omega} = 0$), and initial and final conditions $\omega(t_0) = \mu_0$, $\omega(t_1) = \mu_1$. The minimum in (2.2) is always achieved although there might be multiple minimizers. In particular, one can use formula (2.2) to deduce that the Wasserstein space is a geodesic space and the minimizers ω are geodesics.

By the optimality conditions of problem (2.2), a curve ω is a geodesic if and only if there exists a potential $\phi : [t_0, t_1] \times \mathbb{R}^d \to \mathbb{R}$ that verifies:

(1) $\phi(t_0, \cdot)$ is a continuous $(-(t_1 - t_0)^{-1})$-convex function, i.e. such that the so-called Brenier potential

$$x \mapsto u(x) \coloneqq (t_1 - t_0)\phi(t_0, x) + \frac{|x|^2}{2} \quad \text{is convex}\,; \tag{2.4}$$

(2) the potential ϕ is the unique viscosity solution of the Hamilton-Jacobi equation

$$\partial_t \phi + \frac{|\nabla \phi|^2}{2} = 0\,, \tag{2.5}$$

or equivalently, it verifies the Hopf-Lax representation formula,

$$\phi(t, x) = \inf_{y \in \mathbb{R}^d} \frac{|x - y|^2}{2(t - t_0)} + \phi(t_0, y)\,; \tag{2.6}$$

(3) $\nabla\phi(t, \cdot) \in L^2(\omega(t); \mathbb{R}^d)$ for a.e. $t \in [t_0, t_1]$ and $(\omega, \nabla\phi) \in \mathcal{C}$.

We say that a function ϕ verifying these condition is an optimal potential from μ_0 to μ_1 on the time interval $[t_0, t_1]$. Furthermore, for any optimal potential ϕ, it holds:

$$\frac{W_2^2(\mu_0, \mu_1)}{2(t_1 - t_0)} = \int \phi(t_1, \cdot)\mu_1 - \int \phi(t_0, \cdot)\mu_0\,. \tag{2.7}$$

Because of the semi-convexity of $\phi(t_0, \cdot)$, the maps $X(t, \cdot)$, defined a.e. by

$$X(t, \cdot) \coloneqq \mathrm{Id} + (t - t_0)\nabla\phi(t_0, \cdot) \tag{2.8}$$

are injective for all $t \in [t_0, t_1)$ (as the gradient of a strongly convex function), and the resulting curve of maps $X : [t_0, t_1] \times \mathbb{R}^d \to \mathbb{R}^d$ is the Lagrangian flow of the time-dependent vector field $\nabla\phi(t, \cdot)$, i.e., for a.e. $x \in \mathbb{R}^d$, $X(\cdot, x)$ solves the flow equation

$$\frac{\mathrm{d}}{\mathrm{d}t} X(t, x) = \nabla\phi(t, X(t, x)), \quad X(t_0, x) = x\,.$$

If μ_0 is absolutely continuous, given an optimal potential ϕ and the associated Lagrangian flow X defined by (2.8), one can easily verify that the curve

$$\omega(t) = X(t, \cdot)_{\#}\mu_0 \tag{2.9}$$

solves the continuity equation with velocity $\nabla\phi$ and boundary conditions $\omega(0) = \mu_0$ and $\omega(1) = \mu_1$ (in distributional sense), and therefore it is a geodesic. Moreover, using the absolute continuity of μ_0, one can also show that the initial potential $\phi(t_0, \cdot)$ is uniquely defined μ_0-a.e., and no other geodesic curve exists connecting μ_0 and μ_1. Note also that from (2.9), one can recover Brenier's result (1.10) with the Brenier potential u as in (2.4), and also verify the equivalence with formulation (2.1). As a matter of fact, in this case the optimal transport plan is also unique and is given by $\gamma^* = (\mathrm{Id}, \nabla u)_{\#}\mu_0$, where the map ∇u is the so-called optimal transport map from μ_0 to μ_1. On the other hand, for any convex function u, setting $\phi(0, \cdot)$ via (2.4), the curve ω defined in (2.9) is a geodesic between μ_0 and $(\nabla u)_{\#}\mu_0$ (and the unique one, if μ_0 is absolutely continuous).

3. Analysis of the EVBDF2 scheme

In this section we collect the main properties of the EVBDF2 discretization (1.8), and in particular we prove Theorem 1.2, which establishes the convergence of the discrete flow generated by the scheme to the linear Fokker-Planck equation. Throughout the section, $(\rho_n)_n$ denotes a sequence of measures generated by the EVBDF2 scheme (1.8), where E_α is a θ-dissipative extrapolation, with $\theta < 1/2$.

3.1. Well-posedness and classical estimate.

We start by stating some a priori bounds, which are valid for a general class of energies. In particular, in this paragraph, we only assume that $\mathcal{E}$ is lower semi-continuous with respect to the weak-* topology. Since $\mathcal{P}(\Omega)$ is compact for this topology (we recall that we assume Ω compact) this also implies that $\mathcal{E}$ is bounded from below. Problem (1.8) therefore admits a minimizer at each step n.

Lemma 3.1. *At each step n, the solution ρ_n satisfies the following inequality*

$$(1-\theta)\frac{W_2^2(\rho_n,\rho_{n-1})}{2(1-\beta)\tau} + \mathcal{E}(\rho_n) \le \theta\frac{W_2^2(\rho_{n-1},\rho_{n-2})}{2(1-\beta)\tau} + \mathcal{E}(\rho_{n-1})\,. \tag{3.1}$$

Proof. Due to the optimality of ρ_n and using (1.11), we can write

$$\begin{aligned}\frac{W_2^2(\rho_n,\rho_{n-1}^\alpha)}{2(1-\beta)\tau} + \mathcal{E}(\rho_n) &\le \frac{W_2^2(\rho_{n-1},\rho_{n-1}^\alpha)}{2(1-\beta)\tau} + \mathcal{E}(\rho_{n-1})\\ &\le \frac{\theta^2}{2(1-\beta)\tau}W_2^2(\rho_{n-1},\rho_{n-2}) + \mathcal{E}(\rho_{n-1})\,.\end{aligned}$$

If $\theta = 0$ this coincides with (3.1). If $\theta > 0$, observe that by the triangular and Young's inequalities, for any $c > 0$,

$$W_2^2(\rho_n,\rho_{n-1}) \le \Big(1+\frac{1}{c}\Big)W_2^2(\rho_n,\rho_{n-1}^\alpha) + (1+c)W_2^2(\rho_{n-1},\rho_{n-1}^\alpha)\,.$$

Setting $c = \theta^{-1}-1$ in this last inequality and using again (1.11), we can estimate the left-hand side from below using

$$\begin{aligned}\frac{W_2^2(\rho_n,\rho_{n-1}^\alpha)}{2(1-\beta)\tau} &\ge \frac{1}{2(1-\beta)\tau}\left(\frac{c}{c+1}W_2^2(\rho_n,\rho_{n-1}) - cW_2^2(\rho_{n-1},\rho_{n-1}^\alpha)\right)\\ &\ge \frac{1-\theta}{2(1-\beta)\tau}W_2^2(\rho_n,\rho_{n-1}) - \frac{(1-\theta)\theta}{2(1-\beta)\tau}W_2^2(\rho_{n-1},\rho_{n-2})\,.\end{aligned}$$

Rearranging, we obtain (3.1). □

Note that if we take $\beta = 0$, i.e. we remove the extrapolation step, we can take $\theta = 0$ in (3.1) and recover the standard dissipation estimate for the JKO scheme.

Lemma 3.2. *Let $C_1 > 0$ be a constant such that $W_2^2(\rho_1,\rho_0) \le C_1\tau$ and $\mathcal{E}(\rho_1) \le \mathcal{E}(\rho_0)$. Then, it holds:*

$$\frac{1}{\tau}\sum_{n=0}^{N_\tau} W_2^2(\rho_n,\rho_{n-1}) \le C \tag{3.2}$$

for a constant $C > 0$ depending only on C_1, β, θ, $\mathcal{E}$ and ρ_0.

Proof. Summing over n the inequality (3.1) we obtain

$$\frac{1-2\theta}{2(1-\beta)\tau}\sum_{n=0}^{N} W_2^2(\rho_n,\rho_{n-1}) \le \mathcal{E}(\rho_1) - \mathcal{E}(\rho_n) + \frac{\theta}{(1-\beta)\tau}W_2^2(\rho_1,\rho_0)\,, \tag{3.3}$$

Then, since $\theta < 1/2$ and thanks to the lower bound on the energy and the assumption $\mathcal{E}(\rho_1) \leq \mathcal{E}(\rho_0)$, we have

$$\frac{1}{\tau}\sum_{n=0}^{N} W_2^2(\rho_n, \rho_{n-1}) \leq \frac{2(1-\beta)}{1-2\theta}\Big(\mathcal{E}(\rho_0) - \inf \mathcal{E}\Big) + \frac{2\theta}{1-2\theta} C_1 \,.$$

□

Remark 3.3. *For a given ρ_0, one can always choose ρ_1 so that the constant C_1 above is independent of τ and $\mathcal{E}(\rho_1) \leq \mathcal{E}(\rho_0)$, which are also the assumptions in the statements of Theorems 1.2 and 1.3. For example, it is sufficient to take ρ_1 as the solution obtained after a finite number $N_0 \in \mathbb{N}$ of JKO steps with time step τ/N_0 and initial condition given by ρ_0, with $\mathcal{E}(\rho_0) < \infty$. In fact, in this case, by the same proof as for Lemma 3.2 (with $\beta = \theta = 0$), one can take $C_1 = 2(\mathcal{E}(\rho_0) - \inf \mathcal{E})$.*

3.2. **Convergence towards the Fokker-Planck equation.** Given a Lipschitz continuous exterior potential $V \in W^{1,\infty}(\Omega)$, the Fokker-Planck equation is given by

$$\partial_t \varrho = \Delta \varrho + \operatorname{div}(\varrho \nabla V) \quad \text{in } (0,T)\times\Omega \,, \tag{3.4}$$

complemented with no-flux boundary conditions $(\nabla \varrho + \varrho \nabla V)\cdot n_{\partial\Omega} = 0$ on $\partial\Omega$ and an initial condition $\varrho(0,\cdot) = \rho_0 \in \mathcal{P}(\Omega)$. Equation (3.4) can be interpreted as a Wasserstein gradient flow with respect to the energy functional $\mathcal{E} : \mathcal{P}(\Omega) \to \mathbb{R}$ given by

$$\mathcal{E}(\rho) = \mathcal{U}(\rho) + \int_\Omega \rho V \,, \tag{3.5}$$

where the internal energy $\mathcal{U} : \mathcal{P}(\Omega) \to \mathbb{R}$ (the entropy) is defined by

$$\mathcal{U}(\rho) \coloneqq \begin{cases} \displaystyle\int_\Omega \log\left(\frac{\mathrm{d}\rho}{\mathrm{d}x}\right)\mathrm{d}\rho & \text{if } \rho \ll \mathrm{d}x \llcorner \Omega \,, \\ +\infty & \text{otherwise} \,, \end{cases} \tag{3.6}$$

where $\mathrm{d}x \llcorner \Omega$ denotes the restriction of the Lebesgue measure to the domain Ω. Since the function $x \mapsto x \log x$ is strictly convex and superlinear, the energy $\mathcal{E}$ is also strictly convex on its domain (with respect to the linear structure of $\mathcal{P}(\Omega)$) and lower semi-continuous (with respect to the weak-* topology: see, e.g., Proposition 7.7 in [34]). Since W_2^2 is continuous and convex in its arguments, there exists a unique solution ρ_n to problem (1.8) at each step n, and this is furthermore absolutely continuous with respect to $\mathrm{d}x \llcorner \Omega$. Moreover, both Lemmas 3.1 and 3.2 apply.

As in the previous paragraph, we assume that E_α is a θ-dissipative extrapolation with $\theta < 1/2$, and $(\rho_n)_n$ denotes a sequence of measures generated by the associated EVBDF2 scheme (1.8). Although the discrete flow does not move by strictly minimizing the energy at each step (see Lemma 3.1), we will show that it converges to the maximal slope curve of $\mathcal{E}$. For this, we will rely on the same arguments as in the original work of Jordan, Kinderlehrer, and Otto [22] for the JKO scheme.

Relying on the estimate (3.2), the compactness arguments for obtaining a limit curve are rather standard. We introduce two density curves on the interval $[0,T]$, given by

$$\begin{aligned} \varrho_\tau(t) &= \sum_{n=1}^{N} \rho_{n-1} \mathbb{1}_{(t_{n-1},t_n]}\,, \quad \rho_\tau(0) = \rho_0\,, \\ \tilde{\varrho}_\tau(t) &= \sum_{n=1}^{N} \tilde{\varrho}_n(t) \mathbb{1}_{(t_{n-1},t_n]}\,, \quad \tilde{\rho}_\tau(0) = \rho_0\,, \end{aligned} \tag{3.7}$$

with $t \mapsto \tilde{\varrho}_n(t)$ being the geodesic curve between ρ_{n-1} and ρ_n on the time interval $[t_{n-1},t_n]$ (i.e. the minimizer of problem (2.2) on this interval). Let $\tilde{v}_n$ be the associated optimal vector field as in problem (2.2) for all $1 \le n \le N$. By definition of $\tilde{\varrho}_\tau$, we have that

$$\partial_t \tilde{\varrho}_\tau + \operatorname{div}(\tilde{\varrho}_\tau \tilde{v}_\tau) = 0$$

in the distributional sense on $(0,T)\times\Omega$, where $\tilde{v}_\tau$ is the vector field defined by $\tilde{v}_\tau|_{(t_{n-1},t_n]} = \tilde{v}_n$ for all $1 \le n \le N$. Moreover, on each interval $[t_{n-1},t_n]$ it holds:

$$W_2^2(\rho_n,\rho_{n-1}) = \tau \int_{t_{n-1}}^{t_n} \int_\Omega \tilde{\varrho}_\tau |\tilde{v}_\tau|^2\,.$$

The curve ϱ_τ is a piecewise constant measure-valued curve whereas $\tilde{\varrho}_\tau$ is a (absolutely) continuous one, interpolating the discrete densities.

Proposition 3.4. *For a given ρ_0 and any given $N \ge 1$, let ρ_τ be the curve defined as in equation (3.7), with ρ_1 being such that $W_2^2(\rho_0,\rho_1) \le C\tau$, for a constant $C > 0$ independent of τ, and $\mathcal{E}(\rho_1) \le \mathcal{E}(\rho_0)$. Then, the sequence $(\varrho_\tau)_\tau$ converges uniformly in the W_2 distance to an absolutely continuous curve $\varrho : [0,T] \to \mathcal{P}(\Omega)$.*

Proof. The sequence of curves $(\tilde{\varrho}_\tau)_{\tau\in\mathbb{R}_+}$, defined from $[0,T]$ to the (compact) space $\mathcal{P}(\Omega)$ equipped with the Wasserstein distance, is uniformly Hölder continuous. Indeed, for any $r,s \in (0,T], s > r$, denote N_r, N_s the two integers such that $r \in (t_{N_r}, t_{N_r+1}], s \in (t_{N_s}, t_{N_s+1}]$. By the dynamical formulation of the Wasserstein distance (2.2), it holds

$$\begin{aligned} W_2(\tilde{\varrho}_\tau(s),\tilde{\varrho}_\tau(r)) &\le |s-r|^{\frac12} \left(\int_r^s \int_\Omega \tilde{\varrho}_\tau |\tilde{v}_\tau|^2 \right)^{\frac12} \le |s-r|^{\frac12} \left(\sum_{n=N_r}^{N_s} \int_{t_n}^{t_{n+1}} \int_\Omega \tilde{\varrho}_\tau |\tilde{v}_\tau|^2 \right)^{\frac12} \\ &= |s-r|^{\frac12} \left(\sum_{n=N_r}^{N_s} \frac{1}{\tau} W_2^2(\rho_n,\rho_{n+1}) \right)^{\frac12} \le C|s-r|^{\frac12} \end{aligned} \tag{3.8}$$

where in the last inequality we used the estimate (3.2). By the generalized Ascoli-Arzelà theorem, the sequence converges uniformly in W_2, up to a subsequence, to a limit curve ϱ. As the inequality (3.8) passes to the limit, ϱ is also an absolutely continuous curve with respect to the Wasserstein metric. Finally, for any $r \in [0,T]$,

$$W_2(\varrho_\tau(r),\tilde{\varrho}_\tau(r)) = W_2(\tilde{\varrho}_\tau(t_{N_r}),\tilde{\varrho}_\tau(r)) \le \sqrt{\tau} \left(\int_{t_{N_r}}^{t_{N_r+1}} \int_\Omega \tilde{\varrho}_\tau |\tilde{v}_\tau|^2 \right)^{1/2} \le C\sqrt{\tau}\,,$$

by the same computations. Therefore, 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 will rely on the optimality conditions of the minimization problem in (1.8), which is equivalent to a single JKO step. Consider an absolutely continuous measure ρ and a smooth vector field ξ tangent to the boundary of Ω. We define ω as the absolutely continuous curve solution to

$$\partial_s \omega + \operatorname{div}(\omega \xi) = 0\,, \quad \text{in } (-\delta, \delta) \times \Omega\,, \quad \omega(0) = \rho\,, \tag{3.9}$$

for $\delta > 0$. The variations of the energy and the Wasserstein distance along curves defined in this way can be computed explicitly as follows.

Lemma 3.5. *Consider two measures $\rho \in \mathcal{P}(\Omega)$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$, with ρ absolutely continuous, and denote by γ the optimal transport plan from ρ to ν. For any $\xi \in C_c^\infty(\mathbb{R}^d; \mathbb{R}^d)$ with $\xi \cdot n_{\partial\Omega} = 0$ on $\partial\Omega$, let ω be the curve of measures defined by (3.9) with $\omega(0) = \rho$. It holds:*

$$\frac{dW_2^2(\omega(s), \nu)}{ds}\Big|_{s=0} = 2 \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - y) \cdot \xi(x)\, d\gamma(x, y)\,, \tag{3.10}$$

$$\frac{d\mathcal{E}(\omega(s))}{ds}\Big|_{s=0} = -\int_\Omega \operatorname{div}(\xi(x)) \mathrm{d}\rho(x) + \int_\Omega \nabla V(x) \cdot \xi(x) \mathrm{d}\rho(x)\,. \tag{3.11}$$

Proof. See [2, Corollary 10.2.7] and [38, Theorem 5.30]. □

We are now ready to prove Theorem 1.2 which states the convergence of the sequence of curves $(\varrho_\tau)_\tau$ towards a distributional solution of equation (3.4). Specifically, we need to prove that, for all $\varphi \in C_c^\infty([0, T) \times \mathbb{R}^d)$ such that $\nabla\varphi \cdot n_{\partial\Omega} = 0$ on $\partial\Omega$, the limit curve ϱ satisfies:

$$-\int_0^T \int_\Omega \partial_t \varphi \varrho - \int_\Omega \varphi(0)\varrho(0) - \int_0^T \int_\Omega \Delta\varphi \varrho + \int_0^T \int_\Omega \nabla V \cdot \nabla\varphi \varrho = 0\,. \tag{3.12}$$

Proof of Theorem 1.2. Let us define for all $\rho \in \mathcal{P}(\Omega)$,

$$\mathcal{G}(\rho_{n-1}, \rho_{n-2}; \rho) \coloneqq \frac{W_2^2(\rho, \rho_{n-1}^\alpha)}{2(1-\beta)\tau} + \mathcal{E}(\rho)\,, \tag{3.13}$$

which is minimized by ρ_n, by the definition of the scheme (1.8). Consider a smooth function $\varphi \in C_c^\infty([0, T) \times \mathbb{R}^d)$ such that $\nabla\varphi \cdot n_{\partial\Omega} = 0$ on $\partial\Omega$. We define the sequence $(\varphi_n)_n \subset C_c^\infty(\mathbb{R}^d)$ as $\varphi_n = \varphi(t_n, \cdot)$. Consider then a curve ω defined as in (3.9) with $\omega(0) = \rho_n$ and $\xi = \nabla\varphi_{n-2}$. Denoting by γ_n the optimal transport plan from ρ_n to ρ_{n-1}^α, and using (3.10)-(3.11) as well as the optimality of ρ_n, we obtain

$$\begin{aligned} \frac{\mathrm{d}\mathcal{G}(\rho_{n-1}, \rho_{n-2}; \omega(s))}{\mathrm{d}s}\Big|_{s=0} &= \frac{1}{(1-\beta)\tau} \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - x_\alpha) \cdot \nabla\varphi_{n-2}(x) \mathrm{d}\gamma_n(x, x_\alpha) \\ &\quad - \int_\Omega \Delta\varphi_{n-2}(x) \mathrm{d}\rho_n(x) + \int_\Omega \nabla V(x) \cdot \nabla\varphi_{n-2}(x) \mathrm{d}\rho_n(x) = 0\,. \end{aligned} \tag{3.14}$$

Thanks to Proposition 3.4 and the regularity of φ, we immediately have

$$\left| \sum_{n=2}^N \tau \left(-\int_\Omega \Delta\varphi_{n-2}\rho_n + \int_\Omega \nabla V \cdot \nabla\varphi_{n-2}\rho_n \right) - \left(-\int_0^T \int_\Omega \Delta\varphi\varrho + \int_0^T \int_\Omega \nabla V \cdot \nabla\varphi\varrho \right) \right| \longrightarrow 0\,,$$

for $\tau \to 0$. In order to prove that the measure ϱ is a distributional solution of equation (3.4) we need to show that

$$I_1 \coloneqq \left| \sum_{n=2}^N \frac{1}{1-\beta} \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - x_\alpha) \cdot \nabla\varphi_{n-2}(x) \mathrm{d}\gamma_n(x, x_\alpha) - \left(-\int_0^T \int_\Omega \partial_t\varphi\varrho - \int_\Omega \varphi(0)\varrho(0) \right) \right| \longrightarrow 0\,,$$

as well. We can bound the latter quantity as $I_1 \leq I_2 + I_3$, where $I_2 = \sum_{n=2}^N I_2^n$ with

$$I_2^n := \left| \frac{1}{1-\beta} \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - x_\alpha) \cdot \nabla \varphi_{n-2}(x) \mathrm{d}\gamma_n(x, x_\alpha) - \frac{1}{1-\beta} \int_{\mathbb{R}^d} (\rho_n - \alpha\rho_{n-1} + \beta\rho_{n-2})\varphi_{n-2} \right|,$$

and

$$I_3 := \left| \sum_{n=2}^N \frac{1}{1-\beta} \int_{\mathbb{R}^d} (\rho_n - \alpha\rho_{n-1} + \beta\rho_{n-2})\varphi_{n-2} - \left(- \int_0^T \int_\Omega \partial_t \varphi \varrho - \int_\Omega \varphi(0)\varrho(0) \right) \right|.$$

Integrating by parts the discrete derivative in this last term,

$$\begin{aligned} &\sum_{n=2}^N \frac{1}{1-\beta} \int_{\mathbb{R}^d} (\rho_n - \alpha\rho_{n-1} + \beta\rho_{n-2})\varphi_{n-2} = \\ &\qquad = \sum_{n=2}^N \frac{1}{1-\beta} \int_{\mathbb{R}^d} (\varphi_{n-2} - (\alpha\varphi_{n-1} - \beta\varphi_n))\rho_n + \frac{1}{1-\beta} \int_{\mathbb{R}^d} \beta\varphi_0\rho_0 + (\beta\varphi_1 - \alpha\varphi_0)\rho_1 \,. \end{aligned}$$

Then, since $\alpha = 1 + \beta$, and thanks to the smoothness of the function φ and Proposition 3.4, we obtain $I_3 \leq C\tau$ for some constant C independent of τ.

Let us focus then on the term I_2. Adding and subtracting $(1-\beta)^{-1} \int_{\mathbb{R}^d} (\rho_n - \rho_{n-1}^\alpha)\varphi_{n-2}$ at each step n, we obtain

$$\begin{aligned} I_2^n \leq \frac{1}{1-\beta} &\left| \int_{\mathbb{R}^d \times \mathbb{R}^d} (x - x_\alpha) \cdot \nabla \varphi_{n-2}(x) \mathrm{d}\gamma_n(x, x_\alpha) - \int_{\mathbb{R}^d} (\rho_n - \rho_{n-1}^\alpha)\varphi_{n-2} \right| \\ &+ \frac{1}{1-\beta} \left| \int_{\mathbb{R}^d} (\alpha\rho_{n-1} - \beta\rho_{n-2} - \rho_{n-1}^\alpha)\varphi_{n-2} \right| =: \frac{1}{1-\beta}(I_4^n + I_5^n) \,. \end{aligned} \tag{3.15}$$

Rewriting

$$\int_{\mathbb{R}^d} (\rho_n - \rho_{n-1}^\alpha)\varphi_{n-2} = \int_{\mathbb{R}^d \times \mathbb{R}^d} (\varphi_{n-2}(x) - \varphi_{n-2}(x_\alpha)) \mathrm{d}\gamma_n(x, x_\alpha) \,,$$

we can bound I_4^n as

$$\begin{aligned} I_4^n &= \left| \int_{\mathbb{R}^d \times \mathbb{R}^d} \varphi_{n-2}(x) - \varphi_{n-2}(x_\alpha) - (x - x_\alpha) \cdot \nabla \varphi_{n-2}(x) \mathrm{d}\gamma_n(x, x_\alpha) \right| \\ &\leq \frac{1}{2} ||\mathrm{Hess}(\varphi_{n-2})||_\infty \left(\int_{\mathbb{R}^d \times \mathbb{R}^d} |x - x_\alpha|^2 \mathrm{d}\gamma_n(x, x_\alpha) \right) \\ &= \frac{1}{2} ||\mathrm{Hess}(\varphi_{n-2})||_\infty W_2^2(\rho_n, \rho_{n-1}^\alpha) \\ &\leq ||\mathrm{Hess}(\varphi_{n-2})||_\infty \left(W_2^2(\rho_n, \rho_{n-1}) + W_2^2(\rho_{n-1}, \rho_{n-1}^\alpha) \right) \\ &\leq ||\mathrm{Hess}(\varphi_{n-2})||_\infty \Big(W_2^2(\rho_n, \rho_{n-1}) + \theta^2 W_2^2(\rho_{n-1}, \rho_{n-2}) \Big) \,, \end{aligned}$$

where we used the dissipation estimate (1.11). Similarly by the consistency assumption (1.12) on the extrapolation, there exists a constant C_φ only depending on φ and Ω such that

$$I_5^n \leq C_\varphi W_2^2(\rho_{n-1}, \rho_{n-2}) \,.$$

Using the bound (3.2), the estimates above imply that there exists a constant $C > 0$ such that $I_2 \leq C\tau$. The whole term I_1 is therefore converging to zero and ϱ satisfies equation (3.12).

□

Remark 3.6. *The θ-dissipativity and consistency assumptions play different roles in our proof of convergence. One the one hand, θ-dissipativity is essentially used to get a stable scheme (Lemma 3.1) and obtain compactness (Lemma 3.2). On the other hand, the consistency assumption is necessary to obtain a consistent discretization of the time derivative (appearing in I_5^n in (3.15)) and recover the correct PDE in the limit.*

4. Extrapolation in Wasserstein space

In this section we consider the issue of defining geodesic extrapolations in the Wasserstein space. In particular, we propose several notions of extrapolation operators E_α, which in some cases verify the assumptions of Theorem 1.2, and discuss their relationship. We consider the extrapolation problem on the whole space $\mathcal{P}_2(\mathbb{R}^d)$. This allows us to be more general and to simplify the exposition, in particular avoiding issues with the boundary. On the other hand, some of the proposed definitions may be adapted so that the extrapolation of two measures in $\mathcal{P}(\Omega)$ stays in $\mathcal{P}(\Omega)$ (see Remark 4.7). We stress that this last property is not required in our definition of the EVBDF2 scheme (1.8), but it can be useful to produce a fully-discrete scheme (see Section 6.3) or an intrinsic formulation. See Section 4.4 for more considerations on this issue.

As recalled in the introduction, a globally-minimizing geodesic with respect to the W_2 metric is a curve $\omega : [t_0, t_1] \to \mathcal{P}_2(\mathbb{R}^d)$ such that

$$W_2(\omega(s_0), \omega(s_1)) = \frac{|s_1 - s_0|}{|t_1 - t_0|} W_2(\omega(t_0), \omega(t_1)) \,, \tag{4.1}$$

for all $s_0, s_1 \in (t_0, t_1)$. We say that $\omega : [t_0, t_1] \to \mathcal{P}_2(\mathbb{R}^d)$ is a locally-minimizing geodesic if for all $t \in (t_0, t_1)$ there exists an open interval $J \ni t$ such that (4.1) holds for all $s_0, s_1 \in J \cap (t_0, t_1)$. From the discussion in Section 2, given two measures $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, if μ_0 is absolutely continuous there exists a unique globally length-minimizing geodesic connecting the two, which is given by

$$\omega(t) = ((1-t)\mathrm{Id} + t\nabla u)_{\#}\mu_0 \tag{4.2}$$

for $t \in [0, 1]$, where u is a uniquely defined convex function μ_0-a.e. (up to an additive constant). As a matter of fact, we have for all $s_0, s_1 \in (0, 1)$,

$$\begin{aligned} W_2^2(\omega(s_0), \omega(s_1)) &\le \int_{\mathbb{R}^d} |(1-s_0)x + s_0 \nabla u(x) - (1-s_1)x - s_1 \nabla u(x)|^2 \mathrm{d}\mu_0(x) \\ &= |s_1 - s_0|^2 W_2^2(\mu_0, \mu_1), \end{aligned} \tag{4.3}$$

where for the first inequality we used as competitor the plan $((1-s_0)\mathrm{Id} + s_0\nabla u, (1-s_1)\mathrm{Id} + s_1\nabla u)_{\#}\mu_0$, and for the second equality the optimality of the plan $(\mathrm{Id}, \nabla u)_{\#}\mu_0$ for the transport problem from μ_0 to μ_1. On the other hand, for $s_1 > s_0$, by the triangular inequality and (4.3)

$$\begin{aligned} W_2(\mu_0, \mu_1) &\le W_2(\mu_0, \omega(s_0)) + W_2(\omega(s_0), \omega(s_1)) + W_2(\omega(s_1), \mu_1) \\ &\le (s_0 + 1 - s_1) W_2(\mu_0, \mu_1) + W_2(\omega(s_0), \omega(s_1)) \,, \end{aligned}$$

and therefore the inequality in (4.3) is an equality. Moreover, by similar calculations one can verify that for any $\alpha \ge 1$ the curve $t \in [0, \alpha] \mapsto \omega(t)$, still defined as in (4.2), is a globally length-minimizing geodesic if and only if u is β/α-convex, i.e. the function

$$x \mapsto \alpha u(x) - \beta \frac{|x|^2}{2} \quad \text{is convex,} \tag{4.4}$$

with $\beta = \alpha - 1$. However, in general, there is no guarantee that u is strongly-convex even if μ_0 and μ_1 have smooth and strictly positive densities and for arbitrarily small β, as shown by the following example.

Example 4.1 (Contraction flow)**.** Take $u = \frac{\beta}{2\alpha}|\cdot|^2$, for $\alpha > 1$ and $\beta = \alpha - 1$. Then, for any absolutely continuous $\mu_0 \in \mathcal{P}_2(\mathbb{R}^2)$ and $\mu_1 = (\nabla u)_{\#}\mu_0$, there exists a unique globally length-minimizing geodesic on $(-\infty, \alpha]$ such that $\omega(0) = \mu_0$ and $\omega(1) = \mu_1$, which is given by (4.2). On the other hand, since all trajectories cross at time α (i.e. $(1-\alpha)\mathrm{Id} + \alpha\nabla u = 0$), there exists no geodesic on $(-\infty, \alpha']$ (either local or global) with $\alpha' > \alpha$ satisfying the same property.

In general, globally length-minimizing geodesic extensions may not exist even if particle trajectories do not cross. In this case, however, locally length-minimizing extensions may still exist as shown in the next example.

Example 4.2 (Shear flow)**.** For $d = 2$, let

$$\mu_0 = \frac{1}{2}(\delta_z + \delta_{-z})\,, \quad \mu_1 = \frac{1}{2}(\delta_{z-v} + \delta_{-z+v})$$

where $z = (1,1)$ and $v = (1,0)$. In this case, there exists a unique geodesic $\omega : \mathbb{R} \to \mathcal{P}_2(\mathbb{R}^2)$ which is locally length-minimizing, and such that $\omega(0) = \mu_0$ and $\omega(1) = \mu_1$, which is given by

$$\omega(t) = \frac{1}{2}(\delta_{z-tv} + \delta_{-z+tv}). \tag{4.5}$$

However, ω is globally length-minimizing only when restricted on $(-\infty, 2]$.

In order to define our scheme, we need an extrapolation operator which is well-defined even when the geodesic extension (either globally or locally length-minimizing) does not exist. In the following we will introduce different possible definitions and describe their properties.

4.1. **Free-flow extrapolations.** One possible strategy for defining an extrapolation consists in disregarding the convexity condition on the Brenier potential in (4.4), and allowing particles to cross each other while keeping their straight trajectories at constant speed. If $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ is absolutely continuous, this amounts to defining, for any $\mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ and $\alpha > 1$,

$$\mathsf{E}_\alpha(\mu_0, \mu_1) = ((1-\alpha)\mathrm{Id} + \alpha\nabla u)_{\#}\mu_0\,, \tag{4.6}$$

where u is a Brenier potential from μ_0 to μ_1 (uniquely defined μ_0-a.e.). If μ_0 is not absolutely continuous, there may exist multiple geodesics and optimal transport plans from μ_0 to μ_1. In general, we say that an extrapolation operator E_α yields a *free-flow extrapolation* if, denoting by $\Gamma(\mu_0, \mu_1)$ the set of optimal plans from μ_0 to μ_1, one has:

$$\forall \mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)\,,\ \exists \gamma^* \in \Gamma(\mu_0, \mu_1)\ :\ \mathsf{E}_\alpha(\mu_0, \mu_1) = (\pi_\alpha)_{\#}\gamma^*, \tag{4.7}$$

where $\pi_\alpha : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$ is the map defined by $\pi_\alpha(x, y) = x + \alpha(y - x)$. By construction, when the geodesic induced by γ^* in (4.7) admits a locally (or globally) length-minimizing geodesic extension, the resulting free-flow extrapolation is always consistent with it (for example, free-flow extrapolations yield the curve (4.5) in the case of Example 4.2). Furthermore, such extrapolation operators are admissible for our scheme in the sense of Theorem 1.2, as shown by the following proposition.

Proposition 4.3. *Any free-flow extrapolation operator $\mathsf{E}_\alpha : \mathcal{P}_2(\mathbb{R}^d) \times \mathcal{P}_2(\mathbb{R}^d) \to \mathcal{P}_2(\mathbb{R}^d)$, i.e. any map satisfying (4.7), is β-dissipative with $\beta = \alpha - 1$, and in addition it verifies the consistency assumption (1.12) for all $\varphi \in C_c^\infty(\mathbb{R}^d)$.*

Proof. For simplicity, we only consider the case where μ_0 is absolutely continuous. Let ∇u the optimal transport map from μ_0 to μ_1. To prove the dissipativity, let $\bar\gamma = (\nabla u, (1-\alpha)\mathrm{Id} + \alpha\nabla u)_\# \mu_0$. Then $\bar\gamma \in \Pi(\mu_1, \mathsf{E}_\alpha(\mu_0,\mu_1))$ and by equation (2.1),

$$W_2^2(\mu_1, \mathsf{E}_\alpha(\mu_0,\mu_1)) \le \int |x-y|^2 \mathrm{d}\bar\gamma(x,y) = (1-\alpha)^2 \int |\mathrm{Id} - \nabla u|^2 \mu_1 = \beta^2 W_2^2(\mu_0,\mu_1)\,.$$

For the consistency, let $\varphi \in C_c^\infty(\mathbb{R}^d)$ and observe that, by the definition of pushforward,

$$\int \varphi\,(\mathsf{E}_\alpha(\mu_0,\mu_1) - \alpha\mu_1 + \beta\mu_0) = \int \big[\varphi\big((1-\alpha)x + \alpha\nabla u(x)\big) - \alpha\varphi\big(\nabla u(x)\big) + \beta\varphi(x)\big]\,\mathrm{d}\mu_0(x)\,.$$

Using the Taylor expansion of φ around the point x in the integral on the right-hand side, we find

$$\left|\int \varphi\,(\mathsf{E}_\alpha(\mu_0,\mu_1) - \alpha\mu_1 + \beta\mu_0)\right| \le \frac{\alpha\beta}{2}\|\mathrm{Hess}(\varphi)\|_\infty W_2^2(\mu_0,\mu_1)\,.$$

In the general case where μ_0 is not absolutely continuous, the proof is analogous replacing transport maps by optimal plans. □

4.2. **Extrapolation with collisions.** Free-flow extrapolations are the simplest way to extend geodesics after their maximal time of existence, but they are purely Lagrangian and they cannot be easily implemented in an Eulerian setting. Here we describe an alternative route to construct an extrapolation operator which prevents particles to cross, and which is based on viscosity solutions of the Hamilton-Jacobi equation. The resulting operator can be implemented in a robust way, but unfortunately it falls outside the hypotheses of the convergence results presented in this work. In Section 6, we will describe a possible implementation (in the case of a compact domain Ω) and verify numerically that it leads to a second-order scheme.

Given $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, let us suppose that the optimal potential ϕ for the transport from μ_0 to a given measure μ_1 on the time interval $[0,1]$, is such that

$$\phi(0,\cdot) \text{ is globally Lipschitz.} \tag{4.8}$$

Then, the curve $\omega : [0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ satisfying

$$\omega(t) = \left[\nabla \mathrm{co}\left((1-t)\frac{|\cdot|^2}{2} + tu\right)\right]_\# \mu_0\,, \tag{4.9}$$

where $u = |\cdot|^2/2 + \phi(0,\cdot)$ is a Brenier potential from μ_0 to μ_1, and where co denotes the convex hull, is well-defined. We remark that (4.9) coincides at time $t = \alpha$ with the free-flow extrapolation (4.6) as long as the convexity condition (4.4) holds. On the other hand, if such condition is not verified, taking the convex envelope in (4.9) guarantees that the flow stays monotone and particles cannot cross.

If (4.8) holds, one also has that the Hamilton-Jacobi equation (2.5) with initial condition $\phi(0,\cdot)$ has a unique viscosity solution, which is given by the Hopf-Lax formula

$$\phi(t,\cdot) = \mathcal{H}_t(\phi(0,\cdot)), \quad \mathcal{H}_t(\phi(0,\cdot))(x) \coloneqq \inf_{y\in\mathbb{R}^d} \frac{|x-y|^2}{2t} + \phi(0,y)\,. \tag{4.10}$$

Note that the evolution of the density transported by the velocity field $\nabla\phi(t,\cdot)$ (via the continuity equation) is also well-defined since so is its Lagragian flow [23, 7]. In the following lemma we show that equations (4.10) and (4.9) are closely related.

Lemma 4.4. *Let $\phi : [0,\infty)\times\mathbb{R}^d \to \mathbb{R}$ be the unique viscosity solution to the Hamilton-Jacobi equation, or equivalently verifying (4.10) for $t > 0$, with $\phi(0,\cdot)$ being a Lipschitz function,*

and denote $u := \phi(0,\cdot) + \frac{|\cdot|^2}{2}$. *Let* $\mu_0 \in \mathcal{P}_2(\mathbb{R}^d)$ *be an absolutely continuous measure and* $\omega : [0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$ *be the curve defined by* (4.9) *for all* $t \geq 0$. *Then,*

(1) for all $t \geq 0$, $\omega(t)$ *solves*

$$\min_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \frac{W_2^2(\mu_0,\mu)}{2t} - \int \phi(t,\cdot)\mu\,; \tag{4.11}$$

(2) if $d = 1$, ω *is a weak solution to the continuity equation with velocity* $\nabla\phi(t,\cdot)$.

Proof. Concerning the first point, by the optimality conditions of problem (4.11) [34, Example 7.21] one can verify that:

$$\frac{W_2^2(\mu_0,\mu)}{2t} = \int \phi(t,\cdot)\mu - \int \mathcal{H}_t(-\phi(t,\cdot))\mu_0\,.$$

Therefore, the optimal transport map from μ_0 to the optimal measure μ is the gradient of $\frac{|\cdot|^2}{2} - t\mathcal{H}_t(-\phi(t,\cdot)) = \frac{|\cdot|^2}{2} - t\mathcal{H}_t(-\mathcal{H}_t(\phi(0,\cdot)))$. Noting that for any function ψ it holds

$$\begin{aligned} \frac{|y|^2}{2} - t\mathcal{H}_t(\psi)(y) &= \frac{|y|^2}{2} - \inf_x \frac{|x-y|^2}{2} + t\psi(x) \\ &= \sup_x y\cdot x - \left(\frac{|x|^2}{2} + t\psi(x)\right) = \left(\frac{|\cdot|^2}{2} + t\psi(\cdot)\right)^*(y), \quad \forall y\,, \end{aligned} \tag{4.12}$$

we conclude by applying twice (4.12).

For the second part, we refer to Proposition 4.1 in [3], where an explicit expression for the measure transported by the flow is provided. □

Remark 4.5. *For* $d > 1$, *the curve* (4.9) *does not coincide in general with the solution of the continuity equation with velocity* $\nabla\phi(t,\cdot)$. *This is because* (4.9) *completely disregards the dynamics of mass within the shocks, which may be non-trivial* [3, 7].

There are two main problems with using (4.9) to define an extrapolation operator, i.e. setting $\mathsf{E}_\alpha(\mu_0,\mu_1) = \omega(\alpha)$. First, the initial potential $\phi(0,\cdot)$ is uniquely defined only μ_0-a.e., however the value of the potential outside the support of μ_0 does affect the final measure $\omega(\alpha)$ for $\alpha > 1$. Second, because of the same reason one can easily construct solutions that are not dissipative in the sense of Definition 1.1: for example, one can take $\mu_0 = \mu_1$ with compact support and select an initial potential outside the support in such a way that $\omega(\alpha)$ (defined as in the previous lemma) is different from μ_1.

Remark 4.6 (Extrapolation via pressureless fluids)**.** *With the same notation as above, one could construct geodesic continuations also by looking for solutions* $\omega : [0,\infty) \to \mathcal{P}_2(\mathbb{R}^d)$, $v : [0,\infty) \to L^2(\omega(t);\mathbb{R}^d)$, *of the following system of PDEs:*

$$\begin{cases} \partial_t\omega + \operatorname{div}(\omega v) = 0\,, \\ \partial_t(\omega v) + \operatorname{div}(\omega v \otimes v) = 0\,, \end{cases} \tag{4.13}$$

with initial conditions given by

$$\omega(0) = \mu_0\,, \quad v(0,\cdot) = \nabla\phi(0,\cdot)\,.$$

System (4.13) *describes the evolution of a pressureless fluid with given initial density and velocity. In fact, any sufficiently regular solution* (ω, v) *of problem* (2.2) *on the time interval* $[0,1]$ *also solves* (4.13), *since the absence of shocks implies that the Hamilton-Jacobi equation is equivalent to the conservation of momentum, i.e. the second equation in* (4.13). *Moreover,*

dissipative solutions to such system, i.e. for which the kinetic energy $\mathcal{K} : [0,\infty) \to \mathbb{R}_+$ *given by*

$$\mathcal{K}(t) \coloneqq \int \omega(t)|v(t)|^2$$

is nonincreasing, provide a dissipative notion of extrapolation, since by equation (2.2)*, for any* $\alpha = 1+\beta > 1$

$$W_2^2(\mu_1, \omega(\alpha)) \le \beta \int_1^\alpha \mathrm{d}t \int \omega(t)|u(t)|^2 \le \beta^2 \int_0^1 \mathrm{d}t \int \omega(t)|u(t)|^2 = \beta^2 W_2^2(\mu_0,\mu_1)\,.$$

Such solutions can be constructed by requiring a sticky collision *condition, which enforces particles to share the same position after their collision. In dimension higher than one, few results exist on the well-posedness of system* (4.13)*, so we will not consider this case in detail. On the other hand, in dimension one, sticky solutions to system* (4.13) *have been widely studied in the literature. In particular, Brenier and Grenier* [8] *showed that one can construct solutions to* (4.13) *using the unique entropy solution of a scalar conservation law, and in particular a solution to* (4.13) *is given by the curve*

$$\omega(t) = \tilde{X}(t,\cdot)_{\#}\mu_0,$$

with

$$\tilde{X}(t,x) \coloneqq (\partial_x \mathrm{co}\,\psi(t,\cdot)) \circ F_0(x)\,, \quad \psi(t,s) \coloneqq \int_0^s X(t, F_0^{[-1]}(s'))\,\mathrm{d}s'\,, \tag{4.14}$$

where X *is defined as in* (2.9)*, and* $F_0^{[-1]} : [0,1] \to \overline{\mathbb{R}}$ *is the quantile function of* μ_0*, i.e. the pseudo-inverse of its cumulative distribution function* $F_0 : x \to \int_{-\infty}^x \mathrm{d}\mu_0(x)$*. Note that as long as the geodesic can be extended* $\psi(t,\cdot)$ *stays convex (as it is the integral of a monotone function) and therefore the definitions for* $X(t,\cdot)$ *and* $\tilde{X}(t,\cdot)$*, respectively in* (2.9) *and* (4.14)*, coincide. We will show that in this case the resulting notion of extrapolation coincides with that provided by the metric extrapolation, which is discussed in detail in the next section.*

4.3. **Metric extrapolation.** In analogy with the Euclidean case (see equation (1.13)), one can adopt a variational definition for the extrapolation, which we refer to as *metric extrapolation*, and which is defined for all $\alpha > 1$ and for all $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$\mathsf{E}_\alpha(\mu_0,\mu_1) \coloneqq \operatorname*{argmin}_{\rho\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu_0,\mu_1;\rho)\,, \quad \mathcal{F}(\mu_0,\mu_1;\rho) \coloneqq \alpha W_2^2(\rho,\mu_1) - \beta W_2^2(\rho,\mu_0)\,, \tag{4.15}$$

where $\beta = \alpha - 1$. In Proposition 4.10 we will show that problem (4.15) admits indeed a unique solution, which justifies the definition of the metric extrapolation.

Remark 4.7. *Alternatively, one can define the metric extrapolation as in equation* (4.15) *via a minimization on probability measures in* $\mathcal{P}(\Omega)$ *over a given compact domain* Ω*. In this case, differently from the free-flow case* (4.6)*, the support of the extrapolated measures is always contained in* Ω*. The results of this section hold also in this case without major changes.*

First of all, we observe that by the triangular and Young's inequalities, for any $\rho, \mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$

$$W_2^2(\rho,\mu_0) \le \left(1+\frac{1}{\beta}\right) W_2^2(\rho,\mu_1) + (1+\beta) W_2^2(\mu_0,\mu_1)$$

and therefore

$$\mathcal{F}(\mu_0,\mu_1;\rho) \ge -\alpha\beta W_2^2(\mu_0,\mu_1). \tag{4.16}$$

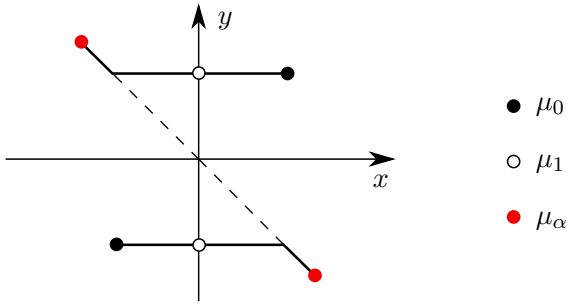


FIGURE 3. Metric extrapolation in the setting of Example 4.2. The black solid line connecting the support of the three measures represents the trajectory followed by the extrapolated measure for different values of the parameter α.

Then, if there exists a unique geodesic (4.2) from μ_0 to μ_1 and this can be continued up to time α, i.e. if the associated Brenier potential u is β/α-convex, then the lower bound is attained only by $\rho = \omega(\alpha)$ with

$$\omega(\alpha) = ((1-\alpha)\mathrm{Id} + \alpha\nabla u)_{\#}\mu_0,$$

since by equation (1.9)

$$W_2^2(\mu_0, \omega(\alpha)) = \alpha^2 W_2^2(\mu_0, \mu_1)\,, \quad W_2^2(\mu_1, \omega(\alpha)) = \beta^2 W_2^2(\mu_0, \mu_1)\,.$$

Remark 4.8. *Note that if the geodesic extension is only locally (but not globally) minimizing, then it may not be recovered as a solution of problem (4.15): for instance, this is the case for the shear flow example 4.2, in which case one can compute the explicit solution to the metric extrapolation problem, which is represented in Figure 3.*

Existence and uniqueness for minimizers of problem (4.15) actually hold in general due to the fact that the functional $\mathcal{F}$ is strongly convex along particular curves known as generalized geodesics. To describe such curves, consider three measures $\nu_0, \nu_1, \nu_2 \in \mathcal{P}_2(\mathbb{R}^d)$, let $\gamma_{0,1} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ and $\gamma_{0,2} \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d)$ optimal transport plans from ν_0 to ν_1 and from ν_0 to ν_2, respectively. A generalized geodesic from ν_1 to ν_2 with base ν_0 is a curve $\omega : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ satisfying, for all $\varphi \in C_b^0(\mathbb{R}^d)$,

$$\int \varphi\omega(t) = \int \varphi(x_1(1-t) + x_2 t)\mathrm{d}\gamma(x_0, x_1, x_2)$$

where $\gamma \in \mathcal{P}_2(\mathbb{R}^d \times \mathbb{R}^d \times \mathbb{R}^d)$ is a plan verifying

$$\begin{aligned} \int \psi(x_0, x_1)\mathrm{d}\gamma(x_0, x_1, x_2) &= \int \psi(x_0, x_1)\mathrm{d}\gamma_{0,1}(x_0, x_1)\,, \\ \int \psi(x_0, x_2)\mathrm{d}\gamma(x_0, x_1, x_2) &= \int \psi(x_0, x_2)\mathrm{d}\gamma_{0,2}(x_0, x_2)\,, \end{aligned} \tag{4.17}$$

for all $\psi \in C_b^0(\mathbb{R}^d \times \mathbb{R}^d)$. The existence of such a plan is a consequence of the so-called gluing lemma (Lemma 5.3.2 in [2]). In the case where ν_0 is absolutely continuous, denoting by $T_{0,1}$ and $T_{0,2}$ the optimal transport plans from ν_0 to ν_1 and from ν_0 to ν_2 respectively, there exists a unique generalized geodesic from ν_1 to ν_2 with base ν_0 which is given by

$$\omega(t) = ((1-t)T_{0,1} + tT_{0,2})_{\#}\nu_0\,. \tag{4.18}$$

A functional $\mathcal{J} : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ is λ-convex along generalized geodesics based in ν_0, if for all ν_1 to ν_2 and for all generalized geodesics $\omega : [0,1] \to \mathcal{P}_2(\mathbb{R}^d)$ from ν_1 to ν_2 with base ν_0,

$$\mathcal{J}(\omega(t)) \leq (1-t)\mathcal{J}(\nu_1) + t\mathcal{J}(\nu_2) - \lambda \frac{t(1-t)}{2} \int |x_1 - x_2|^2 \mathrm{d}\gamma(x_0, x_1, x_2) \tag{4.19}$$

with γ satisfying equation (4.17). We say that the functional $\mathcal{J}$ is λ-convex along generalized geodesics if the previous definition holds true for any $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$.

The following result was proven in [29] and provides the strong convexity of the functional $\mathcal{F}$ along generalized geodesics.

Lemma 4.9 (Theorem 3.4 in [29]). *For any $\mu_0, \mu_1 \in \mathcal{P}_2(\mathbb{R}^d)$, the functional $\mathcal{F}(\mu_0, \mu_1; \cdot) : \mathcal{P}_2(\mathbb{R}^d) \to \mathbb{R}$ defined in (4.15) is 2-convex along generalized geodesics based in μ_1. In particular, for any $\mu_2, \mu_3 \in \mathcal{P}_2(\mathbb{R}^d)$ there exists a curve $\omega : [0,1] \to \mathcal{P}_2(\mathbb{R}^d), \omega(0) = \mu_2, \omega(1) = \mu_3$, such that for all $t \in [0,1]$, it holds:*

$$\mathcal{F}(\mu_0, \mu_1; \omega(t)) \leq (1-t)\mathcal{F}(\mu_0, \mu_1; \mu_2) + t\mathcal{F}(\mu_0, \mu_1; \mu_3) - t(1-t)W_2^2(\mu_2, \mu_3). \tag{4.20}$$

Lemma 4.9 is the main ingredient to prove the following proposition.

Proposition 4.10. *The metric extrapolation problem (4.15) admits a unique solution μ_α. Moreover, the metric extrapolation is β-dissipative, i.e.*

$$W_2(\mu_1, \mu_\alpha) \leq \beta W_2(\mu_0, \mu_1)\,, \tag{4.21}$$

and for all $\mu \in \mathcal{P}_2(\mathbb{R}^d)$,

$$W_2^2(\mu, \mu_\alpha) + \mathcal{F}(\mu_0, \mu_1; \mu_\alpha) \leq \mathcal{F}(\mu_0, \mu_1; \mu)\,. \tag{4.22}$$

Proof. The functional $\mathcal{F}$ is strongly convex along generalized geodesics by Lemma 4.9, which implies uniqueness of the solution. Regarding existence, let $(\mu^n)_n$ be a minimizing sequence. We denote $m = \inf_{\mu \in \mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}(\mu_0, \mu_1; \mu)$, which is finite due to (4.16), and we introduce $\mathcal{G}(\mu) = \mathcal{F}(\mu_0, \mu_1; \mu) - m$. Consider two measures μ^{n_1}, μ^{n_2} of the sequence and the generalized geodesic ω based in μ_1 connecting them, as in Lemma 4.9. The inequality (4.20) for $t = \frac{1}{2}$ provides

$$\frac{1}{4} W_2^2(\mu^{n_1}, \mu^{n_2}) \leq \frac{1}{2}\mathcal{G}(\mu^{n_1}) + \frac{1}{2}\mathcal{G}(\mu^{n_2})\,,$$

which implies that the sequence is Cauchy in the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$. The Wasserstein space being complete [2, Proposition 7.1.5], the sequence converges to a measure μ_α, which is the minimizer since $\mathcal{F}$ is continuous.

Inequality (4.22) derives again from Lemma 4.9. For a given $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, consider a generalized geodesic ω as in Lemma 4.9, with $\omega(0) = \mu_\alpha$ and $\omega(1) = \mu$. By optimality of μ_α, it holds

$$\begin{aligned} 0 &\leq \mathcal{F}(\mu_0, \mu_1; \omega(t)) - \mathcal{F}(\mu_0, \mu_1; \mu_\alpha) \\ &\leq t\big(\mathcal{F}(\mu_0, \mu_1; \mu) - \mathcal{F}(\mu_0, \mu_1; \mu_\alpha)\big) - t(1-t)W_2^2(\mu, \mu_\alpha)\,, \end{aligned}$$

which, dividing by t and taking the limit $t \to 0$, gives (4.22). Using (4.16) on the left-hand side of (4.22) and then taking $\mu = \mu_1$, we obtain the estimate (4.21). □

In order to prove the consistency assumption we will use the following optimality conditions for problem (4.15).

Lemma 4.11. *Let μ_α be the unique solution to problem (4.15). There exist two optimal transport plans $\gamma_{0,\alpha}$ and $\gamma_{1,\alpha}$ from μ_0 to μ_α and from μ_1 to μ_α, respectively, such that*

$$\alpha \int (x_\alpha - x_1)\cdot \xi(x_\alpha) d\gamma_{1,\alpha}(x_1, x_\alpha) - \beta \int (x_\alpha - x_0)\cdot \xi(x_\alpha) d\gamma_{0,\alpha}(x_0, x_\alpha) = 0\,, \tag{4.23}$$

for any $\xi \in C_c^\infty(\mathbb{R}^d;\mathbb{R}^d)$.

Proof. Note that we cannot use directly Lemma 3.5 because μ_α is not necessarily absolutely continuous. Therefore, 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 for $\varepsilon > 0$,

$$\mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu) \coloneqq \mathcal{F}(\mu_0,\mu_1;\mu) + \varepsilon \mathcal{U}(\mu|\nu)\,, \tag{4.24}$$

where $\mathcal{U}(\cdot\,|\nu)$ denotes the relative entropy

$$\mathcal{U}(\mu|\nu) \coloneqq \begin{cases} \displaystyle\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu & \text{if } \mu \ll \nu\,,\\ +\infty & \text{otherwise}\,, \end{cases} \tag{4.25}$$

and $\nu = (2\pi)^{-d/2}\exp(-|x|^2/2)\mathrm{d}x \in \mathcal{P}_2(\mathbb{R}^d)$. We introduce the regularized problem

$$\inf_{\mu\in\mathcal{P}_2(\mathbb{R}^d)} \mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu)\,. \tag{4.26}$$

Let $(\mu^n)_n$ be a minimizing sequence for (4.26). Due to Jensen's inequality the relative entropy is positive. Furthermore, it is convex along generalized geodesics [2, Theorem 9.4.11]. Hence, reasoning as in Proposition 4.10, we obtain convergence in W_2 of μ^n to a measure μ_α^ε. The relative entropy is lower semi-continuous on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$ [1, Theorem 15.4] and therefore μ_α^ε is the unique minimizer.

Note that

$$\int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\mathrm{d}\mu = \int \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}x}\right)\mathrm{d}\mu + \int \frac{|x|^2}{2}\mathrm{d}\mu(x) + \frac{d}{2}\log(2\pi)\,, \quad \text{for } \mu \ll \nu\,.$$

Therefore, by applying Lemma 3.5 (adapted to the case where $\Omega = \mathbb{R}^d$), we can write down the necessary optimality conditions of problem (4.26):

$$\begin{aligned} \frac{\mathrm{d}\mathcal{F}_\varepsilon(\mu_0,\mu_1;\omega(s))}{\mathrm{d}s}\bigg|_{s=0} &= 2\alpha \int (x_\alpha - x_1)\cdot\xi(x_\alpha)\mathrm{d}\gamma_{1,\alpha}^\varepsilon(x_1,x_\alpha)\\ &- 2\beta\int (x_\alpha - x_0)\cdot\xi(x_\alpha)\mathrm{d}\gamma_{0,\alpha}^\varepsilon(x_0,x_\alpha) + \varepsilon\int \big(x_\alpha\cdot\xi(x_\alpha) - \mathrm{div}(\xi(x_\alpha))\big)\mathrm{d}\mu_\alpha^\varepsilon(x_\alpha) = 0\,, \end{aligned} \tag{4.27}$$

for any $\xi \in C_c^\infty(\mathbb{R}^d;\mathbb{R}^d)$, where $\omega : (-\delta,\delta) \to \mathcal{P}_2(\mathbb{R}^d)$ is the curve of measures defined by (3.9) with $\omega(0) = \mu_\alpha$, and where we denote now by $\gamma_{0,\alpha}^\varepsilon$ and $\gamma_{1,\alpha}^\varepsilon$ the optimal transport plans from μ_0 to μ_α^ε and from μ_1 to μ_α^ε, respectively.

We want to show that the regularized functionals $\mathcal{F}_\varepsilon(\mu_0,\mu_1;\cdot)$, interpreted as functionals on the Wasserstein space $(\mathcal{P}_2(\mathbb{R}^d), W_2)$, Γ-converges towards $\mathcal{F}(\mu_0,\mu_1;\cdot)$, in order to pass to the limit in the optimality conditions of problem (4.26). Since $\mathcal{F}$ is continuous with respect to W_2 convergence and $\mathcal{U}$ is positive, the Γ-lim inf is obvious,

$$\mathcal{F}(\mu_0,\mu_1;\mu) \le \liminf_\varepsilon \mathcal{F}(\mu_0,\mu_1;\mu_\varepsilon) \le \liminf_\varepsilon \mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu_\varepsilon)\,,$$

for any $\mu_\varepsilon \to \mu$ in the Wasserstein sense. Concerning the Γ-lim sup, if $\mathcal{U}(\mu|\nu) < \infty$ we can take $\mu_\varepsilon = \mu$ as recovery sequence. Otherwise, since the set of absolutely continuous measures

is dense in $\mathcal{P}_2(\mathbb{R}^d)$, we can take a sequence of absolutely continuous measures μ_ε converging to μ with respect to the Wasserstein metric. Since $\mathcal{U}(\mu|\nu) = \infty$, up to a reparametrization we can assume that the relative entropy is increasing and that

$$\mathcal{U}(\mu_\varepsilon|\nu) \leq \frac{C}{\sqrt{\varepsilon}}\,,$$

for a constant C independent of ε. Then it holds:

$$\limsup_\varepsilon \mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu_\varepsilon) = \lim_\varepsilon \mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu_\varepsilon) = \mathcal{F}(\mu_0,\mu_1;\mu)\,.$$

Therefore $\mathcal{F}_\varepsilon(\mu_0,\mu_1;\cdot)$ Γ-converges to $\mathcal{F}(\mu_0,\mu_1;\cdot)$. Let us show that the sequence of minimizer $(\mu_\alpha^\varepsilon)_\varepsilon$ is Cauchy. For this we observe that $(\mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu_\alpha^\varepsilon))_\varepsilon$ is monotonically decreasing as $\varepsilon \to 0$ since, for $\varepsilon_2 > \varepsilon_1$:

$$\mathcal{F}_{\varepsilon_2}(\mu_0,\mu_1;\mu_\alpha^{\varepsilon_2}) = (\varepsilon_2-\varepsilon_1)\mathcal{U}(\mu_\alpha^{\varepsilon_2}|\nu) + \mathcal{F}_{\varepsilon_1}(\mu_0,\mu_1;\mu_\alpha^{\varepsilon_2}) \geq \mathcal{F}_{\varepsilon_1}(\mu_0,\mu_1;\mu_\alpha^{\varepsilon_1}). \tag{4.28}$$

Since $\mathcal{F}_\varepsilon(\mu_0,\mu_1;\cdot)$ are uniformly bounded from below, $\mathcal{F}_\varepsilon(\mu_0,\mu_1;\mu_\alpha^\varepsilon)$ converges to a value m as $\varepsilon \to 0$. Hence, we can define $\mathcal{G}_\varepsilon(\cdot) := \mathcal{F}_\varepsilon(\mu_0,\mu_1;\cdot) - m \geq 0$. By the same arguments as in the proof of Proposition 4.10 and the strong convexity of $\mathcal{G}_{\varepsilon_1}$ along generalized geodesics, for any $\varepsilon_2 > \varepsilon_1$,

$$\frac{1}{4}W_2^2(\mu_\alpha^{\varepsilon_1},\mu_\alpha^{\varepsilon_2}) \leq \frac{1}{2}\mathcal{G}_{\varepsilon_1}(\mu_\alpha^{\varepsilon_1}) + \frac{1}{2}\mathcal{G}_{\varepsilon_1}(\mu_\alpha^{\varepsilon_2}) \leq \frac{1}{2}\mathcal{G}_{\varepsilon_1}(\mu_\alpha^{\varepsilon_1}) + \frac{1}{2}\mathcal{G}_{\varepsilon_2}(\mu_\alpha^{\varepsilon_2})$$

where the second inequality is a consequence of (4.28). Since $\mathcal{G}_\varepsilon(\mu_\alpha^\varepsilon) \to 0$ as $\varepsilon \to 0$ we can conclude that $(\mu_\alpha^\varepsilon)_\varepsilon$ is Cauchy and by the Γ-convergence showed above, $\mu_\alpha^\varepsilon \to \mu_\alpha$ in W_2.

Finally, by the stability of optimal transport plans [39, Theorem 5.20], there exist optimal plans $\gamma_{0,\alpha}$ and $\gamma_{1,\alpha}$ from μ_0 to μ_α and from μ_1 to μ_α, respectively, such that (up to the extraction of a subsequence)

$$\gamma_{0,\alpha}^\varepsilon \rightharpoonup \gamma_{0,\alpha}\,, \quad \gamma_{1,\alpha}^\varepsilon \rightharpoonup \gamma_{1,\alpha}\,,$$

weakly, i.e. in duality with continuous bounded functions (and also in the Wasserstein sense; in fact, the second moments of $\gamma_{0,\alpha}^\varepsilon$ and $\gamma_{1,\alpha}^\varepsilon$ converge to those of $\gamma_{0,\alpha}$ and $\gamma_{1,\alpha}$ since $\mu_\alpha^\varepsilon \to \mu_\alpha$ in the Wasserstein sense). As the vector field ξ is smooth, passing to the limit in (4.27) we obtain (4.23).

□

Proposition 4.12. *The metric extrapolation defined via (4.15) verifies the consistency assumption (1.12) for all $\varphi \in C_c^\infty(\mathbb{R}^d)$.*

Proof. Using the same notation as in the statement of Lemma 4.11, we have that for all $\varphi \in C_c^\infty(\mathbb{R}^d)$

$$\int \varphi\,(\mu_\alpha - \alpha\mu_1 + \beta\mu_0) = \alpha\int(\varphi(x_1)-\varphi(x_\alpha))\mathrm{d}\gamma_{1,\alpha}(x_1,x_\alpha) - \beta\int(\varphi(x_0)-\varphi(x_\alpha))\mathrm{d}\gamma_{0,\alpha}(x_0,x_\alpha)\,.$$

Using the Taylor expansion of φ at x_α in both integrals on the right-hand side, Lemma 4.11 and the dissipation property (4.21), we obtain

$$\begin{aligned}\left|\int \varphi\,(\mu_\alpha - \alpha\mu_1 + \beta\mu_0)\right| &\leq \frac{1}{2}\|\mathrm{Hess}\varphi\|_\infty(\alpha W_2^2(\mu_1,\mu_\alpha) + \beta W_2^2(\mu_0,\mu_\alpha)) \\ &\leq \alpha\beta\|\mathrm{Hess}\varphi\|_\infty W_2^2(\mu_0,\mu_1).\end{aligned}$$

□

Remark 4.13 (Relation with pressureless fluids)**.** *In dimension one, the Wasserstein distance W_2 coincides with the L^2 distance between the quantile functions. In particular, the metric extrapolation μ_α is given by*

$$\mu_\alpha = (G_\alpha)_\# \mathrm{d}x|_{[0,1]}\,, \quad G_\alpha \coloneqq \operatorname*{argmin}_{\substack{G\in L^2([0,1],\mathbb{R})\\ \text{monotone}}} \alpha\|G - F_1^{[-1]}\|_{L^2}^2 - \beta\|G - F_0^{[-1]}\|_{L^2}^2\,,$$

where $F_0^{[-1]}$ and $F_1^{[-1]}$ are the quantiles of μ_0 and μ_1, respectively. The solution to this problem coincides with the sticky particle model described in Remark 4.6, i.e. $G_\alpha = \tilde{X}(\alpha,\cdot)$ with $\tilde{X}$ as in (4.14).

Remark 4.14 (Dual formulation of the metric extrapolation)**.** *Let us recall that the optimal transport problem (2.1) admits the following dual formulation* [39, Theorem 5.10]*:*

$$\frac{W^2(\mu_0,\mu_1)}{2} = \sup_{\phi_0}\left\{\int \mathcal{H}_1(\phi_0)\mu_1 - \int \phi_0\mu_0 \;:\; \frac{|\cdot|^2}{2} + \phi_0(\cdot) \;\text{ is convex}\right\}, \tag{4.29}$$

and if μ_0 is absolutely continuous, this admits a unique maximiser ϕ_0, and $u(\cdot) \coloneqq \frac{|\cdot|^2}{2} + \phi_0(\cdot)$ is the Brenier potential from μ_0 to μ_1. However, the associated geodesic from μ_0 to μ_1 can be extended up to time $\alpha > 1$ only if (4.4) holds, or equivalently if

$$x \mapsto \frac{|x|^2}{2} + \alpha\phi_0(x) \quad \text{is convex.} \tag{4.30}$$

Therefore, in order to construct an extrapolation, one can instead consider the problem

$$\sup_{\phi_0}\left\{\int \mathcal{H}_1(\phi_0)\mu_1 - \int \phi_0\mu_0 \;:\; \frac{|\cdot|^2}{2} + \alpha\phi_0(\cdot) \;\text{ is convex}\right\}, \tag{4.31}$$

and, if μ_0 is absolutely continuous, set

$$\mathsf{E}_\alpha(\mu_0,\mu_1) = (\nabla u_\alpha)_\#\mu_0,$$

where $u_\alpha(\cdot) \coloneqq \frac{|\cdot|^2}{2} + \alpha\phi_0(\cdot)$ and ϕ_0 solves (4.31). This extrapolation is well defined and it turns out to be a dual formulation for the metric extrapolation in the spirit of [13]*. However, even if very natural, this dual point of view was not needed for the results presented here, and therefore it will be developed in a future work.*

4.4. **Extrapolation on bounded domains.** So far we only discussed the extrapolation problem on the whole space $\mathcal{P}_2(\mathbb{R}^d)$. However, even if the EVBDF2 scheme is well-defined using such extrapolations, it can be convenient for numerical reasons to use an extrapolation operator mapping two measures on $\mathcal{P}(\Omega)$ to an extrapolated one still in $\mathcal{P}(\Omega)$. As mentioned in Remark 4.7, this can be achieved easily in the case of the metric extrapolation, since one can simply perform the minimization problem (4.15) over $\mathcal{P}(\Omega)$ rather than $\mathcal{P}_2(\mathbb{R}^d)$. It is not difficult to check that all the properties discussed in the previous section hold also with this modification.

In general, a straightforward way of defining an extrapolation operator $\mathsf{E}_\alpha^\Omega : \mathcal{P}(\Omega)\times\mathcal{P}(\Omega) \to \mathcal{P}(\Omega)$ is to compose with a W_2 projection. Specifically, given an operator E_α and $\mu_0,\mu_1 \in \mathcal{P}(\Omega)$ we can define:

$$\mathsf{E}_\alpha^\Omega(\mu_0,\mu_1) \coloneqq \operatorname*{argmin}_{\rho\in\mathcal{P}(\Omega)} W_2^2(\rho, \mathsf{E}_\alpha(\mu_0,\mu_1)) = P_\#\mathsf{E}_\alpha(\mu_0,\mu_1)\,,$$

where $P : \mathbb{R}^d \to \Omega$ is the Euclidean projection on the convex set Ω. Then, if E_α is θ-dissipative and satisfies the consistency assumption (1.12), also E_α^Ω does. In fact, denoting by

γ^* the optimal plan from μ_1 to $\mathsf{E}_\alpha(\mu_0,\mu_1)$, $(\mathrm{Id},P)_\#\gamma^* \in \Pi(\mu_1, \mathsf{E}^\Omega_\alpha(\mu_0,\mu_1))$, and therefore one has

$$\begin{aligned} W_2^2(\mu_1, \mathsf{E}^\Omega_\alpha(\mu_0,\mu_1)) &\le \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-P(y)|^2 \mathrm{d}\gamma^*(x,y) \\ &\le \int_{\mathbb{R}^d\times\mathbb{R}^d} |x-y|^2 \mathrm{d}\gamma^*(x,y) = W_2^2(\mu_1, \mathsf{E}_\alpha(\mu_0,\mu_1))\,, \end{aligned}$$

which implies that E^Ω_α is θ-dissipative if so is E_α. Moreover, $\forall \varphi \in C^\infty_c(\mathbb{R}^d)$ with $\nabla\varphi\cdot n_{\partial\Omega} = 0$ on $\partial\Omega$

$$\begin{aligned} \left| \int_{\mathbb{R}^d} \varphi \left(\mathsf{E}^\Omega_\alpha(\mu_0,\mu_1) - \mathsf{E}_\alpha(\mu_0,\mu_1) \right) \right| &= \left| \int_{\mathbb{R}^d} (\varphi\circ P - \varphi) \mathsf{E}_\alpha(\mu_0,\mu_1) \right| \\ &\le \frac{1}{2} ||\mathrm{Hess}(\varphi)||_\infty W_2^2(\mathsf{E}^\Omega_\alpha(\mu_0,\mu_1), \mathsf{E}_\alpha(\mu_0,\mu_1)) \\ &\le \frac{1}{2} ||\mathrm{Hess}(\varphi)||_\infty W_2^2(\mu_1, \mathsf{E}_\alpha(\mu_0,\mu_1))\,, \end{aligned}$$

where to pass from the first to the second line we used a Taylor expansion of φ together with the fact that $\nabla\varphi(P(x))\cdot(P(x)-x) = 0$ on $\mathbb{R}^d$. Hence, using the θ-dissipativity property, we find that if E_α verifies the consistency assumption for all $\varphi \in C^\infty_c(\mathbb{R}^d)$, then E^Ω_α also verifies it for all $\varphi \in C^\infty_c(\mathbb{R}^d)$ such that $\nabla\varphi\cdot n_{\partial\Omega} = 0$ on $\partial\Omega$. As a consequence, the convergence result of Theorem 1.2 holds also when the operator E^Ω_α is used in the extrapolation step.

5. Convergence in the EVI sense

In this section, we make a further assumption on the energy functional $\mathcal{E}$. Besides lower semi-continuity, which ensures well-posedness of the scheme (see Section 3) we assume that $\mathcal{E}$ is λ-convex in the generalized geodesic sense on $\mathcal{P}(\Omega)$, for $\lambda \in \mathbb{R}_+$ (see equation (4.19), and recall that Ω is supposed to be convex, so generalized geodesics with endpoints in $\mathcal{P}(\Omega)$ are well-defined as curves on $\mathcal{P}(\Omega)$). 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

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2} W_2^2(\varrho(t),\nu) \le \mathcal{E}(\nu) - \mathcal{E}(\varrho(t)) - \frac{\lambda}{2} W_2^2(\varrho(t),\nu), \quad \forall t \in (0,T)\,, \tag{5.1}$$

or, equivalently, if for all $r, s \in (0,T)$ with $r \le s$ it holds

$$\frac{1}{2} W_2^2(\varrho(s),\nu) - \frac{1}{2} W_2^2(\varrho(r),\nu) \le \mathcal{E}(\nu)(s-r) - \int_r^s \left(\mathcal{E}(\varrho(t)) + \frac{\lambda}{2} W_2^2(\varrho(t),\nu) \right) \mathrm{d}t\,. \tag{5.2}$$

In this section, we show that the limit curve extracted from the time discretization (1.8) using the metric extrapolation (4.15) (defined on either $\mathcal{P}(\Omega)$ or $\mathcal{P}_2(\mathbb{R}^d)$) satisfies the inequality (5.2).

We first show that for scheme (1.8)-(4.15) a discrete version of the inequality (5.2) holds. As the Wasserstein distance $W_2^2(\cdot,\rho^\alpha_{n-1})$ is 2-convex along any generalized geodesic based in ρ^α_{n-1} (see, e.g., the proof of Lemma 4.9), the overall functional

$$\mathcal{G}(\rho_{n-1},\rho_{n-2};\rho) = \frac{W_2^2(\rho,\rho^\alpha_{n-1})}{2(1-\beta)\tau} + \mathcal{E}(\rho)\,, \tag{5.3}$$

is $\frac{1}{(1-\beta)\tau} + \lambda > 0$ convex along any generalized geodesic on $\mathcal{P}(\Omega)$ based in ρ^α_{n-1}. Note that in order to consider the case $\lambda < 0$ one should explicitly add a restriction on the time step τ so that $\frac{1}{(1-\beta)\tau} + \lambda > 0$.

Lemma 5.1. *At each step n, for all $\nu \in \mathcal{P}(\Omega)$, the following inequality holds:*

$$\Big(\frac{1}{2(1-\beta)\tau}+\frac{\lambda}{2}\Big)W_2^2(\rho_n,\nu)-\alpha\frac{W_2^2(\nu,\rho_{n-1})}{2(1-\beta)\tau}+\beta\frac{W_2^2(\nu,\rho_{n-2})}{2(1-\beta)\tau} \\ \leq \mathcal{E}(\nu)-\mathcal{E}(\rho_n)+\alpha\beta\frac{W_2^2(\rho_{n-1},\rho_{n_2})}{2(1-\beta)\tau}-\frac{W_2^2(\rho_n,\rho_{n-1}^\alpha)}{2(1-\beta)\tau}\,. \tag{5.4}$$

Proof. By the discussion above, considering the generalized geodesic ω between ν and ρ_n with base ρ_{n-1}^α, and using the optimality of ρ_n, we obtain

$$\begin{aligned}0&\leq \mathcal{G}(\rho_{n-1},\rho_{n-2};\omega(t))-\mathcal{G}(\rho_{n-1},\rho_{n-2};\rho_n)\\&\leq t(\mathcal{G}(\rho_{n-1},\rho_{n-2};\nu)-\mathcal{G}(\rho_{n-1},\rho_{n-2};\rho_n))-\frac{1}{2}\Big(\frac{1}{(1-\beta)\tau}+\lambda\Big)t(1-t)W_2^2(\rho_n,\nu).\end{aligned}$$

Dividing by t and taking the limit $t\to 0$, this yields

$$\Big(\frac{1}{2(1-\beta)\tau}+\frac{\lambda}{2}\Big)W_2^2(\rho_n,\nu)-\frac{W_2^2(\nu,\rho_{n-1}^\alpha)}{2(1-\beta)\tau}\leq \mathcal{E}(\nu)-\mathcal{E}(\rho_n)-\frac{W_2^2(\rho_n,\rho_{n-1}^\alpha)}{2(1-\beta)\tau}\,.$$

Adding on both side the term $-\frac{1}{2(1-\beta)\tau}\mathcal{F}(\rho_{n-1},\rho_{n-2};\rho_{n-1}^\alpha)$, using (4.22) on the left-hand side, we obtain

$$\Big(\frac{1}{2(1-\beta)\tau}+\frac{\lambda}{2}\Big)W_2^2(\rho_n,\nu)-\alpha\frac{W_2^2(\nu,\rho_{n-1})}{2(1-\beta)\tau}+\beta\frac{W_2^2(\nu,\rho_{n-2})}{2(1-\beta)\tau} \\ \leq \mathcal{E}(\nu)-\mathcal{E}(\rho_n)-\frac{1}{2(1-\beta)\tau}\mathcal{F}(\rho_{n-1},\rho_{n-2};\rho_{n-1}^\alpha)-\frac{W_2^2(\rho_n,\rho_{n-1}^\alpha)}{2(1-\beta)\tau}\,.$$

Finally, using (4.16) on the right-hand side we conclude. □

Proof of Theorem 1.3. We recall that thanks to the classical estimate (3.2) (Lemma 3.2), 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 W_2 distance to an absolutely continuous limit curve $\varrho:[0,T]\to\mathcal{P}(\Omega)$ (see Proposition 3.4). In order to prove convergence of the scheme in the EVI sense, we show that this curve satisfies inequality (5.2). Thanks to the uniform convergence in time, the procedure is the same as in [29, Theorem 5.1].

For simplicity, assume that given $r,s\in(0,T), r\leq s$, there exist $N_\tau, M_\tau\in\mathbb{N}, N_\tau\leq M_\tau$, such that $r=N_\tau\tau, s=M_\tau\tau$, $\forall\tau$. We multiply by τ inequality (5.4) and sum over n from N_τ to M_τ to obtain the discrete integral form of the EVI:

$$\frac{1}{2(1-\beta)}\sum_{n=N_\tau}^{M_\tau}\big(W_2^2(\rho_n,\nu)-\alpha W_2^2(\nu,\rho_{n-1})+\beta W_2^2(\nu,\rho_{n-2})\big) \\ \leq \mathcal{E}(\nu)(t-s)-\sum_{n=N_\tau}^{M_\tau}\tau\Big(\mathcal{E}(\rho_n)+\frac{\lambda}{2}W_2^2(\rho_n,\nu)\Big) \\ +\frac{1}{2(1-\beta)}\sum_{n=N_\tau}^{M_\tau}\big(\alpha\beta W_2^2(\rho_{n-1},\rho_{n-2})-W_2^2(\rho_n,\rho_{n-1}^\alpha)\big)\,. \tag{5.5}$$

By canceling out terms, the left-hand side is equal to

$$\frac{1}{2(1-\beta)}\left(-\alpha W_2^2(\nu,\rho_{N_\tau-1})+\beta W_2^2(\nu,\rho_{N_\tau-2})+\beta W_2^2(\nu,\rho_{N_\tau-1})\right. \\ \left.+W_2^2(\rho_{M_\tau-1},\nu)+W_2^2(\rho_{M_\tau},\nu)-\alpha W_2^2(\nu,\rho_{M_\tau-1})\right), \tag{5.6}$$

and thanks to the uniform convergence in the Wasserstein distance, (5.6) converges to

$$\frac{1}{2}W_2^2(\varrho(s),\nu)-\frac{1}{2}W_2^2(\varrho(r),\nu),$$

for $\tau\to 0$, where we recall $\alpha-\beta=1$. 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

$$\limsup_{n\to\infty}-\sum_{n=N_\tau}^{M_\tau}\tau\Big(\mathcal{E}(\rho_n)+\frac{\lambda}{2}W_2^2(\rho_n,\nu)\Big)\le-\int_r^s\Big(\mathcal{E}(\varrho(t))+\frac{\lambda}{2}W_2^2(\varrho(t),\nu)\Big)\mathrm{dt}\,.$$

Finally, owing to bound (3.2), we estimate the last contribution of (5.5) as

$$\sum_n\alpha\beta W_2^2(\rho_{n-1},\rho_{n-2})-W_2^2(\rho_n,\rho_{n-1}^\alpha)\le\sum_n\alpha\beta W_2^2(\rho_{n-1},\rho_{n-2})\le C\tau,$$

which converges to zero. As a consequence, we recover the continuous inequality (5.2).

□

6. Finite volume discretization

In this section we describe a space-time discretization of the proposed approach which yields numerically second-order accuracy both in space and time. We consider a discretization in the Eulerian framework of finite volumes. In this setting, neither the free-flow extrapolation nor the metric one have a straightforward implementation. For this reason, we will construct a discrete extrapolation operator based on formula (4.11): in this way the extrapolation step is cast in a variational way allowing for a robust implementation. Although not satisfying the hypotheses of theorem (1.2), this choice leads to a convergent and second order accurate scheme, as we will show numerically. As explained in Remark 4.5, the variational step (4.11) differs from the direct forward integration of the continuity equation. This latter is a viable alternative to define a discrete extrapolation and leads to second order accuracy as well (see [37]), but it is not clear how to discretize this in a robust way.

The fundamental tool is the solution of JKO steps, which requires the expensive problem of computing the Wasserstein distance. Following [12, 30], we linearize the Wasserstein distance obtaining LJKO steps, a more affordable problem to solve. Remarkably, this approach preserves the second order accuracy in time of our time discretization. The discretization in space is based instead on Two-Point Flux Approximation (TPFA) finite volumes with a centered choice for the mobility, which leads to simple and flexible schemes which are second order accurate in space.

6.1. Discrete setting.

TPFA finite volumes require a sufficiently regular partitioning of the domain Ω, according to [18, Definition 9.1]. For simplicity, we describe the methodology in two dimensions only, although generalizations to arbitrary dimensions are possible, and for $\Omega\subset\mathbb{R}^2$ being a polygonal domain. The discretization of Ω consists of three sets: the set of cells $K\in\mathcal{T}$; the set of edges $\sigma\in\overline{\Sigma}$, which is composed of the two subsets of internal edges Σ and external edges $\overline{\Sigma}\setminus\Sigma$; the set of cell centers $(\mathbf{x}_K)_{K\in\mathcal{T}}$. We will denote the

finite volume mesh as $\big(\mathcal{T}, \overline{\Sigma}, (\mathbf{x}_K)_{K\in\mathcal{T}}\big)$. The fundamental regularity hypothesis we need to construct TPFA schemes is the orthogonality between each internal edge $\sigma = K|L \in \Sigma$ and the segment $\mathbf{x}_L - \mathbf{x}_K$. Typical example of meshes that can be used to this end are Cartesian grids, Voronoi tessellations and Delaunay triangulations, by taking the circumcenters of the polygonal cells as cell centers.

For each cell $K \in \mathcal{T}$, we denote $\overline{\Sigma}_K$ and Σ_K the subsets of edges and internal edges belonging to K, and by m_K the measure of the cell. The mesh size h is the largest among all cells' diameters, $h \coloneqq \max_{K\in\mathcal{T}} \operatorname{diam}(K)$, and characterizes the refinement of the mesh. For every internal edge, the diamond cell Δ_σ is the quadrilateral with vertices given by the cell centers, $\mathbf{x}_K$ and $\mathbf{x}_L$, and the vertices of the edge. Denoting by $d_\sigma \coloneqq |\mathbf{x}_L - \mathbf{x}_K|$ and m_σ the measure of the edge, the measure of the diamond cell is equal to $m_{\Delta_\sigma} = \frac{m_\sigma d_\sigma}{d}$, where d stands for the space dimension. Finally, we denote by $d_{K,\sigma}$ the Euclidean distance between the cell center $\mathbf{x}_K$ and the midpoint of the edge $\sigma \in \overline{\Sigma}_K$, and by $\boldsymbol{n}_{K,\sigma}$ the outward unit normal of the cell K on the edge σ.

The finite volume methodology introduces two levels of discretization, on cells and edges. The first one is used to discretize scalar quantities whereas the second one for vectorial ones. To this end, we introduce three discrete inner product spaces $(\mathbb{R}^{\mathcal{T}}, \langle\cdot,\cdot\rangle_{\mathcal{T}})$, $(\mathbb{R}^{\Sigma}, \langle\cdot,\cdot\rangle_{\Sigma})$ and $(\mathbb{F}_{\mathcal{T}}, \langle\cdot,\cdot\rangle_{\mathbb{F}_{\mathcal{T}}})$. The scalar products $\langle\cdot,\cdot\rangle_{\mathcal{T}}$ and $\langle\cdot,\cdot\rangle_{\Sigma}$ are defined as

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

The space $\mathbb{F}_{\mathcal{T}}$ is the space of conservative fluxes, it is defined by

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

and its scalar product is

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

Note that the space $\mathbb{F}_{\mathcal{T}}$ is defined on internal edges only. This is sufficient, since we are dealing with no flux boundary value problems, and therefore we can neglect the flux variables on the boundary. We denote $F_\sigma = |F_{K,\sigma}| = |F_{L,\sigma}|$ the modulus of the flux on each internal edge $\sigma = K|L \in \Sigma$ 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 $\boldsymbol{F} \in \mathbb{F}_{\mathcal{T}}$.

According to finite volumes, the discrete divergence operator $\operatorname{div}_{\mathcal{T}} : \mathbb{F}_{\mathcal{T}} \to \mathbb{R}^{\mathcal{T}}$ is defined in an integral sense as

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

that is, for each cell, the discrete divergence is computed as the sum of the fluxes across its boundary. The discrete gradient $\nabla_\Sigma : \mathbb{R}^{\mathcal{T}} \to \mathbb{F}_{\mathcal{T}}$ is defined by duality, requiring that $\langle\nabla_\Sigma \boldsymbol{a}, \boldsymbol{F}\rangle_{\mathbb{F}_{\mathcal{T}}} = -\langle \boldsymbol{a}, \operatorname{div}_{\mathcal{T}}\boldsymbol{F}\rangle_{\mathcal{T}}$, for all $\boldsymbol{a} \in \mathbb{R}^{\mathcal{T}}$ and $\boldsymbol{F} \in \mathbb{F}_{\mathcal{T}}$. Then, it holds

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

Both the discrete divergence and gradient operators automatically inherit the zero flux boundary condition from the definition of $\mathbb{F}_{\mathcal{T}}$.

The space $(\mathbb{R}^\Sigma, \langle\cdot,\cdot\rangle_\Sigma)$ is introduced in order to match the two different discretizations on cells and edges. In order to reconstruct variables defined on cells to the edges, and vice-versa, we need two reconstruction operators. We use a centered reconstruction for the mobility in order to attain the second order accuracy in space. To this end, we use 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):

$$(\mathcal{L}_\Sigma \boldsymbol{a})_\sigma \coloneqq \lambda_{K,\sigma} a_K + \lambda_{L,\sigma} a_L\,, \quad (\mathcal{L}_\Sigma^* \boldsymbol{u})_K \coloneqq \sum_{\sigma\in\Sigma_K} \lambda_{K,\sigma} u_\sigma \frac{m_\sigma d_\sigma}{m_K}\,, \tag{6.1}$$

for $\boldsymbol{a} \in \mathbb{R}^\mathcal{T}$ and $\boldsymbol{u} \in \mathbb{R}^\Sigma$, with $\lambda_{K,\sigma} + \lambda_{L,\sigma} = 1, \forall \sigma = K|L \in \Sigma$. Two possible choices for the weights are $(\lambda_{K,\sigma}, \lambda_{L,\sigma}) = (\frac{d_{K,\sigma}}{d_\sigma}, \frac{d_{L,\sigma}}{d_\sigma})$ or $(\frac12, \frac12)$, both leading to second order accurate schemes in space [30]. The former choice is possible only if $\mathbf{x}_K \in K$, which may not be always the case for arbitrary admissible meshes.

Remark 6.1. *The definition of the reconstruction operators and the choice of weights may be delicate in general for the discretization of dynamical optimal transport, depending on the discretization chosen for Ω. See [21, 31] for details. Notice in particular that the choice $(\lambda_{K,\sigma}, \lambda_{L,\sigma}) = (\frac12, \frac12)$ may lead to convergence failure in very simple settings [21, Section 5]. Nevertheless, in the context of the discretization of Wasserstein gradient flows the definition of the reconstruction is more flexible, see [12, 20].*

6.2. **Discrete $\dot{H}^{-1}$ norm.** As suggested in [25, 17, 31], a convenient choice for the time discretization of the Wasserstein distance (2.2) is to use a staggered time discretization for the velocity and the density on subintervals of the time interval $[0,1]$, and reconstruct the density on intermediate steps via arithmetic average. It has been shown numerically in [12, 30] that a single step discretization on the whole interval is sufficient in order to preserve the first-order accuracy of the JKO scheme (1.3). Following the same ideas, here we approximate the Wasserstein distance between two measures $\mu, \nu \in \mathcal{P}(\Omega)$ as

$$\frac12 W_2^2(\mu,\nu) \approx \sup_\phi \int_\Omega \phi(\mu-\nu) - \frac12 \int_\Omega \Big(\frac{\mu+\nu}{2}\Big)|\nabla\phi|^2\,. \tag{6.2}$$

Formula (6.2) is obtained by discretizing in one step problem (2.2) and by applying a duality result thanks to the change of variables $(\omega, v) \mapsto (\omega, \omega v)$. For more details on this construction see [12, 30]. This approximation consists in replacing the Wasserstein distance with the weighted dual norm $\frac12 ||\mu-\nu||_{\dot{H}^{-1}_{\frac{\mu+\nu}{2}}}$. The choice of the arithmetic average of the two measures as weight is fundamental in order to achieve second order accuracy in time for the scheme we will propose in the following.

Using the finite volume discretization introduced above we can provide a discrete analogous of the weighted norm. Given the discrete measures $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}_+^\mathcal{T}$ and for any $\boldsymbol{h} \in \mathbb{R}^\mathcal{T}$, the discrete counterpart of the weighted $\dot{H}^{-1}$ norm squared is

$$\mathcal{A}_\mathcal{T}\Big(\frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}; \boldsymbol{h}\Big) \coloneqq \sup_{\boldsymbol{\phi}\in\mathbb{R}^\mathcal{T}} \langle \boldsymbol{h}, \boldsymbol{\phi}\rangle_\mathcal{T} - \frac12 \Big\langle \mathcal{L}_\Sigma\Big(\frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}\Big), |\nabla_\Sigma \boldsymbol{\phi}|^2 \Big\rangle_\Sigma\,. \tag{6.3}$$

A few remarks are in order about such a discretization.

- For any $\boldsymbol{\rho} \in \mathbb{R}_+^\mathcal{T}$, the function $\mathcal{A}_\mathcal{T}(\boldsymbol{\rho};\cdot)$ is proper, convex and lower semi-continuous as supremum of convex and lower semi-continuous functions.
- The supremum is unbounded if the condition $\langle \boldsymbol{h}, \mathbf{1}\rangle_\mathcal{T} = 0$ is not satisfied. On other hand, if $\langle \boldsymbol{h}, \mathbf{1}\rangle_\mathcal{T} = 0$, there exists a maximizer $\boldsymbol{\phi}$, which is however not uniquely

defined, since the function maximised in (6.2) is invariant with respect to addition of a global constant or perturbations sufficiently far from the support of $\boldsymbol{h}$, $\boldsymbol{\mu}$, and $\boldsymbol{\nu}$.

- Setting $\boldsymbol{h} = \boldsymbol{\nu} - \boldsymbol{\mu}$ in (6.3), with $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ being a discrete approximation of two measures μ and ν, we obtain a discrete version of $W_2^2(\mu, \nu)/2$. In this case the optimal potential $\boldsymbol{\phi}$ can be interpreted as a discrete counterpart of a continuous optimal potential ϕ, satisfying the Hamilton-Jacobi equation on the time interval $[0, 1]$, evaluated at time $1/2$.
- The total kinetic energy is discretized on the diamond cells. Notice that due to the definition of the scalar product $\langle \cdot, \cdot \rangle_\Sigma$, the measure of each diamond cell is taken $m_\sigma d_\sigma = dm_{\Delta_\sigma}$, i.e. d times the actual measure. This is done in order to compensate for the unidirectional discretization, since each term $|\nabla_{K,\sigma}\boldsymbol{\phi}|$ is meant as an approximation of the quantity $|\nabla\phi \cdot \boldsymbol{n}_{K,\sigma}|$, and have a consistent discretization. See [31] for more details on this construction.

6.3. **Discrete extrapolation.** We now construct a discrete version of the extrapolation operator E_α at time α, by discretizing the procedure described in Section 4.2, and in particular of equation (4.11). The proposed strategy requires three subsequent steps: i) compute the interpolation between the two measures; ii) integrate forward in time the optimal potential; and finally iii) solve a JKO step.

Let us consider two discrete densities $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}_+^{\mathcal{T}}$ with the same total discrete mass $\langle \boldsymbol{\mu}, \mathbf{1} \rangle_{\mathcal{T}} = \langle \boldsymbol{\nu}, \mathbf{1} \rangle_{\mathcal{T}}$. The first step requires to solve problem (6.3) for $\boldsymbol{h} = \boldsymbol{\nu} - \boldsymbol{\mu}$ in order to find an optimal potential $\boldsymbol{\phi}$, which approximates the continuous one, solution to the Hamilton-Jacobi equation (2.5), at the midpoint of the time interval $[0, 1]$.

In the second step, we evolve the optimal potential according to the Hamilton-Jacobi equation until the final time α, that is considering a temporal step of length $\frac{1}{2} + \beta = \frac{\alpha+\beta}{2}$. This can be done with an explicit Euler step as follows:

$$\boldsymbol{\phi}^\alpha = \boldsymbol{\phi} - \frac{2}{\alpha + \beta} \frac{1}{2} \mathcal{L}_\Sigma^* |\nabla_\Sigma \boldsymbol{\phi}|^2 \,. \tag{6.4}$$

Note that we use the operator $\mathcal{L}_\Sigma^*$ to reconstruct the square of the gradient of the potential. However, as this step is not variational, it is not mandatory to use the adjoint of the reconstruction $\mathcal{L}_\Sigma$ and any other (second order) strategy can be adopted.

Finally, for the third step, we approximate problem (4.11) using again the discrete weighted $\dot{H}^{-1}$ norm. Specifically, we define a discrete extrapolation operator as a map $\mathsf{E}_\alpha^{\mathcal{T}} : \mathbb{R}_+^{\mathcal{T}} \times \mathbb{R}_+^{\mathcal{T}} \to \mathbb{R}_+^{\mathcal{T}}$ verifying

$$\mathsf{E}_\alpha^{\mathcal{T}}(\boldsymbol{\mu}, \boldsymbol{\nu}) \in \operatorname*{argmin}_{\boldsymbol{\rho} \in \mathbb{R}_+^{\mathcal{T}}} \frac{1}{\alpha} \mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho} + \boldsymbol{\mu}}{2}; \boldsymbol{\mu} - \boldsymbol{\rho}\Big) - \langle \boldsymbol{\phi}^\alpha, \boldsymbol{\rho} \rangle_{\mathcal{T}} \,, \tag{6.5}$$

for all $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}_+^{\mathcal{T}}$ and where $\boldsymbol{\phi}^\alpha$ is given by equation (6.4). Due to the definition of $\mathcal{A}_{\mathcal{T}}$, any solution $\boldsymbol{\rho}$ satisfies $\langle \boldsymbol{\rho}, \mathbf{1} \rangle_{\mathcal{T}} = \langle \boldsymbol{\nu}, \mathbf{1} \rangle_{\mathcal{T}}$. However, since $\boldsymbol{\phi}$ is in general not unique, in order to specify a discrete extrapolation operator one needs to select a specific optimal potential for any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \mathbb{R}_+^{\mathcal{T}}$.

6.4. **A space-time discrete EVBDF2 scheme.** We can finally formulate our second order finite volume scheme. Consider a convex discrete energy function $\mathcal{E}_{\mathcal{T}} : \mathbb{R}^{\mathcal{T}} \to \mathbb{R}$ and the two initial densities $\boldsymbol{\rho}_0, \boldsymbol{\rho}_1 \in \mathbb{R}_+^{\mathcal{T}}$, with the same total discrete mass. We define the subspace of discrete probability measures $\mathbb{P}_{\mathcal{T}} \subset \mathbb{R}^{\mathcal{T}}$ as

$$\mathbb{P}_{\mathcal{T}} = \{ \boldsymbol{\rho} \in \mathbb{R}_+^{\mathcal{T}} : \langle \boldsymbol{\rho}, \mathbf{1} \rangle_{\mathcal{T}} = \langle \boldsymbol{\rho}_0, \mathbf{1} \rangle_{\mathcal{T}} \} \,.$$

For the 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:

$$\begin{cases} \boldsymbol{\rho}_{n-1}^{\alpha} = \mathsf{E}_{\alpha}^{\mathcal{T}}(\boldsymbol{\rho}_{n-2}, \boldsymbol{\rho}_{n-1})\,, \\ \boldsymbol{\rho}_n \in \underset{\boldsymbol{\rho}\in\mathbb{R}_+^{\mathcal{T}}}{\operatorname{argmin}} \dfrac{1}{\tau(1-\beta)} \mathcal{A}_{\mathcal{T}}\Big(\dfrac{\boldsymbol{\rho}+\boldsymbol{\rho}_{n-1}^{\alpha}}{2}; \boldsymbol{\rho}_{n-1}^{\alpha} - \boldsymbol{\rho}\Big) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})\,. \end{cases} \tag{6.6}$$

The LJKO step in (6.6) is a well posed convex optimization problem. Uniqueness of the solution at each step is guaranteed if $\mathcal{E}_{\mathcal{T}}$ is strictly convex. Moreover, due to the definition of $\mathcal{A}_{\mathcal{T}}$, any solution $\boldsymbol{\rho}$ belongs to $\mathbb{P}_{\mathcal{T}}$.

Remark 6.2 (Efficient implementation via the interior method)**.** *Problem* (6.5) *and the LJKO step in* (6.6) *can be solved efficiently thanks to an interior point algorithm, as suggested in* [30] *(see also* [31, 19]*). This implies that the density will be always strictly greater than zero, up to the tolerance set for the solver. Hence, one can compute the solution $\boldsymbol{\phi}$, required to define $\mathsf{E}_{\alpha}^{\mathcal{T}}$, solving directly the linear system given by the optimality condition of problem* (6.3)*:*

$$\boldsymbol{\nu} - \boldsymbol{\mu} + \operatorname{div}_{\mathcal{T}}(\mathcal{L}\Big(\frac{\boldsymbol{\mu}+\boldsymbol{\nu}}{2}\Big) \odot \nabla\boldsymbol{\phi}) = 0\,, \tag{6.7}$$

where $\odot$ denotes the component-wise product, which has then a unique solution defined up to a global additive constant.

6.5. **Other implementations.** We now propose a discrete version of the extrapolation-based version of the VIM scheme (1.19) and the BDF2 scheme (1.16) within the same TPFA finite volume setting introduced above. We will study these numerically in Section 7.2.1 by comparing their solutions to the solutions provided by scheme (6.6) on one-dimensional test cases.

Our formulation of the VIM scheme (1.19) requires solving a JKO step with time step $\frac{\tau}{2}$ and then computing the 2-extrapolation. Using the tools introduced above, in the discrete setting this can be formulated as follows. Given the initial density $\boldsymbol{\rho}_0 \in \mathbb{P}_{\mathcal{T}}$ and a time step $\tau > 0$, construct the sequence of densities $(\boldsymbol{\rho}_n)_{n\geq 1} \subset \mathbb{P}_{\mathcal{T}}$ by solving at each step n

$$\begin{cases} \boldsymbol{\rho}_{n-\frac{1}{2}} \in \underset{\boldsymbol{\rho}\in\mathbb{R}_+^{\mathcal{T}}}{\operatorname{argmin}} \dfrac{2}{\tau} \mathcal{A}_{\mathcal{T}}\Big(\dfrac{\boldsymbol{\rho}+\boldsymbol{\rho}_{n-1}}{2}; \boldsymbol{\rho}_{n-1} - \boldsymbol{\rho}\Big) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})\,, \\ \boldsymbol{\rho}_n = \mathsf{E}_2^{\mathcal{T}}(\boldsymbol{\rho}_{n-1}, \boldsymbol{\rho}_{n-\frac{1}{2}})\,. \end{cases} \tag{6.8}$$

As before, the discrete LJKO steps can be computed thanks to an interior point algorithm. From a computational point of view, this scheme is cheaper to compute than (6.6), as in this case the value of the optimal potential in the discrete weighted $\dot{H}^{-1}$ norm from $\boldsymbol{\rho}_{n-1}$ to $\boldsymbol{\rho}_{n-\frac{1}{2}}$ is already known from the LJKO step and does not need to be computed. However, in the next section, we will show numerically that the solutions produced by the VIM scheme (6.8) are much more oscillatory than those obtained with the EVBDF2 scheme.

We can also propose a naive discretization of the BDF2 scheme (1.16) by replacing the Wasserstein distances with discrete weighted $\dot{H}^{-1}$ norms. 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, compute $\boldsymbol{\rho}_n$ as solution to

$$\inf_{\boldsymbol{\rho}\in\mathbb{R}_+^{\mathcal{T}}} \frac{\alpha}{(1-\beta)\tau} \mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}_{n-1}}{2}; \boldsymbol{\rho}_{n-1} - \boldsymbol{\rho}\Big) - \frac{\beta}{(1-\beta)\tau} \mathcal{A}_{\mathcal{T}}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}_{n-2}}{2}; \boldsymbol{\rho}_{n-2} - \boldsymbol{\rho}\Big) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})\,. \tag{6.9}$$

Problem (6.9) is not a convex optimization problem. Notice that it is not even bounded from below in general. Indeed, the function $\mathcal{A}_{\mathcal{T}}(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}_{n-2}}{2}; \boldsymbol{\rho}_{n-2} - \boldsymbol{\rho})$ is not bounded from above if

the density $\boldsymbol{\rho}_{n-2}$ is not supported everywhere. We can nevertheless try to compute stationary points of the objective function in (6.9) using again an interior point algorithm. Despite not being a robust and completely meaningful strategy, in some cases it is possible to solve the problem, which enables us to compare it to our implementation.

Remark 6.3. *In one dimension, as pointed out in Remark 4.13, both the metric extrapolation (1.14) and the BDF2 scheme (1.16) can be recast as convex optimization problems. In this case it is possible then to design effective discretizations for these (as originally done in [29]). Nevertheless, this approach requires, at least in the Eulerian framework, to be able to switch between discrete densities and discrete quantile functions, and it does not appear obvious how to achieve this while preserving the second order accuracy of the space discretization.*

7. Numerical validation of the EVBDF2 scheme

The objective of this section is to validate our numerical scheme (6.6). We will first show qualitatively its behavior with simple one-dimensional examples and compare it to the schemes (6.8) and (6.9). We then show that all these three approaches lead to a second order accurate discretization in both time and space. 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 3.2 and the porous medium equation. This latter writes

$$\partial_t \varrho = \Delta \varrho^\delta + \operatorname{div}(\varrho \nabla V)\,, \tag{7.1}$$

and it is a Wasserstein gradient flow with respect to the energy

$$\mathcal{E}(\rho) = \int_\Omega \frac{1}{\delta - 1}\rho^\delta + \rho V\,, \tag{7.2}$$

for a given $\delta > 1$ and with $V \in W^{1,\infty}(\Omega)$ a Lipschitz continuous exterior potential [32]. The energy functionals (3.5) and (7.2) are both of the form $\mathcal{E}(\rho) = \int_\Omega E(\rho)\mathrm{d}x$ for a strictly convex function $E : \mathbb{R}_+ \to \mathbb{R}$. They can be straightforwardly discretized as $\mathcal{E}_\mathcal{T} = \sum_{K\in\mathcal{T}} E(\rho_K)m_K$. Finally, we will test scheme (6.6) on a more challenging application in order to show its flexibility and robustness, that is an incompressible immiscible multiphase flow in a porous medium.

We remark that when two initial conditions $\boldsymbol{\rho}_0, \boldsymbol{\rho}_1$ are needed, we compute first $\boldsymbol{\rho}_1$ from $\boldsymbol{\rho}_0$ via an LJKO step:

$$\boldsymbol{\rho}_1 = \operatorname*{argmin}_{\boldsymbol{\rho}\in\mathbb{R}_+^\mathcal{T}} \frac{1}{\tau}\mathcal{A}_\mathcal{T}\Big(\frac{\boldsymbol{\rho}+\boldsymbol{\rho}_0}{2}; \boldsymbol{\rho}_0 - \boldsymbol{\rho}\Big) + \mathcal{E}_\mathcal{T}(\boldsymbol{\rho})\,.$$

In the ODE setting, computing the second initial condition via a first step of implicit Euler scheme ensures the overall second order accuracy [16]. This strategy reveals to be numerically effective also in this setting.

7.1. Comparison of the three approaches.

We compare the three different approaches on simple one dimensional tests for the diffusion equation and the porous medium equation. For both system we set $\Omega = [0, 1]$, discretized in subintervals of equal length $m_K = 0.02$.

We first consider the diffusion equation, which is problem (3.4) with zero external potential V. We take as initial condition

$$\rho_0 = \exp\Big(-50\Big(x - \frac{1}{2}\Big)^2\Big)\,,$$

which we discretize as $\boldsymbol{\rho}_0 = (\rho_0(\mathbf{x}_K))_{K\in\mathcal{T}}$, and the time step $\tau = 0.01$. In Figure 4, we show the density obtained with the three schemes at three different times. Using the VIM scheme

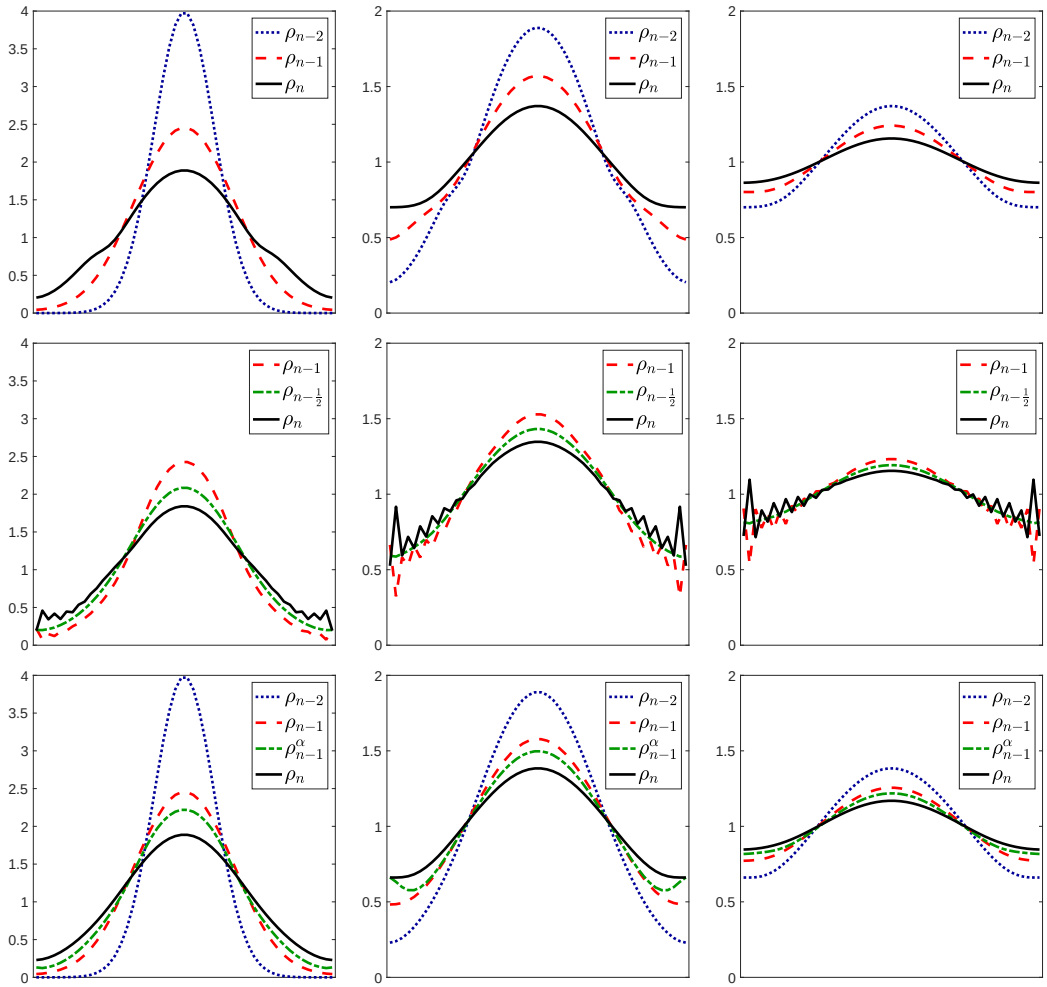


FIGURE 4. Comparison between the three schemes for the diffusion equation. From top to bottom, the BDF2 scheme (6.9), the VIM scheme (6.8) and the EVBDF2 scheme (6.6). From left to right, three different time steps: $t = 0.02, 0.04, 0.06$.

(6.8), spurious oscillations appear in the solution and these persist along the integration in time. Such oscillations can be explained as the result of the interaction of the extrapolation step, causing the mass to exit the domain, and the boundary conditions, forcing the mass to stay within Ω. Neither the EVBDF2 scheme (6.6) nor the BDF2 scheme (6.9) suffer from this problem. However, notice that in both cases the dynamics slightly differ from pure diffusion due to the presence of bumps in the solution.

Consider now the porous medium equation (7.1) with $\delta = 2$ and the external potential $V(x) = -x$, which causes the mass to drift towards the positive direction. We take as initial condition

$$\rho_0(x) = \mathbb{1}_{x \leq \frac{3}{10}},$$

discretized again as $\boldsymbol{\rho}_0 = (\rho_0(\mathbf{x}_K))_{K \in \mathcal{T}}$, and the time step $\tau = 0.002$. In this case, the naive implementation we proposed for the BDF2 scheme does not converge, which is not surprising

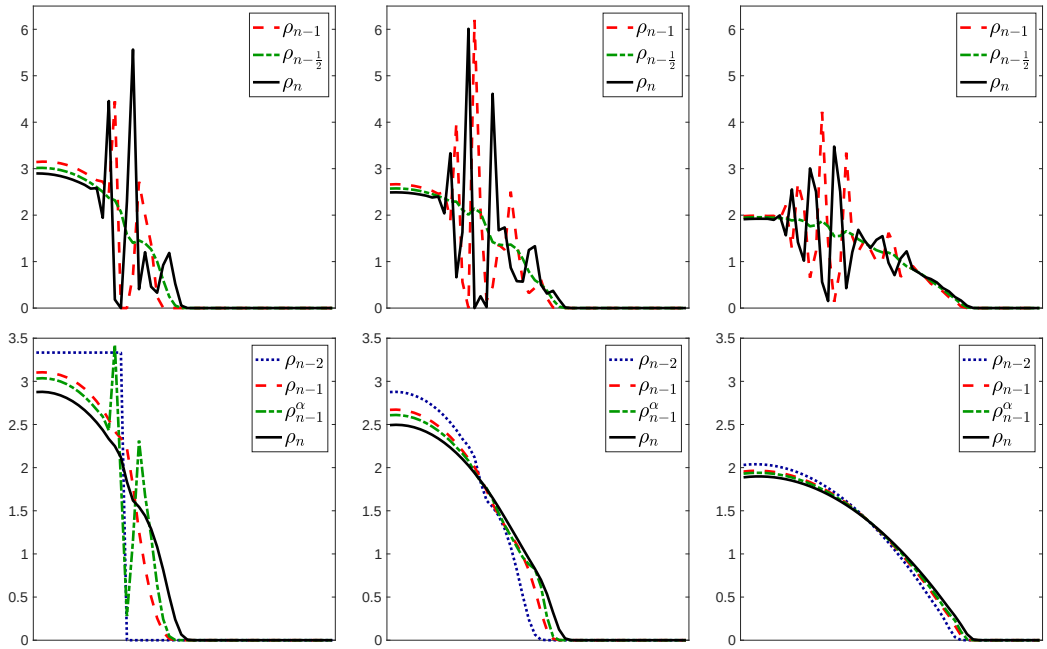


FIGURE 5. Comparison between the VIM scheme (6.8) (top row) and the EVBDF2 scheme (6.6) (bottom row) for the porous medium equation. The BDF2 scheme (6.9) does not converge in this case. From left to right, three different time steps: $t = 0.004, 0.008, 0.020$.

since the objective function in (6.9) is unbounded from below. The results for the VIM scheme (6.8) and the EVBDF2 scheme (6.6) are shown in Figure 5. Again, the VIM scheme is unstable whereas the EVBDF2 scheme controls and smooths the oscillations generated by the extrapolation step. Note that in this case the oscillations are due to the compact support of the density 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.

Finally, we observe that, as in the continuous setting, we cannot expect any regularity on the measure obtained after the extrapolation, and the JKO step is the only source of regularity for both the EVBDF2 and the VIM scheme. One may argue that the two schemes perform the same operations up to a temporal shift, which should contradict the different behavior shown in Figure 4. However, notice that scheme (6.6) performs a smaller extrapolation and a bigger JKO step with respect to scheme (6.8). Furthermore, in (6.6) one needs to compute an extrapolation between two minimizers of the JKO step, whereas in (6.8) the extrapolation is between an extrapolated measure and a JKO minimizer.

7.2. **Convergence tests.** We now compare the three schemes in terms of order of convergence with respect to an exact one-dimensional solution of the Fokker-Planck equation (3.4). For the EVBDF2 scheme (6.6), we will also perform two dimensional tests using the porous medium equation (7.1). For all tests, we consider a sequence of meshes $\left(\mathcal{T}_m, \overline{\Sigma}_m, (\mathbf{x}_K)_{K\in\mathcal{T}_m}\right)$ with decreasing meshsize h_m and a sequence of decreasing time steps τ_m such that $\frac{h_{m+1}}{h_m} = \frac{\tau_{m+1}}{\tau_m}$. We solve the discrete problem for each couple (h_m, τ_m) and evaluate the convergence

TABLE 1. Errors and convergence rates for the three schemes for the Fokker-Planck equation in one dimension. Integration time $[0, 0.25]$ for the first three cases, $[0.05, 0.25]$ for the last one.

		BDF2 (6.9)		EVBDF2 (6.6)		VIM (6.8)		VIM (6.8)	
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

with respect to the discrete $L^1((0,T);L^1(\Omega))$ error:

$$\epsilon_m = \sum_n \tau \sum_{K\in\mathcal{T}_m} |\rho_{K,n} - \varrho(\mathbf{x}_K, n\tau)| m_K \,.$$

We compute the rate of convergence as:

$$\frac{\log(\epsilon_{m-1}) - \log(\epsilon_m)}{\log(\tau_{m-1}) - \log(\tau_m)} \,.$$

7.2.1. *One-dimensional tests.* On the domain $\Omega = [0,1]$ and for the external potential $V(x) = -gx$, we consider the following exact solution to the Fokker-Planck equation (3.4):

$$\varrho(t,x) = \exp\left(-\Big(\pi^2 + \frac{g^2}{4}\Big)t + \frac{g}{2}x\right)\Big(\pi\cos(\pi x) + \frac{g}{2}\sin(\pi x)\Big) + \pi\exp\Big(g\Big(x - \frac{1}{2}\Big)\Big). \tag{7.3}$$

We consider the value $g = 1$. For each mesh $\big(\mathcal{T}_m, \overline{\Sigma}_m, (\mathbf{x}_K)_{K\in\mathcal{T}_m}\big)$ and time step τ_m, we compute then the discrete solution using the three schemes, starting from the initial condition $\boldsymbol{\rho}_0 = (\varrho(0,\mathbf{x}_K))_{K\in\mathcal{T}}$. The results are presented in Table 1. Both the BDF2 and the EVBDF2 schemes are second order accurate, whereas the order of convergence is less than one for the VIM scheme. This is due to the presence of oscillations in the solutions obtained with the VIM scheme, which are however only present at the beginning of the time interval $[0, 0.25]$. Repeating the test on the interval $[0.05, 0.25]$, the convergence significantly improves and attains second order accuracy as well.

7.2.2. *Two-dimensional tests.* We now estimate the order of convergence of the EVBDF2 scheme on two-dimensional test cases. Here, we set $\Omega = [0,1]^2$ and use the same sequence of grids that have been used in [12, 30], which allows for a direct comparison of the results therein.

We repeat first the test on the Fokker-Planck equation in two dimensions using the same solution (7.3) on the domain $\Omega = [0,1]^2$. The results are shown in Table 2 and confirm the second order accuracy of the scheme.

We also perform a convergence test with respect to an explicit solution of the porous medium equation (7.1) with zero exterior potential V. This equation admits a solution called Barenblatt profile [32]:

$$\varrho(t,x) = \frac{1}{t^{d\lambda}}\Big(\frac{\delta-1}{\delta}\Big)^{\frac{1}{\delta-1}} \max\Big(M - \frac{\lambda}{2}\Big|\frac{x-x_0}{t^\lambda}\Big|^2, 0\Big)^{\frac{1}{\delta-1}}, \tag{7.4}$$

TABLE 2. Errors and convergence rate for the EVBDF2 scheme (6.6) for the Fokker-Planck equation in two dimensions.

h_m	τ_m	ϵ_m	rate
0.2986	0.0500	2.111e-02	/
0.1493	0.0250	6.800e-03	1.634
0.0747	0.0125	2.017e-03	1.754
0.0373	0.0063	5.669e-04	1.831
0.0187	0.0031	1.535e-04	1.884

TABLE 3. Errors and convergence rates for the EVBDF2 scheme (6.6) 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

where $\lambda = \frac{1}{d(\delta-1)+2}$, d standing for the space dimension, and 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 (7.4) solves (7.1) on the domain $\Omega = [0,1]^d$, with $\mathbf{x}_0$ in the interior of Ω, starting from $t_0 > 0$ and for a sufficiently small time horizon T, such that the mass does not reach the boundary of the domain. We consider the two-dimensional case and $x_0 = (0.5, 0.5)$. We solve the problem for $\delta = 2, 3, 4$, with initial condition $\boldsymbol{\rho}_0 = (\varrho(t_0, \mathbf{x}_K))_{K\in\mathcal{T}}$, starting respectively from $t_0 = 10^{-4}, 10^{-5}, 10^{-6}$ and up to time $T = t_0 + 10^{-3}$. The results are presented in Table 3. The convergence profile is not clean, probably 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 approaching order two with refinement. In the cases $\delta = 3, 4$, where the solution is less regular, the order tends to 1.5.

7.3. **Incompressible immiscible multiphase flows in porous media.** Incompressible immiscible multiphase flows in porous media can be described as Wasserstein gradient flows, as shown in [10]. We recall quickly the model problem in a simplified way. In the porous medium Ω, $N+1$ phases are flowing and we denote by $\boldsymbol{s} = (s_0, ..., s_N)$ the saturations of each phase, i.e. the portion of volume occupied by each phase in each point. The evolution of each

saturation obeys the following equations:

$$\begin{cases} \dfrac{\partial s_i}{\partial t} + \operatorname{div}(s_i v_i) = 0\,, \\ v_i = -\dfrac{1}{\mu_i}(\nabla p_i - \rho_i g)\,, \\ p_i - p_0 = \pi_i(\boldsymbol{s}, x)\,, \end{cases} \tag{7.5}$$

$i \in \{0, ..., N\}$ for the first two equations, $i \in \{1, ..., N\}$ for the third one, plus the total saturation condition $\sum_{i=0}^{N} s_i(t,x) = 1$ and the no-flux boundary conditions. The densities ρ_i and the viscosities μ_i, both constant in the whole domain, are characteristic of each phase. In (7.5) the porosity of the medium is considered constant and neglected. The term $\rho_i g$ reflects the influence of the potential energy on the motion (g is the gravitational acceleration), but other types of potential energy could be considered. The model is completed specifying the N capillary pressure relations, described by the functions π_i.

We introduce the probability spaces

$$\mathcal{P}_i = \left\{ s_i \in \mathcal{P}(\Omega) : s_i(\Omega) = c_i \right\}, \quad i \in \{0, ..., N\},$$

with the constant c_i denoting the total mass of each phase. Each space $\mathcal{P}_i$ is endowed with the following quadratic Wasserstein distance,

$$W_{2,i}^2(s_i^1, s_i^2) = \min_{\gamma \in \Pi(s_i^1, s_i^2)} \int \mu_i |x - y|^2 \mathrm{d}\gamma(x,y)\,,$$

for $s_i^1, s_i^2 \in \mathcal{P}_i$ and we can define the global quadratic Wasserstein distance $\boldsymbol{W}_2$ on $\boldsymbol{\mathcal{P}} := \mathcal{P}_0 \times ... \times \mathcal{P}_N$ by setting

$$\boldsymbol{W}_2^2(\boldsymbol{s}^1, \boldsymbol{s}^2) = \sum_{i=0}^{N} W_{2,i}^2(s_i^1, s_i^2), \quad \forall \boldsymbol{s}^1, \boldsymbol{s}^2 \in \boldsymbol{\mathcal{P}}.$$

Problem (7.5) can then be represented as the gradient flow in the space $\boldsymbol{\mathcal{P}}$ with respect to the (strictly convex) energy functional

$$\mathcal{E}(\boldsymbol{s}) = \int_\Omega \boldsymbol{\Psi} \cdot \boldsymbol{s} + \int_\Omega \Pi(\boldsymbol{s}, x) + i_{\boldsymbol{\mathcal{S}}}(\boldsymbol{s})\,, \tag{7.6}$$

where $\boldsymbol{\Psi} = (\Psi_0, \dots, \Psi_N)$ is the exterior gravitational potential given by

$$\Psi_i(x) = -\rho_i g \cdot x, \quad \forall x \in \Omega\,,$$

$\Pi(s,x)$ is a strictly convex potential such that

$$\pi_i(\boldsymbol{s}, x) = \frac{\partial \Pi(\boldsymbol{s}, x)}{\partial s_i}, \quad i \in \{1, ..., N\},$$

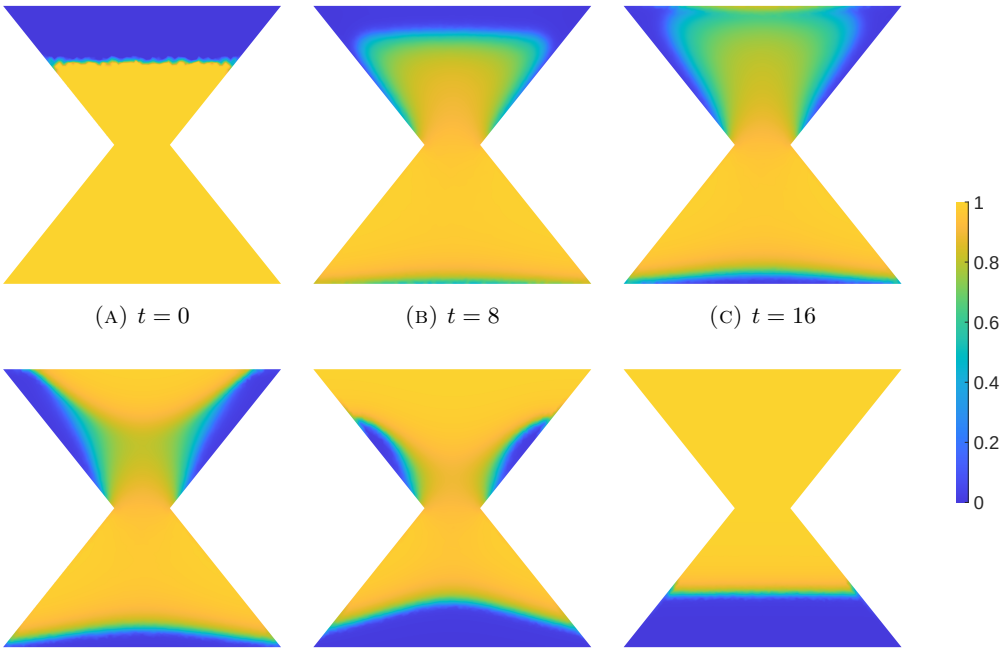
and $i_{\boldsymbol{\mathcal{S}}}$ is the indicator function of the set

$$\boldsymbol{\mathcal{S}} = \left\{ \boldsymbol{s} \in \boldsymbol{\mathcal{P}} : \sum_{i=0}^{N} s_i(x) = 1, \text{for a.e. } x \in \Omega \right\}.$$

When applying the EVBDF2 scheme to such gradient flow, the extrapolation may be taken in each space $\mathcal{P}_i$ independently, i.e. we define the extrapolation in the space $\boldsymbol{\mathcal{P}}$ as

$$\mathsf{E}_\alpha(\boldsymbol{s}^1, \boldsymbol{s}^2) \coloneqq (\mathsf{E}_\alpha(s_i^1, s_i^2))_{i=0}^N\,,$$

for all $\boldsymbol{s}^1, \boldsymbol{s}^2 \in \boldsymbol{\mathcal{P}}$. This does not guarantee at all that at each step n of the scheme the extrapolation is a feasible point for $\mathcal{E}(\boldsymbol{s})$, that is $\mathsf{E}_\alpha(\boldsymbol{s}^1, \boldsymbol{s}^2) \notin \boldsymbol{\mathcal{S}}$ in general even though



(A) $t = 0$ (B) $t = 8$ (C) $t = 16$

(D) $t = 24$ (E) $t = 48$ (F) $t = 120$

FIGURE 6. Evolution of the saturation of the oil phase in the hourglass. The evolution of the water is complementary. As expected, the water, the denser phase, flows down the hourglass under the effect of gravity up until reaching the bottom.

$\boldsymbol{s}^1, \boldsymbol{s}^2 \in \boldsymbol{\mathcal{S}}$. Nevertheless, the resulting scheme is well defined as well as the numerical approach (6.6). In our implementation, we linearize each Wasserstein distances independently. The energy functional can be discretized straightforwardly.

As a specific instance of problem (7.5), we consider a two-phase flow, where water (s_0) and oil (s_1) are competing in the porous medium. We choose the classical Brooks-Corey capillary pressure model,

$$p_1 - p_0 = \pi_1(s_1) = \lambda(1 - s_1)^{-\frac{1}{2}},$$

and take g acting along the negative direction of the y axis, $|g| = 9.81$. We set the model parameter $\lambda = 0.05$. The densities and the viscosities of the two fluids are, respectively, $\rho_0 = 1$ and $\rho_1 = 0.87$, $\mu_0 = 1$ and $\mu_1 = 100$. We consider a non convex domain Ω shaped as an hourglass and set an initial condition where the water is distributed uniformly in a layer in the upper part, whereas the oil takes the complementary space (see Figure 6a). The evolution of the oil saturation s_1 is presented in Figure 6.

8. Conclusion

In this work we proposed and analyzed different notions of extrapolation in the Wasserstein space. We showed how these can be used to construct a second-order time discretization of Wasserstein gradient flows, based on a two-step reformulation of the classical BDF2 scheme. According to the specific notion considered, we could prove different types of convergence

TABLE 4. Summary of the different types of extrapolation proposed in the present work.

	Free-flow extrapolation (4.7)	Viscosity extrapolation (4.9)	Metric extrapolation (4.15)
Fokker-Planck conv.	✓	?	✓
EVI conv.	?	?	✓
Implementation	?	✓	?
Second order	?	✓	?

guarantees for the scheme. We also proposed a fully-discrete version of the method, and demonstrated numerically its second-order accuracy in space and time. The possibility to provide an implementable scheme is in fact the main advantage of our approach compared to previous works also based on the BDF2 scheme [29], or on the midpoint rule [27]. The different type of extrapolations and their properties are summarized in Table 4.

In order to provide our fully discrete scheme, we worked in the framework of Eulerian discretizations and considered an extrapolation based on viscosity solutions of the Hamilton-Jacobi equation. The resulting scheme is robust and allows to achieve second order of accuracy both in space and time, but it does not verify the hypotheses of our convergence results. The free-flow extrapolation could be implemented straightforwardly in the framework of Lagrangian discretizations (see, e.g., [28, 9] for Lagrangian discretizations of Wasserstein gradient flows), although in this setting it would be challenging to achieve second order accuracy in space. The metric extrapolation enjoys the nicest mathematical structure, and in principle one could exploit its dual formulation (4.31), which is a convex optimization problem, to implement it numerically. However, dealing with the strong-convexity constraint on the Brenier potential requires the development of dedicated tools. We will investigate this direction in a future work.

ACKNOWLEDGEMENTS

This work was partly supported by the Labex CEMPI (ANR-11-LABX-0007-01). TOG acknowledges the support of the french Agence Nationale de la Recherche through the project MAGA (ANR-16-CE40-0014). GT acknowledges that 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. The authors would like to thank Clément Cancès and Guillaume Carlier for fruitful discussions and suggestions on the topic.

REFERENCES

[1] Luigi Ambrosio, Elia Brué, Daniele Semola, et al. *Lectures on optimal transport*. Springer, 2021.

[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
[3] B Ben Moussa and GT Kossioris. On the system of hamilton–jacobi and transport equations arising in geometrical optics. 2003.
[4] 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.
[5] 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.
[6] Adrien Blanchet. A gradient flow approach to the Keller-Segel systems. RIMS Kokyuroku's lecture notes, vol. 1837, pp. 52–73, June 2013.
[7] Ilya A Bogaevsky. Matter evolution in Burgulence. *arXiv preprint math-ph/0407073*, 2004.
[8] Yann Brenier and Emmanuel Grenier. Sticky particles and scalar conservation laws. *SIAM Journal on Numerical Analysis*, 35(6):2317–2328, 1998.
[9] Vincent Calvez and Thomas Gallouët. Particle approximation of the one dimensional keller-segel equation, stability and rigidity of the blow-up. *arXiv preprint arXiv:1404.0139*, 2014.
[10] 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.
[11] Clément Cancès, Daniel Matthes, and Flore Nabet. A two-phase two-fluxes degenerate cahn–hilliard model as constrained wasserstein gradient flow. *Archive for Rational Mechanics and Analysis*, 233(2):837–866, 2019.
[12] Clément Cancès, Thomas Gallouët, and Gabriele Todeschi. A variational finite volume scheme for wasserstein gradient flows. *Numerische Mathematik*, 146:437–480, 10 2020.
[13] Guillaume Carlier. Remarks on toland's duality, convexity constraint and optimal transport. 2008.
[14] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM Journal on Mathematical Analysis*, 49(2):1385–1418, 2017.
[15] Jose A. Carrillo, Katy Craig, Li Wang, and Chaozhen Wei. Primal dual methods for wasserstein gradient flows, 2019.
[16] Peter Deuflhard and Folkmar Bornemann. *Scientific computing with ordinary differential equations*, volume 42. Springer Science & Business Media, 2002.
[17] 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.
[18] Robert Eymard, Thierry Gallouët, and Raphaèle Herbin. Finite volume methods. In *Handbook of Numerical Analysis*, volume 7, pages 713–1020.
[19] Enrico Facca, Gabriele Todeschi, Andrea Natale, and Michele Benzi. Efficient preconditioners for solving dynamical optimal transport via interior point methods. *arXiv preprint arXiv:2209.00315*, 2022.
[20] Dominik Forkert, Jan Maas, and Lorenzo Portinale. Evolutionary Γ-convergence of entropic gradient flow structures for fokker-planck equations in multiple dimensions. *arXiv preprint arXiv:2008.10962*, 2020.
[21] Peter Gladbach, Eva Kopfer, and Jan Maas. Scaling limits of discrete optimal transport. *arXiv preprint arXiv:1809.01092*, 2018.
[22] Richard Jordan, David Kinderlehrer, and Felix Otto. The Variational Formulation of the Fokker–Planck Equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.
[23] Konstantin Khanin and Andrei Sobolevski. Particle dynamics inside shocks in Hamilton–Jacobi equations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 368(1916):1579–1593, 2010.
[24] Philippe Laurençot and Bogdan-Vasile Matioc. A gradient flow approach to a thin film approximation of the muskat problem. *Calculus of Variations and Partial Differential Equations*, 47(1):319–341, 2013.
[25] 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.
[26] Hugo Leclerc, Quentin Mérigot, Filippo Santambrogio, and Federico Stra. Lagrangian discretization of crowd motion and linear diffusion. *SIAM Journal on Numerical Analysis*, 58(4):2093–2118, 2020.
[27] 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.
[28] Daniel Matthes and Horst Osberger. Convergence of a variational lagrangian scheme for a nonlinear drift diffusion equation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 48(3):697–726, 2014.

[29] 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.
[30] 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.
[31] Andrea Natale and Gabriele Todeschi. Computation of optimal transport with finite volumes. *ESAIM: Mathematical Modelling and Numerical Analysis*, 55(5):1847–1871, 2021.
[32] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[33] Simon Plazotta. A bdf2-approach for the non-linear fokker-planck equation, 2018.
[34] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.
[35] Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.
[36] Filippo Santambrogio. Crowd motion and evolution PDEs under density constraints. *ESAIM: Proceedings and Surveys*, 64:137–157, 2018.
[37] Gabriele Todeschi. *Finite volume approximation of optimal transport and Wasserstein gradient flows*. PhD thesis, PSL Université Paris Dauphine, 2021.
[38] C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.
[39] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Thomas O. Gallouët (thomas.gallouet@inria.fr), Team Mokaplan, Inria Paris 75012 Paris, CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016 Paris, France

Andrea Natale (andrea.natale@inria.fr), Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France

Gabriele Todeschi (gabriele.todeschi@univ-grenoble-alpes.fr), Univ. Grenoble-Alpes, ISTerre, F-38058 Grenoble, France

2.2 Wasserstein Gradient flows

2.2.1 Incompressible immiscible multiphase flows in porous media

Articles:

- **The gradient flow structure for incompressible immiscible two-phase flows in porous media.** *C. R. Acad. Sci. Paris, Ser. I(353) :985– 989* (2015). https://hal.science/hal-01122770. Cancès C., Gallouët T.O., Monsaingeon L.
- **Incompressible immiscible multiphase flows in porous media: a variational approach.** *Analysis and PDE Vol. 10* (2017), No. 8, 1845–1876 https://arxiv.org/abs/1607.04009. Cancès C., Gallouët T.O., Monsaingeon L.
- **Simulation of multiphase porous media flows with minimizing movement and finite volume schemes.)** *European Journal of Applied Mathematics, Cambridge University Press (CUP), 30 (6), pp.1123-1152* (2019). https://arxiv.org/abs/arXiv:1802.01321. Cancès C., Gallouët T.O., Laborde M., Monsaingeon L.

Collaborators: All this three papers are written in collaboration with Clément Cancès and Léonard Monsaingeon. We also worked with Maxime Laborde for the numerical paper.

Main contributions:

- We underline the variational structure (Wasserstein Gradient flow) for the system of N Incompressible, immiscible flows in a porous media. It means a couple (Energy, Metric) under which the system of PDE becomes a gradient flow.
- We prove the existence of solutions for this system. this result is new and was not achieved at that time by classical PDE tools other than for two phases. The proof is based on the convergence of a JKO scheme. It uses, among other, flow interchange and duality methods.
- We proposed implemented and compared two numerical methods which are both designed to decrease the natural energy.

Research directions: One natural research direction is to understand what happens if the energy vanishes. This is the object of Erwan Stampli's Phd that I co-supervised with Y. Brenier. We have two works in progress on this subject, one proving the convergence towards the Muskat problem on the torus in dimension 1. This proof is based on a modulated energy method.

The gradient flow structure for incompressible immiscible two-phase flows in porous media

Clément Cancès [a], Thomas O. Gallouët [b] Léonard Monsaingeon [c]

[a] *Team RAPSODI, Inria Lille - Nord Europe, 40, av. Halley, 59650 Villeneuve d'Ascq, France*

[b] *CMLS, UMR 7640, Ecole Polytechnique, FR-91128 Palaiseau Cedex*

[c] *CAMGSD, Instituto Superior Técnico, Universidade de Lisboa, Av. Rovisco Pais, 1049-001 Lisboa, Portugal*

Received on March 4, 2015; accepted after revision on September 24, 2015

Presented by ...

Abstract

We show that the widely used model governing the motion of two incompressible immiscible fluids in a possibly heterogeneous porous medium has a formal gradient flow structure. More precisely, the fluid composition is governed by the gradient flow of some non-smooth energy. Starting from this energy together with a dissipation potential, we recover the celebrated Darcy-Muskat law and the capillary pressure law governing the flow thanks to the steepest descent condition for the energy. Our interpretation does not require the introduction of any algebraic transformation like, e.g., the global pressure or the Kirchhoff transform, and can be transposed to the case of more phases. *To cite this article: A. Name1, A. Name2, C. R. Acad. Sci. Paris, Ser. I 340 (2005).*

Résumé

La structure de flot gradient pour les écoulements incompressibles immiscible en milieux poreux. Nous montrons qu'un modèle très couramment utilisé dans l'industrie pour décrire un écoulement diphasique incompressible et immiscible dans un milieux poreux possiblement hétérogène possède une structure de flot gradient. Plus précisément, la composition du fluide est gouvernée par flot gradient d'une énergie singulière. En partant de cette énergie et d'un potentiel de dissipation, nous retrouvons les lois de Darcy-Muskat et de pression capillaire gouvernant l'écoulement à l'aide d'un principe de moindre dissipation de l'énergie. Notre interprétation ne nécessite pas l'introduction de transformation algébrique du type pression globale ou transformée de Kirchhoff, ce qui permet son extension à un nombre plus grand de phases. *Pour citer cet article : A. Name1, A. Name2, C. R. Acad. Sci. Paris, Ser. I 340 (2005).*

Email addresses: clement.cances@inria.fr (Clément Cancès), thomas.gallouet@polytechnique.edu (Thomas O. Gallouët), leonard.monsaingeon@tecnico.ulisboa.pt (Léonard Monsaingeon).

1. Introduction

1.1. *General motivations*

The models for multiphase porous media flows have been widely studied in the last decades since they are of great interest in several fields of applications, like e.g. oil-engineering, carbon dioxide sequestration, or nuclear waste repository management. We refer to the monographs [5,6] for an extensive discussion on the derivation of models for porous media flows, and to [4,11,3,13] for numerical and mathematical studies.

More recently, F. Otto showed in his seminal work [17] that the so-called *porous medium equation*:

$$\partial_t \rho - \Delta \rho^m = 0 \quad \text{for } (\boldsymbol{x}, t) \in \mathbb{R}^N \times \mathbb{R}_+ \text{ and } m > 1,$$

which is a very simplified model corresponding to the case of an isentropic gas flowing within a porous medium, can be reinterpreted in a physically relevant way as the gradient flow of the free energy with respect to some Wasserstein metric in the space of Borel probability measures. Extensions to more general degenerate parabolic equations were then proposed for example in [1,15].

In this note, we will focus on the model governing the motion of an incompressible immiscible two-phase flow in a possibly heterogeneous porous medium, that will appear in the sequel as (3) and (11)–(13). This model is relevant for instance for describing the flow of oil and water, whence the subscripts o and w appearing in the sequel of this note, within a rock that is possibly made of several rock-types. Our goal is to show that, at least formally, this model can be reinterpreted as the gradient flow of some singular energy. This will motivate the use of structure-preserving numerical methods inspired from [9] to this model in the future.

Our approach is inspired from the one of A. Mielke [16] and, more closely, to the one of M. A. Peletier [18]. The basic recipe for variational modeling is recalled in §1.2, then its ingredients are identified in §2. This approach is purely formal, but it can be made rigorous under some unphysical strict positivity assumption on the phase mobilities η_o, η_w defined below. We will remain sloppy about regularity issues all along this note.

1.2. *The recipe of getting formal variational models*

Here we recall very briefly the main ingredients needed for defining a formal gradient flow.

- **i.** The *state space* $\mathcal{M}$ is the set where the solution of the gradient flow can evolve.
- **ii.** At a point $\boldsymbol{s} \in \mathcal{M}$, the tangent space $T_{\boldsymbol{s}}\mathcal{M}$, to whom would belong $\partial_t \boldsymbol{s}$, is identified in a non-unique way with a so-called *process space* $\mathcal{Z}_{\boldsymbol{s}}$ (that might depend on $\boldsymbol{s}$). More precisely, we assume that for each $\boldsymbol{s} \in \mathcal{M}$ there exists an onto linear application $\mathcal{P}(\boldsymbol{s}) : \mathcal{Z}_{\boldsymbol{s}} \to T_{\boldsymbol{s}}\mathcal{M}$.
- **iii.** The *energy functional* $\mathcal{E} : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ admits a (local) sub-differential $\partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s}) \subset (T_{\boldsymbol{s}}\mathcal{M})^*$ at $\boldsymbol{s} \in \mathcal{M}$.
- **iv.** The *dissipation potential* $\mathcal{D}$ is such that, for all $\boldsymbol{s} \in \mathcal{M}$ and all $\mathbf{V} \in \mathcal{Z}_{\boldsymbol{s}}$, one has $\mathcal{D}(\boldsymbol{s}; \mathbf{V}) \geq 0$. It is supposed to be convex and coercive w.r.t. to its second variable.
- **v.** The initial data $\boldsymbol{s}^0$ belongs to $\mathcal{M}$.

All these ingredient being defined, we obtain from the *steepest descent condition* that $\boldsymbol{s} : \mathbb{R}_+ \to \mathcal{M}$ is the gradient flow of the energy $\mathcal{E}$ for the dissipation $\mathcal{D}$ if

$$\partial_t \boldsymbol{s} = \mathcal{P}(\boldsymbol{s})\mathbf{V} \quad \text{where} \quad \mathbf{V} \in \operatorname*{argmin}_{\widehat{\mathbf{V}} \in \mathcal{Z}_{\boldsymbol{s}}} \left(\max_{\boldsymbol{h} \in \partial_{\boldsymbol{s}} \mathcal{E}(\boldsymbol{s})} \left(\mathcal{D}\big(\boldsymbol{s}(t); \widehat{\mathbf{V}}(t)\big) + \left\langle \boldsymbol{h}\,,\, \mathcal{P}(\boldsymbol{s})\widehat{\mathbf{V}} \right\rangle_{(T_{\boldsymbol{s}}\mathcal{M})^*, T_{\boldsymbol{s}}\mathcal{M}} \right) \right). \tag{1}$$

The formula (1) means that a gradient flow is lazy and smart: the motion aims to minimize the dissipation while maximizing the decay of the energy. We refer to [16,18] for additional material on such a formal modeling and to [2] for an extensive (and rigorous) discussion on gradient flows in metric spaces.

2. Variational modeling for two-phase flows in porous media

2.1. *State space and process space*

Let Ω be an open subset of $\mathbb{R}^N$ representing a (possibly heterogeneous) *porous medium*, let $\phi : \Omega \to (0,1)$ be a measurable function (called *porosity*) such that $\underline{\phi} \leq \phi(\boldsymbol{x}) \leq \overline{\phi}$ for a.e. $\boldsymbol{x} \in \Omega$ for some constants $\underline{\phi}, \overline{\phi} \in (0,1)$, and let $\underline{s}_o, \underline{s}_w : \Omega \to [0,1)$ be two measurable functions (so-called *residual saturations*) such that $\underline{s}_o(\boldsymbol{x}) + \underline{s}_w(\boldsymbol{x}) < 1$ for a.e. $\boldsymbol{x} \in \Omega$. In what follows, we denote by

$$\overline{s}_o(\boldsymbol{x}) = 1 - \underline{s}_w(\boldsymbol{x}), \qquad \overline{s}_w(\boldsymbol{x}) = 1 - \underline{s}_o(\boldsymbol{x}), \qquad \text{for a.e. } \boldsymbol{x} \in \Omega.$$

For almost all $\boldsymbol{x} \in \Omega$, we denote by

$$\Delta_{\boldsymbol{x}} = \Big\{ \boldsymbol{s} = (s_o, s_w) \in \mathbb{R}^2 \;\Big|\; s_o + s_w = 1 \text{ with } \underline{s}_\alpha(\boldsymbol{x}) \leq s_\alpha \leq \overline{s}_\alpha(\boldsymbol{x}) \text{ for } \alpha \in \{o, w\} \Big\}.$$

Let $\boldsymbol{s}^0 = (s_o^0, s_w^0)$ be a given initial saturation profile, we denote by m_α ($\alpha \in \{o, w\}$) the volume occupied by the phase α in the porous medium, i.e.,

$$m_o = \int_\Omega \phi(\boldsymbol{x}) s_o^0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \quad \text{and} \quad m_w = \int_\Omega \phi(\boldsymbol{x}) s_w^0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}.$$

For simplicity, we restrict our attention to the case where the volume of each phase is preserved: no source term and no-flux boundary conditions (otherwise, non-autonomous gradient flows should be considered). Hence the saturation profile lies at each time in the so-called state space $\mathcal{M}$, defined by

$$\mathcal{M} = \left\{ \boldsymbol{s} = (s_o, s_w) \;\middle|\; s_\alpha : \Omega \to \mathbb{R}_+ \text{ with } \int_\Omega \phi(\boldsymbol{x}) s_\alpha(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = m_\alpha \text{ for } \alpha \in \{o, w\} \right\}.$$

Let us now describe the processes that allow to transform the saturation profile. We denote by

$$\mathcal{Z}_{\boldsymbol{s}} = \Big\{ \mathbf{V} = (\boldsymbol{v}_o, \boldsymbol{v}_w) \;\Big|\; \boldsymbol{v}_\alpha : \Omega \to \mathbb{R}^N \text{ with } \boldsymbol{v}_\alpha \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega \Big\}$$

the *process space* of the admissible processes for modifying a saturation profile $\boldsymbol{s} \in \mathcal{M}$. The identification between $\mathbf{V} = (\boldsymbol{v}_o, \boldsymbol{v}_m) \in \mathcal{Z}_{\boldsymbol{s}}$ and $\dot{\boldsymbol{s}} = (\dot{s}_o, \dot{s}_w) \in T_{\boldsymbol{s}}\mathcal{M}$ is made through the onto operator $\mathcal{P}(\boldsymbol{s}) : \mathcal{Z}_{\boldsymbol{s}} \to T_{\boldsymbol{s}}\mathcal{M}$ defined by

$$\mathcal{P}(\boldsymbol{s})\mathbf{V} = \left(-\frac{1}{\phi} \boldsymbol{\nabla} \cdot \boldsymbol{v}_o \,;\, -\frac{1}{\phi} \boldsymbol{\nabla} \cdot \boldsymbol{v}_w \right), \qquad \forall \mathbf{V} \in \mathcal{Z}_{\boldsymbol{s}}. \tag{2}$$

Since $\partial_t \boldsymbol{s} \in T_{\boldsymbol{s}}\mathcal{M}$, the relation (2) yields the existence of some *phase filtration speeds* $(\boldsymbol{v}_o, \boldsymbol{v}_w) \in \mathcal{Z}_{\boldsymbol{s}}$ such that the following *continuity equations* hold:

$$\phi \partial_t s_\alpha + \boldsymbol{\nabla} \cdot \boldsymbol{v}_\alpha = 0, \qquad \alpha \in \{o, w\}. \tag{3}$$

The relation (3) must be understood as the local volume conservation of each phase $\alpha \in \{o, w\}$. Finally, the duality bracket $\langle \cdot\,,\, \cdot \rangle_{(T_{\boldsymbol{s}}\mathcal{M})^*, T_{\boldsymbol{s}}\mathcal{M}}$ is given by

$$\langle \boldsymbol{h}, \dot{\boldsymbol{s}} \rangle_{(T_{\boldsymbol{s}}\mathcal{M})^*, T_{\boldsymbol{s}}\mathcal{M}} = \sum_{\alpha \in \{o,w\}} \int_\Omega \phi h_\alpha \dot{s}_\alpha = - \sum_{\alpha \in \{o,w\}} \int_\Omega h_\alpha \boldsymbol{\nabla} \cdot \boldsymbol{v}_\alpha = \sum_{\alpha \in \{o,w\}} \int_\Omega \boldsymbol{\nabla} h_\alpha \cdot \boldsymbol{v}_\alpha.$$

2.2. *About the energy*

For a.e. $\boldsymbol{x} \in \Omega$, we assume the *capillary pressure* graph $\pi(\cdot, \boldsymbol{x}) : [\underline{s}_o(\boldsymbol{x}), \overline{s}_o(\boldsymbol{x})] \to 2^{\mathbb{R}}$ to be a maximal monotone graph whose restriction $\pi_{|_{(\underline{s}_o, \overline{s}_o)}}(\cdot, \boldsymbol{x})$ to the open interval $(\underline{s}_o(\boldsymbol{x}), \overline{s}_o(\boldsymbol{x}))$ is an increasing (single-valued) function belonging to $L^1(\underline{s}_o(\boldsymbol{x}), \overline{s}_o(\boldsymbol{x}))$. In particular, $\pi^{-1}(\cdot, \boldsymbol{x}) : \mathbb{R} \to [\underline{s}_o(\boldsymbol{x}), \overline{s}_o(\boldsymbol{x})]$ is a single valued function.

We denote by $\Pi : \mathbb{R} \times \Omega \to \mathbb{R} \cup \{+\infty\}$ the (strictly convex w.r.t. its first variable) function defined by

$$\Pi(s_o, \boldsymbol{x}) = \begin{cases} \displaystyle\int_{\sigma(\boldsymbol{x})}^{s_o} \pi(a, \boldsymbol{x}) \mathrm{d}a - (\rho_o - \rho_w) s g z & \text{if } s_o \in [\underline{s}_o(\boldsymbol{x}), \overline{s}_o(\boldsymbol{x})], \\ +\infty & \text{otherwise,} \end{cases}$$

where, denoting by $\boldsymbol{e}_z$ the downward unit normal vector of $\mathbb{R}^N$, we have set $z = \boldsymbol{x} \cdot \boldsymbol{e}_z$, and where g and ρ_α denote the gravity constant and the density of the phase α respectively, and where σ is such that $\boldsymbol{x} \mapsto \pi(\sigma(\boldsymbol{x}), \boldsymbol{x}) - (\rho_o - \rho_w) g z$ is constant. Since $\pi_{|_{(\underline{s}_o, \overline{s}_o)}}(\cdot, \boldsymbol{x}) \in L^1(\underline{s}_o(\boldsymbol{x}), \overline{s}_o(\boldsymbol{x}))$, we get that $\Pi(\underline{s}_o(\boldsymbol{x}), \boldsymbol{x})$ and $\Pi(\overline{s}_o(\boldsymbol{x}), \boldsymbol{x})$ are finite for a.e. $\boldsymbol{x} \in \Omega$.

The *volume energy* function $E : \mathbb{R}^2 \times \Omega \to \mathbb{R} \cup \{+\infty\}$ is defined by

$$E(\boldsymbol{s}, \boldsymbol{x}) = \begin{cases} \Pi(s_o, \boldsymbol{x}) & \text{if } \boldsymbol{s} = (s_o, s_w) \in \Delta_{\boldsymbol{x}}, \\ +\infty & \text{otherwise.} \end{cases} \tag{4}$$

The function $E(\cdot, \boldsymbol{x})$ is convex and finite on $\Delta_{\boldsymbol{x}}$ for a.e. $\boldsymbol{x} \in \Omega$. Its sub-differential is given by

$$\partial_{\boldsymbol{s}} E(\boldsymbol{s}, \boldsymbol{x}) = \begin{cases} \Big\{ (h_o, h_w) \in \mathbb{R}^2 \;\Big|\; h_o - h_w + (\rho_o - \rho_w) g z \in \pi(s_o, \boldsymbol{x}) \Big\} & \text{if } \boldsymbol{s} \in \Delta_{\boldsymbol{x}}, \\ \emptyset & \text{otherwise.} \end{cases}$$

Finally, we can define the so-called *global energy* $\mathcal{E} : \mathcal{M} \to \mathbb{R} \cup \{+\infty\}$ by

$$\mathcal{E}(\boldsymbol{s}) = \int_\Omega \phi(\boldsymbol{x}) E(\boldsymbol{s}(\boldsymbol{x}), \boldsymbol{x}) \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{s} = (s_o, s_w) \in \mathcal{M}. \tag{5}$$

The saturation profile $\boldsymbol{s} \in \mathcal{M}$ is of finite energy $\mathcal{E}(\boldsymbol{s}) < \infty$ if and only if $\boldsymbol{s}(\boldsymbol{x}) \in \Delta_{\boldsymbol{x}}$ for a.e. $\boldsymbol{x} \in \Omega$. For $\boldsymbol{s} \in \mathcal{M}$ with finite energy one can check that the local sub-differential $\partial_{\boldsymbol{s}} \mathcal{E}(\boldsymbol{s})$ of $\mathcal{E}$ at $\boldsymbol{s}$ is given by

$$\partial_{\boldsymbol{s}} \mathcal{E}(\boldsymbol{s}) = \Big\{ \boldsymbol{h} = (h_o, h_w) : \Omega \to \mathbb{R}^2 \,\Big|\, h_o - h_w + (\rho_o - \rho_w) g z \in \pi(s_o, \boldsymbol{x}) \;\text{ for a.e. } \boldsymbol{x} \in \Omega \Big\}. \tag{6}$$

2.3. *About the dissipation*

The *permeability tensor* field $\boldsymbol{\Lambda} \in L^\infty(\Omega; \mathbb{R}^{N \times N})$ is assumed to be such that $\boldsymbol{\Lambda}(\boldsymbol{x})$ is a symmetric and positive matrix for a.e. $\boldsymbol{x} \in \Omega$. Moreover, we assume that there exist $\lambda_\star, \lambda^\star \in \mathbb{R}_+^*$ such that

$$\lambda_\star |\boldsymbol{u}|^2 \le \boldsymbol{\Lambda}(\boldsymbol{x}) \boldsymbol{u} \cdot \boldsymbol{u} \le \lambda^\star |\boldsymbol{u}|^2, \qquad \text{for all } \boldsymbol{u} \in \mathbb{R}^N \text{ and a.e. } \boldsymbol{x} \in \Omega.$$

This ensures that $\boldsymbol{\Lambda}(\boldsymbol{x})$ is invertible for a.e. $\boldsymbol{x} \in \Omega$. Its inverse is denoted by $\boldsymbol{\Lambda}^{-1}(\boldsymbol{x})$.

We also need the two Carathéodory functions $\eta_o, \eta_w : \mathbb{R} \times \Omega \to \mathbb{R}_+$ — the so-called *phase mobilities* — such that $\eta_\alpha(\cdot, \boldsymbol{x})$ are Lipschitz continuous and nondecreasing on $\mathbb{R}_+$ for a.e. $\boldsymbol{x} \in \Omega$ and $\alpha \in \{o, w\}$. Moreover, we assume that $\eta_\alpha(s, \boldsymbol{x}) = 0$ if $s \le \underline{s}_\alpha(\boldsymbol{x})$ and that $\eta_\alpha(s, \boldsymbol{x}) > 0$ if $s > \underline{s}_\alpha(\boldsymbol{x})$.

Given $\boldsymbol{s} = (s_o, s_w) \in \mathcal{M}$ and $\mathbf{V} = (\boldsymbol{v}_o, \boldsymbol{v}_w) \in \mathcal{Z}_{\boldsymbol{s}}$, we define the *dissipation potential* $\mathcal{D}$ by

$$\mathcal{D}(\boldsymbol{s}, \mathbf{V}) = \frac{1}{2} \sum_{\alpha \in \{o,w\}} \int_\Omega \frac{\boldsymbol{\Lambda}^{-1} \boldsymbol{v}_\alpha \cdot \boldsymbol{v}_\alpha}{\eta_\alpha(s_\alpha)} \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{s} \in \mathcal{M},\ \forall \mathbf{V} \in \mathcal{Z}_{\boldsymbol{s}}.$$

The finiteness of the dissipation, i.e., $\mathcal{D}(\boldsymbol{s}, \mathbf{V}) < \infty$, implies $\boldsymbol{v}_\alpha = \mathbf{0}$ a.e. on $\{\boldsymbol{x} \in \Omega \mid s_\alpha(\boldsymbol{x}) \le \underline{s}_\alpha(\boldsymbol{x})\}$.

2.4. *Steepest descent condtition and resulting equations*

Let us consider the gradient flow governed by the energy $\mathcal{E}$, the continuity equation (3), and the dissipation $\mathcal{D}$. Let $\boldsymbol{s} \in \mathcal{M}$ be a finite energy saturation profile, then because of the *steepest descent condition* (1) and of the definition (2) of the operator $\mathcal{P}(\boldsymbol{s}) : \mathcal{Z}_{\boldsymbol{s}} \to T_{\boldsymbol{s}}\mathcal{M}$, the process $\mathbf{V} = (\boldsymbol{v}_o, \boldsymbol{v}_w) \in \mathcal{Z}_{\boldsymbol{s}}$ and the *hydrostatic phase pressures* $\boldsymbol{h} = (h_o, h_w)$ must be chosen so that $(\mathbf{V}, \boldsymbol{h})$ is the $\min - \max$ saddle-point of the functional

$$(\widehat{\mathbf{V}}, \widehat{\boldsymbol{h}}) \mapsto \mathcal{D}(\boldsymbol{s}, \widehat{\mathbf{V}}) - \sum_{\alpha \in \{o,w\}} \int_\Omega \widehat{h}_\alpha \boldsymbol{\nabla} \cdot \widehat{\boldsymbol{v}}_\alpha \mathrm{d}\boldsymbol{x}. \tag{7}$$

One can first fix $\widehat{\boldsymbol{h}} \in \partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$ and minimize w.r.t. $\mathbf{V}$. This provides

$$\operatorname*{argmin}_{\widehat{\mathbf{V}} \in \mathcal{Z}} \left(\mathcal{D}(\boldsymbol{s}, \widehat{\mathbf{V}}) - \sum_{\alpha \in \{o,w\}} \int_\Omega \widehat{h}_\alpha \boldsymbol{\nabla} \cdot \widehat{\boldsymbol{v}}_\alpha \mathrm{d}\boldsymbol{x} \right) = \left(-\eta_o(s_o) \boldsymbol{\Lambda} \boldsymbol{\nabla} \widehat{h}_o, -\eta_w(s_w) \boldsymbol{\Lambda} \boldsymbol{\nabla} \widehat{h}_w \right). \tag{8}$$

Injecting this expression in (7) and maximizing w.r.t. $\widehat{\boldsymbol{h}} \in \partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$, that is minimizing

$$\boldsymbol{h} = \operatorname*{argmin}_{\widehat{\boldsymbol{h}} \in \partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})} \left(\frac{1}{2} \int_\Omega \eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \nabla \hat{h}_\alpha \cdot \nabla \hat{h}_\alpha \right) \tag{9}$$

among all elements $\widehat{\boldsymbol{h}}$ in the subdifferential $\partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$, yields

$$-\boldsymbol{\nabla} \cdot \Big(\boldsymbol{v}_o + \boldsymbol{v}_w \Big) = 0, \qquad \boldsymbol{v}_\alpha = -\eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \boldsymbol{\nabla} h_\alpha. \tag{10}$$

In (10) the first condition follows from the constraint $\hat{h} \in \partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$ in (9), and the second one from (8).

Define the *phase pressures* $\boldsymbol{p} = (p_o, p_w)$ by $p_\alpha(\boldsymbol{x}) = h_\alpha(\boldsymbol{x}) + \rho_\alpha g z$, for a.e. $\boldsymbol{x} \in \Omega$ and $\alpha \in \{o, w\}$, then we recover the classical *Darcy-Muskat law*:

$$\boldsymbol{v}_\alpha = -\eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \boldsymbol{\nabla} \left(p_\alpha - \rho_\alpha g z \right), \quad \alpha \in \{o, w\}. \tag{11}$$

Moreover, it follows from (6) that the following *capillary pressure relation* holds:

$$p_o(\boldsymbol{x}) - p_w(\boldsymbol{x}) \in \pi(s_o(\boldsymbol{x}), \boldsymbol{x}) \quad \text{ a.e. in } \Omega. \tag{12}$$

We recover here the multivalued capillary pressure relation proposed in [19,7,8,10].

Combining (3) and (10) easily gives $\partial_t(s_o + s_w) = 0$, so that the condition

$$s_o + s_w = 1 \quad \text{a.e. in } \Omega, \tag{13}$$

is preserved along time and the whole pore volume remains saturated by the two fluids.

Gathering (3), (11), (12) and (13) gives the usual system of equations governing immiscible incompressible two-phase flows in porous media [5,11,3,12,10].

Remark 1 By similarity with the classical Wasserstein distance used in optimal mass transport [17] one could here endow the tangent space $T_{\boldsymbol{s}}\mathcal{M}$ at $\boldsymbol{s} \in \mathcal{M}$ with a weighted $\dot{H}^{-1}$-scalar product

$$\left(\dot{\boldsymbol{s}}_1, \dot{\boldsymbol{s}}_2 \right)_{T_{\boldsymbol{s}}\mathcal{M}} = \sum_{\alpha \in \{o,w\}} \int_\Omega \eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \boldsymbol{\nabla} h_{1,\alpha} \cdot \boldsymbol{\nabla} h_{2,\alpha} \mathrm{d}\boldsymbol{x},$$

where, for $i \in \{1, 2\}$ and $\alpha \in \{o, w\}$, we have set $\dot{\boldsymbol{s}}_i = (\dot{s}_{i,o}, \dot{s}_{i,w})$ and where $h_{i,\alpha}$ solves

$$-\boldsymbol{\nabla} \cdot \left(\eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \boldsymbol{\nabla} h_{i,\alpha} \right) = \dot{s}_{i,\alpha} \text{ in } \Omega, \qquad \eta_\alpha(s_\alpha) \boldsymbol{\Lambda} \boldsymbol{\nabla} h_{i,\alpha} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega.$$

Under some conditions on the functions η_α (see [14]), this should allow us to consider $\mathcal{M}$ as a metric space endowed with the corresponding distance, but $\mathcal{E}$ is not locally λ-convex for this Riemannian structure. The minimization (9) *then consists in the selection of the subgradient with minimal norm.*

Acknowledgements

This work was supported by the French National Research Agency ANR through grant ANR-13-JS01-0007-01 (Geopor project). TG acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2014-2019 Grant Agreement QUANTHOM 335410). LM was supported by the Portuguese Science Fundation through FCT fellowship SFRH/BPD/88207/2012.

References

[1] M. Agueh. Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory. *Adv. Differential Equations*, 10(3):309–360, 2005.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[3] S. N. Antontsev, A. V. Kazhikhov, and V. N. Monakhov. *Boundary value problems in mechanics of nonhomogeneous fluids*, volume 22 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1990. Translated from the Russian.

[4] K. Aziz and A. Settari. *Petroleum Reservoir Simulation*. Elsevier Applied Science Publishers, Londres, 1979.

[5] J. Bear. *Dynamic of Fluids in Porous Media*. American Elsevier, New York, 1972.

[6] J. Bear and Y. Bachmat. *Introduction to modeling of transport phenomena in porous media*, volume 4. Springer, 1990.

[7] F. Buzzi, M. Lenzinger, and B. Schweizer. Interface conditions for degenerate two-phase flow equations in one space dimension. *Analysis*, 29:299–316, 2009.

[8] C. Cancès, T. Gallouët, and A. Porretta. Two-phase flows involving capillary barriers in heterogeneous porous media. *Interfaces Free Bound.*, 11(2):239–258, 2009.

[9] C. Cancès and C. Guichard. Numerical analysis of a robust entropy-diminishing finite volume scheme for degenerate parabolic equations. HAL: hal-01119735, submitted for publication.

[10] C. Cancès and M. Pierre. An existence result for multidimensional immiscible two-phase flows with discontinuous capillary pressure field. *SIAM J. Math. Anal.*, 44(2):966–992, 2012.

[11] G. Chavent and J. Jaffré. *Mathematical Models and Finite Elements for Reservoir Simulation*, volume 17. North-Holland, Amsterdam, stud. math. appl. edition, 1986.

[12] Z. Chen. Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution. *J. Differential Equations*, 171(2):203–232, 2001.

[13] Z. Chen, G. Huan, and Y. Ma. *Computational methods for multiphase flows in porous media*, volume 2. SIAM, 2006.

[14] J. Dolbeault, B. Nazaret, and G. Savaré A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009.

[15] S. Lisini. Nonlinear diffusion equations with variable coefficients as gradient flows in Wasserstein spaces. *ESAIM Control Optim. Calc. Var.*, 15(3):712–740, 2009.

[16] A. Mielke. A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.

[17] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[18] M. A. Peletier. Variational modelling: Energies, gradient flows, and large deviations. Lecture Notes, Würzburg. Available at `http://www.win.tue.nl/~mpeletie`, Feb. 2014.

[19] B. Schweizer. Homogenization of degenerate two-phase flow equations with oil trapping. *SIAM J. Math. Anal.*, 39(6):1740–1763, 2008.

INCOMPRESSIBLE IMMISCIBLE MULTIPHASE FLOWS IN POROUS MEDIA: A VARIATIONAL APPROACH

CLÉMENT CANCÈS, THOMAS O. GALLOUËT, AND LÉONARD MONSAINGEON

Abstract. We describe the competitive motion of $(N+1)$ incompressible immiscible phases within a porous medium as the gradient flow of a singular energy in the space of non-negative measures with prescribed masses, endowed with some tensorial Wasserstein distance. We show the convergence of the approximation obtained by a minimization scheme *à la* [R. Jordan, D. Kinderlehrer & F. Otto, SIAM J. Math. Anal, 29(1):1–17, 1998]. This allow to obtain a new existence result for a physically well-established system of PDEs consisting in the Darcy-Muskat law for each phase, N capillary pressure relations, and a constraint on the volume occupied by the fluid. Our study does not require the introduction of any global or complementary pressure.

Keywords. Multiphase porous media flows, Wasserstein gradient flows, constrained parabolic system, minimizing movement scheme

AMS subjects classification. 35K65, 35A15, 49K20, 76S05

Contents

1. Introduction

1.1. **Equations for multiphase flows in porous media.** We consider a convex open bounded set $\Omega \subset \mathbb{R}^d$ representing a porous medium. $N+1$ incompressible and immiscible phases, labeled by subscripts $i \in \{0, \dots, N\}$ are supposed to flow within the pores. Let us present now some classical equations that describe the motion of such a mixture. The physical justification of these equations can be found for instance in [10, Chapter 5]. We denote by $s_i : \Omega \times (0,T) =: Q \to [0,1]$ the content of the phase i, i.e., the volume ratio of the phase i compared to all the phases and the solid matrix, and by $\boldsymbol{v}_i$ the filtration speed of the phase i. Then the conservation of the volume of each phase writes

$$\partial_t s_i + \boldsymbol{\nabla} \cdot (s_i \boldsymbol{v}_i) = 0 \quad \text{in } Q, \quad \forall i \in \{0, \dots, N\}, \tag{1}$$

where $T > 0$ is an arbitrary finite time horizon. The filtration speed of each phase is assumed to be given by Darcy's law

$$\boldsymbol{v}_i = -\frac{1}{\mu_i} \, \mathbb{K} \, (\boldsymbol{\nabla} p_i - \rho_i \mathbf{g}) \quad \text{in } \ Q, \quad \forall i \in \{0, \dots, N\}. \tag{2}$$

In the above relation, $\mathbf{g}$ is the gravity vector, μ_i denotes the constant viscosity of the phase i, p_i its pressure, and ρ_i its density. The intrinsic permeability tensor $\mathbb{K} : \overline{\Omega} \to \mathbb{R}^{d \times d}$ is supposed to be smooth, symmetric $\mathbb{K} = \mathbb{K}^T$, and uniformly positive definite: there exist $\kappa_\star, \kappa^\star > 0$ such that:

$$\kappa_\star |\boldsymbol{\xi}|^2 \leq \mathbb{K}(\boldsymbol{x}) \boldsymbol{\xi} \cdot \boldsymbol{\xi} \leq \kappa^\star |\boldsymbol{\xi}|^2, \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \ \forall \boldsymbol{x} \in \overline{\Omega}. \tag{3}$$

The pore volume is supposed to be saturated by the fluid mixture

$$\sigma := \sum_{i=0}^{N} s_i = \omega(\boldsymbol{x}) \quad \text{a.e. in } Q, \tag{4}$$

where the porosity $\omega : \overline{\Omega} \to (0,1)$ of the surrounding porous matrix is assumed to be smooth. In particular, there exists $0 < \omega_\star \leq \omega^\star$ such that $\omega_\star \leq \omega(\boldsymbol{x}) \leq \omega^\star$ for all $\boldsymbol{x} \in \overline{\Omega}$. In what follows, we denote by $\boldsymbol{s} = (s_0, \dots, s_N)$, by

$$\Delta(\boldsymbol{x}) = \left\{ \boldsymbol{s} \in (\mathbb{R}_+)^{N+1} \,\middle|\, \sum_{i=0}^{N} s_i = \omega(\boldsymbol{x}) \right\},$$

and by

$$\boldsymbol{\mathcal{X}} = \left\{ \boldsymbol{s} \in L^1(\Omega; \mathbb{R}_+^{N+1}) \,\middle|\, \boldsymbol{s}(\boldsymbol{x}) \in \Delta(\boldsymbol{x}) \text{ a.e. in } \Omega \right\}.$$

There is an obvious one-to-one mapping between the sets $\Delta(\boldsymbol{x})$ and

$$\Delta^*(\boldsymbol{x}) = \left\{ \boldsymbol{s}^* = (s_1, \dots, s_N) \in (\mathbb{R}_+)^N \,\middle|\, \sum_{i=1}^{N} s_i \leq \omega(\boldsymbol{x}) \right\},$$

and consequently also between $\boldsymbol{\mathcal{X}}$ and

$$\boldsymbol{\mathcal{X}}^* = \left\{ \boldsymbol{s}^* \in L^1(\Omega; \mathbb{R}_+^{N}) \,\middle|\, \boldsymbol{s}^*(\boldsymbol{x}) \in \Delta^*(\boldsymbol{x}) \text{ a.e. in } \Omega \right\}.$$

In what follows, we denote by $\boldsymbol{\Upsilon} = \bigcup_{\boldsymbol{x} \in \overline{\Omega}} \Delta^*(\boldsymbol{x}) \times \{\boldsymbol{x}\}$.

In order to close the system, we impose N capillary pressure relations

$$p_i - p_0 = \pi_i(\boldsymbol{s}^*, \boldsymbol{x}) \quad \text{a.e in } Q, \quad \forall i \in \{1, \dots, N\}, \tag{5}$$

where the capillary pressure functions $\pi_i : \boldsymbol{\Upsilon} \to \mathbb{R}$ are assumed to be continuously differentiable and to derive from a strictly convex potential $\Pi : \boldsymbol{\Upsilon} \to \mathbb{R}_+$:

$$\pi_i(\boldsymbol{s}^*, \boldsymbol{x}) = \frac{\partial \Pi}{\partial s_i}(\boldsymbol{s}^*, \boldsymbol{x}) \qquad \forall i \in \{1, \dots, N\}.$$

We assume that Π is uniformly convex w.r.t. its first variable. More precisely, we assume that there exist two positive constants $\varpi_\star$ and $\varpi^\star$ such that, for all $\boldsymbol{x} \in \overline{\Omega}$ and all $\boldsymbol{s}^*, \widehat{\boldsymbol{s}}^* \in \Delta^*(\boldsymbol{x})$, one has

$$\frac{\varpi^\star}{2}|\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*|^2 \geq \Pi(\widehat{\boldsymbol{s}}^*, \boldsymbol{x}) - \Pi(\boldsymbol{s}^*, \boldsymbol{x}) - \boldsymbol{\pi}(\boldsymbol{s}^*, \boldsymbol{x}) \cdot (\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*) \geq \frac{\varpi_\star}{2}|\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*|^2, \tag{6}$$

where we introduced the notation

$$\boldsymbol{\pi} : \begin{cases} \boldsymbol{\Upsilon} \to \mathbb{R}^N \\ (\boldsymbol{s}^*, \boldsymbol{x}) \mapsto \boldsymbol{\pi}(\boldsymbol{s}^*, \boldsymbol{x}) = (\pi_1(\boldsymbol{s}^*, \boldsymbol{x}), \dots, \pi_N(\boldsymbol{s}^*, \boldsymbol{x})) \,. \end{cases}$$

The relation (6) implies that $\boldsymbol{\pi}$ is monotone and injective w.r.t. its first variable. Denoting by

$$\boldsymbol{z} \mapsto \boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{x}) = (\phi_1(\boldsymbol{z}, \boldsymbol{x}), \dots, \phi_N(\boldsymbol{z}, \boldsymbol{x})) \in \Delta^*(\boldsymbol{x})$$

the inverse of $\boldsymbol{\pi}(\cdot, \boldsymbol{x})$, it follows from (6) that

$$0 < \frac{1}{\varpi^\star} \leq \mathbb{J}_{\boldsymbol{z}} \boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{x}) \leq \frac{1}{\varpi_\star} \qquad \text{for all } \boldsymbol{x} \in \overline{\Omega} \text{ and all } \boldsymbol{z} \in \boldsymbol{\pi}(\Delta^*(\boldsymbol{x}), \boldsymbol{x}), \tag{7}$$

where $\mathbb{J}_{\boldsymbol{z}}$ stands for the Jacobian with respect to $\boldsymbol{z}$ and the above inequality should be understood in the sense of positive definite matrices. Moreover, due to the regularity of $\boldsymbol{\pi}$ w.r.t. the space variable, there exists $M_{\boldsymbol{\phi}} > 0$ such that

$$|\boldsymbol{\nabla}_{\boldsymbol{x}} \boldsymbol{\phi}(\boldsymbol{z}, \boldsymbol{x})| \leq M_{\boldsymbol{\phi}} \qquad \text{for all } \boldsymbol{x} \in \overline{\Omega} \text{ and all } \boldsymbol{z} \in \boldsymbol{\pi}(\Delta^*(\boldsymbol{x}), \boldsymbol{x}), \tag{8}$$

where $\boldsymbol{\nabla}_{\boldsymbol{x}}$ denote the gradient w.r.t. to the second variable only.

The problem is complemented with no-flux boundary conditions

$$\boldsymbol{v}_i \cdot \boldsymbol{n} = 0 \quad \text{on } \partial\Omega \times (0, T), \quad \forall i \in \{0, \dots, N\}, \tag{9}$$

and by the initial content profile $\boldsymbol{s}^0 = (s_0^0, \dots, s_N^0) \in \boldsymbol{\mathcal{X}}$:

$$s_i(\cdot, 0) = s_i^0 \quad \forall i \in \{0, \dots, N\}, \quad \text{with} \quad \sum_{i=0}^{N} s_i^0 = \omega \text{ a.e. in } \Omega. \tag{10}$$

Since we did not consider sources, and since we imposed no-flux boundary conditions, the volume of each phase is conserved along time

$$\int_\Omega s_i(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} = \int_\Omega s_i^0(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} =: m_i > 0, \qquad \forall i \in \{0, \dots, N\}. \tag{11}$$

We can now give a proper definition of what we call a weak solution to the problem (1)–(2), (4)–(5), and (9)–(10).

Definition 1.1 (Weak solution)**.** *A measurable function $\boldsymbol{s} : Q \to (\mathbb{R}_+)^{N+1}$ is said to be a weak solution if $\boldsymbol{s} \in \Delta$ a.e. in Q, if there exists $\boldsymbol{p} = (p_0, \dots, p_N) \in L^2((0,T); H^1(\Omega))^{N+1}$ such that the relations (5) hold, and such that, for all $\phi \in C_c^\infty(\overline{\Omega} \times [0, T))$ and all $i \in \{0, \dots, N\}$, one has*

$$\iint_Q s_i \partial_t \phi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega s_i^0 \phi(\cdot, 0) \mathrm{d}\boldsymbol{x} - \iint_Q \frac{s_i}{\mu_i} \mathbb{K} \left(\boldsymbol{\nabla} p_i - \rho_i \mathbf{g} \right) \cdot \boldsymbol{\nabla} \phi \mathrm{d}\boldsymbol{x}\mathrm{d}t = 0. \tag{12}$$

1.2. **Wasserstein gradient flow of the energy.**

1.2.1. *Energy of a configuration.* First, we extend the convex function $\Pi : \boldsymbol{\Upsilon} \to [0,+\infty]$, called *capillary energy density*, to a convex function (still denoted by) $\Pi : \mathbb{R}^{N+1} \times \overline{\Omega} \to [0,+\infty]$ by setting

$$\Pi(\boldsymbol{s},\boldsymbol{x}) = \begin{cases} \Pi\left(\omega\frac{\boldsymbol{s}^*}{\sigma},\boldsymbol{x}\right) = \Pi\left(\omega\frac{s_1}{\sigma},\dots,\omega\frac{s_N}{\sigma},\boldsymbol{x}\right) & \text{if } \boldsymbol{s}\in\mathbb{R}^{N+1}_+ \text{ and } \sigma \le \omega(\boldsymbol{x}), \\ +\infty & \text{otherwise,} \end{cases}$$

σ being defined by (4). The extension of Π by $+\infty$ where $\sigma > \omega$ is natural because of the incompressibility of the fluid mixture. The extension to $\{\sigma < \omega\} \cup \mathbb{R}^{N+1}_+$ is designed so that the energy density only depends on the relative composition of the fluid mixture. However, this extension is somehow arbitrary, and, as it will appear in the sequel, it has no influence on the flow since the solution $\boldsymbol{s}$ remains in $\boldsymbol{\mathcal{X}}$ (i-e $\sum_{i=0}^N s_i = \omega$). In our previous note [15] the appearance of void $\sigma < \omega$ was directly prohibited by a penalization in the energy.

The second part in the energy comes from the gravity. In order to lighten the notations, we introduce the functions

$$\Psi_i : \left\{ \begin{array}{ccc} \overline{\Omega} & \to & \mathbb{R}_+, \\ \boldsymbol{x} & \mapsto & -\rho_i \mathbf{g}\cdot\boldsymbol{x}, \end{array} \right. \qquad \forall i \in \{0,\dots,N\},$$

and

$$\boldsymbol{\Psi} : \left\{ \begin{array}{ccc} \overline{\Omega} & \to & \mathbb{R}^{N+1}_+, \\ \boldsymbol{x} & \mapsto & (\Psi_0(\boldsymbol{x}),\dots,\Psi_N(\boldsymbol{x}))\,. \end{array} \right.$$

The fact that Ψ_i can be supposed to be positive come from the fact that Ω is bounded. Even though the physically relevant potentials are indeed the gravitational $\Psi_i(\boldsymbol{x}) = -\rho_i \mathbf{g}\cdot\boldsymbol{x}$, the subsequent analysis allows for a broader class of external potentials and for the sake of generality we shall therefore consider arbitrary $\Psi_i \in \mathcal{C}^1(\overline{\Omega})$ in the sequel.

We can now define the convex energy functional $\mathcal{E} : L^1(\Omega,\mathbb{R}^{N+1}) \to \mathbb{R}\cup\{+\infty\}$ by adding the capillary energy to the gravitational one:

$$\mathcal{E}(\boldsymbol{s}) = \int_\Omega \left(\Pi(\boldsymbol{s},\boldsymbol{x}) + \boldsymbol{s}\cdot\boldsymbol{\Psi}\right)\mathrm{d}\boldsymbol{x} \ge 0, \qquad \forall \boldsymbol{s}\in L^1(\Omega;\mathbb{R}^{N+1}). \tag{13}$$

Note moreover that $\mathcal{E}(\boldsymbol{s}) < \infty$ iff $\boldsymbol{s} \ge 0$ and $\sigma \le \omega$ a.e. in Ω. It follows from the mass conservation (11) that

$$\int_\Omega \sigma(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{i=0}^N m_i = \int_\Omega \omega(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

Assume that there exists a non-negligible subset A of Ω such that $\sigma < \omega$ on A, then necessarily, there must be a non-negligible subset B of Ω such that $\sigma > \omega$ so that the above equation holds, hence $\mathcal{E}(\boldsymbol{s}) = +\infty$. Therefore,

$$\mathcal{E}(\boldsymbol{s}) < \infty \quad \Leftrightarrow \quad \boldsymbol{s} \in \boldsymbol{\mathcal{X}}. \tag{14}$$

Let $\boldsymbol{p} = (p_0,\dots,p_N) : \Omega \to \mathbb{R}^{N+1}$ be such that $\boldsymbol{p} \in \partial_{\boldsymbol{s}}\Pi(\boldsymbol{s},\boldsymbol{x})$ for a.e. $\boldsymbol{x}$ in Ω, then, defining $h_i = p_i + \Psi_i(\boldsymbol{x})$ for all $i \in \{0,\dots,N\}$ and $\boldsymbol{h} = (h_i)_{0\le i\le N}$, $\boldsymbol{h}$ belongs to the subdifferential $\partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$ of $\mathcal{E}$ at $\boldsymbol{s}$, i.e.,

$$\mathcal{E}(\widehat{\boldsymbol{s}}) \ge \mathcal{E}(\boldsymbol{s}) + \sum_{i=0}^N \int_\Omega h_i(\widehat{s}_i - s_i)\mathrm{d}\boldsymbol{x}, \qquad \forall \widehat{\boldsymbol{s}} \in L^1(\Omega;\mathbb{R}^{N+1}).$$

The reverse inclusion also holds, hence

$$\partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s}) = \left\{\boldsymbol{h} : \Omega \to \mathbb{R}^{N+1} \;\middle|\; h_i - \Psi_i(\boldsymbol{x}) \in \partial_{\boldsymbol{s}}\Pi(\boldsymbol{s},\boldsymbol{x}) \text{ for a.e. } \boldsymbol{x} \in \Omega\right\}. \tag{15}$$

Thanks to (14), we know that a configuration $\boldsymbol{s}$ has finite energy iff $\boldsymbol{s} \in \boldsymbol{\mathcal{X}}$. Since we are interested in finite energy configurations, it is relevant to consider the restriction of $\mathcal{E}$ to $\boldsymbol{\mathcal{X}}$. Then using the one-to-one mapping between $\boldsymbol{\mathcal{X}}$ and $\boldsymbol{\mathcal{X}}^*$, we define the energy of a configuration $\boldsymbol{s}^* \in \boldsymbol{\mathcal{X}}^*$, that we denote by $\mathcal{E}(\boldsymbol{s}^*)$ by setting $\mathcal{E}(\boldsymbol{s}^*) = \mathcal{E}(\boldsymbol{s})$ where $\boldsymbol{s}$ is the unique element of $\boldsymbol{\mathcal{X}}$ corresponding to $\boldsymbol{s}^* \in \boldsymbol{\mathcal{X}}^*$.

1.2.2. *Geometry of* Ω *and Wasserstein distance.* Inspired by the paper of Lisini [36], where heterogeneous anisotropic degenerate parabolic equations are studied from a variational point of view, we introduce $(N+1)$ distances on Ω that take into account the permeability of the porous medium and the phase viscosities. Given two points $\boldsymbol{x}, \boldsymbol{y}$ in Ω, we denote by

$$P(\boldsymbol{x},\boldsymbol{y}) = \left\{\boldsymbol{\gamma} \in C^1([0,1];\Omega) \,\middle|\, \boldsymbol{\gamma}(0) = \boldsymbol{x} \text{ and } \boldsymbol{\gamma}(1) = \boldsymbol{y}\right\}$$

the set of the smooth paths joining $\boldsymbol{x}$ to $\boldsymbol{y}$, and we introduce distances d_i, $i \in \{0,\dots,N\}$ between elements on Ω by setting

$$d_i(\boldsymbol{x},\boldsymbol{y}) = \inf_{\boldsymbol{\gamma}\in P(\boldsymbol{x},\boldsymbol{y})} \left(\int_0^1 \mu_i \mathbb{K}^{-1}(\boldsymbol{\gamma}(\tau))\boldsymbol{\gamma}'(\tau)\cdot\boldsymbol{\gamma}'(\tau)\mathrm{d}\tau\right)^{1/2}, \qquad \forall (\boldsymbol{x},\boldsymbol{y}) \in \overline{\Omega}. \tag{16}$$

It follows from (3) that

$$\sqrt{\frac{\mu_i}{\kappa^\star}}|\boldsymbol{x}-\boldsymbol{y}| \le d_i(\boldsymbol{x},\boldsymbol{y}) \le \sqrt{\frac{\mu_i}{\kappa_\star}}\,|\boldsymbol{x}-\boldsymbol{y}|, \qquad \forall (\boldsymbol{x},\boldsymbol{y}) \in \overline{\Omega}^2. \tag{17}$$

For $i \in \{0,\dots,N\}$ we define

$$\mathcal{A}_i = \left\{ s_i \in L^1(\Omega;\mathbb{R}_+) \;\middle|\; \int_\Omega s_i \mathrm{d}\boldsymbol{x} = m_i \right\}.$$

Given $s_i, \widehat{s}_i \in \mathcal{A}_i$, the set of admissible transport plans between s_i and $\widehat{s}_i$ is given by

$$\Gamma_i(s_i,\widehat{s}_i) = \left\{\theta_i \in \mathcal{M}_+(\Omega\times\Omega) \,\middle|\, \theta_i(\Omega\times\Omega) = m_i,\ \theta_i^{(1)} = s_i \text{ and } \theta_i^{(2)} = \widehat{s}_i \right\},$$

where $\mathcal{M}_+(\Omega\times\Omega)$ stands for the set of Borel measures on $\Omega\times\Omega$ and $\theta_i^{(k)}$ is the k^{th} marginal of the measure θ_i. We define the quadratic Wasserstein distance W_i on $\mathcal{A}_i$ by setting

$$W_i(s_i,\widehat{s}_i) = \left(\inf_{\theta_i\in\Gamma(s_i,\widehat{s}_i)} \iint_{\Omega\times\Omega} d_i(\boldsymbol{x},\boldsymbol{y})^2 \mathrm{d}\theta_i(\boldsymbol{x},\boldsymbol{y})\right)^{1/2}. \tag{18}$$

Due to the permeability tensor $\mathbb{K}(\boldsymbol{x})$, the porous medium Ω might be heterogeneous and anisotropic. Therefore, some directions and areas might me privileged by the fluid motions. This is encoded in the distances d_i we put on Ω. Moreover, the more viscous the phase is, the more costly are its displacements, hence the μ_i in the definition (16) of d_i. But it follows from (17) that

$$\sqrt{\frac{\mu_i}{\kappa^\star}}W_{\text{ref}}(s_i,\widehat{s}_i) \le W_i(s_i,\widehat{s}_i) \le \sqrt{\frac{\mu_i}{\kappa_\star}}W_{\text{ref}}(s_i,\widehat{s}_i)., \qquad \forall s_i,\widehat{s}_i \in \mathcal{A}_i, \tag{19}$$

where W_{ref} denotes the classical quadratic Wasserstein distance defined by

$$W_{\mathrm{ref}}(s_i,\widehat{s}_i)=\left(\inf_{\theta_i\in\Gamma(s_i,\widehat{s}_i)}\iint_{\Omega\times\Omega}|\boldsymbol{x}-\boldsymbol{y}|^2\mathrm{d}\theta_i(\boldsymbol{x},\boldsymbol{y})\right)^{1/2}. \tag{20}$$

With the phase Wasserstein distances $(W_i)_{0\leq i\leq N}$ at hand, we can define the global Wasserstein distance $\boldsymbol{W}$ on $\boldsymbol{\mathcal{A}}:=\mathcal{A}_0\times\cdots\times\mathcal{A}_N$ by setting

$$\boldsymbol{W}(\boldsymbol{s},\widehat{\boldsymbol{s}})=\left(\sum_{i=0}^{N}W_i(s_i,\widehat{s}_i)^2\right)^{1/2},\quad\forall\boldsymbol{s},\widehat{\boldsymbol{s}}\in\boldsymbol{\mathcal{A}}.$$

Finally for technical reasons we also assume that there exist smooth extensions $\widetilde{\mathbb{K}}$ and $\widetilde{\omega}$ to $\mathbb{R}^d$ of the tensor and the porosity, respectively, such that (3) holds on $\mathbb{R}^d$ for $\widetilde{\mathbb{K}}$, and such that $\tilde{\omega}$ is strictly bounded from below. This allows to define distances $\widetilde{d}_i$ on the whole $\mathbb{R}^d$by

$$\widetilde{d}_i(\boldsymbol{x},\boldsymbol{y})=\inf_{\boldsymbol{\gamma}\in\widetilde{P}(\boldsymbol{x},\boldsymbol{y})}\left(\int_0^1\mu_i\widetilde{\mathbb{K}}^{-1}(\boldsymbol{\gamma}(\tau))\boldsymbol{\gamma}'(\tau)\cdot\boldsymbol{\gamma}'(\tau)\mathrm{d}\tau\right)^{1/2},\qquad\forall\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^d \tag{21}$$

where $\widetilde{P}(\boldsymbol{x},\boldsymbol{y})=\left\{\boldsymbol{\gamma}\in C^1([0,1];\mathbb{R}^d)\,\middle|\,\boldsymbol{\gamma}(0)=\boldsymbol{x}\text{ and }\boldsymbol{\gamma}(1)=\boldsymbol{y}\right\}$. In the sequel, we assume that the extension $\widetilde{\mathbb{K}}$ of $\mathbb{K}$ is such that

$$\Omega\text{ is geodesically convex in }\mathcal{M}_i=(\mathbb{R}^d,\widetilde{d}_i)\text{ for all }i. \tag{22}$$

In particular $\widetilde{d}_i=d_i$ on $\Omega\times\Omega$. Since $\widetilde{\mathbb{K}}^{-1}$ is smooth, at least $C^2_b(\mathbb{R}^d)$, the Ricci curvature of the smooth complete Riemannian manifold $\mathcal{M}_i$ is uniformly bounded, i.e., there exists C depending only on $(\mu_i)_{0\leq i\leq N}$ and $\widetilde{\mathbb{K}}$ such that

$$|\mathrm{Ric}_{\mathcal{M}_i,\boldsymbol{x}}(\boldsymbol{v})|\leq C\mu_i\mathbb{K}^{-1}\boldsymbol{v}\cdot\boldsymbol{v},\qquad\forall\boldsymbol{x}\in\mathbb{R}^d,\ \forall\boldsymbol{v}\in\mathbb{R}^d. \tag{23}$$

Combined with the assumptions on $\widetilde{\omega}$ we deduce that $\mathcal{H}_{\widetilde{\omega}}$ is $\widetilde{\lambda}_i$ displacement convex on $\mathcal{P}_2^{ac}(\mathcal{M}_i)$ for some $\widetilde{\lambda}_i\in\mathbb{R}$. Then (22) and mass scaling implies that $\mathcal{H}_\omega$ is λ_i displacement convex on $(\mathcal{A}_i,W_i)$ for some $\lambda_i\in\mathbb{R}$. We refer to [46, Chap. 14 & 17] for further details on the Ricci curvature and its links with optimal transportation.

In the homogeneous and isotropic case $\mathbb{K}(\boldsymbol{x})=\mathrm{Id}$, Condition (22) simply amounts to assuming that Ω is convex. A simple sufficient condition implying (22) is given in Appendix A in the isotropic but heterogeneous case $\mathbb{K}(\boldsymbol{x})=\kappa(\boldsymbol{x})\mathbb{I}_d$.

1.2.3. *Gradient flow of the energy.* The content of this section is formal. Our aim is to write the problem as a gradient flow, i.e.

$$\frac{\mathrm{d}\boldsymbol{s}}{\mathrm{d}t}\in-\mathbf{grad}_{\boldsymbol{W}}\,\mathcal{E}(\boldsymbol{s})=-\left(\mathrm{grad}_{W_0}\mathcal{E}(\boldsymbol{s}),\dots,\mathrm{grad}_{W_N}\mathcal{E}(\boldsymbol{s})\right) \tag{24}$$

where $\mathbf{grad}_{\boldsymbol{W}}\,\mathcal{E}(\boldsymbol{s})$ denotes the full Wasserstein gradient of $\mathcal{E}(\boldsymbol{s})$, and $\mathrm{grad}_{W_i}\mathcal{E}(\boldsymbol{s})$ stands for the partial gradient of $s_i\mapsto\mathcal{E}(\boldsymbol{s})$ with respect to the Wasserstein distance W_i. The Wasserstein distance W_i was built so that $\dot{\boldsymbol{s}}=(\dot{s}_i)_i\in\mathbf{grad}_{\boldsymbol{W}}\,\mathcal{E}(\boldsymbol{s})$ iff there exists $\boldsymbol{h}\in\partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$ such that

$$\partial_t s_i=-\boldsymbol{\nabla}\cdot\left(s_i\frac{\mathbb{K}}{\mu_i}\boldsymbol{\nabla}h_i\right),\qquad\forall i\in\{0,\dots,N\}.$$

Such a construction was already performed by Lisini in the case of a single equation. Owing to the definitions (13) and (15) of the energy $\mathcal{E}(\boldsymbol{s})$ and its subdifferential $\partial_{\boldsymbol{s}}\mathcal{E}(\boldsymbol{s})$, the partial differential equations can be (at least formally) recovered. This was roughly speaking to purpose of our note [15].

In order to define rigorously the gradient $\mathbf{grad}_{\boldsymbol{W}}\mathcal{E}$ in (24), $\boldsymbol{\mathcal{A}}$ has to be a Riemannian manifold. The so-called Otto's calculus (see [42] and [46, Chapter 15]) allows to put a formal Riemannian structure on $\boldsymbol{\mathcal{A}}$. But as far as we know, this structure cannot be made rigorous and $\boldsymbol{\mathcal{A}}$ is a mere metric space. This leads us to consider generalized gradient flows in metric spaces (cf. [5]). We won't go deep into details in this direction, but we will prove that weak solutions can be obtained as limits of a minimizing movement scheme presented in the next section. This characterizes the gradient flow structure of the problem.

1.3. **Minimizing movement scheme and main result.**

1.3.1. *The scheme and existence of a solution.* For a fixed time-step $\tau > 0$, the so-called minimizing movement scheme [24, 5] or JKO scheme [30] consists in computing recursively $(\boldsymbol{s}^n)_{n\geq 1}$ as the solution to the minimization problem

$$\boldsymbol{s}^n = \underset{\boldsymbol{s}\in\boldsymbol{\mathcal{A}}}{\mathrm{Argmin}} \left(\frac{\boldsymbol{W}(\boldsymbol{s}, \boldsymbol{s}^{n-1})^2}{2\tau} + \mathcal{E}(\boldsymbol{s}) \right), \tag{25}$$

the initial data $\boldsymbol{s}^0$ being given (10).

1.3.2. *Approximate solution and main result.* Anticipating that the JKO scheme (25) is well posed (this is the purpose of Proposition 2.1 below), we can now define the piecewise constant interpolation $\boldsymbol{s}^\tau \in L^\infty((0,T);\boldsymbol{\mathcal{X}}\cap\boldsymbol{\mathcal{A}})$ by

$$\boldsymbol{s}^\tau(0,\cdot) = \boldsymbol{s}^0, \quad \text{and} \quad \boldsymbol{s}^\tau(t,\cdot) = \boldsymbol{s}^n \quad \forall t \in ((n-1)\tau, n\tau],\ \forall n \geq 1. \tag{26}$$

The main result of our paper is the following.

Theorem 1.2. *Let $(\tau_k)_{k\geq 1}$ be a sequence of time steps tending to 0, then there exists one weak solution $\boldsymbol{s}$ in the sense of Definition 1.1 such that, up to an unlabeled subsequence, $(\boldsymbol{s}^{\tau_k})_{k\geq 1}$ converges a.e. in Q towards s as k tends to ∞.*

As a direct by-product of Theorem 1.2, the continuous problem admits (at least) one solution in the sense of Definition 1.1. As far as we know, this existence result is new.

Remark 1.3. *It is worth stressing that our final solution will satisfy* a posteriori *$\partial_t s_i \in L^2((0,T);H^1(\Omega)')$, $s_i \in L^2((0,T);H^1(\Omega))$, and thus $s_i \in \mathcal{C}([0,T];L^2(\Omega))$. This regularity is enough to retrieve the so-called* Energy-Dissipation-Equality

$$\frac{d}{dt}\mathcal{E}(\boldsymbol{s}(t)) = -\sum_{i=0}^{N} \int_\Omega \mathbb{K}\frac{s_i(t)}{\mu_i}\nabla(p_i(t)+\Psi_i)\cdot\nabla(p_i(t)+\Psi_i)\mathrm{d}\boldsymbol{x} \leq 0 \quad \textit{for a.e. } t\in(0,T),$$

which is another admissible formulation of gradient flows in metric spaces [5].

1.4. **Goal and positioning of the paper.**

The aims of the paper are twofolds. First, we aim to provide rigorous foundations to the formal variational approach exposed in the authors' recent note [15]. This gives new insights into the modeling of complex porous media flows and their numerical approximation. Our approach appears to be very natural since only physically motivated quantities appear in the study. Indeed, we manage to avoid the introduction of the so-called Kirchhoff transform and global pressure, which classically appear in the mathematical study of multiphase flows in porous media (see for instance [18, 9, 20, 26, 27, 22, 19, 2, 3]).

Second, the existence result that we deduce from the convergence of the variational scheme is new as soon as there are at least three phases ($N \geq 2$). Indeed, since our study does not require the introduction of any global pressure, we get rid of many structural assumptions on the data among which the so-called *total differentiability condition*, see for instance Assumption **(H3)** in the paper by Fabrie and Saad [26]. This structural condition is not naturally satisfied by the models, and suitable algorithms have to be employed in order to adapt the data to this constraint [21]. However, our approach suffers from another technical difficulty: we are stuck to the case of linear relative permeabilities. The extension to the case of nonlinear concave relative permeabilities, i.e., where (1) is replaced by

$$\partial_t s_i + \boldsymbol{\nabla} \cdot (k_i(s_i)\boldsymbol{v}_i) = 0,$$

may be reachable thanks to the contributions of Dolbeault, Nazaret, and Savaré [25] (see also [48]), but we did not push in this direction since the relative permeabilities k_i are in general supposed to be convex in models coming from engineering.

Since the seminal paper of Jordan, Kinderlehrer, and Otto [30], gradient flows in metric spaces (and particularly in the space of probability measures endowed with the quadratic Wasserstein distance) were the object of many studies. Let us for instance refer to the monograph of Ambrosio, Gigli, and Savaré [5] and to Villani's book [46, Part II] for a complete overview. Applications are numerous. We refer for instance to [41] for an application to magnetic fluids, to [43, 7, 6] for applications to supra-conductivity, to [12, 11, 47] for applications to chemotaxis, to [37] for phase field models, to [39] for a macroscopic model of crowd motion, to [13] for an application to granular media, to [17] for aggregation equations, or to [31] for a model of ionic transport that applies in semi-conductors. In the context of porous media flows, this framework has been used by Otto [42] to study the asymptotic behavior of the porous medium equation, that is a simplified model for the filtration of a gas in a porous medium. The gradient flow approach in Wasserstein metric spaces was used more recently by Laurençot and Matioc [34] on a thin film approximation model for two-phase flows in porous media. Finally, let us mention that similar ideas were successfully applied for multicomponent systems, see e.g. [16, 32, 48, 49].

The variational structure of the system governing incompressible immiscible two-phase flows in porous media was recently depicted by the authors in their note [15]. Whereas the purpose of [15] is formal, our goal is here to give a rigorous foundation to the variational approach for complex flows in porous media. Finally, let us mention the work of Gigli and Otto [28] where it was noticed that multiphase linear transportation with saturation constraint (as we have here thanks to (1) and (4)) yields nonlinear transport with mobilities that appear naturally in the two-phase flow context.

The paper is organized as follows. In Section 2, we derive estimates on the solution $\boldsymbol{s}^\tau$ for a fixed τ. Beyond the classical energy and distance estimates detailed in §2.1, we obtain enhanced regularity estimates thanks to an adaptation of the so-called *flow interchange* technique of Matthes, McCann, and Savaré [38] to our inhomogeneous context in §2.2. Because of the constraint on the pore volume (4), the auxiliary flow we use is no longer the heat flow, and a drift term has to be added. An important effort is then done in §3 to derive the Euler-Lagrange equations that follow from the optimality of $\boldsymbol{s}^n$. Our proof is inspired from the work of

Maury, Roudneff-Chupin, and Santambrogio [39]. It relies on an intensive use of the dual characterization of the optimal transportation problem and the corresponding Kantorovitch potentials. However, additional difficulties arise from the multiphase aspect of our problem, in particular when there are at least three phases (i.e., $N \geq 2$). These are overpassed using a generalized multicomponent bathtub principle (Theorem B.1 in Appendix) and computing the associated Lagrange multipliers in §3.1. This key step then allows to define the notion of discrete phase and capillary pressures in §3.2. Then Section 4 is devoted to the convergence of the approximate solutions $(\boldsymbol{s}^{\tau_k})_k$ towards a weak solution $\boldsymbol{s}$ as τ_k tends to 0. The estimates we obtained in Section 2 are integrated w.r.t. time in §4.1. In §4.2, we show that these estimates are sufficient to enforce the relative compactness of $(\boldsymbol{s}^{\tau_k})_k$ in the strong $L^1(Q)^{N+1}$ topology. Finally, it is shown in §4.3 that any limit $\boldsymbol{s}$ of $(\boldsymbol{s}^{\tau_k})_k$ is a weak solution in the sense of Definition 1.1.

2. One-step regularity estimates

The first thing to do is to show that the JKO scheme (25) is well-posed. This is the purpose of the following Proposition.

Proposition 2.1. *Let $n \geq 1$ and $\boldsymbol{s}^{n-1} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$, then there exists a unique solution $\boldsymbol{s}^n$ to the scheme (25). Moreover, one has $\boldsymbol{s}^n \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$.*

Proof. Any $\boldsymbol{s}^{n-1} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ has finite energy thanks to (14). Let $(\boldsymbol{s}^{n,k})_k \subset \boldsymbol{\mathcal{A}}$ be a minimizing sequence in (25). Testing $\boldsymbol{s}^{n-1}$ in (25) it is easy to see that $\mathcal{E}(\boldsymbol{s}^{n,k}) \leq \mathcal{E}(\boldsymbol{s}^{n-1}) < \infty$ for large k, thus $(\boldsymbol{s}^{n,k})_k \subset \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ thanks to (14). Hence, one has $0 \leq s_i^{n,k}(\boldsymbol{x}) \leq \omega(\boldsymbol{x})$ for all k. By Dunford-Pettis theorem, we can therefore assume that $s_i^{n,k} \rightharpoonup s_i^n$ weakly in $L^1(\Omega)$. It is then easy to check that the limit $\boldsymbol{s}^n$ of $\boldsymbol{s}^{n,k}$ belongs to $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$. The lower semi-continuity of the Wasserstein distance with respect to weak L^1 convergence is well known (see, e.g., [44, Prop. 7.4]), and since the energy functional is convex thus l.s.c., we conclude that $\boldsymbol{s}^n$ is indeed a minimizer. Uniqueness follows from the strict convexity of the energy as well as from the convexity of the Wasserstein distances (w.r.t. linear interpolation $\boldsymbol{s}_\theta = (1-\theta)\boldsymbol{s}_0 + \theta \boldsymbol{s}_1$). □

The rest of this section is devoted to improving the regularity of the successive minimizers.

2.1. Energy and distance estimates.

Testing $\boldsymbol{s} = \boldsymbol{s}^{n-1}$ in (25) we obtain

$$\frac{\boldsymbol{W}(\boldsymbol{s}^n, \boldsymbol{s}^{n-1})^2}{2\tau} + \mathcal{E}(\boldsymbol{s}^n) \leq \mathcal{E}(\boldsymbol{s}^{n-1}), \tag{27}$$

As a consequence we have the monotonicity

$$\ldots \leq \mathcal{E}(\boldsymbol{s}^n) \leq \mathcal{E}(\boldsymbol{s}^{n-1}) \leq \ldots \leq \mathcal{E}(\boldsymbol{s}^0) < \infty$$

at the discrete level, thus $\boldsymbol{s}^n \in \boldsymbol{\mathcal{X}}$ for all $n \geq 0$ thanks to (14). Summing (27) over n we also obtain the classical *total square distance* estimate

$$\frac{1}{\tau} \sum_{n \geq 0} \boldsymbol{W}^2(\boldsymbol{s}^{n+1}, \boldsymbol{s}^n) \leq 2\mathcal{E}(\boldsymbol{s}^0) \leq C\left(\Omega, \Pi, \boldsymbol{\Psi}\right), \tag{28}$$

the last inequality coming from the fact that $\boldsymbol{s}^0$ is uniformly bounded since it belongs to $\boldsymbol{\mathcal{X}}$, thus so is $\mathcal{E}(\boldsymbol{s}^0)$. This readily gives the approximate 1/2-Hölder

estimate

$$\boldsymbol{W}(\boldsymbol{s}^{n_1}, \boldsymbol{s}^{n_2}) \leq C\sqrt{|n_2 - n_1|\tau}. \tag{29}$$

2.2. **Flow interchange, entropy estimate and enhanced regularity.** The goal of this section is to obtain some additional Sobolev regularity on the capillary pressure field $\boldsymbol{\pi}(\boldsymbol{s}^{n*}, \boldsymbol{x})$, where $\boldsymbol{s}^{n*} = (s_1^n, \dots, s_N^n)$ is the unique element of $\boldsymbol{\mathcal{X}}^*$ corresponding to the minimizer $\boldsymbol{s}^n$ of (25). In what follows, we denote by

$$\pi_i^n : \left\{ \begin{array}{ccc} \Omega & \to & \mathbb{R}, \\ \boldsymbol{x} & \mapsto & \pi_i(\boldsymbol{s}^{n*}(\boldsymbol{x}), \boldsymbol{x}), \end{array} \right. \qquad \forall i \in \{1, \dots, N\}$$

and $\boldsymbol{\pi}^n = (\pi_1^n, \dots, \pi_N^n)$. Bearing in mind that $\omega(\boldsymbol{x}) \geq \omega_\star > 0$ in $\overline{\Omega}$, we can define the relative Boltzmann entropy $\mathcal{H}_\omega$ with respect to ω by

$$\mathcal{H}_\omega(s) := \int_\Omega s(\boldsymbol{x}) \log\left(\frac{s(\boldsymbol{x})}{\omega(\boldsymbol{x})}\right) d\boldsymbol{x}, \quad \text{for all measurable } s : \Omega \to \mathbb{R}_+.$$

Lemma 2.2. *There exists C depending only on $\Omega, \Pi, \omega, \mathbb{K}, (\mu_i)_i$, and $\boldsymbol{\Psi}$ such that, for all $n \geq 1$ and all $\tau > 0$, one has*

$$\sum_{i=0}^{N} \|\nabla \pi_i^n\|_{L^2(\Omega)}^2 \leq C\left(1 + \frac{\boldsymbol{W}^2(\boldsymbol{s}^n, \boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N} \frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(s_i^n)}{\tau}\right). \tag{30}$$

Proof. The argument relies on the *flow interchange* technique introduced by Matthes, McCann, and Savaré in [38]. Throughout the proof, C denotes a fluctuating constant that depends on the prescribed data $\Omega, \Pi, \omega, \mathbb{K}, (\mu_i)_i$, and $\boldsymbol{\Psi}$, but neither on t, τ, nor on n. For $i = 0 \dots N$ consider the auxiliary flows

$$\left\{ \begin{array}{ll} \partial_t \check{s}_i = \operatorname{div}(\mathbb{K}\nabla \check{s}_i - \check{s}_i \mathbb{K}\nabla \log \omega), & t > 0,\ \boldsymbol{x} \in \Omega, \\ \mathbb{K}(\nabla \check{s}_i - \check{s}_i \nabla \log \omega) \cdot \nu = 0, & t > 0,\ \boldsymbol{x} \in \partial\Omega, \\ \check{s}_i|_{t=0} = s_i^n, & \boldsymbol{x} \in \Omega \end{array} \right. \tag{31}$$

for each $i \in \{0, \dots, N\}$. By standard parabolic theory (see for instance [33, Chapter III, Theorem 12.2]), these Initial-Boundary value problems are well-posed, and their solutions $\check{s}_i(\boldsymbol{x})$ belong to $\mathcal{C}^{1,2}((0,1] \times \overline{\Omega}) \cap \mathcal{C}([0,1]; L^p(\Omega))$ for all $p \in (1, \infty)$ if $\omega \in \mathcal{C}^{2,\alpha}(\overline{\Omega})$ and $\mathbb{K} \in \mathcal{C}^{1,\alpha}(\overline{\Omega})$ for some $\alpha > 0$. Therefore, $t \mapsto \check{s}_i(\cdot, t)$ is absolutely continuous in $L^1(\Omega)$, thus in $\mathcal{A}_i$ endowed with the usual quadratic distance W_{ref} (20) thanks to [44, Prop. 7.4]. Because of (19), the curve $t \mapsto \check{s}_i(\cdot, t)$ is also absolutely continuous in $\mathcal{A}_i$ endowed with W_i.

From Lisini's results [36], we know that the evolution $t \mapsto \check{s}_i(\cdot, t)$ can be interpreted as the gradient flow of the relative Boltzmann functional $\frac{1}{\mu_i}\mathcal{H}_\omega$ with respect to the metric W_i, the scaling factor $\frac{1}{\mu_i}$ appearing due to the definition (18) of the distance W_i. As a consequence of (23), The Ricci curvature of (Ω, d_i) is bounded, hence bounded from below. Since $\omega \in \mathcal{C}^2(\overline{\Omega})$ and with our assumption (22) we also have that $\frac{1}{\mu_i}\mathcal{H}_\omega$ is λ_i-displacement convex with respect to W_i for some $\lambda_i \in \mathbb{R}$ depending on ω and the geometry of (Ω, d_i), see [46, Chapter 14]. Therefore, we can use the so-called *Evolution Variational Inequality* characterization of gradient flows (see for instance [4, Definition 4.5]) centered at s_i^{n-1}, namely

$$\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t} W_i^2(\check{s}_i(t), s_i^{n-1}) + \frac{\lambda_i}{2} W_i^2(\check{s}_i(t), s_i^{n-1}) \leq \frac{1}{\mu_i}\mathcal{H}_\omega(s_i^{n-1}) - \frac{1}{\mu_i}\mathcal{H}_\omega(\check{s}_i(t)).$$

Denote by $\check{\boldsymbol{s}} = (\check{s}_0, \dots, \check{s}_N)$, and by $\check{\boldsymbol{s}}^* = (\check{s}_1, \dots, \check{s}_N)$. Summing the previous inequality over $i \in \{0, \dots, N\}$ leads to

(32)
$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{1}{2\tau}\boldsymbol{W}^2(\check{\boldsymbol{s}}(t), \boldsymbol{s}^{n-1})\right) \leq C\left(\frac{\boldsymbol{W}^2(\check{\boldsymbol{s}}(t), \boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(\check{s}_i(t))}{\tau}\right).$$

In order to estimate the internal energy contribution in (25), we first note that $\sum s_i^n(\boldsymbol{x}) = \omega(\boldsymbol{x})$ for all $\boldsymbol{x} \in \overline{\Omega}$, thus by linearity of (31) and since ω is a stationary solution we have $\sum \check{s}_i(\boldsymbol{x}, t) = \omega(\boldsymbol{x})$ as well. Moreover, the problem (31) is monotone, thus order preserving, and admits 0 as a subsolution. Hence $\check{s}_i(\boldsymbol{x}, t) \geq 0$, so that $\check{\boldsymbol{s}}(t) \in \boldsymbol{\mathcal{A}} \cap \boldsymbol{\mathcal{X}}$ is an admissible competitor in (25) for all $t > 0$. The smoothness of $\check{\boldsymbol{s}}$ for $t > 0$ allows to write

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_\Omega \Pi(\check{\boldsymbol{s}}^*(\boldsymbol{x}, t), \boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) = \sum_{i=1}^{N}\int_\Omega \check{\pi}_i(\boldsymbol{x}, t)\partial_t \check{s}_i(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x} = I_1(t) + I_2(t), \tag{33}$$

where $\check{\pi}_i := \pi_i(\check{\boldsymbol{s}}^*, \cdot)$, and where, for all $t > 0$, we have set

$$I_1(t) = -\sum_{i=1}^{N}\int_\Omega \boldsymbol{\nabla}\check{\pi}_i(t) \cdot \mathbb{K}\boldsymbol{\nabla}\check{s}_i(t)\mathrm{d}\boldsymbol{x}, \quad I_2(t) = -\sum_{i=1}^{N}\int_\Omega \frac{\check{s}_i(t)}{\omega}\boldsymbol{\nabla}\check{\pi}_i(t) \cdot \mathbb{K}\boldsymbol{\nabla}\omega\mathrm{d}\boldsymbol{x}.$$

To estimate I_1, we first use the invertibility of $\boldsymbol{\pi}$ to write

$$\check{\boldsymbol{s}}(\boldsymbol{x}, t) = \boldsymbol{\phi}(\check{\boldsymbol{\pi}}(\boldsymbol{x}, t), \boldsymbol{x}) =: \check{\boldsymbol{\phi}}(\boldsymbol{x}, t),$$

yielding

$$\boldsymbol{\nabla}\check{\boldsymbol{s}}(\boldsymbol{x}, t) = \mathbb{J}_{\boldsymbol{z}}\boldsymbol{\phi}(\check{\boldsymbol{\pi}}(\boldsymbol{x}, t), \boldsymbol{x})\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(\boldsymbol{x}, t) + \boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{\phi}(\check{\boldsymbol{\pi}}(\boldsymbol{x}, t), \boldsymbol{x}). \tag{34}$$

Combining (3), (7), (8) and the elementary inequality

$$ab \leq \delta\frac{a^2}{2} + \frac{b^2}{2\delta} \quad \text{with } \delta > 0 \text{ arbitrary}, \tag{35}$$

we get that for all $t > 0$, there holds

$$I_1(t) \leq -\frac{\kappa_\star}{\varpi^\star}\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2\mathrm{d}\boldsymbol{x} + \kappa^\star\left(\delta\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2\mathrm{d}\boldsymbol{x} + \frac{1}{\delta}\int_\Omega |\boldsymbol{\nabla}_{\boldsymbol{x}}\boldsymbol{\phi}(\check{\boldsymbol{\pi}}(t))|^2\mathrm{d}\boldsymbol{x}\right).$$

Choosing $\delta = \frac{\kappa_\star}{4\kappa^\star\varpi^\star}$, we get that

$$I_1(t) \leq -\frac{3\kappa_\star}{4\varpi^\star}\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2\mathrm{d}\boldsymbol{x} + C, \qquad \forall t > 0. \tag{36}$$

In order to estimate I_2, we use that $\check{\boldsymbol{s}}(t) \in \boldsymbol{\mathcal{X}}$ for all $t > 0$, so that $0 \leq \check{s}_i(\boldsymbol{x}, t) \leq \omega(\boldsymbol{x})$, hence we deduce that $\sum_{i=1}^{N}\left(\frac{\check{s}_i}{\omega}\right)^2 \leq 1$. Therefore, using (35) again, we get

$$I_2(t) \leq \delta\kappa^\star\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2\mathrm{d}\boldsymbol{x} + \frac{\kappa^\star}{\delta}\int_\Omega |\boldsymbol{\nabla}\omega|^2\mathrm{d}\boldsymbol{x}.$$

Choosing again $\delta = \frac{\kappa_\star}{4\kappa^\star\varpi^\star}$ yields

$$I_2(t) \leq \frac{\kappa_\star}{4\varpi^\star}\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2\mathrm{d}\boldsymbol{x} + C. \tag{37}$$

Taking (36)–(37) into account in (33) provides

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\int_\Omega \Pi(\check{\boldsymbol{s}}^*(\boldsymbol{x}, t), \boldsymbol{x})\mathrm{d}\boldsymbol{x}\right) \leq -\frac{\kappa_\star}{2\varpi^\star}\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2\mathrm{d}\boldsymbol{x} + C, \qquad \forall t > 0. \tag{38}$$

Let us now focus on the potential (gravitational) energy. Since $\check{\boldsymbol{s}}(t)$ belongs to $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ for all $t > 0$, we can make use of the relation

$$\check{s}_0(\boldsymbol{x}, t) = \omega(\boldsymbol{x}) - \sum_{i=1}^{N} \check{s}_i(\boldsymbol{x}, t), \qquad \text{for all } (\boldsymbol{x}, t) \in \Omega \times \mathbb{R}_+,$$

to write: for all $t > 0$,

$$\sum_{i=0}^{N} \int_\Omega \check{s}_i(\boldsymbol{x}, t)\Psi_i(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \sum_{i=1}^{N} \int_\Omega \check{s}_i(\boldsymbol{x}, t)(\Psi_i - \Psi_0)(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \omega(\boldsymbol{x})\Psi_0(\boldsymbol{x})\mathrm{d}\boldsymbol{x}.$$

This leads to

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\sum_{i=0}^{N} \int_\Omega \check{s}_i(t)\Psi_i \mathrm{d}\boldsymbol{x}\right) = \sum_{i=1}^{N} \int_\Omega (\Psi_i(\boldsymbol{x}) - \Psi_0(\boldsymbol{x}))\partial_t s_i(\boldsymbol{x}, t)\mathrm{d}\boldsymbol{x} = J_1(t) + J_2(t), \tag{39}$$

where, using the equations (31), we have set

$$J_1(t) = -\sum_{i=1}^{N} \int_\Omega \boldsymbol{\nabla}(\Psi_i - \Psi_0) \cdot \mathbb{K}\boldsymbol{\nabla}\check{s}_i(t)\mathrm{d}\boldsymbol{x},$$

$$J_2(t) = \sum_{i=1}^{N} \int_\Omega \frac{\check{s}_i(t)}{\omega} \boldsymbol{\nabla}(\Psi_i - \Psi_0) \cdot \mathbb{K}\boldsymbol{\nabla}\omega \mathrm{d}\boldsymbol{x}.$$

The term J_1 can be estimated using (35). More precisely, for all $\delta > 0$, we have

$$J_1(t) \leq \kappa^\star \left(\delta \|\boldsymbol{\nabla}\check{\boldsymbol{s}}^*(t)\|_{L^2}^2 + \frac{1}{\delta}\sum_{i=1}^{N} \|\nabla(\Psi_i - \Psi_0)\|_{L^2}^2\right). \tag{40}$$

Using (34) together with (7)–(8), we get that

$$\|\boldsymbol{\nabla}\check{\boldsymbol{s}}^*\|_{L^2}^2 \leq \left(\frac{1}{\varpi_\star}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}\|_{L^2} + |\Omega| M_{\boldsymbol{\phi}}\right)^2 \leq \frac{2}{(\varpi_\star)^2}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}\|_{L^2}^2 + 2\left(|\Omega| M_{\boldsymbol{\phi}}\right)^2.$$

Therefore, choosing $\delta = \frac{(\varpi_\star)^2 \kappa_\star}{8\kappa^\star \varpi^\star}$ in (40), we infer from the regularity of $\boldsymbol{\Psi}$ that

$$J_1(t) \leq \frac{\kappa_\star}{4\varpi^\star}\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2 \mathrm{d}\boldsymbol{x} + C, \qquad \forall t > 0. \tag{41}$$

Finally, it follows from the fact that $\sum_{i=1}^{N} \check{s}_i \leq \omega$, from the Cauchy-Schwarz inequality, and from the regularity of $\boldsymbol{\Psi}, \omega$ that

$$J_2(t) \geq -\kappa^\star \sum_{i=1}^{N} \|\boldsymbol{\nabla}\Psi_i - \boldsymbol{\nabla}\Psi_0\|_{L^2}\|\boldsymbol{\nabla}\omega\|_{L^2} = C. \tag{42}$$

Combining (39), (41), and (42) with (38), we get that

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}(\check{\boldsymbol{s}}(t)) \leq -\frac{\kappa_\star}{4\varpi^\star}\int_\Omega |\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)|^2 \mathrm{d}\boldsymbol{x} + C, \qquad \forall t > 0. \tag{43}$$

Denote by

$$\mathcal{F}_\tau^n(\boldsymbol{s}) := \frac{1}{2\tau}\boldsymbol{W}^2(\boldsymbol{s}, \boldsymbol{s}^{n-1}) + \mathcal{E}(\boldsymbol{s}) \tag{44}$$

the functional to be minimized in (25), then gathering (32) and (43) provides

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\tau^n(\check{\boldsymbol{s}}(t)) + \frac{\kappa_\star}{4\varpi^\star}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}\|_{L^2}^2 \\ \leq C\left(1 + \frac{\boldsymbol{W}^2(\check{\boldsymbol{s}}(t),\boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(\check{s}_i(t))}{\tau}\right) \qquad \forall t>0.$$

Since $\check{\boldsymbol{s}}(0) = \boldsymbol{s}^n$ is a minimizer of (25) we must have

$$0 \leq \limsup_{t\to 0^+}\left(\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_\tau^n(\check{\boldsymbol{s}}(t))\right),$$

otherwise $\check{\boldsymbol{s}}(t)$ would be a strictly better competitor than $\boldsymbol{s}^n$ for small $t>0$. As a consequence, we get

$$\liminf_{t\to 0^+}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)\|_{L^2}^2 \leq C\limsup_{t\to 0^+}\left(1 + \frac{\boldsymbol{W}^2(\check{\boldsymbol{s}}(t),\boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(\check{s}_i(t))}{\tau}\right).$$

Since $\check{s}_i$ belongs to $C([0,1];L^p(\Omega))$ for all $p\in[1,\infty)$ (see for instance [14]), the continuity of the Wasserstein distance and of the Boltzmann entropy with respect to strong L^p-convergence imply that

$$\boldsymbol{W}^2(\check{\boldsymbol{s}}(t),\boldsymbol{s}^{n-1}) \underset{t\to 0^+}{\longrightarrow} \boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1}) \quad\text{and}\quad \mathcal{H}_\omega(\check{s}_i(t)) \underset{t\to 0^+}{\longrightarrow} \mathcal{H}_\omega(s_i^n).$$

Therefore, we obtain that

$$\liminf_{t\to 0^+}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)\|_{L^2}^2 \leq C\left(1 + \frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(s_i^n)}{\tau}\right). \tag{45}$$

It follows from the regularity of $\boldsymbol{\pi}$ that

$$\boldsymbol{\pi}(\check{\boldsymbol{s}}^*(t),\boldsymbol{x}) = \check{\boldsymbol{\pi}}(t) \underset{t\to 0^+}{\longrightarrow} \boldsymbol{\pi}^n = \boldsymbol{\pi}(\boldsymbol{s}^{n*},\boldsymbol{x}) \quad \text{in } L^p(\Omega).$$

Finally, let $(t_\ell)_{\ell\geq 1}$ be a decreasing sequence tending to 0 realizing the lim inf in (45), then the sequence $(\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t_\ell))_{\ell\geq 1}$ converges weakly in $L^2(\Omega)^{N\times d}$ towards $\boldsymbol{\nabla}\boldsymbol{\pi}^n$. The lower semi-continuity of the norm w.r.t. the weak convergence leads to

$$\sum_{i=1}^{N}\|\boldsymbol{\nabla}\pi_i^n\|_{L^2}^2 \leq \lim_{\ell\to\infty}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t_\ell)\|_{L^2}^2 = \liminf_{t\to 0^+}\|\boldsymbol{\nabla}\check{\boldsymbol{\pi}}(t)\|_{L^2}^2 \\ \leq C\left(1 + \frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(s_i^n)}{\tau}\right)$$

and the proof is complete. □

3. The Euler-Lagrange equations and pressure bounds

The goal of this section is to extract informations coming from the optimality of $\boldsymbol{s}^n$ in the JKO minimization (25). The main difficulty consists in constructing the phase and capillary pressures from this optimality condition. Our proof is inspired from [39] and makes an extensive use of the Kantorovich potentials. Therefore, we first recall their definition and some useful properties. We refer to [44, §1.2] or [46, Chapter 5] for details.

Let $(\nu_1,\nu_2)\in\mathcal{M}_+(\Omega)^2$ be two nonnegative measures with same total mass. A pair of Kantorovich potentials $(\varphi_i,\psi_i)\in L^1(\nu_1)\times L^1(\nu_2)$ associated to the measures

ν_1 and ν_2 and to the cost function $\frac{1}{2}d_i^2$ defined by (16), $i \in \{0, \dots, N\}$, is a solution of the Kantorovich *dual problem*

$$DP_i(\nu_1, \nu_2) = \max_{\substack{(\varphi_i, \psi_i) \in L^1(\nu_1) \times L^1(\nu_2) \\ \varphi_i(\boldsymbol{x}) + \psi_i(\boldsymbol{y}) \leq \frac{1}{2} d_i^2(\boldsymbol{x}, \boldsymbol{y})}} \int_\Omega \varphi_i(\boldsymbol{x}) \nu_1(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} + \int_\Omega \psi_i(\boldsymbol{y}) \nu_2(\boldsymbol{y}) \mathrm{d}\boldsymbol{y}.$$

We will use the three following important properties of the Kantorovich potentials:

(a) There is always duality

$$DP_i(\nu_1, \nu_2) = \frac{1}{2} W_i^2(\nu_1, \nu_2), \qquad \forall i \in \{0, \dots, N\}.$$

(b) A pair of Kantorovich potentials (φ_i, ψ_i) is $\mathrm{d}\nu_1 \otimes \mathrm{d}\nu_2$ unique, up to additive constants.

(c) The Kantorovich potentials φ_i and ψ_i are $\frac{1}{2}d_i^2$-conjugate, that is

$$\varphi_i(\boldsymbol{x}) = \inf_{\boldsymbol{y} \in \Omega} \frac{1}{2} d_i^2(\boldsymbol{x}, \boldsymbol{y}) - \psi_i(\boldsymbol{y}), \quad \forall\, \boldsymbol{x} \in \Omega,$$

$$\psi_i(\boldsymbol{y}) = \inf_{\boldsymbol{x} \in \Omega} \frac{1}{2} d_i^2(\boldsymbol{x}, \boldsymbol{y}) - \varphi_i(\boldsymbol{x}), \quad \forall\, \boldsymbol{y} \in \Omega.$$

Remark 3.1. *Since Ω is bounded, the cost functions $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \frac{1}{2}d_i^2(\boldsymbol{x}, \boldsymbol{y})$, $i \in \{1, \dots, N\}$, are globally Lipschitz continuous, see (17). Thus item (c) shows that φ_i and ψ_i are also Lipschitz continuous.*

3.1. **A decomposition result.** The next lemma is an adaptation of [39, Lemma 3.1] to our framework. It essentially states that, since $\boldsymbol{s}^n$ is a minimizer of (25), it is also a minimizer of the linearized problem.

Lemma 3.2. *For $n \geq 1$ and $i = 0, \dots, N$ there exist some (backward, optimal) Kantorovich potentials φ_i^n from s_i^n to s_i^{n-1} such that, using the convention $\pi_0^n = \frac{\partial \Pi}{\partial s_0}(s_1^n, \dots, s_N^n, \boldsymbol{x}) = 0$, setting*

$$F_i^n := \frac{\varphi_i^n}{\tau} + \pi_i^n + \Psi_i, \quad \forall i \in \{0, \dots, N\}, \tag{46}$$

and denoting $\mathbf{F}^n = (F_i^n)_{0 \leq i \leq N}$, there holds

$$\boldsymbol{s}^n \in \operatorname*{Argmin}_{\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}} \int_\Omega \mathbf{F}^n(\boldsymbol{x}) \cdot \boldsymbol{s}(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}. \tag{47}$$

Moreover, $F_i^n \in L^\infty \cap H^1(\Omega)$ for all $i \in \{0, \dots, N\}$.

Proof. We assume first that $s_i^{n-1}(\boldsymbol{x}) > 0$ everywhere in Ω for all $i \in \{1, \dots, N\}$, so that the Kantorovich potentials (φ_i^n, ψ_i^n) from s_i^n to s_i^{n-1} are uniquely determined after normalizing $\varphi_i^n(\boldsymbol{x}_{\mathrm{ref}}) = 0$ for some arbitrary point $\boldsymbol{x}_{\mathrm{ref}} \in \Omega$ (cf. [44, Proposition 7.18]). Given any $\boldsymbol{s} = (s_i)_{1 \leq 0 \leq N} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ and $\varepsilon \in (0, 1)$ we define the perturbation

$$\boldsymbol{s}^\varepsilon := (1 - \varepsilon)\boldsymbol{s}^n + \varepsilon \boldsymbol{s}.$$

Note that $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ is convex, thus $\boldsymbol{s}^\varepsilon$ is an admissible competitor for all $\varepsilon \in (0, 1)$. Let $(\varphi_i^\varepsilon, \psi_i^\varepsilon)$ be the unique Kantorovich potentials from s_i^ε to s_i^{n-1}, similarly normalized

as $\varphi_i^\varepsilon(\boldsymbol{x}_{\rm ref}) = 0$. Then by characterization of the squared Wasserstein distance in terms of the dual Kantorovich problem we have

$$\begin{cases} \dfrac{1}{2}W_i^2(s_i^\varepsilon, s_i^{n-1}) = \displaystyle\int_\Omega \varphi_i^\varepsilon(\boldsymbol{x}) s_i^\varepsilon(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \psi_i^\varepsilon(\boldsymbol{y}) s_i^{n-1}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}, \\ \dfrac{1}{2}W_i^2(s_i^n, s_i^{n-1}) \geq \displaystyle\int_\Omega \varphi_i^\varepsilon(\boldsymbol{x}) s_i^n(\boldsymbol{x})\mathrm{d}\boldsymbol{x} + \int_\Omega \psi_i^\varepsilon(\boldsymbol{y}) s_i^{n-1}(\boldsymbol{y})\mathrm{d}\boldsymbol{y}. \end{cases}$$

By definition of the perturbation $\boldsymbol{s}^\varepsilon$ it is easy to check that $s_i^\varepsilon - s_i^n = \varepsilon(s_i - s_i^n)$. Subtracting the previous inequalities we get

$$\frac{W_i^2(s_i^\varepsilon, s_i^{n-1}) - W_i^2(s_i^n, s_i^{n-1})}{2\tau} \leq \frac{\varepsilon}{\tau}\int_\Omega \varphi_i^\varepsilon(s_i - s_i^n)\mathrm{d}\boldsymbol{x}. \tag{48}$$

Denote by $\boldsymbol{s}^{\varepsilon *} = (s_1^\varepsilon, \dots, s_N^\varepsilon)$, $\boldsymbol{\pi}^\varepsilon = \boldsymbol{\pi}(\boldsymbol{s}^{\varepsilon *}, \cdot)$, and extend to the zero-th component $\overline{\boldsymbol{\pi}}^\varepsilon = (0, \boldsymbol{\pi}^\varepsilon)$. The convexity of Π as a function of $s_1, \dots, s_N$ implies that

$$\begin{aligned} \int_\Omega \left(\Pi(\boldsymbol{s}^{n*}, \boldsymbol{x}) - \Pi(\boldsymbol{s}^{\varepsilon *}, \boldsymbol{x})\right)\, \mathrm{d}\boldsymbol{x} &\geq \int_\Omega \boldsymbol{\pi}^\varepsilon \cdot (\boldsymbol{s}^{n*} - \boldsymbol{s}^{\varepsilon *})\, \mathrm{d}\boldsymbol{x} \\ &= \int_\Omega \overline{\boldsymbol{\pi}}^\varepsilon \cdot (\boldsymbol{s}^n - \boldsymbol{s}^\varepsilon)\, \mathrm{d}\boldsymbol{x} = -\varepsilon \int_\Omega \overline{\boldsymbol{\pi}}^\varepsilon \cdot (\boldsymbol{s} - \boldsymbol{s}^n)\, \mathrm{d}\boldsymbol{x}. \end{aligned} \tag{49}$$

For the potential energy, we obtain by linearity that

$$\int_\Omega (\boldsymbol{s}^\varepsilon - \boldsymbol{s}^n) \cdot \boldsymbol{\Psi}\, \mathrm{d}\boldsymbol{x} = \varepsilon \int_\Omega (\boldsymbol{s} - \boldsymbol{s}^n) \cdot \boldsymbol{\Psi} \mathrm{d}\boldsymbol{x}. \tag{50}$$

Summing (48)–(50), dividing by ε, and recalling that $\boldsymbol{s}^n$ minimizes the functional $\mathcal{F}_\tau^n$ defined by (44), we obtain

$$0 \leq \frac{\mathcal{F}_\tau^n(\boldsymbol{s}^\varepsilon) - \mathcal{F}_\tau^n(\boldsymbol{s}^n)}{\varepsilon} \leq \sum_{i=0}^N \int_\Omega \left(\frac{\varphi_i^\varepsilon}{\tau} + \overline{\pi}_i^\varepsilon + \Psi_i\right)(s_i - s_i^n)\mathrm{d}\boldsymbol{x} \tag{51}$$

for all $\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ and all $\varepsilon \in (0,1)$. Because Ω is bounded, any Kantorovich potential is globally Lipschitz with bounds uniform in ε (see for instance the proof of [44, Theorem 1.17]). Since $\boldsymbol{s}^\varepsilon$ converges uniformly towards $\boldsymbol{s}^n$ when ε tends to 0, we infer from [44, Theorem 1.52] that φ_i^ε converges uniformly towards φ_i^n as ε tends to 0, where φ_i^n is a Kantorovich potential form s_i^n to s_i^{n-1}. Moreover, since $\boldsymbol{\pi}$ is uniformly continuous in $\boldsymbol{s}$, we also know that $\boldsymbol{\pi}^\varepsilon$ converges uniformly towards $\boldsymbol{\pi}^n$ and thus the extension to the zero-th component $\overline{\boldsymbol{\pi}}^\varepsilon \to \overline{\boldsymbol{\pi}}^n = (0, \boldsymbol{\pi}^n)$ as well. Then we can pass to the limit in (51) and infer that

$$0 \leq \int_\Omega \mathbf{F}^n \cdot (\boldsymbol{s} - \boldsymbol{s}^n)\mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}} \tag{52}$$

and (47) holds.

If $s_i^{n-1} > 0$ does not hold everywhere we argue by approximation. Running the flow (31) for a short time $\delta > 0$ starting from $\boldsymbol{s}^{n-1}$, we construct an approximation $\boldsymbol{s}^{n-1,\delta} = (s_0^{n-1,\delta}, \dots, s_N^{n-1,\delta})$ converging to $\boldsymbol{s}^{n-1} = (s_0^{n-1}, \dots, s_N^{n-1})$ in $L^1(\Omega)$ as δ tends to 0. By construction $\boldsymbol{s}^{n-1,\delta} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$, and it follows from the strong maximum principle that $s_i^{n-1,\delta} > 0$ in $\overline{\Omega}$ for all $\delta > 0$. By Proposition 2.1 there exists a unique minimizer $\boldsymbol{s}^{n,\delta}$ to the functional

$$\mathcal{F}_\tau^{n,\delta} : \begin{cases} \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}} \to \mathbb{R}_+ \\ \boldsymbol{s} \mapsto \frac{1}{2\tau}\boldsymbol{W}^2(\boldsymbol{s}, \boldsymbol{s}^{n-1,\delta}) + \mathcal{E}(\boldsymbol{s}) \end{cases}$$

Since $\boldsymbol{s}^{n-1,\delta} > 0$, there exist unique Kantorovich potentials $(\varphi_i^{n,\delta}, \psi_i^{n,\delta})$ from $s_i^{n,\delta}$ to $s_i^{n-1,\delta}$. This allows to construct $\mathbf{F}^{n,\delta}$ using (46) where φ_i^n (resp. π_i^n) has been replaced by $\varphi_i^{n,\delta}$ (resp. $\pi_i^{n,\delta}$). Thanks to the above discussion,

$$0 \leq \int_\Omega \mathbf{F}^{n,\delta *} \cdot (\boldsymbol{s}^* - \boldsymbol{s}^{n,\delta *}) \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{s}^* \in \boldsymbol{\mathcal{X}}^* \cap \boldsymbol{\mathcal{A}}^*. \tag{53}$$

We can now let δ tend to 0. Because of the time continuity of the solutions to (31), we know that $\boldsymbol{s}^{n-1,\delta}$ converges towards $\boldsymbol{s}^{n-1}$ in $L^1(\Omega)$. On the other hand, from the definition of $\boldsymbol{s}^{n,\delta}$ and Lemma 2.2 (in particular (30) with $s^{n-1,\delta}, s^{n,\delta}, \boldsymbol{\pi}^{n,\delta}$ instead of $s^{n-1}, s^n, \boldsymbol{\pi}^n$) we see that $\boldsymbol{\pi}^{n,\delta}$ is bounded in $H^1(\Omega)^{N+1}$ uniformly in $\delta > 0$. Using next the Lipschitz continuous (8) of $\boldsymbol{\phi}$, one deduces that $\boldsymbol{s}^{n,\delta}$ is uniformly bounded in $H^1(\Omega)^{N+1}$. Then, thanks to Rellich's compactness theorem, we can assume that $\boldsymbol{s}^{n,\delta}$ converges strongly in $L^2(\Omega)^{N+1}$ as δ tends to 0. By the strong convergence $\boldsymbol{s}^{n-1,\delta} \to \boldsymbol{s}^{n-1}$ and standard properties of the squared Wasserstein distance, one readily checks that $\mathcal{F}_\tau^{n,\delta}$ Γ-converges towards $\mathcal{F}_\tau^n$, and we can therefore identify the limit of $\boldsymbol{s}^{n,\delta}$ as the unique minimizer $\boldsymbol{s}^n$ of $\mathcal{F}_\tau^n$. Thanks to Lebesgue's dominated convergence theorem, we also infer that $\pi_i^{n,\delta}$ converges in $L^2(\Omega)$ towards π_i^n. Using once again the stability of the Kantorovich potentials [44, Theorem 1.52], we know that $\varphi_i^{n,\delta}$ converges uniformly towards some Kantorovich potential φ_i^n. Then we can pass to the limit in (53) and claim that (52) is satisfied even when some coordinates of $\boldsymbol{s}^{n-1}$ vanish on some parts of Ω.

Finally, note that since the Kantorovich potentials φ_i^n are Lipschitz continuous and because $\pi_i^n \in H^1$ (cf. Lemma 2.2) and $\boldsymbol{\Psi}$ is smooth, we have $F_i^n \in H^1$. Since the phases are bounded $0 \leq s_i^n(\boldsymbol{x}) \leq \omega(\boldsymbol{x})$ and $\boldsymbol{\pi}$ is continuous we have $\boldsymbol{\pi}^n \in L^\infty$, thus $F_i^n \in L^\infty$ as well and the proof is complete. □

We can now suitably decompose the vector field $\mathbf{F}^n = (F_i^n)_{0 \leq i \leq N}$ defined by (46).

Corollary 3.3. *Let $\boldsymbol{F}^n = (F_0^n, \dots, F_N^n)$ be as in Lemma 3.2. There exists $\boldsymbol{\alpha}^n \in \mathbb{R}^{N+1}$ such that, setting $\lambda^n(\boldsymbol{x}) := \min_j (F_j^n(\boldsymbol{x}) + \alpha_j^n)$, there holds $\lambda^n \in H^1(\Omega)$ and*

$$F_i^n + \alpha_i^n = \lambda^n \ \ \mathrm{d}s_i^n - a.e.\ in\ \Omega, \qquad \forall i \in \{0, \dots, N\}, \tag{54}$$

$$\boldsymbol{\nabla} F_i^n = \boldsymbol{\nabla} \lambda^n \ \ \mathrm{d}s_i^n - a.e.\ in\ \Omega, \qquad \forall i \in \{0, \dots, N\}. \tag{55}$$

Proof. By Lemma 3.2 we know that $\boldsymbol{s}^n$ minimizes $\boldsymbol{s} \mapsto \int \boldsymbol{F}^n \cdot \boldsymbol{s}$ among all admissible $\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$. Applying the multicomponent bathtub principle, Theorem B.1 in appendix, we infer that there exists $\boldsymbol{\alpha}^n = (\alpha_0^n, \dots, \alpha_N^n) \in \mathbb{R}^{N+1}$ such that $F_i^n + \alpha_i^n = \lambda^n$ for $\mathrm{d}s_i^n$-a.e. $\boldsymbol{x} \in \Omega$ and $\lambda^n = \min_j (F_j^n + \alpha_j^n)$ as in our statement. Note first that $\lambda^n \in H^1(\Omega)$ as the minimum of finitely many H^1 functions $F_0, \dots, F_N \in H^1(\Omega)$. From the usual Serrin's chain rule we have moreover that

$$\boldsymbol{\nabla} \lambda^n = \boldsymbol{\nabla} \min_j (F_j^n + \alpha_j^n) = \boldsymbol{\nabla} F_i \cdot \chi_{[F_i^n + \alpha_i^n = \lambda^n]},$$

and since $s_i^n = 0$ inside $[F_i^n + \alpha_i^n \neq \lambda^n]$ the proof is complete. □

3.2. **The discrete capillary pressure law and pressure estimates.** In this section, some calculations in the Riemannian settings (Ω, d_i) will be carried out. In order to make them as readable as possible, we have to introduce a few basics. We refer to [46, Chapter 14] for a more detail presentation.

Let $i \in \{0,\dots,N\}$, then consider the Riemannian geometry (Ω, d_i), and let $\boldsymbol{x} \in \Omega$, then we denote by $g_{i,\boldsymbol{x}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ the local metric tensor defined by

$$g_{i,\boldsymbol{x}}(\boldsymbol{v},\boldsymbol{v}) = \mu_i \mathbb{K}^{-1}(\boldsymbol{x})\boldsymbol{v}\cdot\boldsymbol{v} = \mathbb{G}_i(\boldsymbol{x})\boldsymbol{v}\cdot\boldsymbol{v}, \qquad \forall \boldsymbol{v} \in \mathbb{R}^d.$$

In this framework, the gradient $\boldsymbol{\nabla}_{g_i}\varphi$ of a function $\varphi \in \mathcal{C}^1(\Omega)$ is defined by

$$\varphi(\boldsymbol{x}+h\boldsymbol{v}) = \varphi(\boldsymbol{x}) + h g_{i,\boldsymbol{x}}\left(\boldsymbol{\nabla}_{g_{i,\boldsymbol{x}}}\varphi(\boldsymbol{x}),\boldsymbol{v}\right) + o(h), \qquad \forall \boldsymbol{v} \in \mathbb{S}^{d-1},\ \forall \boldsymbol{x} \in \Omega.$$

It is easy to check that this leads to the formula

$$\boldsymbol{\nabla}_{g_i}\varphi = \frac{1}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\varphi, \tag{56}$$

where $\boldsymbol{\nabla}\varphi$ stands for the usual (euclidean) gradient. The formula (56) can be extended to Lipschitz continuous functions φ thanks to Rademacher's theorem.

For φ belonging to $\mathcal{C}^2$, we can also define the Hessian $D^2_{g_i}\varphi$ of φ in the Riemannian setting by

$$g_{i,\boldsymbol{x}}\left(D^2_{g_i}\varphi(\boldsymbol{x})\cdot\boldsymbol{v},\boldsymbol{v}\right) = \left.\frac{d^2}{dt^2}\varphi(\boldsymbol{\gamma}_t)\right|_{t=0}$$

for any geodesic $\boldsymbol{\gamma}_t = \exp_{i,\boldsymbol{x}}(t\boldsymbol{v})$ starting from $\boldsymbol{x}$ with initial speed $\boldsymbol{v} \in T_{i,\boldsymbol{x}}\Omega$.

Denote by φ_i^n the backward Kantorovich potential sending s_i^n to s_i^{n-1} associated to the cost $\frac{1}{2}d_i^2$. By the usual definition of the Wasserstein distance through the Monge problem, one has

$$W_i^2(s_i^n, s_i^{n-1}) = \int_\Omega d_i^2(\boldsymbol{x}, \mathbf{t}_i^n(\boldsymbol{x}))s_i^n(\boldsymbol{x})\mathrm{d}\boldsymbol{x},$$

where $\mathbf{t}_i^n$ denotes the optimal map sending s_i^n on s_i^{n-1}. It follows from [46, Theorem 10.41] that

$$\mathbf{t}_i^n(\boldsymbol{x}) = \exp_{i,\boldsymbol{x}}\left(-\boldsymbol{\nabla}_{g_i}\varphi_i^n(\boldsymbol{x})\right), \qquad \forall \boldsymbol{x} \in \Omega. \tag{57}$$

Moreover, using the definition of the exponential and the relation (56), one gets that

$$d_i^2(\boldsymbol{x}, \exp_{i,\boldsymbol{x}}\left(-\boldsymbol{\nabla}_{g_i}\varphi_i^n(\boldsymbol{x})\right) = g_{i,\boldsymbol{x}}\left(\boldsymbol{\nabla}_{g_i}\varphi_i^n(\boldsymbol{x}), \boldsymbol{\nabla}_{g_i}\varphi_i^n(\boldsymbol{x})\right) = \frac{1}{\mu_i}\mathbb{K}(\boldsymbol{x})\boldsymbol{\nabla}\varphi_i^n(\boldsymbol{x})\cdot\boldsymbol{\nabla}\varphi_i^n(\boldsymbol{x}).$$

This yields the formula

$$W_i^2(s_i^n, s_i^{n-1}) = \int_\Omega \frac{s_i^n}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\varphi_i^n\cdot\boldsymbol{\nabla}\varphi_i^n\mathrm{d}\boldsymbol{x}, \qquad \forall i \in \{0,\dots,N\}. \tag{58}$$

We have now introduced the necessary material in order to reconstruct the phase and capillary pressures. This is the purpose of the following Proposition 3.4 and of then Corollary 3.5

Proposition 3.4. *For $n \geq 1$ let $\varphi_i^n : s_i^n \to s_i^{n-1}$ be the (backward) Kantorovich potentials from Lemma 3.2. There exists $\boldsymbol{h} = (h_0^n,\dots,h_N^n) \in H^1(\Omega)^{N+1}$ such that*

(i) $\boldsymbol{\nabla}h_i^n = -\frac{\boldsymbol{\nabla}\varphi_i^n}{\tau}$ *for* $\mathrm{d}s_i^n$*-a.e.* $\boldsymbol{x} \in \Omega$

(ii) $h_i^n(\boldsymbol{x}) - h_0^n(\boldsymbol{x}) = \pi_i^n(\boldsymbol{x}) + \Psi_i(\boldsymbol{x}) - \Psi_0(\boldsymbol{x})$ *for* $\mathrm{d}\boldsymbol{x}$*-a.e.* $\boldsymbol{x} \in \Omega$, $i \in \{1,\dots,N\}$

(iii) there exists C depending only on $\Omega, \Pi, \omega, \mathbb{K}, (\mu_i)_i$, and $\boldsymbol{\Psi}$ such that, for all $n \geq 1$ and all $\tau > 0$, one has

$$\|\boldsymbol{h}^n\|^2_{H^1(\Omega)^{N+1}} \leq C\left(1 + \frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau^2} + \sum_{i=0}^N \frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(s_i^n)}{\tau}\right).$$

Proof. Let φ_i^n be the Kantorovich potentials from Lemma 3.2 and $F_i^n \in L^\infty \cap H^1(\Omega)$ as in (46), as well as $\boldsymbol{\alpha}^n \in \mathbb{R}^{N+1}$ and $\lambda^n = \min_j (F_j^n + \alpha_j^n) \in L^\infty \cap H^1(\Omega)$ as in Corollary 3.3. Setting

$$h_i^n := -\frac{\varphi_i^n}{\tau} + F_i^n - \lambda^n, \qquad \forall i \in \{0, \dots, N\},$$

we have $h_i^n \in H^1(\Omega)$ as the sum of Lipschitz functions (the Kantorovich potentials φ_i^n) and H^1 functions F_i^n, λ^n. Recalling that we use the notation $\pi_0 = \frac{\partial \Pi}{\partial s_0} = 0$, we see from the definition (46) of F_i^n that

(59)
$$h_i^n - h_0^n = \left(F_i^n - \frac{\varphi_i^n}{\tau}\right) - \left(F_0^n - \frac{\varphi_0^n}{\tau}\right) = (\pi_i^n + \Psi_i) - (\pi_0^n + \Psi_0) = \pi_i^n + \Psi_i - \Psi_0$$

for all $i \in \{1, \dots, N\}$ and $\mathrm{d}\boldsymbol{x}$-a.e. x, which is exactly our statement (ii).

For (i), we simply use (55) to compute

(60)
$$\boldsymbol{\nabla} h_i^n = -\frac{\boldsymbol{\nabla}\varphi_i^n}{\tau} + \boldsymbol{\nabla}(F_i^n - \lambda_i^n) = -\frac{\boldsymbol{\nabla}\varphi_i^n}{\tau} \quad \text{for } \mathrm{d}s_i^n\text{-a.e. } \boldsymbol{x} \in \Omega, \qquad \forall i \in \{0, \dots, N\}.$$

In order to establish now the H^1 estimate (iii), let us denote

$$\mathcal{U}_i = \left\{ \boldsymbol{x} \in \Omega \;\middle|\; s_i^n(\boldsymbol{x}) \geq \frac{\omega_\star}{N+1} \right\}.$$

Then since $\sum s_i^n(\boldsymbol{x}) = \omega(\boldsymbol{x}) \geq \omega_\star > 0$, one gets that, up to a negligible set,

$$\bigcup_{i=0}^{N} \mathcal{U}_i = \Omega, \quad \text{hence} \quad (\mathcal{U}_i)^c \subset \bigcup_{j \neq i} \mathcal{U}_j. \tag{61}$$

We first estimate $\boldsymbol{\nabla} h_0^n$. To this end, we write

$$\|\boldsymbol{\nabla} h_0^n\|_{L^2}^2 \leq \frac{1}{\kappa_\star} \int_\Omega \mathbb{K} \boldsymbol{\nabla} h_0^n \cdot \boldsymbol{\nabla} h_0^n \mathrm{d}\boldsymbol{x} \leq A + B, \tag{62}$$

where we have set

$$A = \frac{1}{\kappa_\star} \int_{\mathcal{U}_0} \mathbb{K} \boldsymbol{\nabla} h_0^n \cdot \boldsymbol{\nabla} h_0^n \mathrm{d}\boldsymbol{x}, \qquad B = \frac{1}{\kappa_\star} \int_{(\mathcal{U}_0)^c} \mathbb{K} \boldsymbol{\nabla} h_0^n \cdot \boldsymbol{\nabla} h_0^n \mathrm{d}\boldsymbol{x}.$$

Owing to (60) one has $\boldsymbol{\nabla} h_0^n = -\frac{\boldsymbol{\nabla}\varphi_0}{\tau}$ on $\mathcal{U}_0 \subset \Omega$, where $s_0^n \geq \frac{\omega_\star}{N+1}$. Therefore,

$$A \leq \frac{(N+1)\mu_0}{\omega_\star \kappa_\star} \int_{\mathcal{U}_0} \frac{s_0^n}{\mu_0} \mathbb{K} \boldsymbol{\nabla} h_0^n \cdot \boldsymbol{\nabla} h_0^n \mathrm{d}\boldsymbol{x} \leq \frac{(N+1)\mu_0}{\tau^2 \omega_\star \kappa_\star} \int_\Omega \frac{s_0^n}{\mu_0} \mathbb{K} \boldsymbol{\nabla} \varphi_0^n \cdot \boldsymbol{\nabla} \varphi_0^n \mathrm{d}\boldsymbol{x}.$$

Then it results from formula (58) that

$$A \leq \frac{C}{\tau^2} W_0^2(s_0^n, s_0^{n-1}) \tag{63}$$

where C depends neither on n nor on τ. Combining (61) and (59), we infer

$$B \leq \frac{1}{\kappa_\star} \sum_{i=1}^{N} \int_{\mathcal{U}_i} \mathbb{K} \boldsymbol{\nabla}[h_i^n - (\pi_i^n + \Psi_i - \Psi_0)] \cdot \boldsymbol{\nabla}[h_i^n - (\pi_i^n + \Psi_i - \Psi_0)] \mathrm{d}\boldsymbol{x}.$$

Using $(a+b+c)^2 \leq 3(a^2+b^2+c^2)$ and (3), we get that

$$B \leq \frac{3}{\kappa_\star}\sum_{i=1}^{N}\int_{\mathcal{U}_i}\mathbb{K}\boldsymbol{\nabla}h_i\cdot\boldsymbol{\nabla}h_i\mathrm{d}\boldsymbol{x} + \frac{3\kappa^\star}{\kappa_\star}\sum_{i=1}^{N}\left(\|\boldsymbol{\nabla}\pi_i^n\|_{L^2}^2 + \|\boldsymbol{\nabla}(\Psi_i-\Psi_0)\|_{L^2}^2\right). \tag{64}$$

Similar calculations to those carried out to estimate A yield

$$\int_{\mathcal{U}_i}\mathbb{K}\boldsymbol{\nabla}h_i\cdot\boldsymbol{\nabla}h_i\mathrm{d}\boldsymbol{x} \leq \frac{C}{\tau^2}W_i^2(s_i^n, s_i^{n-1})$$

for some C depending neither on n, i nor on τ. Combining this inequality with Lemma 2.2 and the regularity of $\boldsymbol{\Psi}$, we get from (64) that

$$B \leq C\left(1+\frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau^2}+\sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1})-\mathcal{H}_\omega(s_i^n)}{\tau}\right) \tag{65}$$

for some C not depending on n and τ (here we also used $1/\tau \leq 1/\tau^2$ for small τ in the W^2 terms). Gathering (63) and (65) in (62) provides

$$\|\boldsymbol{\nabla}h_0^n\|_{L^2}^2 \leq C\left(1+\frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau^2}+\sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1})-\mathcal{H}_\omega(s_i^n)}{\tau}\right).$$

Note that (i)(ii) remain invariant under subtraction of the same constant $h_0^n, h_i^n \rightsquigarrow h_0^n - C, h_i^n - C$, as the gradients remain unchanged in (i) and only the differences $h_i^n - h_0^n$ appear in (ii) for $i \in \{1 \dots N\}$. We can therefore assume without loss of generality that $\int_\Omega h_0^n \mathrm{d}\boldsymbol{x} = 0$. Hence by the Poincaré-Wirtinger inequality, we get that

$$\|h_0^n\|_{H^1}^2 \leq C\|\boldsymbol{\nabla}h_0^n\|_{L^2}^2 \leq C\left(1+\frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau^2}+\sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1})-\mathcal{H}_\omega(s_i^n)}{\tau}\right).$$

Finally, from (ii) $h_i^n = h_0^n + \pi_i^n + \Psi_i - \Psi_0$, the smoothness of $\boldsymbol{\Psi}$, and using again the estimate (30) for $\|\boldsymbol{\nabla}\boldsymbol{\pi}^n\|_{L^2}^2$ we finally get that for all $i \in \{1,\dots,N\}$, one has

$$\begin{aligned}\|h_i^n\|_{H^1}^2 &\leq C(\|h_0^n\|_{H^1}^2 + \|\pi_i^n\|_{H^1}^2 + \|\Psi_i\|_{H^1}^2 + \|\Psi_0\|_{H^1}^2)\\ &\leq C\left(1+\frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau^2}+\sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1})-\mathcal{H}_\omega(s_i^n)}{\tau}\right),\end{aligned}$$

and the proof of Proposition 3.4 is complete. □

We can now define the phase pressures $(p_i^n)_{i=0,\dots,N}$ by setting

$$p_i^n := h_i^n - \Psi_i, \qquad \forall i \in \{0,\dots,N\}. \tag{66}$$

The following corollary is a straightforward consequence of Proposition 3.4 and of the regularity of Ψ_i.

Corollary 3.5. *The phase pressures $\boldsymbol{p}^n = (p_i^n)_{0\leq i\leq N} \in H^1(\Omega)^{N+1}$ satisfy*

$$\|\boldsymbol{p}^n\|_{H^1(\Omega)}^2 \leq C\left(1+\frac{\boldsymbol{W}^2(\boldsymbol{s}^n,\boldsymbol{s}^{n-1})}{\tau^2}+\sum_{i=0}^{N}\frac{\mathcal{H}_\omega(s_i^{n-1})-\mathcal{H}_\omega(s_i^n)}{\tau}\right) \tag{67}$$

for some C depending only on $\Omega, \Pi, \omega, \mathbb{K}, (\mu_i)_i$, and $\boldsymbol{\Psi}$ (but neither on n nor on τ), and the capillary pressure relations are fulfilled:

$$p_i^n - p_0^n = \pi_i^n, \qquad \forall i \in \{1,\dots,N\}. \tag{68}$$

Our next result is a first step towards the recovery of the PDEs.

Lemma 3.6. *There exists C depending depending only on $\Omega, \Pi, \omega, \mathbb{K}, (\mu_i)_i$, and $\boldsymbol{\Psi}$ (but neither on n nor on τ) such that, for all $i \in \{0, \dots, N\}$ and all $\xi \in \mathcal{C}^2(\overline{\Omega})$, one has*

$$\left| \int_\Omega \left(s_i^n - s_i^{n-1}\right) \xi \mathrm{d}\boldsymbol{x} + \tau \int_\Omega s_i^n \frac{\mathbb{K}}{\mu_i} \boldsymbol{\nabla} \left(p_i^n + \Psi_i\right) \cdot \boldsymbol{\nabla} \xi \mathrm{d}\boldsymbol{x} \right| \leq C W_i^2(s_i^n, s_i^{n-1}) \|D_{g_i}^2 \xi\|_\infty. \tag{69}$$

This is of course a discrete approximation to the continuity equation $\partial_t s_i = \boldsymbol{\nabla} \cdot (s_i \frac{\mathbb{K}}{\mu_i} \boldsymbol{\nabla} (p_i + \Psi_i))$.

Proof. Let φ_i^n denote the (backward) optimal Kantorovich potential from Lemma 3.2 sending s_i^n to s_i^{n-1}, and let $\mathbf{t}_i^n$ be the corresponding optimal map as in (57). For fixed $\xi \in \mathcal{C}^2(\overline{\Omega})$ let us first Taylor expand (in the g_i Riemannian framework)

$$\left| \xi(\mathbf{t}_i^n(\boldsymbol{x})) - \xi(\boldsymbol{x}) + \frac{1}{\mu_i} \mathbb{K}(\boldsymbol{x}) \boldsymbol{\nabla} \xi(\boldsymbol{x}) \cdot \boldsymbol{\nabla} \varphi_i^n(\boldsymbol{x}) \right| \leq \frac{1}{2} \|D_{g_i}^2 \xi\|_\infty d_i^2(\boldsymbol{x}, \mathbf{t}_i^n(\boldsymbol{x})).$$

Using the definition of the pushforward $s_i^{n-1} = \mathbf{t}_i^n \# s_i^n$, we then compute

$$\begin{aligned}
&\left| \int_\Omega (s_i^n(\boldsymbol{x}) - s_i^{n-1}(\boldsymbol{x})) \xi(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \frac{\mathbb{K}(\boldsymbol{x})}{\mu_i} \boldsymbol{\nabla} \xi(x) \cdot \boldsymbol{\nabla} \varphi_i^n(\boldsymbol{x}) s_i^n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad = \left| \int_\Omega (\xi(\boldsymbol{x}) - \xi(\mathbf{t}_i^n(\boldsymbol{x})) s_i^n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \int_\Omega \frac{\mathbb{K}(\boldsymbol{x})}{\mu_i} \boldsymbol{\nabla} \xi(x) \cdot \boldsymbol{\nabla} \varphi_i^n(\boldsymbol{x}) s_i^n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \right| \\
&\quad\quad \leq \int_\Omega \frac{1}{2} \|D_{g_i}^2 \xi\|_\infty d_i^2(\boldsymbol{x}, \mathbf{t}_i^n(\boldsymbol{x})) s_i^n(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = \frac{1}{2} \|D_{g_i}^2 \xi\|_\infty W_i^2(s_i^n, s_i^{n-1}).
\end{aligned}$$

From Proposition 3.4(i) we have $\boldsymbol{\nabla} \varphi_i^n = -\tau \boldsymbol{\nabla} h_i^n$ for $\mathrm{d}s_i^n$ a.e. $\boldsymbol{x} \in \Omega$, thus by the definition (66) of p_i^n, we get $\boldsymbol{\nabla} \varphi^n = -\tau \boldsymbol{\nabla} (p_i^n + \Psi_i)$. Substituting in the second integral of the left-hand side gives exactly (69) and the proof is complete. □

4. Convergence towards a weak solution

The goal is now to prove the convergence of the piecewise constant inteprolated solutions $\boldsymbol{s}^\tau$, defined by (26), towards a weak solution $\boldsymbol{s}$ as $\tau \to 0$. Similarly, the τ superscript denotes the piecewise constant interpolation of any previous discrete quantity (e.g. $p_i^\tau(t)$ stands for the piecewise constant time interpolation of the discrete pressures p_i^n). In what follows, we will also use the notations $\boldsymbol{s}^{\tau *} = (s_1^\tau, \dots, s_N^\tau) \in L^\infty((0,T); \boldsymbol{\mathcal{X}}^*)$ and $\boldsymbol{\pi}^\tau = \boldsymbol{\pi}(\boldsymbol{s}^{\tau *}, \boldsymbol{x})$.

4.1. Time integrated estimates.

We immediately deduce from (29) that

$$\boldsymbol{W}(\boldsymbol{s}^\tau(t_2), \boldsymbol{s}^\tau(t_1)) \leq C |t_2 - t_1 + \tau|^{\frac{1}{2}}, \qquad \forall\, 0 \leq t_1 \leq t_2 \leq T. \tag{70}$$

From the total saturation $\sum_{i=0}^N s_i^n(\boldsymbol{x}) = \omega(\boldsymbol{x}) \leq \omega^\star$ and $s_i^\tau \geq 0$, we have the L^∞ estimates

$$0 \leq s_i^\tau(\boldsymbol{x}, t) \leq \omega^\star \quad \text{a.e. in } Q \text{ for all } i \in \{0, \dots, N\}. \tag{71}$$

Lemma 4.1. *There exists C depending only on $\Omega, T, \Pi, \omega, \mathbb{K}, (\mu_i)_i$, and $\boldsymbol{\Psi}$ such that*

$$\|\boldsymbol{p}^\tau\|_{L^2((0,T);H^1(\Omega)^{N+1})}^2 + \|\boldsymbol{\pi}^\tau\|_{L^2((0,T);H^1(\Omega)^N)}^2 \leq C. \tag{72}$$

Proof. Summing (67) from $n=1$ to $n=N_\tau := \lceil T/\tau \rceil$, we get

$$\begin{aligned}\|\boldsymbol{p}^\tau\|^2_{L^2(H^1)} &= \sum_{n=1}^{N_\tau} \tau \|\boldsymbol{p}^n\|^2_{H^1} \\ &\le C \sum_{n=1}^{N_\tau} \tau \left(1 + \frac{\boldsymbol{W}^2(\boldsymbol{s}^n, \boldsymbol{s}^{n-1})}{\tau^2} + \sum_{i=0}^{N_\tau} \frac{\mathcal{H}_\omega(s_i^{n-1}) - \mathcal{H}_\omega(s_i^n)}{\tau}\right) \\ &\le C \left((T+1) + \sum_{n=1}^{N_\tau} \frac{\boldsymbol{W}^2(\boldsymbol{s}^n, \boldsymbol{s}^{n-1})}{\tau} + \sum_{i=0}^{N} \left(\mathcal{H}_\omega(s_i^0) - \mathcal{H}_\omega(s_i^{N_\tau})\right)\right).\end{aligned}$$

We use that

$$0 \ge \mathcal{H}_\omega(s) \ge -\frac{1}{e}\|\omega\|_{L^1} \ge -\frac{|\Omega|}{e}, \qquad \forall s \in L^\infty(\Omega) \text{ with } 0 \le s \le \omega$$

together with the total square distance estimate (28) to infer that $\|\boldsymbol{p}\|^2_{L^2(H^1)} \le C$. The proof is identical for the capillary pressure $\boldsymbol{\pi}^\tau$ (simply summing the one-step estimate from Lemma 2.2). □

4.2. **Compactness of approximate solutions.** We denote by $H' = H^1(\Omega)'$.

Lemma 4.2. *For each $i \in \{0, \dots, N\}$, there exists C depending only on Ω, Π, $\boldsymbol{\Psi}$, $\mathbb{K}$, and μ_i (but not on τ) such that*

$$\|s_i^\tau(t_2) - s_i^\tau(t_1)\|_{H'} \le C|t_2 - t_1 + \tau|^{\frac{1}{2}}, \qquad \forall\, 0 \le t_1 \le t_2 \le T.$$

Proof. Thanks to (71), we can apply [39, Lemma 3.4] to get

$$\left|\int_\Omega f\{s_i^\tau(t_2) - s_i^\tau(t_1)\}\mathrm{d}\boldsymbol{x}\right| \le \|\boldsymbol{\nabla} f\|_{L^2(\Omega)} W_{\mathrm{ref}}(s_i^\tau(t_1), s_i^\tau(t_2)), \qquad \forall f \in H^1(\Omega).$$

Thus by duality and thanks to the distance estimate (70) and to the lower bound in (19), we obtain that

$$\|s_i^\tau(t_2) - s_i^\tau(t_1)\|_{H'} \le W_{\mathrm{ref}}(s_i^\tau(t_1), s_i^\tau(t_2)) \le C W_i(s_i^\tau(t_1), s_i^\tau(t_2)) \le C|t_2 - t_1 + \tau|^{\frac{1}{2}}$$

for some C depending only on Ω, Π, $(\rho_i)_i$, $\mathbf{g}$, $(\mu_i)_i$, $\mathbb{K}$. □

From the previous equi-continuity in time, we deduce full compactness of the capillary pressure:

Lemma 4.3. *The family $(\boldsymbol{\pi}^\tau)_{\tau>0}$ is sequentially relatively compact in $L^2(Q)^N$.*

Proof. We use Alt & Luckhaus' trick [1] (an alternate solution would consist in slightly adapting the nonlinear time compactness results [40, 8] to our context). Let $h > 0$ be a small time shift, then by monotonicity and Lipschitz continuity of the capillary pressure function $\boldsymbol{\pi}(., \boldsymbol{x})$

$$\begin{aligned}&\|\boldsymbol{\pi}^\tau(\cdot + h) - \boldsymbol{\pi}^\tau(\cdot)\|^2_{L^2((0,T-h);L^2(\Omega)^N)} \\ &\quad\le \frac{1}{\kappa_\star}\int_0^{T-h}\int_\Omega (\boldsymbol{\pi}^\tau(t+h,\boldsymbol{x}) - \boldsymbol{\pi}^\tau(t,\boldsymbol{x}))\cdot(\boldsymbol{s}^{\tau *}(t+h,\boldsymbol{x}) - \boldsymbol{s}^{\tau *}(t,\boldsymbol{x}))\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ &\qquad\le \frac{2\sqrt{T}}{\kappa_\star}\|\boldsymbol{\pi}^\tau\|_{L^2((0,T);H^1(\Omega)^N)}\|\boldsymbol{s}^{\tau *}(\cdot + h, \cdot) - \boldsymbol{s}^{\tau *}\|_{L^\infty((0,T-h);H')^N}.\end{aligned}$$

Then it follows from Lemmas 4.1 and 4.2 that there exists $C > 0$, depending neither on h nor on τ, such that

$$\|\boldsymbol{\pi}^\tau(\cdot + h, \cdot) - \boldsymbol{\pi}^\tau\|_{L^2((0,T-h);L^2(\Omega)^N)} \leq C|h + \tau|^{1/2}.$$

On the other hand, the (uniform w.r.t. τ) $L^2((0,T);H^1(\Omega)^N)$- and $L^\infty(Q)^N$-estimates on $\boldsymbol{\pi}^\tau$ ensure that

$$\|\boldsymbol{\pi}^\tau(\cdot, \cdot + \boldsymbol{y})) - \boldsymbol{\pi}^\tau\|_{L^2(0,T;L^2)} \leq C\sqrt{|\boldsymbol{y}|}(1 + \sqrt{|\boldsymbol{y}|}), \qquad \forall\, \boldsymbol{y} \in \mathbb{R}^d,$$

where $\boldsymbol{\pi}^\tau$ is extended by 0 outside Ω. This allows to apply Kolmogorov's compactness theorem (see, for instance, [29]) and entails the desired relative compactness. □

4.3. **Identification of the limit.** In this section we prove our main Theorem 1.2, and the proof goes in two steps: we first retrieve strong convergence of the phase contents $\boldsymbol{s}^\tau \to \boldsymbol{s}$ and weak convergence of the pressures $\boldsymbol{p}^\tau \rightharpoonup \boldsymbol{p}$, and then use the strong-weak limit of products to show that the limit is a weak solution. All along this section, $(\tau_k)_{k\geq 1}$ denotes a sequence of times steps tending to 0 as $k \to \infty$.

Lemma 4.4. *There exist $\boldsymbol{s} \in L^\infty(Q)^{N+1}$ with $\boldsymbol{s}(\cdot,t) \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ for a.e. $t \in (0,T)$, and $\boldsymbol{p} \in L^2((0,T);H^1(\Omega)^{N+1})$ such that, up to an unlabeled subsequence, the following convergence properties hold:*

$$\boldsymbol{s}^{\tau_k} \underset{k\to\infty}{\longrightarrow} \boldsymbol{s} \quad \text{a.e. in } Q, \tag{73}$$

$$\boldsymbol{\pi}^{\tau_k} \underset{k\to\infty}{\longrightarrow} \boldsymbol{\pi}(\boldsymbol{s}^*, \cdot) \quad \text{weakly in } L^2((0,T);H^1(\Omega)^N), \tag{74}$$

$$\boldsymbol{p}^{\tau_k} \underset{k\to\infty}{\longrightarrow} \boldsymbol{p} \quad \text{weakly in } L^2((0,T);H^1(\Omega)^{N+1}). \tag{75}$$

Moreover, the capillary pressure relations (5) hold.

Proof. From Lemma 4.3, we can assume that $\boldsymbol{\pi}^{\tau_k} \to \boldsymbol{z}$ strongly in $L^2(Q)^N$ for some limit $\boldsymbol{z}$, thus a.e. up to the extraction of an additional subsequence. Since $\boldsymbol{z} \mapsto \boldsymbol{\phi}(\boldsymbol{z},\boldsymbol{x}) = \boldsymbol{\pi}^{-1}(\boldsymbol{z},\boldsymbol{x})$ is continuous, we have that

$$\boldsymbol{s}^{\tau_k *} = \boldsymbol{\phi}(\boldsymbol{\pi}^{\tau_k}, \boldsymbol{x}) \underset{k\to\infty}{\longrightarrow} \boldsymbol{\phi}(\boldsymbol{\pi}, \boldsymbol{x}) =: \boldsymbol{s}^* \quad \text{a.e. in } Q.$$

In particular, this yields $\boldsymbol{\pi}^{\tau_k} \underset{k\to\infty}{\longrightarrow} \boldsymbol{\pi}(\boldsymbol{s}^*, \cdot)$ a.e. in Q. Since we had the total saturation $\sum_{i=0}^{N} s_i^{\tau_k}(t,\boldsymbol{x}) = \omega(\boldsymbol{x})$, we conclude that the first component $i = 0$ converges pointwise as well. Therefore, (73) holds. Thanks to Lebesgue's dominated convergence theorem, it is easy to check that $\boldsymbol{s}(\cdot,t) \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ for a.e. $t \in (0,T)$. The convergences (74) and (75) are straightforward consequences of Lemma 4.1. Lastly, it follows from (68) that

$$p_i^{\tau_k} - p_0^{\tau_k} = \boldsymbol{\pi}_i(\boldsymbol{s}^{\tau_k *}, \cdot), \qquad \forall i \in \{1,\dots,N\},\ \forall k \geq 1.$$

We can finally pass to the limit $k \to \infty$ in the above relation thanks to (74)–(75) and infer

$$p_i - p_0 = \boldsymbol{\pi}_i(\boldsymbol{s}^*, \boldsymbol{x}) \quad \text{in } L^2((0,T);H^1(\Omega)), \qquad \forall i \in \{1,\dots,N\}.$$

which immediately implies (5) as claimed. □

Lemma 4.5. *Up to the extraction of an additional subsequence, the limit $\boldsymbol{s}$ of $(\boldsymbol{s}^{\tau_k})_{k\geq 1}$ belongs to $\mathcal{C}([0,T];\boldsymbol{\mathcal{A}})$ where $\boldsymbol{\mathcal{A}}$ is equipped with the metric $\boldsymbol{W}$. Moreover, $\boldsymbol{W}(\boldsymbol{s}^{\tau_k}(t),\boldsymbol{s}(t)) \underset{k\to\infty}{\longrightarrow} 0$ for all $t\in[0,T]$.*

Proof. It follows from the bounds (71) on s_i that for all $t\in[0,T]$, the sequence $(s_i^{\tau_k})_k$ is weakly compact in $L^1(\Omega)$. It is also compact in $\mathcal{A}_i$ equipped with the metric W_i due to the continuity of W_i with respect to the weak convergence in $L^1(\Omega)$ (this is for instance a consequence of [44, Theorem 5.10] together with the equivalence of W_i with W_{ref} stated in (19)). Thanks to (70), one has

$$\limsup_{k\to\infty} W_i\left(s_i^{\tau_k}(t_2), s_i^{\tau_k}(t_1)\right) \leq |t_2-t_1|^{1/2}, \qquad \forall t_1,t_2\in[0,T].$$

Applying a refined version of the Arzelà-Ascoli theorem [5, Prop. 3.3.1] then provides the desired result. □

In order to conclude the proof of Theorem 1.2, it only remains to show that $\boldsymbol{s}=\lim \boldsymbol{s}^{\tau_k}$ and $\boldsymbol{p}=\lim \boldsymbol{p}^{\tau_k}$ satisfy the weak formulation (12):

Proposition 4.6. *Let $(\tau_k)_{k\geq 1}$ be a sequence such that the convergences in Lemmas 4.4 and 4.5 hold. Then the limit $\boldsymbol{s}$ of $(\boldsymbol{s}^{\tau_k})_{k\geq 1}$ is a weak solution in the sense of Definition 1.1 (with $-\rho_i\mathbf{g}$ replaced by $+\nabla\Psi_i$ in the general case).*

Proof. Let $0\leq t_1\leq t_2\leq T$, and denote $n_{j,k}=\left\lceil \frac{t_j}{\tau_k}\right\rceil$ and $\tilde{t}_j=n_{j,k}\tau_k$ for $j\in\{1,2\}$. Fixing an arbitrary $\xi\in\mathcal{C}^2(\overline{\Omega})$ and summing (69) from $n=n_{1,k}+1$ to $n=n_{2,k}$ yields

$$\begin{aligned}(76)\quad \int_\Omega (s_i^{\tau_k}(t_2)-s_i^{\tau_k}(t_1))\xi \mathrm{d}\boldsymbol{x} &= \sum_{n=n_{1,k}+1}^{n_{2,k}} \int_\Omega (s_i^n-s_i^{n-1})\xi \mathrm{d}\boldsymbol{x}\\ &= -\int_{\tilde{t}_1}^{\tilde{t}_2}\int_\Omega \frac{s_i^{\tau_k}}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i^{\tau_k}+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mathcal{O}\left(\sum_{n=n_{1,k}+1}^{n_{2,k}} W_i^2(s_i^n,s_i^{n-1})\right).\end{aligned}$$

Since $0\leq \tilde{t}_j-t_j\leq\tau_k$ and $\frac{s_i^{\tau_k}}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i^{\tau_k}+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi$ is uniformly bounded in $L^2(Q)$, one has

$$\begin{aligned}&\int_{\tilde{t}_1}^{\tilde{t}_2}\int_\Omega \frac{s_i^{\tau_k}}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i^{\tau_k}+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t\\ &\qquad = \int_{t_1}^{t_2}\int_\Omega \frac{s_i^{\tau_k}}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i^{\tau_k}+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mathcal{O}(\sqrt{\tau_k}).\end{aligned}$$

Combining the above estimate with the total square distance estimate (28) in (76), we obtain

$$(77)\quad \int_\Omega (s_i^{\tau_k}(t_2)-s_i^{\tau_k}(t_1))\xi \mathrm{d}\boldsymbol{x} + \int_{t_1}^{t_2}\int_\Omega \frac{s_i^{\tau_k}}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i^{\tau_k}+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t = \mathcal{O}\left(\sqrt{\tau_k}\right).$$

Thanks to Lemma 4.5, and since the convergence in $(\mathcal{A}_i,W_i)$ is equivalent to the narrow convergence of measures (i.e., the convergence in $\mathcal{C}(\overline{\Omega})'$, see for instance [44, Theorem 5.10]), we get that

$$(78)\qquad \int_\Omega (s_i^{\tau_k}(t_2)-s_i^{\tau_k}(t_1))\xi \mathrm{d}\boldsymbol{x} \underset{k\to\infty}{\longrightarrow} \int_\Omega (s_i(t_2)-s_i(t_1))\xi \mathrm{d}\boldsymbol{x}.$$

Moreover, thanks to Lemma 4.4, one has

$$\int_{t_1}^{t_2}\int_\Omega \frac{s_i^{\tau_k}}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i^{\tau_k}+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{k\to\infty}{\longrightarrow} \int_{t_1}^{t_2}\int_\Omega \frac{s_i}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{79}$$

Gathering (77)–(79) yields, for all $\xi\in\mathcal{C}^2(\overline{\Omega})$ and all $0\le t_1\le t_2\le T$,

$$\int_\Omega (s_i(t_2)-s_i(t_1))\xi \mathrm{d}\boldsymbol{x}+\int_{t_1}^{t_2}\int_\Omega \frac{s_i}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i+\Psi_i\right)\cdot\boldsymbol{\nabla}\xi \mathrm{d}\boldsymbol{x}\mathrm{d}t=0. \tag{80}$$

In order to conclude the proof, it remains to check that the formulation (80) is stronger the formulation (12). Let $\varepsilon>0$ be a time step (unrelated to that appearing in the minimization scheme (25)), and set $L_\varepsilon=\left\lfloor\frac{T}{\varepsilon}\right\rfloor$. Let $\phi\in\mathcal{C}_c^\infty(\overline{\Omega}\times[0,T))$, one sets $\phi_\ell=\phi(\cdot,\ell\varepsilon)$ for $\ell\in\{0,\dots,L_\varepsilon\}$. Since $t\mapsto\phi(\cdot,t)$ is compactly supported in $[0,T)$, then there exists $\varepsilon^\star>0$ such that $\phi_{L_\varepsilon}\equiv 0$ for all $\varepsilon\in(0,\varepsilon^\star]$. Then define by

$$\phi^\varepsilon:\left\{\begin{array}{rcll}\overline{\Omega}\times[0,T] & \to & \mathbb{R} & \\ (\boldsymbol{x},t) & \mapsto & \phi_\ell(\boldsymbol{x}) & \text{if } t\in[\ell\varepsilon,(\ell+1)\varepsilon).\end{array}\right.$$

Choose $t_1=\ell\varepsilon$, $t_2=(\ell+1)\varepsilon$, $\xi=\phi_\ell$ in (80) and sum over $\ell\in\{0,\dots,L_\varepsilon-1\}$. This provides

$$A(\varepsilon)+B(\varepsilon)=0,\qquad \forall\varepsilon>0. \tag{81}$$

where

$$A(\varepsilon)=\sum_{\ell=0}^{L_\varepsilon-1}\int_\Omega\left(s_i((\ell+1)\varepsilon)-s_i(\ell\varepsilon)\right)\phi^\ell\mathrm{d}\boldsymbol{x},$$
$$B(\varepsilon)=\iint_Q\frac{s_i}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i+\Psi_i\right)\cdot\boldsymbol{\nabla}\phi^\varepsilon\mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

Due to the regularity of ϕ, $\boldsymbol{\nabla}\phi^\varepsilon$ converges uniformly towards ϕ as ε tends to 0, so that

$$B(\varepsilon)\underset{\varepsilon\to0}{\longrightarrow}\iint_Q\frac{s_i}{\mu_i}\mathbb{K}\boldsymbol{\nabla}\left(p_i+\Psi_i\right)\cdot\boldsymbol{\nabla}\phi\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{82}$$

Reorganizing the first term and using that $\phi_{L_\varepsilon}\equiv0$, we get that

$$A(\varepsilon)=-\sum_{\ell=1}^{L_\varepsilon}\varepsilon\int_\Omega s_i(\ell\varepsilon)\frac{\phi_\ell-\phi_{\ell-1}}{\varepsilon}\,\mathrm{d}\boldsymbol{x}-\int_\Omega s_i^0\phi(\cdot,0)\mathrm{d}\boldsymbol{x}.$$

It follows from the continuity of $t\mapsto s_i(\cdot,t)$ in $\mathcal{A}_i$ equipped with W_i and from the uniform convergence of

$$(\boldsymbol{x},t)\mapsto\frac{\phi_\ell(\boldsymbol{x})-\phi_{\ell-1}(\boldsymbol{x})}{\varepsilon}\quad\text{if } t\in[(\ell-1)\varepsilon,\ell\varepsilon)$$

towards $\partial_t\phi$ that

$$A(\varepsilon)\underset{\varepsilon\to0}{\longrightarrow}-\iint_Q s_i\partial_t\phi\mathrm{d}\boldsymbol{x}\mathrm{d}t-\int_\Omega s_i^0\phi(\cdot,0)\mathrm{d}\boldsymbol{x}. \tag{83}$$

Combining (81)–(83) shows that the weak formulation (12) is fulfilled. □

Appendix A. A simple condition for the geodesic convexity of (Ω, d_i)

The goal of this appendix is to provide a simple condition on the permeability tensor in order to ensure that Condition (22) is fulfilled. For the sake of simplicity, we only consider here the case of isotropic permeability tensors

$$\mathbb{K}(\boldsymbol{x}) = \kappa(\boldsymbol{x})\mathbb{I}_d, \qquad \forall \boldsymbol{x} \in \overline{\Omega} \tag{84}$$

with $\kappa_\star \leq \kappa(\boldsymbol{x}) \leq \kappa^\star$ for all $\boldsymbol{x} \in \overline{\Omega}$. Let us stress that the condition we provide is not optimal.

As in the core of the paper, Ω denotes a convex open subset of $\mathbb{R}^d$ with C^2 boundary $\partial\Omega$. For $\overline{\boldsymbol{x}} \in \partial\Omega$, we denote by $\boldsymbol{n}(\overline{\boldsymbol{x}})$ the outward-pointing normal. Since $\partial\Omega$ is smooth, then there exists $\ell_0 > 0$ such that, for all $\boldsymbol{x} \in \Omega$ such that $\mathrm{dist}(\boldsymbol{x}, \partial\Omega) < \ell_0$, there exists a unique $\overline{\boldsymbol{x}} \in \partial\Omega$ such that $\mathrm{dist}(\boldsymbol{x}, \partial\Omega) = |\boldsymbol{x} - \overline{\boldsymbol{x}}|$ (here dist denotes the usual Euclidian distance between sets in $\mathbb{R}^d$). As a consequence, one can rewrite $\boldsymbol{x} = \overline{\boldsymbol{x}} - \ell\boldsymbol{n}(\overline{\boldsymbol{x}})$ for some $\ell \in (0, \ell_0)$.

In what follows, a function $f : \overline{\Omega} \to \mathbb{R}$ is said to be normally nondecreasing (resp. nonincreasing) on a neighborhood of $\partial\Omega$ if there exists $\ell_1 \in (0, \ell_0]$ such that $\ell \mapsto f(\overline{\boldsymbol{x}} - \ell\boldsymbol{n}(\overline{\boldsymbol{x}}))$ is nonincreasing (resp. nondecreasing) on $[0, \ell_1]$.

Proposition A.1. *Assume that:*

(i) the permeability field $\boldsymbol{x} \mapsto \kappa(\boldsymbol{x})$ is normally non-increasing in a neighborhood of $\partial\Omega$;

(ii) for all $\overline{\boldsymbol{x}} \in \partial\Omega$, either $\boldsymbol{\nabla}\kappa(\overline{\boldsymbol{x}})\cdot\boldsymbol{n}(\overline{\boldsymbol{x}}) < 0$, or $\boldsymbol{\nabla}\kappa(\overline{\boldsymbol{x}})\cdot\boldsymbol{n}(\overline{\boldsymbol{x}}) = 0$ and $D^2\kappa(\overline{\boldsymbol{x}})\boldsymbol{n}(\boldsymbol{x})\cdot\boldsymbol{n}(\boldsymbol{x}) = 0$.

Then there exists a C^2 extension $\widetilde{\kappa} : \mathbb{R}^d \to [\frac{\kappa_\star}{2}, \kappa^\star]$ of κ and a Riemannian metric

$$\widetilde{\delta}(\boldsymbol{x}, \boldsymbol{y}) = \inf_{\boldsymbol{\gamma} \in \widetilde{P}(\boldsymbol{x}, \boldsymbol{y})} \left(\int_0^1 \frac{1}{\widetilde{\kappa}(\boldsymbol{\gamma}(\tau))} |\boldsymbol{\gamma}'(\tau)|^2 \mathrm{d}\tau \right)^{1/2}, \qquad \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d \tag{85}$$

with $\widetilde{P}(\boldsymbol{x}, \boldsymbol{y}) = \left\{ \boldsymbol{\gamma} \in C^1([0,1]; \mathbb{R}^d) \,\middle|\, \boldsymbol{\gamma}(0) = \boldsymbol{x} \text{ and } \boldsymbol{\gamma}(1) = \boldsymbol{y} \right\}$, such that $(\Omega, \widetilde{\delta}_i)$ is geodesically convex.

Proof. Since Ω is convex, then for all $\boldsymbol{x} \in \mathbb{R}^d \setminus \Omega$, there exists a unique $\overline{\boldsymbol{x}} \in \partial\Omega$ such that $\mathrm{dist}(\boldsymbol{x}, \Omega) = |\boldsymbol{x} - \overline{\boldsymbol{x}}|$. Then one can extend κ in a C^2 way into the whole $\mathbb{R}^d$ by defining

$$\kappa(\boldsymbol{x}) = \kappa(\overline{\boldsymbol{x}}) + |\boldsymbol{x} - \overline{\boldsymbol{x}}| \boldsymbol{\nabla}\kappa(\overline{\boldsymbol{x}}) \cdot \boldsymbol{n}(\overline{\boldsymbol{x}}) + \frac{|\boldsymbol{x} - \overline{\boldsymbol{x}}|^2}{2} D^2\kappa(\overline{\boldsymbol{x}})\boldsymbol{n}(\overline{\boldsymbol{x}}) \cdot \boldsymbol{n}(\overline{\boldsymbol{x}}), \quad \forall \boldsymbol{x} \in \mathbb{R}^d \setminus \Omega.$$

Thanks to Assumptions (i) and (ii), the function $\ell \mapsto \kappa(\overline{\boldsymbol{x}} - \ell\boldsymbol{n}(\overline{\boldsymbol{x}}))$ is non-decreasing on $(-\infty, \ell_1]$ for all $\overline{\boldsymbol{x}} \in \partial\Omega$. Since $\partial\Omega$ is compact, there exists $\ell_2 > 0$ such that

$$\kappa(\overline{\boldsymbol{x}} - \ell\boldsymbol{n}(\overline{\boldsymbol{x}}) \geq \frac{\kappa_\star}{2}, \qquad \forall \ell \in (-\ell_2, 0].$$

Let $\rho : \mathbb{R}_+ \to \mathbb{R}$ be a non-decreasing C^2 function such that $\rho(0) = 1$, $\rho'(0) = \rho''(0) = 0$ and $\rho(\ell) = 0$ for all $\ell \geq \ell_2$. Then define

$$\widetilde{\kappa}(\boldsymbol{x}) = \rho(\mathrm{dist}(\boldsymbol{x}, \Omega))\kappa(\boldsymbol{x}) + (1 - \rho(\mathrm{dist}(\boldsymbol{x}, \Omega)))\frac{\kappa_\star}{2}, \qquad \forall \boldsymbol{x} \in \mathbb{R}^d,$$

so that the function $\ell \mapsto \widetilde{\kappa}(\overline{\boldsymbol{x}} - \ell\boldsymbol{n}(\overline{\boldsymbol{x}}))$ is non-increasing on $(-\infty, \ell_1)$ and bounded from below by $\frac{\kappa_\star}{2}$.

Let $\boldsymbol{x}, \boldsymbol{y} \in \Omega$, then there exists $\varepsilon > 0$ such that $\mathrm{dist}(\boldsymbol{x}, \partial\Omega) \geq \varepsilon$, $\mathrm{dist}(\boldsymbol{y}, \partial\Omega) \geq \varepsilon$, and κ is normally nonincreasing on $\partial\Omega_\varepsilon := \{\boldsymbol{x} \in \overline{\Omega} \mid \mathrm{dist}(\boldsymbol{x}, \partial\Omega) < \varepsilon\}$. A sufficient condition for $(\Omega, \widetilde{\delta})$ to be geodesic is that the geodesic $\boldsymbol{\gamma}^{\mathrm{opt}}_{\boldsymbol{x},\boldsymbol{y}}$ from $\boldsymbol{x}$ to $\boldsymbol{y}$ is such that

$$\mathrm{dist}\left(\boldsymbol{\gamma}^{\mathrm{opt}}_{\boldsymbol{x},\boldsymbol{y}}(t), \partial\Omega\right) \geq \varepsilon, \quad \forall t \in [0,1]. \tag{86}$$

In order to ease the reading, we denote by $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{\mathrm{opt}}_{\boldsymbol{x},\boldsymbol{y}}$ any geodesic such that

$$\widetilde{\delta}^2(\boldsymbol{x}, \boldsymbol{y}) = \int_0^1 \frac{1}{\widetilde{\kappa}(\boldsymbol{\gamma}(\tau))} |\boldsymbol{\gamma}'(\tau)|^2 \mathrm{d}\tau. \tag{87}$$

We define the continuous and piecewise C^1 path $\boldsymbol{\gamma}_\varepsilon$ from $\boldsymbol{x}$ to $\boldsymbol{y}$ by setting

$$\boldsymbol{\gamma}_\varepsilon(t) = \mathrm{proj}_{\overline{\Omega}_\varepsilon}(\boldsymbol{\gamma}(t)), \qquad \forall t \in [0,1], \tag{88}$$

where $\overline{\Omega}_\varepsilon := \{\boldsymbol{x} \in \Omega \mid \mathrm{dist}(\boldsymbol{x}, \partial\Omega) \geq \varepsilon\}$ is convex, and the orthogonal (w.r.t. the euclidian distance dist) projection $\mathrm{proj}_{\overline{\Omega}_\varepsilon}$ onto $\overline{\Omega}_\varepsilon$ is therefore uniquely defined.

Assume that Condition (86) is violated. Then by continuity there exists a non-empty interval $[a, b] \subset [0, 1]$ such that

$$\mathrm{dist}(\boldsymbol{\gamma}(t), \partial\Omega) < \varepsilon, \qquad \forall t \in (a, b),$$

the geodesic between $\gamma(a)$ and $\gamma(b)$ coincides with the part of the geodesics between $\boldsymbol{x}$ and $\boldsymbol{y}$. Then, changing $\boldsymbol{x}$ into $\boldsymbol{\gamma}(a)$ and $\boldsymbol{y}$ into $\boldsymbol{\gamma}(b)$, we can assume without loss of generality that

$$\mathrm{dist}(\boldsymbol{\gamma}(t), \partial\Omega) < \varepsilon, \qquad \forall t \in (0, 1).$$

It is easy to verify that

$$|\boldsymbol{\gamma}'_\varepsilon(t)| \leq |\boldsymbol{\gamma}'(t)|, \quad \forall t \in [0,1], \qquad \text{and} \quad |\boldsymbol{\gamma}'_\varepsilon(t)| < |\boldsymbol{\gamma}'(t)| \text{ on } (a,b) \tag{89}$$

for some non-empty interval $(a, b) \subset [0, 1]$. It follows from (85) that

$$\widetilde{\delta}^2(\boldsymbol{x}, \boldsymbol{y}) \leq \int_0^1 \frac{1}{\widetilde{\kappa}(\boldsymbol{\gamma}_\varepsilon(\tau))} |\boldsymbol{\gamma}'_\varepsilon(\tau)|^2 \mathrm{d}\tau.$$

Since κ is normally non-increasing, one has

$$\widetilde{\delta}^2(\boldsymbol{x}, \boldsymbol{y}) \leq \int_0^1 \frac{1}{\widetilde{\kappa}(\boldsymbol{\gamma}(\tau))} |\boldsymbol{\gamma}'_\varepsilon(\tau)|^2 \mathrm{d}\tau.$$

Thanks to (89), one obtains that

$$\widetilde{\delta}^2(\boldsymbol{x}, \boldsymbol{y}) < \int_0^1 \frac{1}{\widetilde{\kappa}(\boldsymbol{\gamma}(\tau))} |\boldsymbol{\gamma}'(\tau)|^2 \mathrm{d}\tau,$$

providing a contradiction with the optimality (87) of $\boldsymbol{\gamma}$. Thus Condition (86) holds, hence (Ω, δ) is a geodesic space. □

Appendix B. A multicomponent bathtub principle

The following theorem can be seen as a generalization of the classical scalar bathtub principle (see for instance [35, Theorem 1.14]). In what follows, N is a positive integer and Ω denotes an arbitrary measurable subset of $\mathbb{R}^d$.

Theorem B.1. *Let $\omega \in L^1_+(\Omega)$, and let $\boldsymbol{m} = (m_0, \dots, m_N) \in (\mathbb{R}^*_+)^{N+1}$ be such that $\sum_{i=0}^N m_i = \int_\Omega \omega \, \mathrm{d}\boldsymbol{x}$. We denote by*

$$\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}} = \left\{ \boldsymbol{s} = (s_0, \dots, s_N) \in L^1_+(\Omega)^{N+1} \,\middle|\, \int_\Omega s_i \mathrm{d}\boldsymbol{x} = m_i \text{ and } \sum_{i=0}^N s_i = \omega \text{ a.e. in } \Omega \right\}.$$

Then for any $\boldsymbol{F} = (F_0, \dots, F_N) \in (L^\infty(\Omega))^{N+1}$, *the functional*

$$\mathcal{F} : \boldsymbol{s} \mapsto \int_\Omega \boldsymbol{F} \cdot \boldsymbol{s} \, \mathrm{d}\boldsymbol{x}$$

has a minimizer in $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$. *Moreover, there exists* $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_N) \in \mathbb{R}^{N+1}$ *such that, denoting*

$$\lambda(\boldsymbol{x}) := \min_{0 \le j \le N} \{F_j(\boldsymbol{x}) + \alpha_i\}, \qquad \boldsymbol{x} \in \Omega,$$

any minimizer $\underline{\boldsymbol{s}} = (\underline{s}_0, \dots, \underline{s}_N)$ *satisfies*

$$F_i + \alpha_i = \lambda \qquad \mathrm{d}\underline{s}_i\text{-a.e. in } \Omega, \quad \forall i \in \{0, \dots, N\}.$$

One can think of this as: $\underline{s}_i = 0$ in $\{F_i + \alpha_i > \lambda\}$ and $F_i + \alpha_i \ge \lambda$ everywhere, i.e., $\underline{s}_i > 0$ can only occur in the "contact set" $\left\{ \boldsymbol{x} \,\middle|\, F_i(\boldsymbol{x}) + \alpha_i = \min_j (F_j(\boldsymbol{x}) + \alpha_j) \right\}$.

Proof. For the existence part, note that $\mathcal{F}$ is continuous for the weak L^1 convergence, and that $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ is weakly closed. Since $\sum s_i = \omega$ and $s_i \ge 0$ we have in particular $0 \le s_i \le \omega \in L^1$ for all i and $\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$. This implies that $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ is uniformly integrable, and since the mass $\|s_i\|_{L^1} = \int s_i = m_i$ is prescribed, the Dunford-Pettis theorem shows that $\boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ is L^1-weakly relatively compact. Hence from any minimizing sequence we can extract a weakly-L^1 converging subsequence, and by weak L^1 continuity the weak limit is a minimizer.

Let us now introduce a dual problem: for fixed $\boldsymbol{\alpha} = (\alpha_0, \dots, \alpha_N) \in \mathbb{R}^{N+1}$ we denote

$$\lambda_{\boldsymbol{\alpha}}(\boldsymbol{x}) := \min_i \{F_i(\boldsymbol{x}) + \alpha_i\} \tag{90}$$

and define

$$J(\boldsymbol{\alpha}) := \int_\Omega \lambda_{\boldsymbol{\alpha}}(\boldsymbol{x}) \omega(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \sum_{i=0}^N \alpha_i m_i.$$

We shall prove below that

(i) $\sup_{\boldsymbol{\alpha} \in \mathbb{R}^{N+1}} J(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^{N+1}} J(\boldsymbol{\alpha})$ is achieved,
(ii) $\min_{\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}} \mathcal{F}(\boldsymbol{s}) = \max_{\boldsymbol{\alpha} \in \mathbb{R}^{N+1}} J(\boldsymbol{\alpha})$.

The desired decomposition will then follow from equality conditions in (ii), and $\lambda(\boldsymbol{x}) = \lambda_{\overline{\boldsymbol{\alpha}}}(\boldsymbol{x})$ will be retrieved from any maximizer $\overline{\boldsymbol{\alpha}} \in \operatorname{Argmax} J$.

Remark B.2. *The above dual problem can be guessed by introducing suitable Lagrange multipliers* $\lambda(\boldsymbol{x}), \boldsymbol{\alpha}$ *for the total saturation and mass constraints, respectively, and writing the convex indicator of the constraints as a supremum over these multipliers. Formally exchanging* $\inf \sup = \sup \inf$ *and computing the optimality conditions in the right-most infimum relates* λ *to* $\boldsymbol{\alpha}$ *as in* (90)*, which in turn yields exactly the duality* $\inf_{\boldsymbol{s}} \mathcal{F} = \max_{\boldsymbol{\alpha}} J$. *See also Remark B.3*

Let us first establish property (i). For all $\boldsymbol{\alpha} \in \mathbb{R}^{N+1}$ and all $\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$, we first observe that

$$\begin{aligned} J(\boldsymbol{\alpha}) &= \int_\Omega \min_j \{F_j(\boldsymbol{x}) + \alpha_j\} \omega(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \sum_{i=0}^N \alpha_i m_i \\ &= \int_\Omega \min_j \{F_j(\boldsymbol{x}) + \alpha_j\} \sum_{i=0}^N s_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \sum_{i=0}^N \alpha_i \int_\Omega s_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \\ &= \sum_{i=0}^N \int_\Omega \left(\min_j \{F_j(\boldsymbol{x}) + \alpha_j\} - \alpha_i \right) s_i(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} \leq \int_\Omega \boldsymbol{F} \cdot \boldsymbol{s} \, \mathrm{d}\boldsymbol{x} = \mathcal{F}(\boldsymbol{s}). \end{aligned}$$

In particular J is bounded from above and

$$\sup_{\boldsymbol{\alpha} \in \mathbb{R}^{N+1}} J(\boldsymbol{\alpha}) \leq \min_{\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}} \mathcal{F}(\boldsymbol{s}). \tag{91}$$

Since $\int \omega \mathrm{d}\boldsymbol{x} = \sum m_i$, the function J is invariant under diagonal shifts, i.e., $J(\boldsymbol{\alpha} + c\mathbf{1}) = J(\boldsymbol{\alpha})$ for any constant $c \in \mathbb{R}$. As a consequence we can choose a maximizing sequence $\{\boldsymbol{\alpha}^k\}_{k \geq 1}$ such that $\min_j \alpha_j^k = 0$ for all $k \geq 0$. Let $j(k)$ be an index such that $\alpha_{j(k)}^k = \min_j \alpha_j^k = 0$. Then, since $\boldsymbol{\alpha}^k$ is maximizing and $\omega(\boldsymbol{x}) \geq 0$, we get, for k large enough,

$$\begin{aligned} \sup J - 1 \leq J(\boldsymbol{\alpha}^k) &= \int_\Omega \min_j \{F_j(\boldsymbol{x}) + \alpha_j^k\} \omega(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \sum \alpha_i^k m_i \\ &\leq \int_\Omega \big(F_{j(k)}(\boldsymbol{x}) + \underbrace{\alpha_{j(k)}^k}_{=0}\big) \omega(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \sum \alpha_i^k m_i \leq \|\boldsymbol{F}\|_{L^\infty} \|\omega\|_{L^1} - \sum \alpha_i^k m_i. \end{aligned}$$

Thus $\sum \alpha_i^k m_i \leq C$, and since $\alpha_i^k \geq 0$ and $m_i > 0$ we deduce that $\left(\boldsymbol{\alpha}^k\right)_k$ is bounded. Hence, up to extraction of an unlabelled subsequence, we can assume that $\boldsymbol{\alpha}^k$ converges towards some $\overline{\boldsymbol{\alpha}} \in \mathbb{R}_+^{N+1}$. The map J is continuous, hence $\overline{\boldsymbol{\alpha}}$ is a maximizer.

Let us now focus on property (ii). Note from (91) and (i) it suffices to prove the reverse inequality

$$\max_{\boldsymbol{\alpha} \in \mathbb{R}^{N+1}} J(\boldsymbol{\alpha}) \geq \min_{\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}} \mathcal{F}(\boldsymbol{s}).$$

We show below that, for any maximizer $\overline{\boldsymbol{\alpha}}$ of J, we can always construct a suitable $\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ such that $\mathcal{F}(\boldsymbol{s}) = J(\overline{\boldsymbol{\alpha}})$. This will immediately imply the reverse inequality and thus our claim (ii). In order to do so, we first observe that J is concave, thus the optimality condition at $\overline{\boldsymbol{\alpha}}$ can be written in terms of superdifferentials as $\mathbf{0}_{\mathbb{R}^{N+1}} \in \partial J(\overline{\boldsymbol{\alpha}})$. Denoting by

$$\Lambda(\boldsymbol{\alpha}) = \int_\Omega \lambda_{\boldsymbol{\alpha}} \omega \mathrm{d}\boldsymbol{x} = \int_\Omega \min_j \{F_j(x) + \alpha_j\} \omega(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$$

the first contribution in J, this optimality can be recast as

$$\boldsymbol{m} \in \partial \Lambda(\overline{\boldsymbol{\alpha}}). \tag{92}$$

For fixed $\boldsymbol{x} \in \Omega$ and by usual properties of the min function, the superdifferential $\partial \lambda_{\boldsymbol{\alpha}}(\boldsymbol{x})$ of the concave map $\boldsymbol{\alpha} \mapsto \lambda_{\boldsymbol{\alpha}}(\boldsymbol{x})$ at $\boldsymbol{\alpha} \in \mathbb{R}^{N+1}$ is characterized by

$$\partial \lambda_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \left\{ \boldsymbol{\theta} \in \mathbb{R}_+^{N+1} \;\middle|\; \sum_{i=0}^N \theta_i = 1, \text{ and } \theta_i = 0 \text{ if } F_i(\boldsymbol{x}) + \alpha_i > \lambda_{\boldsymbol{\alpha}}(\boldsymbol{x}) \right\}.$$

Therefore, it follows from the extension of the formula of differentiation under the integral to the non-smooth case (cf. [23, Theorem 2.7.2]) that

(93)
$$\partial\Lambda(\boldsymbol{\alpha}) = \left\{ \boldsymbol{w} \in \mathbb{R}_+^{N+1} \;\middle|\; \boldsymbol{w} = \int_\Omega \boldsymbol{\theta}(\boldsymbol{x})\omega(\boldsymbol{x})\mathrm{d}\boldsymbol{x} \text{ with } \boldsymbol{\theta}(\boldsymbol{x}) \in \partial\lambda_{\boldsymbol{\alpha}}(\boldsymbol{x}) \text{ a.e. in } \Omega \right\}.$$

The optimality criterion (92) at any maximizer $\overline{\boldsymbol{\alpha}}$ gives the existence of some function $\boldsymbol{\theta}$ as in (93) such that

$$m_i = \int_\Omega \theta_i(\boldsymbol{x})\omega(\boldsymbol{x})\mathrm{d}\boldsymbol{x}, \qquad \forall i \in \{0,\dots,N\}.$$

Defining

$$s_i(\boldsymbol{x}) := \theta_i(\boldsymbol{x})\omega(\boldsymbol{x}), \qquad \forall i \in \{0,\dots,N\}, \tag{94}$$

we have by construction that $s_i \geq 0$, $\int s_i = m_i$, and $\sum s_i = (\sum_i \theta_i)\omega = \omega$ a.e, thus $\boldsymbol{s} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$. Exploiting again $\sum s_i = \omega$ as well as the crucial property that $\theta_i = 0$ a.e. in $\{\boldsymbol{x} \mid \ F_i + \overline{\alpha}_i > \lambda_{\overline{\boldsymbol{\alpha}}}\}$, or in other words that $F_i + \overline{\alpha}_i = \lambda_{\overline{\boldsymbol{\alpha}}}$ for $\mathrm{d}s_i$-a.e $\boldsymbol{x} \in \Omega$, we get

$$\begin{aligned} J(\overline{\boldsymbol{\alpha}}) = \int_\Omega \lambda_{\overline{\boldsymbol{\alpha}}}\omega\mathrm{d}\boldsymbol{x} - \sum_{i=0}^N \overline{\alpha}_i m_i &= \sum_{i=0}^N \int_\Omega \lambda_{\overline{\boldsymbol{\alpha}}} s_i \mathrm{d}\boldsymbol{x} - \sum_{i=0}^N \overline{\alpha}_i m_i \\ &= \sum_{i=0}^N \int_\Omega (F_i + \overline{\alpha}_i) s_i \mathrm{d}\boldsymbol{x} - \sum_{i=0}^N \overline{\alpha}_i m_i = \mathcal{F}(\boldsymbol{s}) \end{aligned}$$

as claimed. Therefore $\boldsymbol{s}$ constructed by (94) is a minimizer of $\mathcal{F}$ and

$$J(\overline{\boldsymbol{\alpha}}) = \mathcal{F}(\underline{\boldsymbol{s}}). \tag{95}$$

In order to finally retrieve the desired decomposition, choose any minimizer $\underline{\boldsymbol{s}} \in \boldsymbol{\mathcal{X}} \cap \boldsymbol{\mathcal{A}}$ of $\mathcal{F}$ and any maximizer $\overline{\boldsymbol{\alpha}} \in \mathbb{R}^{N+1}$ of J. Then it follows from (95) that

$$0 = \mathcal{F}(\underline{\boldsymbol{s}}) - J(\overline{\boldsymbol{\alpha}}) = \sum_{i=0}^N \int_\Omega F_i \underline{s}_i \mathrm{d}\boldsymbol{x} - \int_\Omega \lambda_{\overline{\boldsymbol{\alpha}}}\, \omega \mathrm{d}\boldsymbol{x} + \sum_{i=0}^N \overline{\alpha}_i m_i.$$

Using once again that $\int \underline{s}_i = m_i$ and $\sum_i \underline{s}_i = \omega$, we get that

$$\sum_{i=0}^N \int_\Omega (F_i + \overline{\alpha}_i - \lambda_{\overline{\boldsymbol{\alpha}}})\, \underline{s}_i \mathrm{d}\boldsymbol{x} = 0.$$

By definition of $\lambda_{\overline{\boldsymbol{\alpha}}}$ the above integrand is nonnegative, hence $F_i + \overline{\alpha}_i = \lambda_{\overline{\boldsymbol{\alpha}}}$ a.e. in $\{\underline{s}_i > 0\}$. □

Remark B.3. *To understand the dual problem one chan think the function F_i as $N+1$ bathub that can be translated vertically. The translation of each bathtub is given by $\boldsymbol{\alpha}_i$. Once these translations are given one just wants to fill the bathubs starting from the bottom (that is $\lambda_{\boldsymbol{\alpha}}$), while satisfying the global saturation and mass constraints. For an optimal translation vector $\boldsymbol{\alpha}$, each phase i contributes at $\boldsymbol{x}$ with a ratio $\theta_i(\boldsymbol{x})$ as in (94).*

Acknowledgements. This project was supported by the ANR GEOPOR project (ANR-13-JS01-0007-01). CC also aknowledges the support of Labex CEMPI (ANR-11-LABX-0007-01). TOG was supported by the ANR ISOTACE project (ANR-12-MONU-0013). LM was supported by the Portuguese Science Fundation through FCT fellowship SFRH/BPD/88207/2012 and the UT Austin — Portugal CoLab project. Part of this work was carried out during the stay of CC and TOG at CAMGSD, Instituto Superior Técnico, Universidade de Lisboa. The authors whish to thank Quentin Mérigot for fruitful discussion.

References

[1] H. W. Alt and S. Luckhaus. Quasilinear elliptic-parabolic differential equations. *Math. Z.*, 183(3):311–341, 1983.

[2] B. Amaziane, M. Jurak, and A. Vrbaški. Existence for a global pressure formulation of water-gas flow in porous media. *Electron. J. Differential Equations*, 102:1–22, 2012.

[3] B. Amaziane, M. Jurak, and A. Žgaljić Keko. Modeling compositional compressible two-phase flow in porous media by the concept of the global pressure. *Comput. Geosci.*, 18(3-4):297–309, 2014.

[4] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, volume 2062 of *Lecture Notes in Math.*, pages 1–155. Springer, Heidelberg, 2013.

[5] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[6] L. Ambrosio, E. Mainini, and S. Serfaty. Gradient flow of the Chapman-Rubinstein-Schatzman model for signed vortices. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 28(2):217–246, 2011.

[7] L. Ambrosio and S. Serfaty. A gradient flow approach to an evolution problem arising in superconductivity. *Comm. Pure Appl. Math.*, 61(11):1495–1539, 2008.

[8] B. Andreianov, C. Cancès, and A. Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic-elliptic PDEs. HAL: hal-01142499, 2015.

[9] S. N. Antoncev and V. N. Monahov. Three-dimensional problems of transient two-phase filtration in inhomogeneous anisotropic porous media. *Dokl. Akad. Nauk SSSR*, 243(3):553–556, 1978.

[10] J. Bear and Y. Bachmat. *Introduction to modeling of transport phenomena in porous media*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[11] A. Blanchet. A gradient flow approach to the Keller-Segel systems. RIMS Kokyuroku's lecture notes, 2014.

[12] A. Blanchet, V. Calvez, and J. A. Carrillo. Convergence of the mass-transport steepest descent scheme for the subcritical Patlak-Keller-Segel model. *SIAM J. Numer. Anal.*, 46(2):691–721, 2008.

[13] F. Bolley, I. Gentil, and A. Guillin. Uniform convergence to equilibrium for granular media. *Arch. Ration. Mech. Anal.*, 208(2):429–445, 2013.

[14] C. Cancès and T. Gallouët. On the time continuity of entropy solutions. *J. Evol. Equ.*, 11(1):43–55, 2011.

[15] C. Cancès, T. O. Gallouët, and L. Monsaingeon. The gradient flow structure of immiscible incompressible two-phase flows in porous media. *C. R. Acad. Sci. Paris Sér. I Math.*, 353:985–989, 2015.

[16] G. Carlier and M. Laborde. On systems of continuity equations with nonlinear diffusion and nonlocal drifts. *arXiv preprint*, arXiv:1505.01304, 2015.

[17] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, and D. Slepčev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.*, 156(2):229–271, 2011.

[18] G. Chavent. A new formulation of diphasic incompressible flows in porous media. In *Applications of methods of functional analysis to problems in mechanics (Joint Sympos., IUTAM/IMU, Marseille, 1975)*, pages 258–270. Lecture Notes in Math., 503. Springer, Berlin, 1976.

[19] G. Chavent. A fully equivalent global pressure formulation for three-phases compressible flows. *Appl. Anal.*, 88(10-11):1527–1541, 2009.
[20] G. Chavent and J. Jaffré. *Mathematical Models and Finite Elements for Reservoir Simulation*, volume 17. North-Holland, Amsterdam, stud. math. appl. edition, 1986.
[21] G. Chavent and G. Salzano. Un algorithme pour la détermination de perméabilités relatives triphasiques satisfaisant une condition de différentielle totale. Technical Report 355, INRIA, 1985.
[22] Z. Chen. Degenerate two-phase incompressible flow. I. Existence, uniqueness and regularity of a weak solution. *J. Differential Equations*, 171(2):203–232, 2001.
[23] F. H. Clarke *Optimization and nonsmooth analysis*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 1990.
[24] E. De Giorgi. New problems on minimizing movements. In *Boundary value problems for partial differential equations and applications*, volume 29 of *RMA Res. Notes Appl. Math.*, pages 81–98. Masson, Paris, 1993.
[25] J. Dolbeault, B. Nazaret, and G. Savaré. A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009.
[26] P. Fabrie and M. Saad. Existence de solutions faibles pour un modèle d'écoulement triphasique en milieu poreux. *Ann. Fac. Sci. Toulouse Math. (6)*, 2(3):337–373, 1993.
[27] G. Gagneux and M. Madaune-Tort. *Analyse mathématique de modèles non linéaires de l'ingénierie pétrolière*, volume 22 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer-Verlag, Berlin, 1996. With a preface by Charles-Michel Marle.
[28] N. Gigli and F. Otto. Entropic Burgersd5 equation via a minimizing movement scheme based on the Wasserstein metric. *Calc. Var. Partial Differential Equations*, 47(1-2):181–206, 2013.
[29] H. Hance-Olsen and H. Holden. The Kolmogorov-Riesz compactness theorem. *Expo. Math.*, 28:385–394, 2010.
[30] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.
[31] D. Kinderlehrer, L. Monsaingeon, and X. Xu. A Wasserstein gradient flow approach to Poisson-Nernst-Planck equations. *ESAIM: Control Optim. Calc. Var.*, 23(1):137–164, 2017.
[32] M. Laborde. Systèmes de particules en interaction, approche par flot de gradient dans l'espace de Wasserstein. *Ph.D thesis*, Université Paris-Dauphine, 2016.
[33] O. A. Ladyženskaja, V. A. Solonnikov, and N. N. Ural′ceva. *Linear and quasilinear equations of parabolic type*. Translated from the Russian by S. Smith. Translations of Mathematical Monographs, Vol. 23. American Mathematical Society, Providence, R.I., 1968.
[34] P. Laurençot and B.-V. 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.
[35] E. H. Lieb and M. Loss. *Analysis*. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, second edition, 2001.
[36] S. Lisini. Nonlinear diffusion equations with variable coefficients as gradient flows in Wasserstein spaces. *ESAIM: Control Optim. Calc. Var.*, 15(3):712–740, 2009.
[37] S. Lisini, D. Matthes, and G. Savaré. Cahn-Hilliard and thin film equations with nonlinear mobility as gradient flows in weighted-Wasserstein metrics. *J. Differential Equations*, 253(2):814–850, 2012.
[38] D. Matthes, R. J. McCann, and G. Savaré. A family of nonlinear fourth order equations of gradient flow type. *Comm. Partial Differential Equations*, 34(11):1352–1397, 2009.
[39] B. Maury, A. Roudneff-Chupin, and F. Santambrogio. A macroscopic crowd motion model of gradient flow type. *Math. Models Methods Appl. Sci.*, 20(10):1787–1821, 2010.
[40] A. Moussa. Some variants of the classical Aubin-Lions Lemma. *J. Evol. Equ.*, 16(1):65–93, 2016.
[41] F. Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory. *Arch. Rational Mech. Anal.*, 141(1):63–103, 1998.
[42] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.
[43] E. Sandier and S. Serfaty. Gamma-convergence of gradient flows with applications to Ginzburg-Landau. *Comm. Pure Appl. Math.*, 57(12):1627–1672, 2004.
[44] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications 87. Birkhäuser Verlag, Basel, 1 edition, 2015.

[45] S. Simons. *Minimax and monotonicity*, volume 1693 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1998.
[46] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.
[47] J. Zinsl and D. Matthes. Exponential convergence to equilibrium in a coupled gradient flow system modeling chemotaxis. *Anal. PDE*, 8(2):425–466, 2015.
[48] J. Zinsl and D. Matthes. Transport distances and geodesic convexity for systems of degenerate diffusion equations *Calc. Var. Partial Differential Equations*, 54(4):3397–2438, 2015.
[49] J. Zinsl Existence of solutions for a nonlinear system of parabolic equations with gradient flow structure *Monatsh. Math.*, 174(4):653–679, 2014.

Clément Cancès (clement.cances@inria.fr). Team RAPSODI, Inria Lille – Nord Europe, 40 av. Halley, F-59650 Villeneuve d'Ascq, France.

Thomas O. Gallouët (thomas.gallouet@ulg.ac.be). Département de mathématiques, Université de Liège, Allée de la découverte 12, B-4000 Liège, Belgique.

Léonard Monsaingeon (leonard.monsaingeon@univ-lorraine.fr). Institut Élie Cartan de Lorraine, Université de Lorraine, Site de Nancy B.P. 70239, F-54506 Vandoeuvre-lèss-Nancy Cedex

SIMULATION OF MULTIPHASE POROUS MEDIA FLOWS WITH MINIMIZING MOVEMENT AND FINITE VOLUME SCHEMES

C. CANCÈS, T. O. GALLOUET, M. LABORDE, AND L. MONSAINGEON

Abstract. The Wasserstein gradient flow structure of the PDE system governing multiphase flows in porous media was recently highlighted in [C. Cancès, T. O. Gallouët, and L. Monsaingeon, *Anal. PDE* 10(8):1845–1876, 2017]. The model can thus be approximated by means of the minimizing movement (or JKO) scheme, that we solve thanks to the ALG2-JKO scheme proposed in [J.-D. Benamou, G. Carlier, and M. Laborde, *ESAIM Proc. Surveys*, 57:1–17, 2016]. The numerical results are compared to a classical upstream mobility Finite Volume scheme, for which strong stability properties can be established.

Keywords. Multiphase porous media flows; Wasserstein gradient flow; minimizing movement scheme; Augmented Lagrangian method; Finite Volumes

AMS subjects classification. 35K65, 35A15, 49M29, 65M08, 76S05

Contents

1. Multiphase porous media flows as Wasserstein gradient flow

Because of their wide range of interest in the applications, multiphase flows in porous media have been the object of countless scientific studies. In particular, there has been an extensive effort in order to develop reliable and efficient tools for the simulation of such flows. In many practical situations, the characteristic size of the pores (typically of the order the μm for regular sandstones) is much smaller than the characteristic size of the domain of interest. The direct numerical simulation of fluid flows at the pore scale is therefore not tractable. The use of homogenized models of Darcy type is therefore commonly used to simulate porous media flows.

The derivation of such models is the purpose of a very extended literature. We refer for instance to [3] for an extended introduction to the modeling of porous media flows. But let us stress that, as far as we know, there is no rigorous mathematical derivation of homogenized models for multiphase porous media flows.

Because of the very large friction of the fluid with the porous matrix, the energy is dissipated and inertia is often naturally neglected in the Darcy type models. The resulting models therefore have a formal gradient flow structure, as highlighted in [10] for immiscible incompressible multiphase porous media flows. This was then rigorously established in [11] that the equations governing such flows can be reinterpreted as a gradient flow in some appropriate Wasserstein space. The goal of this paper is to explore how this new point of view can be used to simulate multiphase flows in porous media.

1.1. **Incompressible immiscible multiphase flows.** As a first step, let us recall the equations governing multiphase porous media flows. We remain synthetic here and refer to the monograph [3] for a rather complete presentation of the models. The porous medium is represented by a convex bounded open subset Ω of $\mathbb{R}^d$ ($d \leq 3$). Within this porous medium, $N+1$ phases are flowing. Denoting by $\boldsymbol{s} = (s_0, \dots, s_N)$ the saturations, i.e., the volume ratios of the various phases in the fluid, the following total saturation relations has to be fulfilled:

$$s_0 + s_1 + \cdots + s_N = 1. \tag{1a}$$

In what follows, we denote by

$$\boldsymbol{\Delta} = \left\{ \boldsymbol{s} \in \mathbb{R}_+^N \;\middle|\; s_0 + s_1 + \cdots + s_N = 1 \right\},$$

and by

$$\boldsymbol{\mathcal{X}} = \left\{ \boldsymbol{s} : \Omega \to \mathbb{R}^N \;\middle|\; \boldsymbol{s}(\boldsymbol{x}) \in \boldsymbol{\Delta} \text{ for a.e. } \boldsymbol{x} \in \Omega \right\}.$$

As a consequence of (1a), the composition of the fluid is fully characterized by the knowledge of

$$\boldsymbol{s}^* = (s_1, \dots, s_N) \in \boldsymbol{\Delta}^* = \left\{ (s_1, \dots, s_N) \in \mathbb{R}_+^N \;\middle|\; \sum_{i=1}^N s_i \leq 1 \right\}.$$

Concerning the evolution, each phase is convected with its own speed

$$\omega \partial_t s_i + \boldsymbol{\nabla} \cdot (s_i \boldsymbol{v}_i) = 0, \tag{1b}$$

where ω stands for the porosity of the medium Ω and is assumed to be constant in the sequel for simplicity. Then a straightforward rescaling in time allows to choose $\omega = 1$. We further assume a no flux condition across the boundary $\partial\Omega$ for each phase, hence the mass is conserved along time. This motivates the introduction of the set

$$\boldsymbol{\mathcal{A}} = \left\{ \boldsymbol{s} \in L^1_+(\Omega)^N \;\middle|\; \int_\Omega s_i \mathrm{d}\boldsymbol{x} = \int_\Omega s_i^0 \mathrm{d}\boldsymbol{x} \right\},$$

where $\boldsymbol{s}^0 = \left(s_i^0\right) : \Omega \to \boldsymbol{\Delta}$ is a prescribed initial data.

The phase speeds $\boldsymbol{v}_i$ are prescribed by the Darcy law [14]

$$\boldsymbol{v}_i = -\frac{\kappa}{\mu_i} \left(\boldsymbol{\nabla} p_i - \rho_i \mathbf{g} \right), \qquad i \in \{0, \dots, N\}. \tag{1c}$$

In (1c), κ denotes the permeability of the porous medium. For simplicity, it is assumed to be constant and positive. We refer to [11] for the case of space-dependent anisotropic permeability tensors. The fluid viscosity and density are denoted by $\mu_i > 0$ and $\rho_i \geq 0$, respectively, whereas $\mathbf{g} = -g\boldsymbol{e}_z$ denotes the gravity. The unknown phase pressures $\boldsymbol{p} = (p_i)_{0 \leq i \leq N}$ are related to the saturations by N capillary pressure relations

$$p_i - p_0 = \pi_i(\boldsymbol{s}^*), \qquad \forall i \in \{1, \dots, N\}. \tag{1d}$$

The capillary pressure functions $\boldsymbol{\pi} = (\pi_i)_{1\le i\le N}$ are assumed to derive from a strictly convex and ϖ-concave potential $\Pi : \boldsymbol{\Delta}^* \to \mathbb{R}_+$ for some $\varpi > 0$, i.e.,

$$0 < \Pi(\widehat{\boldsymbol{s}}^*) - \Pi(\boldsymbol{s}^*) - \boldsymbol{\pi}(\boldsymbol{s}^*)\cdot(\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*) \le \frac{\varpi}{2}|\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*|^2, \qquad \forall \boldsymbol{s}^*, \widehat{\boldsymbol{s}}^* \in \boldsymbol{\Delta}^*, \text{ with } \boldsymbol{s}^* \neq \widehat{\boldsymbol{s}}^*. \tag{2}$$

This implies that $\boldsymbol{\pi} : \boldsymbol{\Delta}^* \to \mathbb{R}^N$ is strictly monotone (thus one-to-one) and Lipschitz continuous:

$$0 < \big(\boldsymbol{\pi}(\widehat{\boldsymbol{s}}^*) - \boldsymbol{\pi}(\boldsymbol{s}^*)\big)\cdot\big(\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*\big) \le \varpi|\widehat{\boldsymbol{s}}^* - \boldsymbol{s}^*|^2, \qquad \forall \boldsymbol{s}^*, \widehat{\boldsymbol{s}}^* \in \boldsymbol{\Delta}^*, \text{ with } \boldsymbol{s}^* \neq \widehat{\boldsymbol{s}}^*,$$

and thus

$$0 \le D^2\Pi(\boldsymbol{s}^*) \le \varpi \mathbf{I}_N, \qquad \forall \boldsymbol{s}^* \in \boldsymbol{\Delta}^*.$$

The last inequalities have to be understood in the sense of the symmetric matrices. The function Π is extended by $+\infty$ outside of $\boldsymbol{\Delta}^*$.

As established in [11], the problem (1) can be interpreted as the Wasserstein gradient flow of the energy

$$\mathcal{E}(\boldsymbol{s}) = \int_\Omega \left[\Pi(\boldsymbol{s}^*) + \boldsymbol{s}\cdot\boldsymbol{\Psi} + \chi_{\boldsymbol{\Delta}}(\boldsymbol{s})\right] \mathrm{d}\boldsymbol{x}, \qquad \forall \boldsymbol{s} \in \boldsymbol{\mathcal{A}}. \tag{3}$$

In formula (3), the exterior gravitational potential $\boldsymbol{\Psi} = (\Psi_i)_{0\le i\le N}$ is given by

$$\Psi_i(\boldsymbol{x}) = -\rho_i \mathbf{g}\cdot\boldsymbol{x}, \qquad \forall \boldsymbol{x} \in \Omega. \tag{4}$$

Remark 1.1. *In fact in* (3) *one can consider a large class of arbitrary potential* $\boldsymbol{\Psi}$*, see* [11] *for details.*

The constraint (1a) is incorporated in the energy rather than in the geometry thanks to the term

$$\chi_{\boldsymbol{\Delta}}(\boldsymbol{s}) = \begin{cases} 0 & \text{if } \boldsymbol{s} \in \boldsymbol{\Delta}, \\ +\infty & \text{otherwise.} \end{cases}$$

We refer to [8, 5, 28] for a presentation of the multiphase optimal transportation problem for which the constraint (1a) is directly incorporated in the geometry. In order to be more precise in our statements, we need to introduce some extra material concerning the Wasserstein distance to be used to equip $\boldsymbol{\mathcal{A}}$. This is the purpose of the next section.

Remark 1.2. *In the previous work* [11]*, the uniform convexity of the capillary potential* Π *was required. In* (2)*, we relax this assumption into a mere strict convexity requirement. This can be done by slightly adapting the proofs of* [11].

1.2. **Wasserstein distance.** For $i \in \{0, \dots, N\}$ we define

$$\mathcal{A}_i = \left\{ s_i \in L^1(\Omega; \mathbb{R}_+) \;\middle|\; \int_\Omega s_i \mathrm{d}\boldsymbol{x} = m_i \right\}.$$

Given $s_i, \widehat{s}_i \in \mathcal{A}_i$, the set of admissible transport plans between s_i and $\widehat{s}_i$ is given by

$$\Gamma_i(s_i, \widehat{s}_i) = \left\{ \gamma_i \in \mathcal{M}_+(\Omega\times\Omega) \;\middle|\; \gamma_i(\Omega\times\Omega) = m_i,\ \gamma_i^{(1)} = s_i \text{ and } \gamma_i^{(2)} = \widehat{s}_i \right\},$$

where $\mathcal{M}_+(\Omega\times\Omega)$ stands for the set of Borel measures on $\Omega\times\Omega$ and $\gamma_i^{(k)}$ is the k^{th} marginal of the measure γ_i. The quadratic Wasserstein distance W_i on $\mathcal{A}_i$ is then defined as

$$W_i^2(s_i, \widehat{s}_i) = \min_{\gamma_i \in \Gamma(s_i, \widehat{s}_i)} \iint_{\Omega\times\Omega} \frac{\mu_i}{\kappa} |\boldsymbol{x} - \boldsymbol{y}|^2 \mathrm{d}\gamma_i(\boldsymbol{x}, \boldsymbol{y}). \tag{5}$$

Equivalently, the continuity equation (1b) allows to give the following dynamical characterization:

Proposition 1.3 (Benamou-Brenier formula [4]). *For $s_{i,0}, s_{i,1} \in \mathcal{A}_i$ we have*

$$W_i^2(s_{i,0}, s_{i,1}) = \min_{s_i, \boldsymbol{v}} \int_0^1 \int_\Omega \frac{\mu_i}{\kappa} |\boldsymbol{v}_t(\boldsymbol{x})|^2 \mathrm{d}s_{i,t}(\boldsymbol{x})\mathrm{d}t, \tag{6}$$

where the minimum runs over curves of measures $t \mapsto s_{i,t} \in \mathcal{A}_i$ with endpoints $s_{i,0}, s_{i,1}$ and velocity fields $t \mapsto \boldsymbol{v}_t \in \mathcal{M}^d(\Omega)$ such that

$$\partial_t s_{i,t} + \boldsymbol{\nabla} \cdot (s_{i,t} \boldsymbol{v}_t) = 0$$

in the sense of distributions.

Remark 1.4. *As originally developped in [4], the right variables to be used in the Benamou-Brenier formula (6) is not the velocity $\boldsymbol{v}$, but in fact the momentum $\boldsymbol{m} = s\boldsymbol{v}$, since the action $A(s, \boldsymbol{m}) = \frac{|\boldsymbol{m}|^2}{s} = s|\boldsymbol{v}|^2$ becomes then jointly convex in both arguments.*

A third equivalent formulation is the Kantorovich dual problem:

Proposition 1.5. *There holds*

$$\frac{1}{2} W_i^2(s_i, \widehat{s}_i) = \max_{\phi, \psi} \left\{ \int \phi(\boldsymbol{x}) \mathrm{d}s_i(\boldsymbol{x}) + \int \psi(\boldsymbol{y}) \mathrm{d}\widehat{s}_i(\boldsymbol{y}) \right\}, \tag{7}$$

where the maximum runs over all pairs $(\phi, \psi) \in L^1(\mathrm{d}s_i) \times L^1(\mathrm{d}\widehat{s}_i)$ such that $\phi(\boldsymbol{x}) + \psi(\boldsymbol{y}) \leq \frac{\mu_i}{2\kappa} |\boldsymbol{x} - \boldsymbol{y}|^2$. Any maximizer is called a (pair of optimal) Kantorovich potential.

The viscosity μ_i and permeability κ appear in (5)–(7) as scaling factors in the cost function $\mu_i |\boldsymbol{x} - \boldsymbol{y}|^2 / \kappa$, and this is required for consistency with Darcy's law (1c). For more general heterogeneous permeability tensors $\mathbb{K}(\boldsymbol{x})$ one could use instead the intrinsic distance $d_i^2(\boldsymbol{x}, \boldsymbol{y})$ induced on Ω by the Riemannian tensor $\mu_i \mathbb{K}^{-1}(\boldsymbol{x})$, see [27] for a general approach of Wasserstein distances with variable coefficients and [11] in the particular context of multiphase flows in porous media.

With the phase Wasserstein distances $(W_i)_{0 \leq i \leq N}$ at hand, we can define the global Wasserstein distance $\boldsymbol{W}$ on $\boldsymbol{\mathcal{A}} := \mathcal{A}_0 \times \cdots \times \mathcal{A}_N$ by setting

$$\boldsymbol{W}(\boldsymbol{s}, \widehat{\boldsymbol{s}}) = \left(\sum_{i=0}^N W_i(s_i, \widehat{s}_i)^2 \right)^{1/2}, \quad \forall \boldsymbol{s}, \widehat{\boldsymbol{s}} \in \boldsymbol{\mathcal{A}}.$$

1.3. **Approximation by minimization scheme.** As already mentioned, the problem (1) is the Wasserstein gradient flow of our singular energy (3), see our earlier works [10, 11]. Rather than discussing the meaning of gradient flows in the Wasserstein setting, we refer to the monograph [2] for an exposition of gradient flows in abstract metric spaces [2] and to [35, 36] for a detailed overview. As is now well understood from the work of Jordan, Kinderlehrer, and Otto [25], one possible way to formalize this gradient flow structure is to implement the JKO scheme (also referred to as DeGiorgi's minimizing movement, see [15]). Given an initial datum $\boldsymbol{s}^0 \in \boldsymbol{\mathcal{A}}$ with energy $\mathcal{E}(\boldsymbol{s}^0) < \infty$ and a time step $\tau > 0$, the strategy consists in:

(i) construct a time discretization $\boldsymbol{s}^n(.) \approx \boldsymbol{s}(n\tau, .)$ by solving recursively

$$\boldsymbol{s}^{n+1} = \operatorname*{argmin}_{\boldsymbol{s} \in \boldsymbol{\mathcal{A}}} \left\{ \frac{1}{2\tau} \boldsymbol{W}^2(\boldsymbol{s}, \boldsymbol{s}^n) + \mathcal{E}(\boldsymbol{s}) \right\}; \tag{8}$$

(ii) define the piecewise-constant interpolation

$$\boldsymbol{s}_\tau(t) := \boldsymbol{s}^{n+1} \qquad \text{if } t \in (n\tau, (n+1)\tau];$$

(iii) retrieve a continuous solution $\boldsymbol{s}(t) = \lim_{\tau \to 0} \boldsymbol{s}_\tau(t)$ in the limit of small time steps.

This is a variant in the Wasserstein space of the implicit variational Euler scheme: indeed, in Euclidean spaces $x \in \mathbb{R}^d$ and for smooth functions $E : \mathbb{R}^d \to \mathbb{R}$, the Euler-Lagrange equation corresponding to minimizing $x \mapsto \frac{1}{2\tau}|x - x^n|^2 + E(x)$ is nothing but the finite difference approximation $\frac{x^{n+1}-x^n}{\tau} = -\nabla E(x^{n+1})$. We refrain from giving more details at this stage and refer again to [2, 35, 37].

Due to lower semi-continuity and convexity, it is easy to prove that the minimization problem (8) is well-posed, hence the discrete solution $\boldsymbol{s}_\tau$ is uniquely and unambiguously defined. But we still need to construct approximate phase pressures $\boldsymbol{p}_\tau = (p_{1,\tau}, p_{2,\tau})$. Their construction makes use of the backward Kantorovich potentials (see [11, Section 3]).

Lemma 1.6. *There exist pressures p_i^{n+1} and Kantorovich potentials ϕ_i^{n+1} (from s_i^{n+1} to s_i^n) such that*

$$\frac{\phi_i^{n+1}}{\tau} = p_i^{n+1} + \Psi_i \quad \text{a.e. in } \{s_i^{n+1} > 0\} \quad \text{for } i = 0, \dots, N, \tag{9}$$

and

$$p_i^{n+1} - p_0^{n+1} = \pi_i(\boldsymbol{s}^{n+1,*}) \quad \text{a.e. in } \Omega \quad \text{for } i = 1, \dots, N. \tag{10}$$

From classical optimal transport theory [35], $\boldsymbol{v}_i^{n+1} := \frac{\kappa}{\mu_i}\frac{\nabla\phi_i^{n+1}}{\tau}$ should be interpreted as the discrete velocity driving the i-th phase, which will automatically give $\partial_t s_i + \boldsymbol{\nabla}\cdot(s_i\boldsymbol{v}_i) = 0$ in the limit $\tau \to 0$. Hence (9) is a discrete counterpart of Darcy law (1c). The capillary relation (1d) hold as well at the discrete level thanks to relations (10), whereas the total saturation constraint (1a) is automatically enforced in (8) thanks to $\mathcal{E}(\boldsymbol{s}^{n+1}) < \infty$. For the sake of brevity we omit the details and refer again to [11].

1.4. **Main properties of the approximation.** Since our system (1) of PDEs is highly nonlinear, taking the limit $\boldsymbol{s}(t) = \lim_{\tau\to 0} \boldsymbol{s}_\tau(t)$ will require sufficient compactness both in time and space. In this section we sketch the main arguments leading to such compactness.

Compactness in time is derived from the classical *total square distance estimate* below, which is a characteristic feature of any JKO variational discretization. Testing $\boldsymbol{s} = \boldsymbol{s}^n$ as a competitor in (8) gives first

$$\frac{1}{2\tau}\boldsymbol{W}^2(\boldsymbol{s}^{n+1}, \boldsymbol{s}^n) + \mathcal{E}(\boldsymbol{s}^{n+1}) \le \mathcal{E}(\boldsymbol{s}^n).$$

This implies of course the energy monotonicity $\mathcal{E}(\boldsymbol{s}^{n+1}) \le \mathcal{E}(\boldsymbol{s}^n)$, but summing over n, we also get the *total square distance estimate* in the form

$$\frac{1}{\tau}\sum_{n\ge 0} \boldsymbol{W}^2(\boldsymbol{s}^{n+1}, \boldsymbol{s}^n) \le 2\left(\mathcal{E}(\boldsymbol{s}^0) - \inf_{\mathcal{A}} \mathcal{E}\right). \tag{11}$$

By definition of the piecewise-constant interpolation, an easy application of the Cauchy-Schwarz inequality gives then the *approximate equicontinuity*

$$\boldsymbol{W}(\boldsymbol{s}_\tau(t_1), \boldsymbol{s}_\tau(t_2)) \le C|t_2 - t_1 + \tau|^{\frac{1}{2}}, \qquad \forall\, 0 \le t_1 \le t_2,$$

uniformly in τ, which yields the desired compactness in time (see [2, Proposition 3.10] or [18, Theorem C.10]).

Compactness in space will be obtained exploiting the *flow interchange technique* from [29]. Roughly speaking, this amounts to estimating the dissipation of the driving functional $\mathcal{E}$ along a well-behaved auxiliary gradient flow, driven by an auxiliary functional and starting from the

minimizer $\boldsymbol{s}^{n+1}$. More explicitly, we define the ϵ-perturbation $\tilde{\boldsymbol{s}}_\epsilon = (\tilde{s}_{0,\epsilon}, \dots, \tilde{s}_{N,\epsilon})$ as solutions to the independent heat equations

$$\begin{cases} \frac{\partial \tilde{s}_{i,\epsilon}}{\partial \epsilon} = \kappa \Delta \tilde{s}_{i,\epsilon} & \text{for small } \epsilon > 0, \\ \tilde{s}_i|_{\epsilon=0} = s_i^{n+1}. \end{cases}$$

The key observation is that, for each $i = 0 \dots N$, the above heat equation is a gradient flow in the Wasserstein space $(\mathcal{A}_i, W_i)$ with driving functional $\mu_i \mathcal{H}$, where the Boltzmann entropy

$$\mathcal{H}(s) = \int_\Omega s(\boldsymbol{x}) \log\left(s(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x}. \tag{12}$$

In addition to the usual regularizing effects, this heat equation is particularly well-behaved here in the sense that it preserves the total saturation constraint $\sum_{i=0}^N \tilde{s}_{i,\epsilon} = \sum_{i=0}^N s_i^{n+1} = 1$ and, since Ω is convex, the auxiliary driving functional $\mathcal{H}$ is displacement convex in $(\mathcal{A}_i, W_i)$ [37, 31]. If

$$\mathcal{F}_\tau^n(\boldsymbol{s}) = \frac{1}{2\tau} \boldsymbol{W}^2(\boldsymbol{s}, \boldsymbol{s}^n) + \mathcal{E}(\boldsymbol{s})$$

denotes the JKO functional, then by optimality of the minimizer $\boldsymbol{s}^{n+1}$ in (8) we must have

$$\limsup_{\epsilon \to 0^+} \frac{d}{d\epsilon} \mathcal{F}_\tau^n(\tilde{\boldsymbol{s}}_\epsilon) \geq 0.$$

The energy term $\mathcal{E}(\tilde{\boldsymbol{s}}_\epsilon) = \int_\Omega \Pi(\tilde{\boldsymbol{s}}_\epsilon^*)$ can easily be differentiated under the integral sign (with respect to ϵ), while the variation of the first $\boldsymbol{W}^2(\tilde{\boldsymbol{s}}_\epsilon, \boldsymbol{s}^n)$ term can be estimated using the *evolution variational inequality* [2] for the well-behaved $\mathcal{H}$-flow $\tilde{\boldsymbol{s}}_\epsilon$ (this metric characterization precisely requires some displacement convexity of the auxiliary flow, see [19, Theorem 2.23]). Omitting again the details, one gets in the end the dissipation estimate

$$\tau \sum_{i=0}^N \|\nabla \pi_i((\boldsymbol{s}^{n+1})^*)\|_{L^2(\Omega)} \leq C \left(\tau + \boldsymbol{W}^2(\boldsymbol{s}^{n+1}, \boldsymbol{s}^n) + \sum_{i=0}^N \mathcal{H}(s_i^n) - \mathcal{H}(s_i^{n+1}) \right),$$

see [11, Section 2.2] for the details. Exploiting the previous total square distance estimate and summing over $n = 0 \dots \lfloor T/\tau \rfloor$ (or equivalently integrating in time), we control next

$$\|\boldsymbol{\pi}(\boldsymbol{s}_\tau^*)\|_{L^2(0,T;H^1)} \leq C \left(T + \sum_{i=0}^N \mathcal{H}(s_i^0) + 1 \right) = C_T \tag{13}$$

for arbitrary $T > 0$ and fixed initial datum $\boldsymbol{s}^0$. It is worth recalling at this stage that, due to our assumption (2), $\boldsymbol{\pi}(\boldsymbol{s}^*) = \nabla_{\boldsymbol{s}^*} \Pi(\boldsymbol{s}^*)$ is a strictly monotone thus invertible map of $\boldsymbol{s}^*$ due to the strict convexity of Π. The compactness w.r.t. the space variable of $(\boldsymbol{s}_\tau)_{\tau>0}$ then follows from (13).

Remark 1.7. *A formal but more PDE-oriented explanation of the above flow-interchange simply consists in taking* $\log(s_i)$ *as a test function in the weak formulation of system* (1)*. The delicate technical part is to justify this computation and mimic this formal chain rule in the discrete time setting in order to retrieve enhanced regularity of the JKO minimizers.*

Exploiting the above compactness, one can argue as in [11] and finally prove the following convergence results. The existence of a weak solution to the problem (1) is a direct byproduct.

Theorem 1.8. *For any discrete sequence* $\tau_k \to 0$ *and up extraction of a subsequence if needed, we have convergence*

$$\begin{array}{rl} \boldsymbol{s}_{\tau_k} \to \boldsymbol{s} & \text{strongly in all } L^q((0,T)\times\Omega), \\ \boldsymbol{\pi}(\boldsymbol{s}_{\tau_k}^*) \rightharpoonup \boldsymbol{\pi}(\boldsymbol{s}^*) & \text{weakly in } L^2(0,T;H^1(\Omega)), \\ \boldsymbol{p}_\tau \rightharpoonup \boldsymbol{p} & \text{weakly in } L^2(0,T;H^1(\Omega)), \end{array}$$

and the limit $(\boldsymbol{s}, \boldsymbol{p})$ *is a weak solution of* (1).

2. Numerical approximation of the flow

We present here the ALG2-JKO scheme and the upstream mobility finite volume scheme. The first method is based on the variational JKO scheme (8) described in subsection 1.3 whereas the second method is based on the PDE formulation of the problem (1) given by (1a)-(1b)-(1c)-(1d). Both methods are well adapted for gradient flows equations, and more precisely we will check the following key properties for the numerical solutions:

- preservation of the positivity
- conservation of the mass and saturation constraints,
- energy dissipation along solutions.

2.1. The ALG2-JKO scheme.

This algorithm relies on the seminal work of Benamou and Brenier [4] where an augmented Lagrangian approach was used to compute Wasserstein distances. In [6], this approach was extended to the computation of Wasserstein gradient flows. The method is very well suited for computing solutions to constrained gradient flows, as it will appear in the numerical simulations presented in Section 3.

2.1.1. *The augmented Lagrangian formulation.* Roughly speaking, the ALG2-JKO scheme consists in rewriting the single JKO step (8) as a more fashionable (and effectively implementable) convex minimization problem. In order to do so, let us first introduce the convex lower-semicontinuous 1-homogeneous action function given, for all $(s, \boldsymbol{m}) \in \mathbb{R} \times \mathbb{R}^d$, by

$$A(s, \boldsymbol{m}) := \begin{cases} \frac{|\boldsymbol{m}|^2}{2s} & \text{if } s > 0, \\ 0 & \text{if } s = 0 \text{ and } \boldsymbol{m} = \boldsymbol{0}, \\ +\infty & \text{otherwise.} \end{cases} \tag{14}$$

We recall that $\boldsymbol{m} = s\boldsymbol{v}$ is the momentum variable in the continuity equation $\partial_t s + \boldsymbol{\nabla} \cdot (s\boldsymbol{v}) = 0$ and $|\boldsymbol{m}|^2/s = s|\boldsymbol{v}|^2$ is a kinetic energy, see Remark 1.4. As originally observed in [4], the function A can be seen as the support function

$$A(s, \boldsymbol{m}) = \sup_{(a,\boldsymbol{b}) \in K_2} \{as + \boldsymbol{b} \cdot \boldsymbol{m}\} \tag{15}$$

of the convex set K_2, where K_α is defined for $\alpha > 0$ as

$$K_\alpha := \left\{ (a, \boldsymbol{b}) \in \mathbb{R} \times \mathbb{R}^d \, : \, a + \frac{1}{\alpha}|\boldsymbol{b}|^2 \leq 0 \right\}. \tag{16}$$

Taking advantage of the Benamou-Brenier formula (6), and given the previous JKO step $\boldsymbol{s}^n$, (8) can be recast as

$$\min_{\boldsymbol{s},\mathbf{m}} \left\{ \sum_{i=0}^{N} \frac{\mu_i}{\kappa} \int_0^1 \int_\Omega A(s_{i,t}(\boldsymbol{x}), \boldsymbol{m}_{i,t}(\boldsymbol{x})) \, \mathrm{d}\boldsymbol{x}\mathrm{d}t + \tau \mathcal{E}(\boldsymbol{s}_{t=1}) \right\}, \tag{17}$$

where the infimum runs over curves of measures $t \mapsto \boldsymbol{s}_t = (s_{0,t}, \dots, s_{N,t}) \in \boldsymbol{\mathcal{A}}$ and momenta $t \mapsto \boldsymbol{m}_t = (\boldsymbol{m}_{0,t}, \dots, \boldsymbol{m}_{N,t}) \in \mathcal{M}^d(\Omega)^{N+1}$, subject to $N+1$ linear constraints

$$\begin{cases} \partial_t s_{i,t} + \boldsymbol{\nabla} \cdot (\boldsymbol{m}_{i,t}) = 0 & \text{in } \mathcal{D}', \\ \boldsymbol{m}_{i,t} \cdot \nu = 0 & \text{on } \partial\Omega, \\ s_i|_{t=0} = s_i^n, \end{cases} \qquad i = 0, \dots, N. \tag{18}$$

Note that only the initial endpoint $\boldsymbol{s}_{t=0} = \boldsymbol{s}^n$ is prescribed for the curve $(\boldsymbol{s}_t)_{t\in[0,1]}$. The terminal endpoint is free and contributes to the objective functional (17) through the $\mathcal{E}(\boldsymbol{s}_{t=1})$ term, and the JKO minimizer will be retrieved as $\boldsymbol{s}^{n+1} = \boldsymbol{s}_{t=1}$. Note also that the minimizing curve

$(\boldsymbol{s}_t)_{t\in[0,1]}$ in (17)–(18) will automatically be a Wasserstein geodesic between the successive JKO minimizers $\boldsymbol{s}_{t=0} = \boldsymbol{s}^n$ and $\boldsymbol{s}_{t=1} = \boldsymbol{s}^{n+1}$.

As a first step towards a Lagrangian formulation, we rewrite the constraint (18) as a sup problem with multipliers $\phi_i(t,\boldsymbol{x})$

$$\sup_{\boldsymbol{\phi}} \left\{ \sum_{i=0}^{N} \int_\Omega \phi_i(1,\cdot) s_{i,1} - \int_\Omega \phi_i(0,\cdot) s_i^n - \int_0^1 \int_\Omega (\partial_t \phi_i s_{i,t} + \boldsymbol{\nabla}\phi_i \cdot \boldsymbol{m}_{i,t}) \right\} = \begin{cases} 0 & \text{if (18) holds,} \\ +\infty & \text{else,} \end{cases} \tag{19}$$

and minimizing (17) under the constraint (18) can thus be written $\inf_{\boldsymbol{s},\boldsymbol{m}} \sup_{\boldsymbol{\phi}} \{\dots\}$. Swapping $\inf\sup = \sup\inf$ as in [6] and using that the Legendre transform of $\frac{\mu_i}{\kappa} A$ is the charateristic function (convex indicator) of the convex set $K_{2\mu_i/\kappa}$ defined in (16),

$$\left(\frac{\mu_i}{\kappa} A\right)^* (a,\boldsymbol{b}) = \chi_{K_{2\mu_i/\kappa}}(a,\boldsymbol{b}) = \begin{cases} 0 & \text{if } (a,\boldsymbol{b}) \in K_{2\mu_i/\kappa}, \\ +\infty & \text{else,} \end{cases}$$

the problem (17)-(18) finally becomes after a few elementary manipulations

$$\inf_{\boldsymbol{\phi}} \left\{ \sum_{i=0}^{N} \int_\Omega \phi_i(0,\cdot) s_i^n + \mathcal{E}_\tau^*(-\boldsymbol{\phi}(1,\cdot)) \ : \qquad (\partial_t \phi_i, \boldsymbol{\nabla}\phi_i) \in K_{2\mu_i/\kappa} \right\}.$$

Here $\mathcal{E}_\tau^*$ denotes the Legendre transform of $\mathcal{E}_\tau := \tau\mathcal{E}$. This dual problem can be reformulated as

$$\inf_{\boldsymbol{\phi}} \Big\{ F(\boldsymbol{\phi}) + G(\boldsymbol{q}) \ : \qquad \boldsymbol{q} = \Lambda\boldsymbol{\phi} \Big\},$$

where

$$\Lambda\boldsymbol{\phi} = (\partial_t\boldsymbol{\phi}, \boldsymbol{\nabla}\boldsymbol{\phi}, -\boldsymbol{\phi}(1,\cdot)) \qquad \text{and} \qquad \boldsymbol{q} = (\boldsymbol{a},\boldsymbol{b},\boldsymbol{c})$$

are functions with values in $(\mathbb{R}\times\mathbb{R}^d\times\mathbb{R})^{N+1}$,

$$F(\boldsymbol{\phi}) = \sum_{i=0}^{N} \int_\Omega \phi_i(0,\cdot) s_i^n,$$

$$G(\boldsymbol{q}) = \sum_{i=0}^{N} \int_0^1 \int_\Omega \chi_{K_{2\mu_i/\kappa}}(a_i,\boldsymbol{b}_i) + \mathcal{E}_\tau^*(\boldsymbol{c}),$$

and $\chi_{K_{2\mu_i/\kappa}}$ stands again for the characteristic function of $K_{2\mu_i/\kappa}$. Introducing a Lagrange multiplier

$$\boldsymbol{\sigma} = (\boldsymbol{s},\boldsymbol{m},\tilde{\boldsymbol{s}}_1)$$

for the constraint $\Lambda\boldsymbol{\phi} = \boldsymbol{q}$, finding a minimizer $\boldsymbol{s}^{n+1}$ in the JKO scheme (8) is thus equivalent to finding a saddle-point of the Lagrangian

$$L(\boldsymbol{\phi},\boldsymbol{q},\boldsymbol{\sigma}) := F(\boldsymbol{\phi}) + G(\boldsymbol{q}) + \boldsymbol{\sigma}\cdot(\Lambda\boldsymbol{\phi} - \boldsymbol{q}). \tag{20}$$

Here we slightly abuse the notations: $\boldsymbol{s} = (\boldsymbol{s}_t)_{t\in[0,1]}$ and $\boldsymbol{m} = (\boldsymbol{m}_t)_{t\in[0,1]}$ are time-depending curves while $\tilde{\boldsymbol{s}}_1 \in \boldsymbol{\mathcal{A}}$ is independent of time. The scalar product in (20) is

$$\boldsymbol{\sigma}\cdot(\Lambda\boldsymbol{\phi} - \boldsymbol{q}) = \sum_{i=0}^{N} \Big(\int_0^1 \int_\Omega (s_i(\partial_t\phi_i - a_i) + \boldsymbol{m}_i\cdot(\boldsymbol{\nabla}\phi_i - \boldsymbol{b}_i)) - \int_\Omega \tilde{s}_{1,i}(\phi_i(1,\cdot) + c_i) \Big).$$

We stress that the free variable $\tilde{\boldsymbol{s}}_1$ is a priori independent of the curve $(\boldsymbol{s}_t)_{t\in[0,1]}$, but that the saddle-point will ultimately satisfy $\boldsymbol{s}_{t=1} = \tilde{\boldsymbol{s}}_1$. In the Lagrangian (20), the original unknowns $(\boldsymbol{s}, \boldsymbol{m}, \tilde{\boldsymbol{s}}_1)$ become the Lagrange multipliers for the constraint $\boldsymbol{q} = \Lambda\boldsymbol{\phi}$, i.e., respectively

$$\boldsymbol{a} = \partial_t\boldsymbol{\phi}, \qquad \boldsymbol{b} = \boldsymbol{\nabla}\boldsymbol{\phi}, \qquad \text{and} \qquad \boldsymbol{c} = -\boldsymbol{\phi}(1,\cdot).$$

For some fixed regularization parameter $r > 0$, we introduce now the augmented Lagrangian

$$L_r(\boldsymbol{\phi}, \boldsymbol{q}, \boldsymbol{\sigma}) := F(\boldsymbol{\phi}) + G(\boldsymbol{q}) + \boldsymbol{\sigma}\cdot(\Lambda\boldsymbol{\phi} - \boldsymbol{q}) + \frac{r}{2}\|\Lambda\boldsymbol{\phi} - \boldsymbol{q}\|^2, \tag{21}$$

where the extra regularizing term is given by the L^2 norm

$$\frac{r}{2}\|\Lambda\boldsymbol{\phi} - \boldsymbol{q}\|^2 = \frac{r}{2}\sum_{i=0}^{N}\left(\int_0^1\int_\Omega(|\partial_t\phi_i - a_i|^2 + |\boldsymbol{\nabla}\phi_i - \boldsymbol{b}_i|^2) + \int_\Omega|\phi_i(1,\cdot) + c_i|^2\right).$$

Observe that being a saddle-point of (20) is equivalent to being a saddle-point of (21), see for instance [22]. Thus in order to solve one step of the JKO scheme (8), it suffices to find a saddle-point of the augmented Lagrangian L_r.

2.1.2. *Algorithm and discretization.* The augmented Lagrangian algorithm ALG2 aims at finding a saddle-point of L_r and consists in a splitting scheme. Starting from $(\boldsymbol{\phi}^0, \boldsymbol{q}^0, \boldsymbol{\sigma}^0)$, we generate a sequence $(\boldsymbol{\phi}^k, \boldsymbol{q}^k, \boldsymbol{\sigma}^k)_{k\geq 0}$ by induction as follows

Step 1: minimize with respect to $\boldsymbol{\phi}$:

$$\boldsymbol{\phi}^{k+1} = \operatorname*{argmin}_{\boldsymbol{\phi}}\left(F(\boldsymbol{\phi}) + \boldsymbol{\sigma}^k\cdot\Lambda\boldsymbol{\phi} + \frac{r}{2}|\Lambda\boldsymbol{\phi} - \boldsymbol{q}^k|^2\right),$$

Step 2: minimize with respect to $\boldsymbol{q}$:

$$\boldsymbol{q}^{k+1} = \operatorname*{argmin}_{\boldsymbol{q}}\left(G(\boldsymbol{q}) - \boldsymbol{\sigma}^k\cdot\boldsymbol{q} + \frac{r}{2}|\Lambda\boldsymbol{\phi}^{k+1} - \boldsymbol{q}|^2\right),$$

Step 3: maximize with respect to $\boldsymbol{\sigma}$, which amounts here to updating the multiplier by the gradient ascent formula

$$\boldsymbol{\sigma}^{k+1} = \boldsymbol{\sigma}^k + r(\Lambda\boldsymbol{\phi}^{k+1} - \boldsymbol{q}^{k+1}).$$

Since step 3 is a mere pointwise update we only describe in details the first two steps. In order to keep the notations light we sometimes write $s_i(t,\boldsymbol{x}) = s_{i,t}(\boldsymbol{x})$, and likewise for any other variable depending on time.

- The first step corresponds to solving $N+1$ independent linear elliptic problems in time and space, namely

$$-r\boldsymbol{\Delta}_{t,\boldsymbol{x}}\phi_i^{k+1} = \boldsymbol{\nabla}_{t,\boldsymbol{x}}\cdot((s_i^k, \boldsymbol{m}_i^k) - r(a_i^k, \boldsymbol{b}_i^k)) \qquad \text{in } (0,1)\times\Omega$$

  with the boundary conditions

$$\begin{cases} r\partial_t\phi_i^{k+1}(0,\cdot) = s_i^n(\cdot) - s_i^k(0,\cdot) + ra_i^k(0,\cdot) & \text{in } \Omega,\\ r\left(\partial_t\phi_i^{k+1}(1,\cdot) + \phi_i^{k+1}(1,\cdot)\right) = \tilde{s}_{1,i}^k(\cdot) - s_i^k(1,\cdot) + r\left(a_i^k(1,\cdot) - c_i^k(\cdot)\right) & \text{in } \Omega,\\ \left(r\boldsymbol{\nabla}\phi_i^{k+1} + \boldsymbol{m}_i^k - r\boldsymbol{b}_i^k\right)\cdot\nu = 0 & \text{on } \partial\Omega. \end{cases}$$

- The second step splits into two convex pointwise subproblems. The first one corresponds to projections onto the parabolas $K_{2\mu_i/\kappa}$:

$$(a_i^{k+1}, \boldsymbol{b}_i^{k+1})(t,\boldsymbol{x}) = P_{K_{2\mu_i/\kappa}}\left((\partial_t\phi_i^{k+1}, \boldsymbol{\nabla}\phi_i^{k+1})(t,\boldsymbol{x}) + \frac{1}{r}(s_i^k, \boldsymbol{m}_i^k)(t,\boldsymbol{x})\right), \qquad \forall\, i = 0,\dots,N.$$

This projection $P_{K_{2\mu_i/\kappa}}$ onto $K_{2\mu_i/\kappa}$ is explicitly given by (see [33])

$$P_{K_{2\mu_i/\kappa}}(\alpha,\boldsymbol{\beta}) = \begin{cases} (\alpha,\boldsymbol{\beta}), & \text{if } (\alpha,\boldsymbol{\beta}) \in K_{2\mu_i/\kappa}, \\ \left(\alpha-\lambda, \frac{\mu_i \boldsymbol{\beta}}{\kappa\lambda+\mu_i}\right), & \text{otherwise,} \end{cases}$$

where λ is the largest real root of the cubic equation

$$(\alpha-\lambda)(\mu_i/\kappa+\lambda)^2 + \frac{\mu_i}{2\kappa}|\boldsymbol{\beta}|^2 = 0.$$

The second subproblem should update $\boldsymbol{c}$. To this end, we need to solve the pointwise proximal problem: for each $\boldsymbol{x} \in \Omega$

$$\boldsymbol{c}^{k+1}(\boldsymbol{x}) = \underset{\boldsymbol{c}\in\mathbb{R}^{N+1}}{\operatorname{argmin}} \left\{ \frac{r}{2}\sum_{i=0}^{N} |\phi_i^{k+1}(1,\boldsymbol{x}) - \frac{1}{r}\tilde{s}_{1,i}^k(\boldsymbol{x}) + c_i|^2 + E_\tau^*(\boldsymbol{x},\boldsymbol{c}) \right\}, \tag{22}$$

where $E_\tau^*(\boldsymbol{x},\cdot)$ is the Legendre transform of the energy density $E_\tau(\boldsymbol{x},\cdot) = \tau E(\boldsymbol{x},\cdot)$ in its second argument (E being implicitly defined as $\mathcal{E}(\boldsymbol{s}) = \int_\Omega E(\boldsymbol{x},\boldsymbol{s}(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x}$).

Notice that the energy functional $\mathcal{E}$ only plays a role in the minimization with respect to the internal $\boldsymbol{c}$ variable, namely the second subproblem (22) in **Step 2**. In Section 3 we will try to make this step explicit for our two particular applications.

In order to implement this algorithm in a computational setting we use P2 finite elements in time and space for $\boldsymbol{\phi}$, and P1 finite elements for $\boldsymbol{\sigma}$ and $\boldsymbol{q}$. The variables $\boldsymbol{\nabla}_{t,\boldsymbol{x}}\phi_i^{k+1} = (\partial_t\phi_i^{k+1}, \boldsymbol{\nabla}\phi_i^{k+1})$ are understood as the projection onto P1 finite elements and the algorithm was implemented using `FreeFem++` [24]. The convergence of this algorithm is known in finite dimension [22], i.e., the iterates $(\boldsymbol{\phi}^k, \boldsymbol{q}^k, \boldsymbol{\sigma}^k)$ are guaranteed to converge to a saddle point $(\boldsymbol{\phi}, \boldsymbol{q}, \boldsymbol{\sigma})$ as $k \to \infty$. Once the saddle-point is reached, the output $\boldsymbol{\sigma} = (\boldsymbol{s}, \boldsymbol{m}, \tilde{\boldsymbol{s}}_1)$ is a minimizer for the problem (17)-(18) and the solution of the JKO scheme (8) is simply recovered as $\boldsymbol{s}^{n+1} = \tilde{\boldsymbol{s}}_1 = \boldsymbol{s}|_{t=1}$.

Numerically, the Benamou-Brenier formula involves an additional time dimension to be effectively discretized in each elementary JKO step, and this can be seen as a drawback. However the successive JKO densities are close due to the small time step $\tau \to 0$ (indeed $\boldsymbol{W}(\boldsymbol{s}^{n+1}, \boldsymbol{s}^n) = \mathcal{O}(\sqrt{\tau})$ from the total square distance estimate (11)) and, in practice, only a very few inner timesteps are needed.

2.1.3. *Some properties of the approximate solution.* As previously mentioned, the above Lagrangian framework can be practically implemented by simply projecting the (infinite dimensional) problem onto P1/P2 finite elements. Provided that the iteration procedure (Steps 1 to 3 in Section 2.1.2) converges as $k \to \infty$, as guaranteed from [22], the saddle-point $\boldsymbol{\sigma} = (\boldsymbol{s}, \boldsymbol{m}, \tilde{\boldsymbol{s}}_1)$ satisfies by construction:

(i) $(s_i, \boldsymbol{m}_i)$ remains in the domain $\mathrm{Dom}(A)$ of the action functional A defined in (14);
(ii) the continuity equation $\partial_t s_{i,t} + \boldsymbol{\nabla}\cdot(\boldsymbol{m}_{i,t}) = 0$ holds with zero-flux boundary condition.

As a consequence of (i) the scheme preserves the positivity, i.e., $s_i^{n+1} \geq 0$, whereas (ii) ensures the mass conservation $\int_\Omega s_i^{n+1} = \int_\Omega s_i^n$.

Moreover, the fully discrete ALG2-JKO scheme preserves by construction the gradient flow structure, hence the scheme is automatically energy diminishing. Since the energy functional (3) includes the $\chi_{\boldsymbol{\Delta}}$ term accounting for the saturation constraint $\sum s_i = 1$, one can and should include this convex indicator term in the discretized energy. This contraint is then passed on to the proximal operator to be used in the implementation, see Section 3 for details. As a result the saturation constraint is satisfied.

2.2. **Upstream mobility Finite Volume scheme.** The ALG2-JKO scheme described in the previous section will be compared to the widely used upstream mobility Finite Volume scheme [34, 7, 21]. As a first step, let us detail how Ω is discretized.

2.2.1. *The finite volume mesh.* The domain Ω is assumed to be polygonal. Then following [20], an admissible mesh consists in a triplet $\big(\mathfrak{T}, \mathfrak{E}, (\boldsymbol{x}_K)_{K\in\mathfrak{T}}\big)$. The elements K of $\mathfrak{T}$ are open polygonal convex subsets of Ω called *control volumes*. Their boundaries are made of elements $\sigma \in \mathfrak{E}$ of codimension 1 (*edges* if $d = 2$ or *faces* if $d = 3$). Let K, L be two distinct elements of $\mathfrak{T}$, then $\overline{K}\cap\overline{L}$ is either empty, or reduced to a point (a vertex), or there exists $\sigma \in \mathfrak{E}$ denoted by $\sigma = K|L$ such that $\overline{K} \cap \overline{L} = \overline{\sigma}$. In particular, two control volumes share at most one edge. We denote by $\mathfrak{E}_K = \big\{\sigma \in \mathfrak{E} \,\big|\, \bigcup_{\sigma\in\mathfrak{E}_K} \overline{\sigma} = \partial K\big\}$ the set of the edges associated to an element $K \in \mathfrak{T}$, and by $\mathfrak{N}_K = \{L \in \mathfrak{T} \mid \text{there exists } \sigma = K|L \in \mathfrak{E}\}$ the set of the neighboring control volumes to K. We also denote by

$$\mathfrak{E}_{\rm ext} = \{\sigma \in \mathfrak{E} \mid \sigma \subset \partial\Omega\}, \qquad \mathfrak{E}_{\rm int} = \mathfrak{E} \setminus \mathfrak{E}_{\rm ext}, \qquad \mathfrak{E}_{\rm int,K} = \mathfrak{E}_{\rm int} \cap \mathfrak{E}_K, \qquad \forall K \in \mathfrak{T}.$$

The last element $(\boldsymbol{x}_K)_{K\in\mathfrak{T}}$ of the triplet corresponds to the so called *cell-centers*. To each control volume $K \in \mathfrak{T}$, we associate an element $\boldsymbol{x}_K \in \Omega$ such that for all $L \in \mathfrak{N}_K$, the straight line $(\boldsymbol{x}_K, \boldsymbol{x}_L)$ is orthogonal to the edge $K|L$. This implicitly requires that $\boldsymbol{x}_K$ and $\boldsymbol{x}_L$ are distinct, and we denote by $d_\sigma = |\boldsymbol{x}_K - \boldsymbol{x}_L|$ for $\sigma = K|L$ the distance between the cell centers of the neighboring control volumes K and L. For $\sigma \in \mathfrak{E}_K \cap \mathfrak{E}_{\rm ext}$, we denote by $\boldsymbol{x}_\sigma$ the projection of $\boldsymbol{x}_K$ on the hyperplane containing σ, and by $d_\sigma = |\boldsymbol{x}_K - \boldsymbol{x}_\sigma|$. We also require that the vector $\boldsymbol{x}_L - \boldsymbol{x}_K$ is oriented in the same sense as the normal $\boldsymbol{n}_{K,\sigma}$ to $\sigma \in \mathfrak{E}_K$ outward w.r.t. K. We refer to Figure 1 for an illustration of the notations used hereafter.

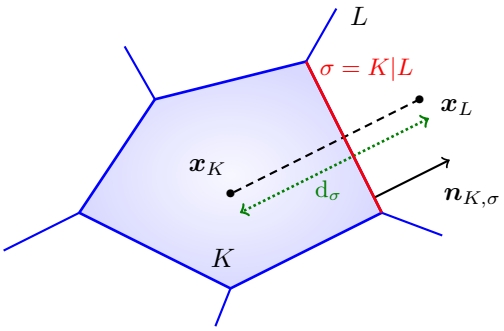


FIGURE 1. Here is an example of *admissible mesh* in the sense of [20]

Beyond cartesian grids, there are two classical ways to construct admissible meshes in the above sense when $d = 2$. The first one consists in the classical Delaunay triangulation, the cell-center $\boldsymbol{x}_K$ of $K \in \mathfrak{T}$ being the center of the circumcircle of K. The second classical construction consists in choosing the cell centers $(\boldsymbol{x}_K)$ at first, and then to construct $\mathfrak{T}$ as the associated Voronoï diagram.

In what follows, we denote by m_K the d-dimensional Lebesgue measure of the control volume $K \in \mathfrak{T}$, while m_σ denotes the $(d-1)$-dimensional Lebesgue measure of the edge $\sigma \in \mathfrak{E}$. We also denote by $a_\sigma = \frac{m_\sigma}{d_\sigma}$ the *transmissivity* of the edge σ.

In order to simplify the presentation, we restrict our presentation to the case of uniform time discretizations with time step $\tau > 0$. The extension to the case of time discretizations with varying time steps does lead to any particular difficulty.

2.2.2. *Definition of the Finite Volume scheme.* The Finite Volume scheme relies on the discretization of the Euler-Lagrange equations (1) rather than on the minimizing movement scheme (8). The main unknowns to the problems are located at the cell centers $(\boldsymbol{x}_K)_{K\in\mathfrak{T}}$. They consist in discrete saturations $s^n_{i,K} \simeq s_i(\boldsymbol{x}_K, n\tau)$ and discrete pressures $p^n_{i,K} \simeq p_i(\boldsymbol{x}_K, n\tau)$. In what follows, we denote by $\boldsymbol{s}^n_K = \left(s^n_{i,K}\right)_{0\le i\le N}$ (resp. $\boldsymbol{p}^n_K = \left(p^n_{i,K}\right)_{0\le i\le N}$) and $\boldsymbol{s}^n_{\mathfrak{T}} = (\boldsymbol{s}^n_K)_{K\in\mathfrak{T}}$.

The first equation of the scheme is a straightforward consequence of (1a), i.e.,

$$\sum_{i=0}^{N} s^n_{i,K} = 1, \qquad \forall K \in \mathfrak{T},\ \forall n \ge 1. \tag{23a}$$

This motivates the introduction of the discrete counterpart $\boldsymbol{\mathcal{X}}_{\mathfrak{T}}$ of $\boldsymbol{\mathcal{X}}$ defined by

$$\boldsymbol{\mathcal{X}}_{\mathfrak{T}} = \left\{\boldsymbol{s}_{\mathfrak{T}} \mid \boldsymbol{s}_K \in \boldsymbol{\Delta} \text{ for all } K \ \in \mathfrak{T}\right\},$$

so that (23a) amounts to requiring that $\boldsymbol{s}^n_{\mathfrak{T}}$ belongs to $\boldsymbol{\mathcal{X}}_{\mathfrak{T}}$ for all n (the nonnegativity of the saturations will be established later on). The capillary pressure relations (1d) are discretized into

$$p^n_{i,K} - p^n_{0,K} = \pi_i(s^n_K), \qquad \forall i \in \{1,\dots,N\},\ \forall K \in \mathfrak{T},\ \forall n \ge 1. \tag{23b}$$

Integrating (1b) over the control volume $K \in \mathfrak{T}$ (recall here that the porosity ω was artificially set to 1) and using Stokes' formula, one gets the natural approximation

$$\frac{s^n_{i,K} - s^{n-1}_{i,K}}{\tau} m_K + \sum_{\sigma\in\mathfrak{E}_K} s^n_{i,\sigma} v^n_{i,K,\sigma} = 0, \qquad \forall i \in \{0,\dots,N\},\ \forall K \in \mathfrak{T},\ \forall n \ge 1. \tag{23c}$$

Here, $v^n_{i,K,\sigma}$ is an approximation of $\int_\sigma \boldsymbol{v}_i(\gamma, n\tau)\cdot\boldsymbol{n}_{K,\sigma}\mathrm{d}\gamma$, where $\boldsymbol{v}_i$ is related to p_i through to Darcy law (1c). Thanks to the orthogonality condition on the mesh, the choice

$$v^n_{i,K,\sigma} = a_\sigma \frac{\kappa}{\mu_i}\left(p^n_{i,K} + \Psi_{i,K} - p^n_{i,L} - \Psi_{i,L}\right), \qquad \forall \sigma = K|L \in \mathfrak{E}_{\mathrm{int}}, \tag{23d}$$

is consistent — we use the shortened notation $\Psi_{i,K} = \Psi_i(\boldsymbol{x}_K)$ —. In accordance with the no-flux boundary conditions, we impose that

$$v^n_{i,K,\sigma} = 0, \qquad \forall \sigma \in \mathfrak{E}_K \cap \mathfrak{E}_{\mathrm{ext}},\ \forall n \ge 1.$$

It remains to define the approximate saturations $s^n_{i,\sigma}$ for $\sigma \in \mathfrak{E}_{\mathrm{int}}$. We use here the very classical upwind choice [34, 7, 21], i.e.,

$$s^n_{i,\sigma} = \begin{cases} \left(s^n_{i,K}\right)^+ & \text{if } v^n_{i,K,\sigma} \ge 0, \\ \left(s^n_{i,L}\right)^+ & \text{otherwise,} \end{cases} \qquad \forall \sigma = K|L \in \mathfrak{E}_{\mathrm{int}}. \tag{23e}$$

Note that even though the mapping $(\boldsymbol{s}^n_{\mathfrak{T}}, \boldsymbol{p}^n_{\mathfrak{T}}) \mapsto \boldsymbol{s}^n_{\mathfrak{E}} = \left(\left(s^n_{i,\sigma}\right)_{0\le i\le n}\right)_{\sigma\in\mathfrak{E}_{\mathrm{int}}}$ is discontinuous, the quantity $s^n_{i,\sigma}v^n_{i,K,\sigma}$ depends in a continuous way of the main unknowns.

The scheme (23) amounts to a nonlinear system of equations to be solved a each time step. This will be practically done thanks to Newton-Raphson method. But before, we establish some properties of the FV scheme, namely the energy decay, the entropy control, the non-negativity of the saturations, or the existence of a solution $(\boldsymbol{s}^n_{\mathfrak{T}}, \boldsymbol{p}^n_{\mathfrak{T}})$ to the scheme.

2.2.3. *Some properties of the approximate solution.* The first key property of the FV scheme that we point out is the non-negativity of the saturations:

$$s_{i,K}^n \geq 0 \qquad \forall i \in \{0,\dots,N\},\ \forall K \in \mathfrak{T},\ \forall n \geq 1.$$

In order to establish this estimate, it suffices to rewrite (23c) as

$$s_{i,K}^n + \frac{\tau}{m_K} \sum_{\substack{\sigma \in \mathfrak{E}_{\mathrm{int},K} \\ \sigma = K|L}} \left[\left(s_{i,K}^n\right)^+ \left(v_{i,K,\sigma}^n\right)^+ - \left(s_{i,L}^n\right)^+ \left(v_{i,K,\sigma}^n\right)^- \right] = s_{i,K}^{n-1}$$

thanks to (23e). In the previous expression, we used the convention $a^- = \max(0,-a) \geq 0$. Assume for contradiction that $s_{i,K}^n$ is negative, then so does the left-hand side, while the right-hand side is nonnegative by induction. Together with (23a), this shows that

$$\boldsymbol{s}_{\mathfrak{T}}^n \in \boldsymbol{\mathcal{X}}_{\mathfrak{T}}, \qquad \forall n \geq 1. \tag{24}$$

The scheme is mass conservative for the $N+1$ phases since

$$v_{i,K,\sigma}^n + v_{i,L,\sigma}^n = 0 \quad \text{hence} \quad s_{i,\sigma}^n v_{i,K,\sigma}^n + s_{i,\sigma}^n v_{i,L,\sigma}^n = 0, \qquad \text{for } \sigma = K|L.$$

Together with the no-flux boundary conditions, this shows that the mass is conserved along time:

$$\sum_{K\in\mathfrak{T}} s_{i,K}^n m_K = \sum_{K\in\mathfrak{T}} s_{i,K}^{n-1} m_K = \sum_{K\in\mathfrak{T}} s_{i,K}^0 m_K, \qquad \forall n \geq 1,\ \forall i \in \{0,\dots,N\}. \tag{25}$$

Then the discrete solution $\boldsymbol{s}_{\mathfrak{T}}^n$ remains in the discrete counterpart $\boldsymbol{\mathcal{A}}_{\mathfrak{T}}$ of $\boldsymbol{\mathcal{A}}$ defined as the elements $\boldsymbol{s}_{\mathfrak{T}}$ of $\mathbb{R}_+^{\mathfrak{T}}$ such that $\sum_{K\in\mathfrak{T}} s_{i,K} m_K = \sum_{K\in\mathfrak{T}} s_{i,K}^0 m_K$ for all $i \in \{0,\dots,N\}$.

Multiplying the scheme (23c) by $\tau\left(p_{i,K}^n + \Psi_{i,K}\right)$ and summing over $K \in \mathfrak{T}$ yields

$$\sum_{i=0}^N \sum_{K\in\mathfrak{T}} \left(s_{i,K}^n - s_{i,K}^{n-1}\right)\left(p_{i,K}^n + \Psi_{i,K}\right) m_K + \tau \sum_{i=0}^N \frac{\kappa}{\mu_i} \sum_{\sigma\in\mathfrak{E}_{\mathrm{int}}} a_\sigma s_{i,\sigma}^n \left(p_{i,K}^n + \Psi_{i,K} - p_{i,L}^n - \Psi_{i,L}\right)^2 = 0.$$

The second term in the above expression is clearly nonnegative. concerning the first term, one can use the constraint (23a) to rewrite as

$$\begin{aligned}\sum_{i=0}^N \sum_{K\in\mathfrak{T}} \left(s_{i,K}^n - s_{i,K}^{n-1}\right) p_{i,K}^n m_K &= \sum_{i=1}^N \sum_{K\in\mathfrak{T}} \left(s_{i,K}^n - s_{i,K}^{n-1}\right)\left(p_{i,K}^n - p_{0,K}^n\right) m_K \\ &\geq \sum_{K\in\mathfrak{T}} \left(\Pi(\boldsymbol{s}_K^{n,*}) - \Pi(\boldsymbol{s}_K^{n-1,*})\right) m_K,\end{aligned}$$

the last inequality being a consequence of the convexity of Π. This establishes that the scheme is energy diminishing: denoting by

$$\mathcal{E}(\boldsymbol{s}_{\mathfrak{T}}^n) = \sum_{K\in\mathfrak{T}} \left(\Pi(\boldsymbol{s}_K^{n,*}) + \sum_{i=0}^N s_{i,K}^n \Psi_{i,K}\right) m_K, \qquad n \geq 0,$$

one has

$$\mathcal{E}(\boldsymbol{s}_{\mathfrak{T}}^n) + \tau \sum_{i=0}^N \frac{\kappa}{\mu_i} \sum_{\sigma\in\mathfrak{E}_{\mathrm{int}}} a_\sigma s_{i,\sigma}^n \left(p_{i,K}^n + \Psi_{i,K} - p_{i,L}^n - \Psi_{i,L}\right)^2 \leq \mathcal{E}(\boldsymbol{s}_{\mathfrak{T}}^{n-1}), \qquad \forall n \geq 1. \tag{26}$$

The last *a priori* estimate we want to point out is the discrete counterpart of the flow interchange estimate. It is obtained by multiplying (23c) by $\tau\mu_i \log(s^n_{i,K})$ and by summing over $i \in \{0,\dots,N\}$ and $K \in \mathfrak{T}$, leading to

(27)
$$\sum_{i=0}^{N} \mu_i \sum_{K\in\mathfrak{T}} \left(s^n_{i,K} - s^{n-1}_{i,K}\right) \log(s^n_{i,K}) m_K + \tau \sum_{i=0}^{N} \kappa \sum_{\sigma\in\mathfrak{E}_{\rm int}} a_\sigma s^n_{i,\sigma} v^n_{i,K,\sigma} \left(\log(s^n_{i,K}) - \log(s^n_{i,L})\right) = 0.$$

As already discussed in Remark 1.7 this corresponds to taking $\log s_i$ as a test-function in the weak formulation of the continuous PDEs. The first term of (27) can be estimated thanks to an elementary convexity inequality

$$\sum_{i=0}^{N} \mu_i \sum_{K\in\mathfrak{T}} \left(s^n_{i,K} - s^{n-1}_{i,K}\right) \log(s^n_{i,K}) m_K \geq \mathcal{H}(\boldsymbol{s}^n_{\mathfrak{T}}) - \mathcal{H}(\boldsymbol{s}^{n-1}_{\mathfrak{T}}), \qquad \forall n \geq 1$$

with

$$\mathcal{H}(\boldsymbol{s}^n_{\mathfrak{T}}) = \sum_{i=0}^{N} \mu_i \sum_{K\in\mathfrak{T}} \left(h(s^n_{i,K}) - h(s^{n-1}_{i,K})\right) m_K, \qquad h(s) = s\log(s) - s + 1 \geq 0.$$

Note that the entropy functional $\mathcal{H}$ is bounded on $\boldsymbol{\mathcal{X}}_{\mathfrak{T}}$. The second term of (27) can be estimated as follows. First, the concavity of $s \mapsto \log(s)$ yields

$$s^n_{i,L}\left(\log(s^n_{i,K}) - \log(s^n_{i,L})\right) \leq s^n_{i,K} - s^n_{i,L} \leq s^n_{i,K}\left(\log(s^n_{i,K}) - \log(s^n_{i,L})\right), \qquad \sigma = K|L,$$

so that the upwind choice (23e) for $s^n_{i,\sigma}$ ensures that

$$a_\sigma s^n_{i,\sigma} v^n_{i,K,\sigma}\left(\log(s^n_{i,K}) - \log(s^n_{i,L})\right) \geq a_\sigma v^n_{i,K,\sigma}(s^n_{i,K} - s^n_{i,L}), \qquad \sigma = K|L.$$

Using the expression (23d) of $v^n_{i,K,\sigma}$ and the relation (23a) on the saturations, one gets that

$$\sum_{i=0}^{N} \sum_{\sigma\in\mathfrak{E}_{\rm int}} a_\sigma s^n_{i,\sigma} v^n_{i,K,\sigma}\left(\log(s^n_{i,K}) - \log(s^n_{i,L})\right) \geq A + B,$$

where

$$A = \sum_{i=1}^{N} \sum_{\substack{\sigma\in\mathfrak{E}_{\rm int}\\ \sigma=K|L}} a_\sigma(\pi_i(\boldsymbol{s}^{n,*}_K) - \pi_i(\boldsymbol{s}^{n,*}_L))(s^n_{i,K} - s^n_{i,L}),$$

$$B = \sum_{i=0}^{N} \sum_{\substack{\sigma\in\mathfrak{E}_{\rm int}\\ \sigma=K|L}} a_\sigma(\Psi_{i,K} - \Psi_{i,L})(s^n_{i,K} - s^n_{i,L}) = \sum_{i=0}^{N} \sum_{K\in\mathfrak{T}} s^n_{i,K} \sum_{L\in\mathfrak{N}_K} a_\sigma(\Psi_{i,K} - \Psi_{i,L}).$$

Recalling the definition (4) of the external potential and denoting by $\Psi_{i,\sigma} = \Psi_i(\boldsymbol{x}_\sigma)$, one has

$$\sum_{L\in\mathfrak{N}_K} a_\sigma(\Psi_{i,K} - \Psi_{i,L}) + \sum_{\sigma\in\mathfrak{E}_{\rm ext}\cap\mathfrak{E}_K} a_\sigma(\Psi_{i,K} - \Psi_{i,\sigma}) = 0.$$

Since $0 \leq s^n_{i,K} \leq 1$, this implies that

$$B \geq \tau\kappa|\partial\Omega||\mathbf{g}|.$$

On the other hand, the assumption (2) on the capillary pressure potential ensures that

$$A \geq \frac{1}{\varpi} \sum_{i=1}^{N} \sum_{\substack{\sigma\in\mathfrak{E}_{\rm int}\\ \sigma=K|L}} a_\sigma \left(\pi_i(\boldsymbol{s}^{n,*}_K) - \pi_i(\boldsymbol{s}^{n,*}_L)\right)^2, \qquad \forall \sigma = K|L \in \mathfrak{E}_{\rm int}.$$

Hence collecting the previous inequalities in (27) provides the following discrete $L^2_{\rm loc}(H^1)$-estimate one the capillary pressures

$$\sum_{n=1}^{M}\tau\sum_{i=1}^{N}\sum_{\substack{\sigma\in\mathfrak{E}_{\rm int}\\ \sigma=K|L}} a_\sigma\left(\pi_i(\boldsymbol{s}_K^{n,*})-\pi_i(\boldsymbol{s}_L^{n,*})\right)^2\leq C(1+M\tau). \tag{28}$$

Clearly, (28) is the discrete counterpart of the estimate (13) obtained thanks to the flow interchange technique. The derivation of a discrete $L^2_{\rm loc}(H^1)$ estimate on the phase pressures from (28) and (26) requires one additional assumption on the capillary pressure functions $(\pi_i)_{1\leq i\leq n}$. More precisely, we assume that

$$\pi_i \text{ only depends on } s_i\colon\ \frac{\partial}{\partial s_j}\,\pi_i(\boldsymbol{s}^*)=0 \text{ if } i\neq j. \tag{29}$$

Since Π is convex, the functions π_i are increasing. Assumption (29) is needed to establish that, at least for fine enough grids, there holds

$$\sum_{i=0}^{N} s_{i,\sigma}^n\geq\alpha>0,\qquad \forall n\geq 1,\ \forall\sigma\in\mathfrak{E}_{\rm int},$$

for some uniform α. Thanks to this estimate, one can follow the lines of [11, Proposition 3.4 & Corollary 3.5] (see also [13]) to derive the estimate

$$\sum_{n=1}^{M}\tau\sum_{0=1}^{N}\sum_{\substack{\sigma\in\mathfrak{E}_{\rm int}\\ \sigma=K|L}} a_\sigma\left(p_{i,K}^n-p_{i,L}^n\right)\Big)^2\leq C(1+M\tau). \tag{30}$$

The phase pressures being defined up to an additive constant (recall that they are related to Kantorovich potentials), one has to fix this degree of freedom. This can be done by enforcing

$$\sum_{K\in\mathfrak{T}} p_{0,K}^n m_K=0,\qquad \forall n\geq 1.$$

Based on the *a priori* estimates (24) and (30), we can make use of a topological degree argument (see for instance [16]) to claim that there exists (at least) one solution to the scheme. Moreover, assuming some classical regularity on the mesh $\mathfrak{T}$ (see for instance [1]), one can prove the piecewise constant approximate solutions converge towards a weak solution when the size of the mesh $\mathfrak{T}$ and the time step τ tend to 0. This convergence results together with the properties (24)–(30) as well as the wide popularity of this scheme in the engineering community makes this scheme a reference for solving (1). In the next section, we show that the ALG2-JKO scheme presented in Section 2.1 produces very similar results: same qualitative results, conservation of the mass of each phase and preservation of the positivity.

3. Numerical experiments

In this section, we compare the numerical results produced by the ALG2-JKO scheme presented in Section 2.1 with the upstream mobility Finite Volume scheme of Section 2.2. In the sequel the regularization parameter r introduced in the augmented Lagrangian formulation (21) is fixed to $r=1$ for simplicity, which gives satisfactory numerical results. The case of a three phase flow (typically water, oil and gas) is presented in Section 3.2, whereas a two-phase flow is simulated in Section 3.1. In both cases, we do not have analytical solutions at hand and the results are compared thanks to snapshots.

Note the both time discretizations are of order 1. The extension to order two methods is a challenging task. Concerning the ALG2-JKO scheme, one possibility could be to use the order 2

approximation based on the midpoint rule proposed in [26], but there is no rigorous foundation to this work up to now as far as we know. An alternative approach would be to use the variational BDF2 approach proposed in [30]. But the variational problem to be solved at each time step is no longer convex-concave, so that its practical resolution becomes more involving. Concerning the finite volume scheme, there is (up to our knowledge) no time integrator of order 2 that ensures the decay of a general energy. Going to higher order time discretizations yields also difficulties concerning the preservation of the positivity. This explains why the backward Euler scheme is very popular in the context of the simulation of multiphase porous media flows.

3.1. **Two-phase flow with Brooks-Corey capillarity.** As a first example we consider a two-phase flow, where water (s_0) and oil (s_1) are competing within the background porous medium. For the capillary pressure, we choose the very classical Brooks-Corey (or Leverett) model

$$p_1 - p_0 = \pi_1(s_1) = \alpha(1-s_1)^{-1/2}. \tag{31}$$

We refer to [3] for an overview of the classical capillary pressure relation for two-phase flows.

As in Section 1.1, the corresponding energy reads explicitly

$$\mathcal{E}(s_0,s_1) = \int_\Omega \Psi_0 s_0 + \int_\Omega \Psi_1 s_1 - 2\alpha \int_\Omega (1-s_1)^{1/2} + \int_\Omega \chi_{\boldsymbol{\Delta}}(s_0,s_1).$$

As already mentioned, only the second subproblem (22) in step 2 of the ALG2-JKO algorithm depends on the choice of the energy functional. For the above particular case, this reads: for each $\boldsymbol{x}\in\Omega$ and setting $\overline{\boldsymbol{c}} := -\boldsymbol{\phi}^{k+1}(1,\boldsymbol{x}) + \tilde{\boldsymbol{s}}_1^k(\boldsymbol{x})$, solve

$$\boldsymbol{c}^{k+1}(\boldsymbol{x}) = \operatorname*{argmin}_{\boldsymbol{c}\in\mathbb{R}^3} \left\{ \frac{1}{2}|\boldsymbol{c}-\overline{\boldsymbol{c}}|^2 + E_\tau^*(\boldsymbol{x},\boldsymbol{c}) \right\},$$

where $E_\tau^*(\boldsymbol{x},\cdot)$ is the Legendre transform of $E_\tau(\boldsymbol{x},\cdot)$ defined by

$$E_\tau(\boldsymbol{x},c_0,c_1) = \tau\Psi_0(\boldsymbol{x})c_0 + \tau\Psi_1(\boldsymbol{x})c_1 - 2\tau\alpha(1-c_1)^{1/2} + \chi_{\boldsymbol{\Delta}}(c_0,c_1) \text{ for all } c_0,c_1\in\mathbb{R}.$$

This minimization problem is equivalent to computing

$$\boldsymbol{c}^{k+1}(\boldsymbol{x}) = \mathrm{Prox}_{E_\tau^*(\boldsymbol{x},\cdot)}(\overline{\boldsymbol{c}}),$$

where the proximal operator Prox_f of a given convex, lower semicontinuous function $f:\mathbb{R}^{N+1}\to\mathbb{R}\cup\{+\infty\}$ is defined by

$$\mathrm{Prox}_f(\overline{\boldsymbol{y}}) := \operatorname*{argmin}_{\boldsymbol{y}\in\mathbb{R}^{N+1}} \left\{ \frac{1}{2}|\boldsymbol{y}-\overline{\boldsymbol{y}}|^2 + f(\boldsymbol{y}) \right\}, \qquad \forall\, \overline{\boldsymbol{y}}\in\mathbb{R}^{N+1}.$$

Thanks to Moreau's identity

$$\mathrm{Prox}_{f^*}(\overline{\boldsymbol{y}}) = \overline{\boldsymbol{y}} - \mathrm{Prox}_f(\overline{\boldsymbol{y}}) \qquad \forall\, \overline{\boldsymbol{y}}\in\mathbb{R}^{N+1}, \tag{32}$$

it suffices to compute Prox_{E_τ} in order to determine $\mathrm{Prox}_{E_\tau^*}$, and we never actually compute the Legendre transform $E_\tau^*(\boldsymbol{x},\cdot)$. Computing the proximal operator $\boldsymbol{c}^{k+1}(\boldsymbol{x}) = \mathrm{Prox}_{E_\tau^*(\boldsymbol{x},\cdot)}(\overline{\boldsymbol{c}})$ thus amounts to evaluating

$$(c_0^{k+1}(\boldsymbol{x}), c_1^{k+1}(\boldsymbol{x})) = (\overline{c}_0,\overline{c}_1) - \mathrm{Prox}_{E_\tau(\boldsymbol{x},\cdot)}(\overline{c}_0,\overline{c}_1).$$

Finally, $(\tilde{c}_0,\tilde{c}_1) := \mathrm{Prox}_{E_\tau(\boldsymbol{x},\cdot)}(\overline{c}_0,\overline{c}_1)$ is computed by solving

$$\tilde{c}_1 = \operatorname*{argmin}_{0\le c_1\le 1} \left\{ \frac{1}{2}|c_1+\overline{c}_0-\tau\Psi_0(\boldsymbol{x})-1|^2 + \frac{1}{2}|c_1-\overline{c}_1+\tau\Psi_1(\boldsymbol{x})|^2 - 2\tau\alpha(1-c_1)^{1/2} \right\}$$

and then setting $\tilde{c}_0 = 1-\tilde{c}_1$. More explicitly, $\tilde{c}_1$ is the positive part of the root on $(-\infty,1)$ of

$$2c - \overline{c}_1 + \tau\Psi_1(\boldsymbol{x}) + \overline{c}_0 - \tau\Psi_0(\boldsymbol{x}) - 1 + \frac{\tau\alpha}{(1-c)^{1/2}} = 0.$$

To conclude, we set $(c_0^{n+1}(\boldsymbol{x}), c_1^{n+1}(\boldsymbol{x})) = (\bar{c}_0 - \tilde{c}_0, \bar{c}_1 - \tilde{c}_1)$.

On Figure 3, we compare the numerical solutions of problem (1) with Brooks-Corey capillarity (31) obtained thanks to the ALG2-JKO scheme and to the upstream mobility finite volume scheme. Simulations with the ALG2-JKO scheme are carried using a structured grid with 5000 triangles and 2601 vertices in space and a single inner time step, and with 200 JKO steps ($\tau = 0.05$). Simulations with the upstream mobility finite volume scheme are performed on the corresponding Cartesian grid with 2500 squares. The time step τ appearing in (23c) can be also set to 0.05 here since Newton's method converges rather easily in this test case.

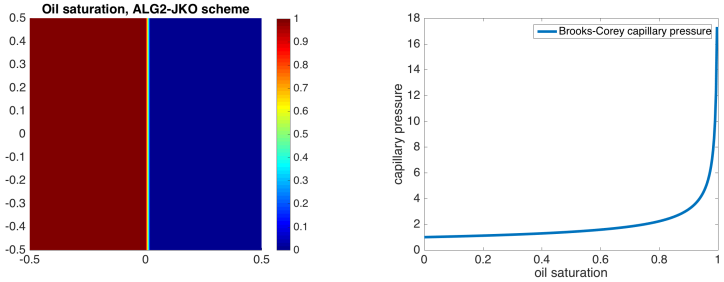


FIGURE 2. Two-phase flow: initial oil saturation profile (left) and Brooks-Corey capillary pressure function (31) with $\alpha = 1$ (right).

As expected, the results produced by the two schemes are very similar. The dense phase (the water) is instantaneously diffused in the whole domain because of the singularity of π_1 near 1. When time goes, oil slowly moves to the top because of buoyancy.

3.2. **Three-phase flow with quadratic capillary potential.** In the second test case, we consider the case of a three-phase flow where water (s_0), oil (s_1), and gas (s_2) are in competition within the porous medium. Here we assume that the capillary pressure functions π_1 and π_2 are linear,

$$p_1 - p_0 = \pi_1(s_1) = \alpha_1 s_1 \qquad \text{and} \qquad p_2 - p_0 = \pi_2(s_2) = \alpha_2 s_2.$$

The corresponding capillary potential Π is then given by

$$\Pi(\boldsymbol{s}^*) = \frac{\alpha_1}{2}(s_1^2) + \frac{\alpha_2}{2}(s_2^2).$$

The Assumption (2) and (29) are fulfilled, so that we are in the theoretical framework of our statements, i.e., convergence of the minimizing movement scheme and of the finite volume scheme. However, the problem is difficult to simulate because of the rather large ratios on the viscosities. Indeed, the phase 0 represents water, the phase 1 corresponds to oil and the phase 2 corresponds to gas, and we set

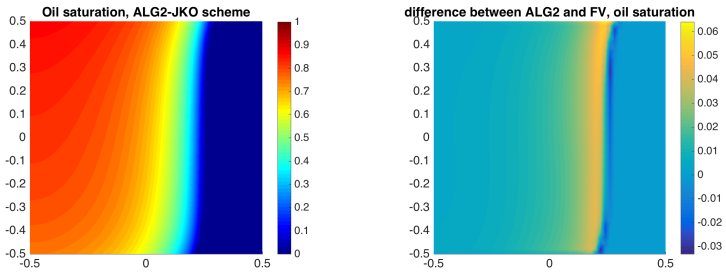
$$\mu_0 = 1, \quad \mu_1 = 50, \quad \mu_2 = 0.1, \qquad \text{and} \qquad \rho_0 = 1, \quad \rho_1 = 0.87, \quad \rho_2 = 0.1.$$

The resulting energy in the JKO scheme (8) is given by

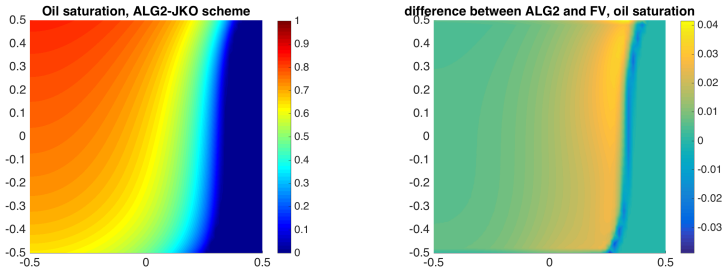
$$\mathcal{E}(s_0, s_1, s_2) := \sum_{i=0}^{2} \int_\Omega \Psi_i s_i + \frac{\alpha_1}{2} \int_\Omega s_1^2 + \frac{\alpha_2}{2} \int_\Omega s_2^2 + \int_\Omega \chi_{\boldsymbol{\Delta}}(s_0, s_1, s_2),$$

and we denote accordingly, for $\boldsymbol{x} \in \Omega$ and $\boldsymbol{c} = (c_0, c_1, c_2) \in \mathbb{R}^3$

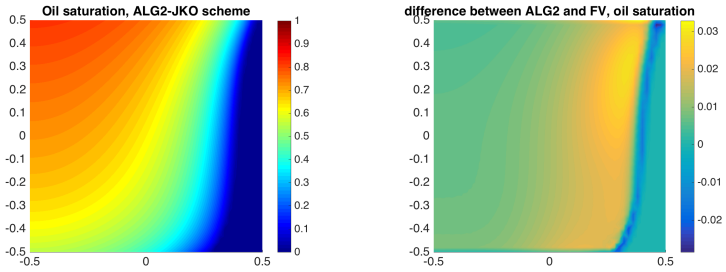
$$E_\tau(\boldsymbol{x}, \boldsymbol{c}) := \sum_{i=0}^{2} \tau \Psi_i(\boldsymbol{x}) c_i + \frac{\tau \alpha_1}{2} c_1^2 + \frac{\tau \alpha_2}{2} c_2^2 + \chi_{\boldsymbol{\Delta}}(\boldsymbol{c}).$$



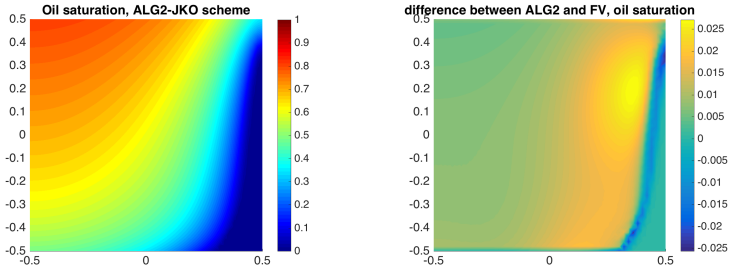
(A) $t = 2.5$



(B) $t = 5$



(C) $t = 7.5$



(D) $t = 10$

FIGURE 3. Oil saturation for the two-phase flow problem with Brooks-Corey capillary pressure function (31), $\alpha = 1$: numerical solution provided by the ALG2-JKO scheme (left) and difference between the ALG2-JKO approximate solution and the upstream mobility finite volume approximation solution (right).

Setting again $\overline{\boldsymbol{c}} = -\boldsymbol{\phi}^{k+1}(1,\boldsymbol{x}) + \tilde{\boldsymbol{s}}_1^k(\boldsymbol{x})$ and taking advantage of Moreau's identity (32), the second subproblem (22) of step 2 is equivalent to, for all $\boldsymbol{x} \in \mathbb{R}^d$,

$$\boldsymbol{c}^{k+1}(\boldsymbol{x}) = \overline{\boldsymbol{c}} - \mathrm{Prox}_{E_\tau(\boldsymbol{x},\cdot)}(\overline{\boldsymbol{c}}).$$

Evaluating the proximal operator $\tilde{\boldsymbol{c}} := \mathrm{Prox}_{E_\tau(\boldsymbol{x},\cdot)}(\overline{\boldsymbol{c}})$ is equivalent to solving

$$(\tilde{c}_1,\tilde{c}_2) = \underset{0\le c_i\le 1,\, 0\le c_1+c_2\le 1}{\mathrm{argmin}} \left\{ \sum_{i=1}^{2} \left(\frac{1}{2}|c_i - \overline{c}_i + \tau\Psi_i(\boldsymbol{x})|^2 + \tau\frac{\alpha_i}{2}c_i^2 \right) + \frac{1}{2}|c_1 + c_2 + \overline{c}_0 - \tau\Psi_0(\boldsymbol{x}) - 1|^2 \right\}, \tag{33}$$

with $\tilde{c}_0 = 1 - \tilde{c}_1 - \tilde{c}_2$. The solution (u_1,u_2) of the unconstrained version of (33) is explicitly given by

$$u_1 = \frac{(2+\tau\alpha_2)\gamma_1 - \gamma_2}{(2+\tau\alpha_1)(2+\tau\alpha_2) - 1} \text{ and } u_2 = \frac{(2+\tau\alpha_1)\gamma_2 - \gamma_1}{(2+\tau\alpha_1)(2+\tau\alpha_2) - 1},$$

where $\gamma_i := \overline{c}_i - \tau\Psi_i(\boldsymbol{x}) - \overline{c}_0 + \tau\Psi_0(\boldsymbol{x}) + 1$. If $(u_1,u_2) \in \boldsymbol{\Delta}^*$ then $(\tilde{c}_1,\tilde{c}_2) = (u_1,u_2)$ is the true solution of (33), and $\tilde{c}_0 = 1 - u_1 - u_2$. Otherwise, one should seek for the minimizer of (33) on the boundary $\partial\boldsymbol{\Delta}^* = \{s_1 = 0, 0 \le s_2 \le 1\} \cup \{0 \le s_1 \le 1, s_2 = 0\} \cup \{s_1 + s_2 = 1\}$. This leads to three easy minimization problems that can be again solved explicitly, and we omit the details. To conclude, the update of $\boldsymbol{c}^{k+1}(\boldsymbol{x})$ is given by $\boldsymbol{c}^{k+1}(\boldsymbol{x}) = \overline{\boldsymbol{c}} - \tilde{\boldsymbol{c}}$.

Figures 5–7 show the evolution of the three phases with quadratic capillarity potential. Again, the simulation with the ALG2-JKO scheme is carried out using a 50×50 discretization in space, with a single inner time step. There are 200 JKO steps ($\tau = 0.05$). The convergence of the augmented Lagrangian iterative method is rather slow: it took around 10 hours on a laptop to produce the results with `FreeFem++`. But because of the large viscosity ratio, Newton's method had severe difficulties to converge for the upstream mobility scheme. A very small time step ($\tau = 10^{-4}$) was needed, so that more that 2 days of computation on a cluster were needed to produce the results with `Matlab`. Concerning the upstream mobility finite volume scheme, we run the scheme on an unstructured Delaunday triangulation made of 5645 triangles. Once again, both methods produce similar results, as highlighted on the figures 5–7 below.

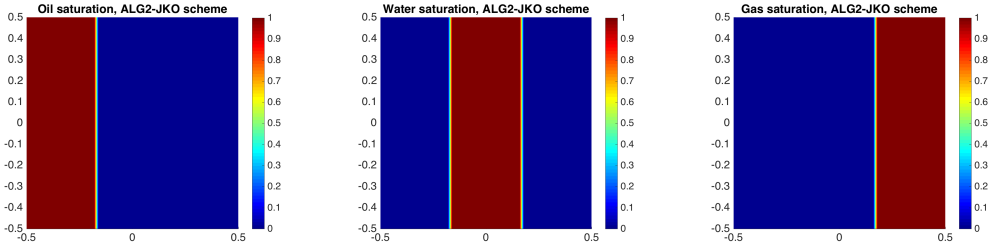
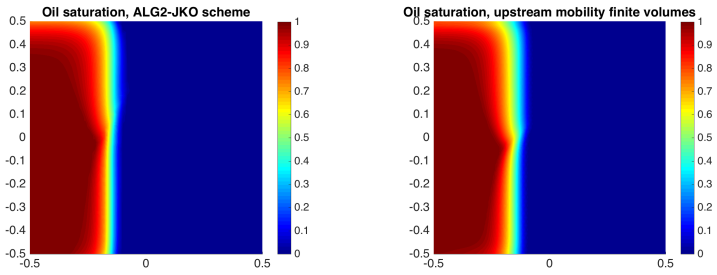
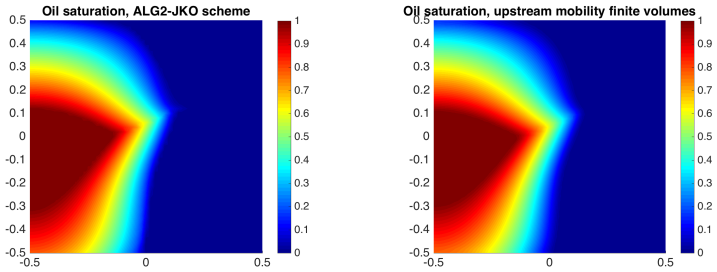


FIGURE 4. Initial oil (left), water (center) and gas (right) saturation profiles.

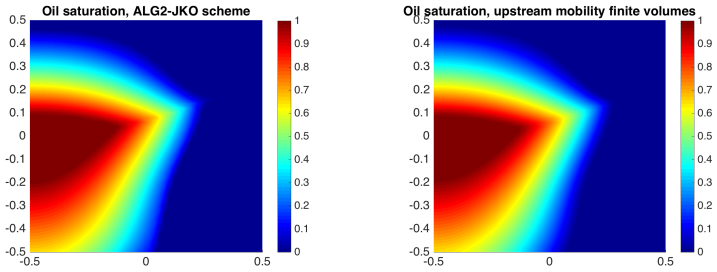
Due to the large viscosity ratios, two distinct time scale appear in the numerical results. Since water and gas have smaller mobilities, they move much faster than oil. This quick phenomenon is not well captured by the ALG2-JKO scheme. The interface between oil and gas is already almost horizontal at $t = 0.1$. This horizontal interface is captured by the finite volume scheme but not by the ALG2-JKO scheme that encounters difficulties to converge for the early time steps. The finite volume scheme also has difficulties to converge, enforcing us to consider very small time steps. Oil is much less mobile and its interface with the two other phases remains



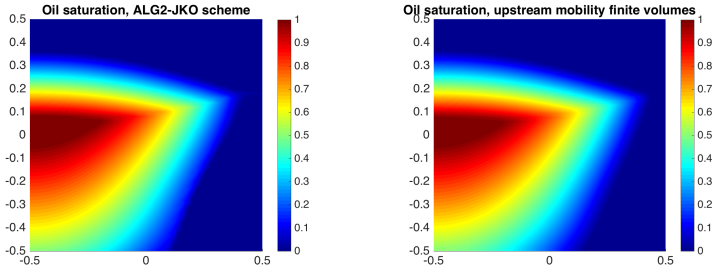
(A) t=0.1



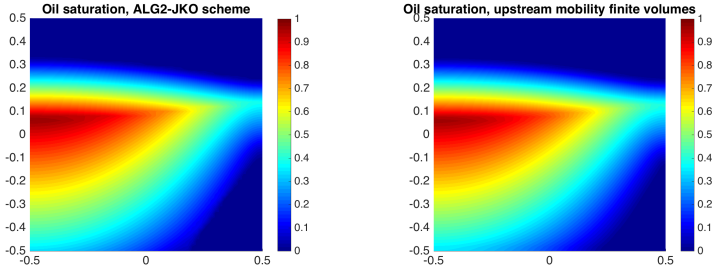
(B) t=1.25



(C) t=2.5

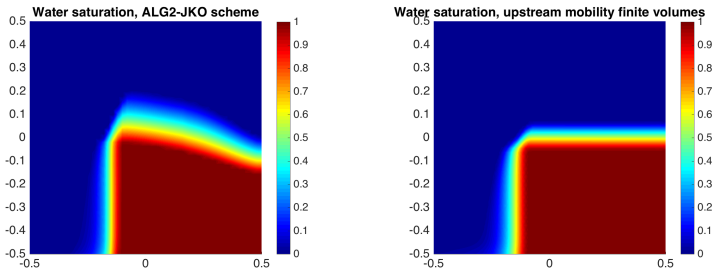


(D) t=5

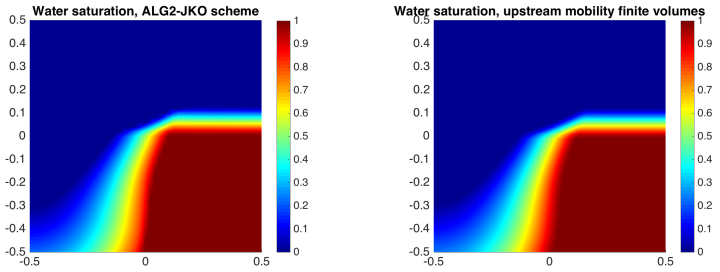


(E) t=10

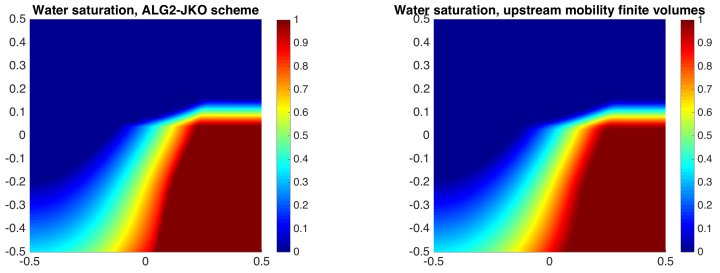
FIGURE 5. Three-phase flow, snapshots of the oil saturation profiles at different times: ALG2-JKO scheme (left) and upstream mobility finites volumes (right).



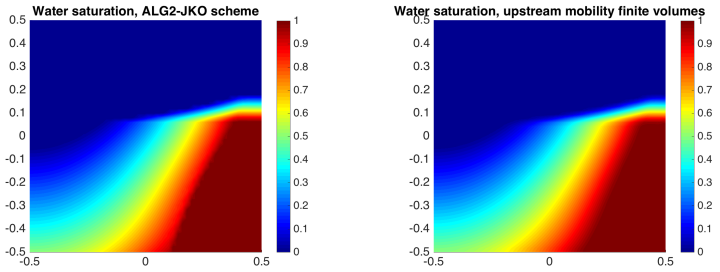
(A) t=0.1



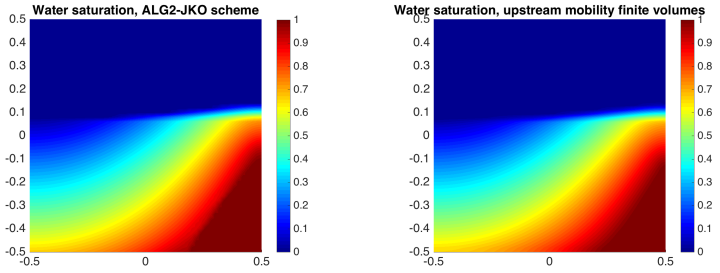
(B) t=1.25



(C) t=2.5

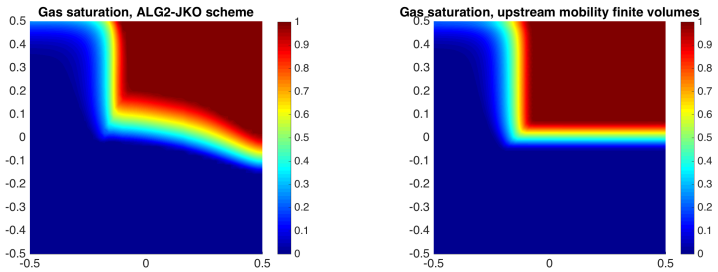


(D) t=5

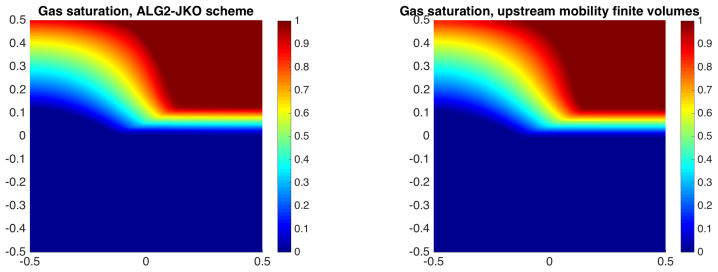


(E) t=10

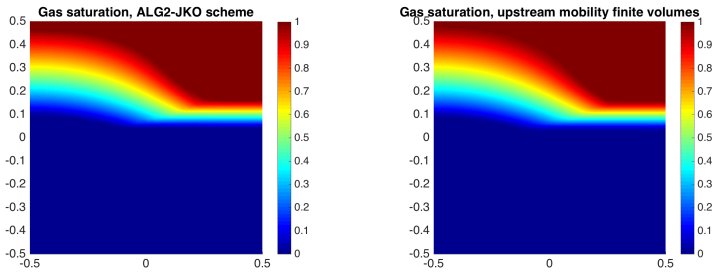
FIGURE 6. Three-phase flow, snapshots of the water saturation profiles at different times: ALG2-JKO scheme (left) and upstream mobility finites volumes (right).



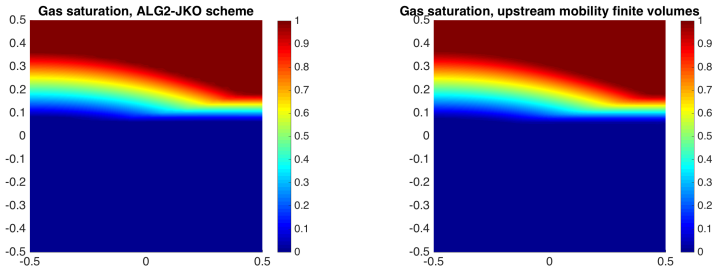
(A) t=0.1



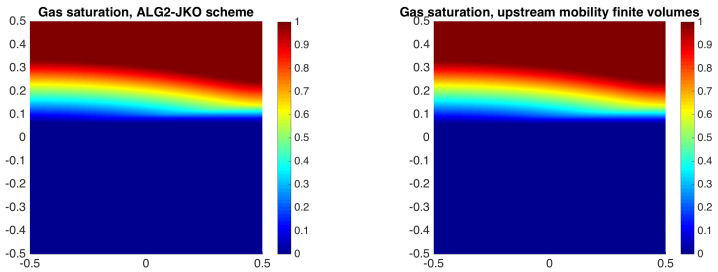
(B) t=1.25



(C) t=2.5



(D) t=5



(E) t=10

FIGURE 7. Three-phase flow, snapshots of the gas saturation profiles at different times: ALG2-JKO scheme (left) and upstream mobility finites volumes (right).

almost vertical at that time. Then oil evolves slowly towards its equilibrium state, that consists in a horizontal layer trapped between gas above and water below. This long time equilibrium is not yet reached for $t = 10$.

3.3. **Energy dissipation.** As already highlighted, both schemes dissipate the energy along time. The goal of this test case is to compare the energy dissipation. To this end, we consider a test case proposed in [9]. We consider a two-phase flow with oil ($i = 1$) and water ($i = 0$) with $\rho_1 = 0.87$, $\rho_0 = 1$, $\mu_1 = 10$ and $\mu_0 = 1$, while $\kappa = 1$ and $\omega = 1$. The capillary pressure law is given by

$$p_1 - p_0 = \pi_1(s_1) = \frac{s_1}{2},$$

so that the energy is defined by

$$\mathcal{E}(s_1) = \int_\Omega \left(\frac{(s_1)^2}{4} + s_1(\rho_0 - \rho_1)\mathbf{g} \cdot \boldsymbol{x} \right).$$

We consider the initial data $s_1^0(\boldsymbol{x}) = e^{-4|\boldsymbol{x}|^2}$. At equilibrium, the saturation s_1^∞ minimizes $\mathcal{E}$ under the constraints $s_1^\infty \in [0, 1]$ and

$$\int_\Omega s_1^\infty = \int_\Omega s_1^0. \tag{34}$$

It is therefore given by

$$\text{either } s_1^\infty \in \{0, 1\} \text{ or } \pi_1(s_1) = (\rho_1 - \rho_0)\mathbf{g} \cdot \boldsymbol{x} + \gamma, \tag{35}$$

the constant γ being fixed thanks to (34). Similar calculations can be performed in the discrete settings, both for the ALG2-JKO scheme and the finite volume scheme. Then one computes

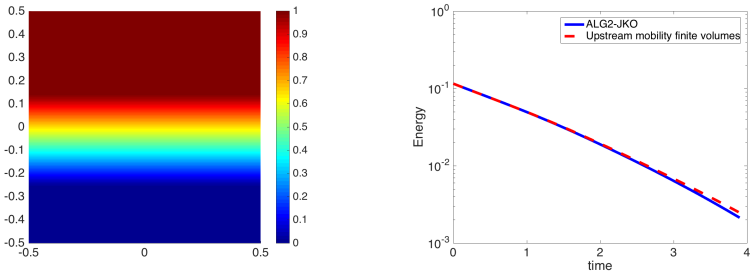


FIGURE 8. Left: The steady state (35). Right: The relative energies computed with the ALG2-JKO scheme (blue) and the finite volume scheme (red).

for both scheme the relative energy $\mathcal{E}(s_1) - \mathcal{E}(s_1^\infty) \geq 0$, that we plot as a function of time on Figure 8. The convergence towards the equilibrium appears to be exponential in both cases.

4. CONCLUSION

We proposed to apply the ALG2-JKO scheme of [6] to simulate multiphase porous media flows. The results have been compared to the widely used upstream mobility finite volume scheme. The ALG2-JKO scheme appears to be robust w.r.t. the capillary pressure function and overall w.r.t. the viscosity ratios. The method is parameter free (the only parameter r has a rather low influence and is chosen equal to 1 in the computations) and is unconditionally converging whatever the time step. This is a great advantage when compared to the Newton method that may require very small time steps in presence of large viscosity ratios. Moreover, the ALG2-JKO scheme preserves the positivity of the saturations, the constraint on the sum

of the saturations, and it is locally conservative. Its main drawback concerns the restriction to linear mobility function so that formulas (15)–(16) hold (this can probably be extended to the non-physical case of concave mobilities [17] but we did not push into this direction). Finally, let us stress that the code depends only at stage (22) of the energy. Therefore, the extension of the ALG2-JKO approach to multiphase models with different energies (like for instance degenerate Cahn-Hilliard models [32, 12]) is not demanding once the code is written. A natural extension to this work would be to add source terms corresponding for instance to production wells. This would for instance require to adapt the material of [23] to our context.

Acknowledgements. CC was supported by the French National Research Agency (ANR) through grant ANR-13-JS01-0007-01 (project GEOPOR) and ANR-11-LABX0007-01 (Labex CEMPI). LM was partially supported by the Portuguese Science Fundation through FCT grant PTDC/MAT-STA/0975/2014. TOG was partially supported by the Fonds de la Recherche Scientifique - FNRS under Grant MIS F.4539.16.

References

[1] A. 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.
[2] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.
[3] J. Bear and Y. Bachmat. *Introduction to modeling of transport phenomena in porous media*. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.
[4] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.
[5] J.-D. Benamou, Y. Brenier, and K. Guittet. Numerical analysis of a multi-phasic mass transport problem. In *Recent advances in the theory and applications of mass transport*, volume 353 of *Contemp. Math.*, pages 1–17. Amer. Math. Soc., Providence, RI, 2004.
[6] J.-D. Benamou, G. Carlier, and M. Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. In *Gradient flows: from theory to application*, volume 54 of *ESAIM Proc. Surveys*, pages 1–17. EDP Sci., Les Ulis, 2016.
[7] Y. Brenier and J. Jaffré. Upstream differencing for multiphase flow in reservoir simulation. *SIAM J. Numer. Anal.*, 28(3):685–696, 1991.
[8] Y. Brenier and M. Puel. Optimal multiphase transportation with prescribed momentum. *ESAIM Control Optim. Calc. Var.*, 8:287–343 (electronic), 2002.
[9] C. Cancès. Energy stable numerical methods for porous media flow type problems. HAL: hal-01719502, February 2018.
[10] C. Cancès, T. O. Gallouët, and L. Monsaingeon. The gradient flow structure of immiscible incompressible two-phase flows in porous media. *C. R. Acad. Sci. Paris Sér. I Math.*, 353:985–989, 2015.
[11] C. Cancès, T. O. Gallouët, and L. Monsaingeon. Incompressible immiscible multiphase flows in porous media: a variational approach. *Anal. PDE*, 10(8):1845–1876, 2017.
[12] C. Cancès, D. Matthes, and F. Nabet. A two-phase two-fluxes degenerate Cahn-Hilliard model as constrained Wasserstein gradient flow. HAL: hal-01665338, December 2017.
[13] C. Cancès and F. Nabet. Finite volume approximation of a degenerate immiscible two-phase flow model of Cahn–Hilliard type. In C. Cancès and P. Omnes, editors, *Finite Volumes for Complex Applications VIII - Methods and Theoretical Aspects : FVCA 8, Lille, France, June 2017*, number 199 in Proceedings in Mathematics and Statistics, pages 431–438, Cham, 2017. Springer International Publishing.
[14] H. Darcy. *Les fontaines publiques de la ville de Dijon*. Dalmont, Paris, 1856.
[15] E. De Giorgi. New problems on minimizing movements. In *Boundary value problems for partial differential equations and applications*, volume 29 of *RMA Res. Notes Appl. Math.*, pages 81–98. Masson, Paris, 1993.
[16] K. Deimling. *Nonlinear functional analysis*. Springer-Verlag, Berlin, 1985.
[17] J. Dolbeault, B. Nazaret, and G. Savaré. A new class of transport distances between measures. *Calc. Var. Partial Differential Equations*, 34(2):193–231, 2009.
[18] J. Droniou, R. Eymard, T. Gallouët, C. Guichard, and R. Herbin. The gradient discretisation method . A framework for the discretisation and numerical analysis of linear and non-linear elliptic and parabolic problems, November 2016.

[19] M. Erbar, K. Kuwada, and K.-T. Sturm. On the equivalence of the entropic curvature-dimension condition and Bochner's inequality on metric measure spaces. *Invent. Math.*, 201(3):993–1071, 2015.
[20] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020, 2000.
[21] R. Eymard, R. Herbin, and A. Michel. Mathematical study of a petroleum-engineering scheme. *M2AN Math. Model. Numer. Anal.*, 37(6):937–972, 2003.
[22] Michel Fortin and Roland Glowinski. *Augmented Lagrangian methods*, volume 15 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 1983. Applications to the numerical solution of boundary value problems, Translated from the French by B. Hunt and D. C. Spicer.
[23] T. Gallouët, M. Laborde, and L. Monsaingeon. An unbalanced Optimal Transport splitting scheme for general advection-reaction-diffusion problems. HAL: hal-01508911, 2017.
[24] F. Hecht. New development in FreeFEM++. *J. Numer. Math.*, 20(3-4):251–265, 2012.
[25] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.
[26] G. Legendre and G. Turinici. Second-order in time schemes for gradient flows in Wasserstein and geodesic metric spaces. *C. R. Acad. Sci. Paris Sér. I Math.*, 353(3):345-353, 2017.
[27] S. Lisini. Nonlinear diffusion equations with variable coefficients as gradient flows in Wasserstein spaces. *ESAIM Control Optim. Calc. Var.*, 15(3):712–740, 2009.
[28] Christian Loeschcke. *On the relaxation of a variational principle for the motion of a vortex sheet in perfect fluid*. PhD thesis, Univ. Bonn, 2012.
[29] D. Matthes, R. J. McCann, and G. Savaré. A family of nonlinear fourth order equations of gradient flow type. *Comm. Partial Differential Equations*, 34(11):1352–1397, 2009.
[30] D. Matthes and S. Plazotta. A Variational Formulation of the BDF2 Method for Metric Gradient Flows. arXiv:1711.02935, 2017.
[31] R. J. McCann. A convexity principle for interacting gases. *Adv. Math.*, 128(1):153–179, 1997.
[32] F. Otto and W. E. Thermodynamically driven incompressible fluid mixtures. *J. Chem. Phys.*, 107(23):10177–10184, 1997.
[33] N. Papadakis, G. Peyré, and E. Oudet. Optimal transport with proximal splitting. *SIAM J. Imaging Sci.*, 7(1):212–238, 2014.
[34] D. W Peaceman. *Fundamentals of numerical reservoir simulation*, volume 6 of *Developments in Petroleum Science*. Elsevier, 1977.
[35] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications 87. Birkhäuser Basel, 1 edition, 2015.
[36] F. Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.
[37] C. Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

Clément Cancès: Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille (clement.cances@inria.fr)

Thomas Gallouët: Inria, Project team MOKAPLAN and Mathematics department, Université de Liège, Belgium, (thomas.gallouet@inria.fr)

Maxime Laborde: Department of Mathematics and Statistics, McGill University, Montreal, CANADA (maxime.laborde@mcgill.ca)

Léonard Monsaingeon: IECL Université de Lorraine, Nancy, FRANCE & GFM Universidade de Lisboa, Lisbon, PORTUGAL (leonard.monsaingeon@univ-lorraine.fr)

2.2.2 Variational finite volume scheme

Articles:

- **A variational finite volume scheme for Wasserstein gradient flows.** *Numerische Mathematik, Springer Verlag, 146 (3), pp 437 - 480* (2020). https://hal.science/hal-02189050. C.Cancès, Gallouët T.O., Todeschi. G
- **From geodesic extrapolation to a variational BDF2 scheme for Wasserstein gradient flows.** Under minor revision for *Mathematics of Computations* (2023) https://hal.science/hal-03790981 Gallouët T.O., Natale A. et Todeschi. G

Collaborators: The first paper was done in collaboration with Clément Cancès and Gabriele Todeschi. It was the starting point of Gabriele Todeschi's PhD Thesis: the goal was to build variational numerical finite volume scheme. The second paper is done with A. Natale and G. Todeschi. It is a follow up of G. Todeschi's PhD Thesis where we aim to build second order in time numerical scheme for Wasserstein Gradient flows.

Main contributions:

- We used a first optimize than discretize approach in order to built, for Wasserstein gradient flows, a finite volume scheme which is exactly the Euler-Lagrange condition of a discretized JKO scheme.
- We prove the convergence of this scheme under some hypothesis on the energy.
- We implemented this scheme for a wider class of energy/system of PDE and gave numerical evidence of convergence.
- In the second work we proposed a second order in time variational finite volume scheme. To do this we had to modify the JKO step of the previous paper.

The second paper has already been included in Section 2.1 since it also contains the work on Wasserstein extrapolation.

Research directions: With G. Todeschi and A. Natale we pursue our investigations into higher order variational numerical scheme for Wassertein gradient flows. Our main focus is now to build a second order in time scheme based on the metric extrapolation of Wasserstein geodesics, see Section 2.1 for more details on this notion.

A VARIATIONAL FINITE VOLUME SCHEME FOR WASSERSTEIN GRADIENT FLOWS

CLÉMENT CANCÈS, THOMAS O. GALLOUËT, AND GABRIELE TODESCHI

Abstract. We propose a variational finite volume scheme to approximate the solutions to Wasserstein gradient flows. The time discretization is based on an implicit linearization of the Wasserstein distance expressed thanks to Benamou-Brenier formula, whereas space discretization relies on upstream mobility two-point flux approximation finite volumes. The scheme is based on a first discretize then optimize approach in order to preserve the variational structure of the continuous model at the discrete level. It can be applied to a wide range of energies, guarantees non-negativity of the discrete solutions as well as decay of the energy. We show that the scheme admits a unique solution whatever the convex energy involved in the continuous problem, and we prove its convergence in the case of the linear Fokker-Planck equation with positive initial density. Numerical illustrations show that it is first order accurate in both time and space, and robust with respect to both the energy and the initial profile.

1. A strategy to approximate Wasserstein gradient flows

1.1. Generalities about Wasserstein gradient flows.

Given a convex and bounded open 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:

$$\begin{cases} \partial_t \varrho - \nabla\cdot(\varrho \nabla \frac{\delta\mathcal{E}}{\delta\rho}[\varrho]) = 0 & \text{in } Q_T = \Omega\times(0,T), \\ \varrho \nabla \frac{\delta\mathcal{E}}{\delta\rho}[\varrho]\cdot \boldsymbol{n} = 0 & \text{on } \Sigma_T = \partial\Omega\times(0,T), \\ \varrho(\cdot,0) = \rho^0 & \text{in } \Omega. \end{cases} \tag{1}$$

Equation (1) expresses the continuity equation for a time evolving density ϱ, starting from the initial condition ρ^0, convected by the velocity field $-\nabla\frac{\delta\mathcal{E}}{\delta\rho}[\varrho]$. The mixed boundary condition the system is subjected to represents a no flux condition across the boundary of the domain for the mass: the total mass is therefore preserved.

It is now well understood since the pioneering works of Otto [34, 52, 53] that equations of the form of (1) can be interpreted as the gradient flow in the Wasserstein space w.r.t. the energy $\mathcal{E}$ [2]. A gradient flow is an evolution stemming from an initial condition and evolving at each time following the steepest decreasing direction of a prescribed functional. Consider the space $\mathbb{P}(\Omega)$ of nonnegative measures defined on the bounded and convex domain Ω with prescribed total mass that are absolutely continuous w.r.t. the Lebesgue measure (hence $\mathbb{P}(\Omega)\subset L^1(\Omega;\mathbb{R}_+)$). The Wasserstein distance W_2 between two densities $\rho,\mu\in\mathbb{P}(\Omega)$ is the cost to transport one into the other in an optimal way with respect to the cost given by the squared euclidean distance, namely the optimization problem

$$W_2^2(\rho,\mu) = \min_{\gamma\in\Gamma(\rho,\mu)} \iint_{\Omega\times\Omega} |\boldsymbol{y}-\boldsymbol{x}|^2 d\gamma(\boldsymbol{x},\boldsymbol{y}), \tag{2}$$

with the set $\Gamma(\rho,\mu)$ of admissible transport plans given by

$$\Gamma(\rho,\mu)=\Big\{\gamma\in\mathbb{P}(\Omega\times\Omega):\gamma^1=\rho,\gamma^2=\mu\Big\},$$

where γ^1,γ^2 denote the first and second marginal measure, respectively.

A typical example of problem entering the framework of (1) is the linear Fokker-Planck equation

$$\partial_t\varrho=\Delta\varrho+\nabla\cdot(\varrho\nabla V)\quad\text{in } Q_T, \tag{3}$$

complemented with no-flux boundary conditions and an initial condition. In (3), $V\in W^{1,\infty}(\Omega)$ denotes a Lipschitz continuous exterior potential. In this case, the energy functional is

$$\mathcal{E}(\rho)=\int_\Omega[\rho\log\frac{\rho}{e^{-V}}-\rho+e^{-V}]\mathrm{d}\boldsymbol{x}. \tag{4}$$

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. Beside this simple example studied for instance in [34, 10], many problems have been proven to exhibit the same variational structure. Porous media flows [53, 38, 15], magnetic fluids [52], superconductivity [4, 3], crowd motions [47], aggregation processes in biology [22, 9], semiconductor devices modelling [36], or multiphase mixtures [18, 33] 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.

1.2. **JKO semi-discretization.** An intriguing question is how to solve numerically a gradient flow. Problem (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 [8, 16, 17, 13, 56, 51, 19]. Nevertheless, the aforementioned methods disregard the fact that the trajectory aims at optimizing the energy decay, in opposition to methods based on minimizing movement scheme (often called JKO scheme after [34]). This scheme can be thought as a generalization to the space $\mathbb{P}(\Omega)$ (the mass being defined by the initial data ρ^0) equipped with the metric W_2 of the backward Euler scheme and writes:

$$\begin{cases}\rho_\tau^0=\rho^0,\\ \rho_\tau^n\in\operatorname{argmin}_\rho\ \frac{1}{2\tau}W_2^2(\rho,\rho_\tau^{n-1})+\mathcal{E}(\rho).\end{cases} \tag{5}$$

The parameter τ is the time discretization step. Scheme (5) generates a sequence of measures $(\rho_\tau^n)_{n\geq1}$. Using this sequence it is possible to construct a time dependent measure by gluing them together in a piecewise constant (in time) fashion: $\rho_\tau(t)=\rho_\tau^n$, for $t\in(t^{n-1}=(n-1)\tau,t^n=n\tau]$. Under suitable assumptions on the functional $\mathcal{E}$, it is possible to prove the uniform convergence in time of this measure to weak solutions ϱ of (1) (see for instance [2] or [55]).

Lagrangian numerical methods appear to be very natural (especially in dimension 1) to approximate the Wasserstein distance and thus the solution to (5). This was already noticed in [37], and motivated numerous contributions, see for instance [45, 12, 46, 35, 23, 20, 39]. 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 [48] and Maas [42, 30, 25, 43, 31]. 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 2.3 does not enter their framework. Last but not least, let us

mention the so-called ALG2-JKO scheme [7, 14] where the optimization problem (5) is discretized and then solved thanks to an augmented Lagrangian iterative method. Our approach is close to the one of [7], with the goal to obtain a faster numerical solver.

Thanks to formal calculations, let us highlight the connection of the minimization problem involved at each step of (5) with a system coupling a forward in time conservation law with a backward in time Hamilton-Jacobi (HJ) equation. The problem can be rewritten thanks to Benamou-Brenier dynamic formulation of optimal transport [6] as

$$\inf_{\rho,\boldsymbol{v}} \frac{1}{2}\int_{t^{n-1}}^{t^n}\int_\Omega \rho|\boldsymbol{v}|^2 \mathrm{d}\boldsymbol{x}\mathrm{d}t + \mathcal{E}(\rho(t^n)), \tag{6}$$

where the density and velocity curves satisfy weakly

$$\begin{cases} \partial_t\rho + \nabla\cdot(\rho\boldsymbol{v}) = 0 & \text{in } \Omega\times(t^{n-1},t^n),\\ \rho\boldsymbol{v}\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega\times(t^{n-1},t^n),\\ \rho(t^{n-1}) = \rho_\tau^{n-1} & \text{in } \Omega. \end{cases} \tag{7}$$

The next value ρ_τ^n is chosen equal to $\rho(t^n)$ for the optimal ρ in (6)–(7). Using the momentum $\boldsymbol{m} = \rho\boldsymbol{v}$ instead of $\boldsymbol{v}$ as a variable, and incorporating the constraint (7) in (6) yields the saddle-point problem

$$\begin{aligned}\inf_{\rho,\boldsymbol{m}}\sup_{\phi} \int_{t^{n-1}}^{t^n}\int_\Omega \frac{|\boldsymbol{m}|^2}{2\rho}\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_{t^{n-1}}^{t^n}\int_\Omega (\rho\partial_t\phi + \boldsymbol{m}\cdot\nabla\phi)\mathrm{d}\boldsymbol{x}\mathrm{d}t \\ + \int_\Omega [\phi(t^{n-1})\rho_\tau^{n-1} - \phi(t^n)\rho(t^n)]\mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho(t^n)).\end{aligned} \tag{8}$$

We will refer to (8) as the primal problem. The dual problem is obtained by exchanging inf and sup in (8). Strong duality can be proven and the problem hence does not change. Optimizing first w.r.t. $\boldsymbol{m}$ leads to $\boldsymbol{m} = -\rho\nabla\phi$, so that the dual problem writes

$$\sup_{\phi}\inf_{\rho} \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\phi - \frac{1}{2}|\nabla\phi|^2)\rho\mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega [\phi(t^{n-1})\rho_\tau^{n-1} - \phi(t^n)\rho(t^n)]\mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho(t^n)). \tag{9}$$

Because of the first term in (9), the infimum is equal to $-\infty$ unless $-\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 \le 0$ a.e. in $\Omega\times(t^{n-1},t^n)$, with equality ρ-almost everywhere since $\rho \ge 0$. Moreover, optimizing w.r.t. $\rho(t^n)$ provides that $\phi(t^n) \le \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n)]$ with equality $\rho(t^n)$-almost everywhere. Hence the dual problem can be rewritten as

$$\sup_{\phi(t^{n-1})} \int_\Omega \phi(t^{n-1})\rho_\tau^{n-1}\mathrm{d}\boldsymbol{x} + \inf_{\rho(t^n)}\left[\mathcal{E}(\rho(t^n)) - \int_\Omega \phi(t^n)\rho(t^n)\mathrm{d}\boldsymbol{x}\right], \tag{10}$$

subject to the constraints

$$\begin{cases} -\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 \le 0 & \text{in } \Omega\times(t^{n-1},t^n),\\ \phi(t^n) \le \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n)] & \text{in } \Omega,\\ \phi(t^n) = \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n)] & \rho(t^n) \text{ a.e.} \end{cases} \tag{11}$$

On the one hand, the monotonicity of the backward HJ equation $-\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 = f$ with respect to its right-hand side $f \le 0$ implies that given $\phi(t^n)$, the solution (which exists) of $-\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 = 0$ gives a bigger value at $\phi(t^{n-1})$ and thus a better competitor for (10). On the other hand, in

order to saturate the final time constraints we use the monotonicity of the backward HJ equation $-\partial_t\phi + \frac{1}{2}|\nabla\phi|^2 = f$ with respect to its final time $\phi(t^n)$. Indeed let $(\bar\rho, \bar\phi)$ be a saddle point of (9) 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 $\varphi(t^n) = \frac{\delta\mathcal{E}}{\delta\rho}[\bar\rho(t^n)] \geq \bar\phi(t^n)$. In particular (11) gives $\bar\phi(t^n) = \varphi(t^n)$ $\rho(t^n)$-almost everywhere and the monotonicity of HJ implies $\varphi(t^{n-1}) \geq \bar\phi(t^{n-1})$. All together this inequalities yields

$$\begin{aligned}
&\int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\varphi - \frac{1}{2}|\nabla\varphi|^2)\bar\rho \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega [\phi(t^{n-1})\rho_\tau^{n-1} - \varphi(t^n)\bar\rho(t^n)]\mathrm{d}\boldsymbol{x} + \mathcal{E}(\bar\rho(t^n))\\
&\geq \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\bar\phi - \frac{1}{2}|\nabla\bar\phi|^2)\bar\rho \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega [\bar\phi(t^{n-1})\rho_\tau^{n-1} - \bar\phi(t^n)\bar\rho(t^n)]\mathrm{d}\boldsymbol{x} + \mathcal{E}(\bar\rho(t^n))\\
&= \sup_\phi \int_{t^{n-1}}^{t^n}\int_\Omega (\partial_t\phi - \frac{1}{2}|\nabla\phi|^2)\bar\rho \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega [\phi(t^{n-1})\rho_\tau^{n-1} - \phi(t^n)\bar\rho(t^n)]\mathrm{d}\boldsymbol{x} + \mathcal{E}(\bar\rho(t^n)).
\end{aligned}$$

Bearing in mind the optimality of $\bar\phi$, this last inequality is then an equality and the strong duality implies that $(\bar\rho, \varphi)$ is also a saddle point of (9). At the end of the day, the primal-dual optimality conditions of problem (5) finally amounts to the mean field game

$$\begin{cases} \partial_t\phi - \frac{1}{2}|\nabla\phi|^2 = 0, \\ \partial_t\rho - \nabla\cdot(\rho\nabla\phi) = 0, \end{cases} \quad \text{in } \Omega\times(t^{n-1},t^n), \text{ with} \quad \begin{cases} \rho(t^{n-1}) = \rho_\tau^{n-1}, \\ \phi(t^n) = \frac{\delta\mathcal{E}}{\delta\rho}[\rho(t^n)], \end{cases} \quad \text{in } \Omega. \tag{12}$$

The optimal ρ_τ^n of (5) is then equal to $\rho(t^n)$. The no-flux boundary condition reduces to $\nabla\phi\cdot\boldsymbol{n} = 0$ on $\partial\Omega\times(t^{n-1},t^n)$.

The approximation of the system (12) is a natural strategy to approximate the solution to (1). This approach was for instance at the basis of the works [7, 21]. These methods require a sub-time stepping to solve system (12) 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.

1.3. **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_τ^n is close to ρ_τ^{n-1}. Then owing to [57, Section 7.6] (see also [54]), the Wasserstein distance between two densities ρ and μ of $\mathbb{P}(\Omega)$ is close to some weighted H^{-1} distance, namely

$$\|\rho-\mu\|_{\dot H_\rho^{-1}} = W_2(\rho,\mu) + o(W_2(\rho,\mu)), \qquad \forall\rho,\mu\in\mathbb{P}(\Omega). \tag{13}$$

In the above formula, we denoted by

$$\|h\|_{\dot H_\rho^{-1}} = \left\{ \sup_\varphi \int_\Omega h\varphi\,\mathrm{d}\boldsymbol{x} \;\middle|\; \|\varphi\|_{\dot H_\rho^1} \leq 1 \right\}, \quad \text{with } \|\varphi\|_{\dot H_\rho^1} = \left(\int_\Omega \rho|\nabla\varphi|^2\mathrm{d}\boldsymbol{x}\right)^{1/2}, \tag{14}$$

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

$$\begin{cases} \rho - \mu - \nabla\cdot(\rho\nabla\psi) = 0 & \text{in } \Omega, \\ \nabla\psi\cdot\boldsymbol{n} = 0 & \text{on } \partial\Omega. \end{cases} \tag{15}$$

Indeed, in view of (14)–(15), there holds

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

with equality if $\varphi = \psi/\|\psi\|_{\dot{H}^1_\rho}$. Equation (15) can be thought as a linearization of the Monge-Ampère equation.

In view of (13), a natural idea is to replace the Wasserstein distance by the weighted $\dot{H}^{-1}_\rho$ norm in (5), leading to what we call the implicitly linearized JKO (or LJKO) scheme:

$$\rho_\tau^n \in \operatorname*{argmin}_{\rho\in\mathbb{P}(\Omega)} \frac{1}{2\tau}\left\|\rho-\rho_\tau^{n-1}\right\|^2_{\dot{H}^{-1}_\rho(\Omega)} + \mathcal{E}(\rho), \qquad n\geq 1. \tag{16}$$

The choice of an implicit weight ρ in (16) appears to be particularly important when $\{\rho_\tau^{n-1}=0\}$ has a non-empty interior set, which can not be properly invaded by the ρ_τ^n if one chooses the explicit (but computationally cheaper) weight ρ_τ^{n-1} as in [50]. Our time discretization is close to the one that was proposed very recently in [41] where the introduction on inner time stepping was also avoided. In [41], 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 $n\geq 1$, (16) 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 (6), and rewrite step n as:

$$\inf_{\rho,\boldsymbol{m}} \int_\Omega \frac{|\boldsymbol{m}|^2}{2\tau\rho}\mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho), \quad \text{subject to: } \begin{cases} \rho-\rho_\tau^{n-1}+\nabla\cdot\boldsymbol{m}=0 & \text{in } \Omega,\\ \boldsymbol{m}\cdot\boldsymbol{n}=0 & \text{on } \partial\Omega. \end{cases} \tag{17}$$

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

$$\inf_{\rho,\boldsymbol{m}}\sup_{\phi} \int_\Omega \frac{|\boldsymbol{m}|^2}{2\tau\rho}\mathrm{d}\boldsymbol{x} - \int_\Omega(\rho-\rho_\tau^{n-1})\phi\,\mathrm{d}\boldsymbol{x} + \int_\Omega \boldsymbol{m}\cdot\nabla\phi\,\mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho), \tag{18}$$

the supremum w.r.t. ϕ being $+\infty$ unless the constraint is satisfied. Problem (18) 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 (18) is equivalent to its dual problem where the inf and the sup have been swapped. Optimizing w.r.t. to $\boldsymbol{m}$ yields the optimality condition $\boldsymbol{m}=-\tau\rho\nabla\phi$, hence the problem reduces to

$$\sup_\phi \int_\Omega \rho_\tau^{n-1}\phi\,\mathrm{d}\boldsymbol{x} + \inf_\rho \int_\Omega(-\phi-\frac{\tau}{2}|\nabla\phi|^2)\rho\,\mathrm{d}\boldsymbol{x} + \mathcal{E}(\rho). \tag{19}$$

The problem is now strictly convex in ρ and concave in ϕ. Optimizing w.r.t. ρ leads to the optimality condition

$$\phi_\tau^n + \frac{\tau}{2}|\nabla\phi_\tau^n|^2 \leq \frac{\delta\mathcal{E}}{\delta\rho}[\rho_\tau^n], \tag{20}$$

with equality on $\{\rho_\tau^n>0\}$. In the above formula, ϕ_τ^n denote the optimal ϕ realizing the sup in (19). Similarly to what has been done in the previous section for the JKO scheme, it is possible to show again that saturating inequality (20) on $\{\rho_\tau^n=0\}$ 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 (16) write

$$\begin{cases} \phi_\tau^n + \frac{\tau}{2}|\nabla\phi_\tau^n|^2 = \frac{\delta\mathcal{E}}{\delta\rho}[\rho_\tau^n],\\ \dfrac{\rho_\tau^n-\rho_\tau^{n-1}}{\tau} - \nabla\cdot(\rho_\tau^n\nabla\phi_\tau^n)=0, \end{cases} \tag{21}$$

set on Ω, complemented with homogeneous Neumann boundary condition $\nabla\phi_\tau^n \cdot \boldsymbol{n} = 0$ on $\partial\Omega$. We can interpret (21) as the one step resolvent of the mean-field game (12). Both the forward in time continuity equation and the backward in time HJ equation are discretized thanks to one step of backward Euler scheme.

1.4. **Goal and organisation of the paper.** As already noted, most of the numerical methods based on backward Euler scheme disregard the optimal character of the trajectory $t \mapsto \varrho(t)$ of the exact solution to (1). Rather than discretizing directly the PDE (1), which can be thought as the Euler-Lagrange equation for the steepest descent of the energy, we propose to first discretize w.r.t. space the functional appearing in the optimization problem (16), and then to optimize. The corresponding Euler-Lagrange equations will then encode the optimality of the trajectory. The choice of the LJKO scheme (16) rather than the classical JKO scheme (5) is motivated by the fact that solving (21) is computationally affordable. Indeed, it merely demands to approximate two functions ρ_τ^n, ϕ_τ^n rather than time depending trajectories in function space as for the JKO scheme (12). This allows in particular to avoid inner time stepping as in [7, 21], 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 (21), 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 $\dot{H}^1_\rho$ norm. The system (21) thus admits a discrete counterpart (36). The derivation of the fully discrete Finite Volume scheme based on the LJKO time discretization is performed in Section 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, we show that our scheme converges in the case of the Fokker-Planck equation (3) under the assumption that the initial density is bounded from below by a positive constant. Even though we do not treat problem (1) in its full generality, this result shows the consistency of the scheme. Finally, Section 4 is devoted to numerical results, where our scheme is tested on several problems, including systems of equations of the type of (1).

2. A variational Finite Volume scheme

The goal of this section is to define the fully discrete scheme to solve (1), and to exhibit some important properties of the scheme. But at first, let us give some assumptions and notations on the mesh.

2.1. **Discretization of Ω.** The domain $\Omega \subset \mathbb{R}^d$ is assumed to be polygonal if $d = 2$ or polyhedral if $d = 3$. The specifications on the mesh are classical for TPFA Finite Volumes [27]. More precisely, 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.

(i) 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 $m_K > 0$.

(ii) 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.

(iii) 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 $\boldsymbol{n}_{KL}$ to $K|L$ outward w.r.t. K.

Cartesian grids, Delaunay triangulations or Voronoï tessellations are typical examples of admissible meshes in the above sense. We refer to [29] for a discussion on the need of such restrictive grids. Since no boundary fluxes appear in our problem, the boundary faces $\Sigma_{\text{ext}} = \{\sigma \subset \partial\Omega\}$ are not involved in our computations. Nonzeros fluxes may only occur across internal faces $\sigma \in \Sigma = \overline{\Sigma} \setminus \Sigma_{\text{ext}}$. 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. The transmissivity of the face $\sigma \in \Sigma$ is defined by $a_\sigma = m_\sigma/d_\sigma$.

The space $\mathbb{R}^{\mathcal{T}}$ is equipped with the scalar product

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

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

2.2. **Upstream weighted dissipation potentials.** Since the LJKO time discretization presented in Section 1.3 relies on weighted $\dot{H}^1_\rho$ and H^{-1}_ρ 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 [48]. 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 [32].

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

$$\mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho}; \boldsymbol{\phi}) = \frac{1}{2} \sum_{\substack{\sigma\in\Sigma \\ \sigma = K|L}} a_\sigma \rho_\sigma \left(\phi_K - \phi_L\right)^2 \geq 0, \tag{22}$$

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

$$\rho_\sigma = \begin{cases} \rho_K & \text{if } \phi_K > \phi_L, \\ \rho_L & \text{if } \phi_K < \phi_L, \end{cases} \qquad \forall \sigma = K|L \in \Sigma. \tag{23}$$

Because of the upwind choice of the mobility (23), the functional (22) 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 introduce the space $\mathbb{F}_{\mathcal{T}} \subset \mathbb{R}^{2\Sigma}$ of conservative fluxes. An element $\boldsymbol{F}$ of $\mathbb{F}_{\mathcal{T}}$ is made of two outward fluxes $F_{K\sigma}, F_{L\sigma}$ for each $\sigma = K|L \in \Sigma$, and one flux $F_{K\sigma}$ per boundary face $\sigma \in \Sigma_K$. We impose

the conservativity across each internal face

$$F_{K\sigma} + F_{L\sigma} = 0, \qquad \forall \sigma = K|L \in \Sigma. \tag{24}$$

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

$$\mathbb{F}_\mathcal{T} = \left\{ \boldsymbol{F} = (F_{K\sigma}, F_{L\sigma})_{\sigma=K|L\in\Sigma} \in \mathbb{R}^{2\Sigma} \;\middle|\; (24) \text{ holds} \right\}.$$

Now, we define the subspace

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

and

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h}) = \inf_{\boldsymbol{F}} \sum_{\sigma\in\Sigma} \frac{(F_\sigma)^2}{2\rho_\sigma} d_\sigma m_\sigma \geq 0, \qquad \forall \boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}, \tag{25}$$

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

$$h_K m_K = \sum_{\sigma\in\Sigma_K} m_\sigma F_{K\sigma}, \qquad \forall K \in \mathcal{T}. \tag{26}$$

In (25), ρ_σ denotes the upwind value w.r.t. $\boldsymbol{F}$, i.e.,

$$\rho_\sigma = \begin{cases} \rho_K & \text{if } F_{K\sigma} > 0, \\ \rho_L & \text{if } F_{L\sigma} > 0, \end{cases} \qquad \forall \sigma = K|L \in \Sigma. \tag{27}$$

In the case where some ρ_σ vanish, we adopt the following convention in (25) and in what follows:

$$\frac{(F_\sigma)^2}{2\rho_\sigma} = \begin{cases} 0 & \text{if } F_\sigma = 0 \text{ and } \rho_\sigma = 0, \\ +\infty & \text{if } F_\sigma > 0 \text{ and } \rho_\sigma = 0, \end{cases} \qquad \forall \sigma \in \Sigma.$$

Remark that this condition is similar to the one implicitly used in (8) and (17). Summing (26) over $K \in \mathcal{T}$ and using the conservativity across the edges (24), one notices that there is no $\boldsymbol{F} \in \mathbb{F}_\mathcal{T}$ satisfying (26) unless $\boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}$. But when $\boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}$, the minimization set in (25) 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 (25) 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 the 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 [24, 26]. The convergence proof carried out in Section 3 somehow shows the non-obvious consistency of this formula.

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

Lemma 2.1. *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*

$$\mathcal{A}_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{h}) = \sup_{\boldsymbol{\phi}} \langle \boldsymbol{h}, \boldsymbol{\phi} \rangle_\mathcal{T} - \mathcal{A}^*_\mathcal{T}(\boldsymbol{\rho}; \boldsymbol{\phi}), \qquad \forall \boldsymbol{h} \in \mathbb{R}_0^\mathcal{T}. \tag{28}$$

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

$$h_K m_K = \sum_{\substack{\sigma\in\Sigma_K\\ \sigma=K|L}} a_\sigma \rho_\sigma(\phi_K-\phi_L), \qquad \forall K\in\mathcal{T}, \tag{29}$$

such that

$$\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{h}) = \mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi}) = \frac{1}{2}\langle \boldsymbol{h},\boldsymbol{\phi}\rangle_{\mathcal{T}}. \tag{30}$$

Proof. Let $\boldsymbol{\rho}\geq \mathbf{0}$ be fixed. Incorporating the constraint (26) in (25), and using the definition of ρ_σ and the twice conservativity constraint (24), we obtain the saddle point primal problem

$$\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{h}) = \inf_{\boldsymbol{F}}\sup_{\boldsymbol{\phi}} \sum_{\substack{\sigma\in\Sigma\\ \sigma=K|L}} \left[\frac{((F_{K\sigma})^+)^2}{2\rho_K} + \frac{((F_{K\sigma})^-)^2}{2\rho_L}\right] m_\sigma d_\sigma \\ + \sum_{K\in\mathcal{T}} h_K\phi_K m_K - \sum_{\substack{\sigma\in\Sigma\\ \sigma=K|L}} m_\sigma F_{K\sigma}(\phi_K-\phi_L).$$

The functional in the right-hand side is convex and coercive w.r.t. $\boldsymbol{F}$ and linear w.r.t. $\boldsymbol{\phi}$, 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 w.r.t. $\boldsymbol{F}$, leading to

$$F_{K\sigma} = \rho_\sigma \frac{\phi_K-\phi_L}{d_\sigma}, \qquad \forall \sigma = K|L\in\Sigma.$$

Substituting $F_{K\sigma}$ by $\rho_\sigma\frac{\phi_K-\phi_L}{d_\sigma}$ in the dual problem leads to (28), while the constraint (26) turns to (29). 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 (28) is achieved, ensuring the existence of the corresponding discrete Kantorovitch potentials $\boldsymbol{\phi}$. Finally, multiplying (29) by the optimal ϕ_K and by 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 (28) 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 $\rho\mapsto\frac{1}{2}W_2^2(\rho,\mu)$ is convex, which is key in the study of Wasserstein gradient flows.

Lemma 2.2. *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 (28) that

$$\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\mu}-\boldsymbol{\rho}) = \sup_{\boldsymbol{\phi}}\langle\boldsymbol{\mu}-\boldsymbol{\rho},\boldsymbol{\phi}\rangle_{\mathcal{T}} - \mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho};\boldsymbol{\phi}). \tag{31}$$

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, whence it is convex. □

2.3. **A variational upstream mobility Finite Volume scheme.** The finite volume discretization replaces the functions ρ_τ^n, ϕ_τ^n at time step $n \geq 1$ defined on Ω with the vectors $\boldsymbol{\rho}^n \in \mathbb{R}_+^{\mathcal{T}}$ and $\boldsymbol{\phi}^n \in \mathbb{R}^{\mathcal{T}}$. In each cell K, the restriction of each of these functions is approximated by a single real number ρ_K^n, ϕ_K^n, which can be thought as its mean value located in the cell center $\boldsymbol{x}_K$. Given $\boldsymbol{\rho}^0 \in \mathbb{R}_+^{\mathcal{T}}$, the space $\mathbb{P}_{\mathcal{T}}$ which is the discrete counterpart of $\mathbb{P}(\Omega)$ is then 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 and 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:

$$\boldsymbol{\rho}^n \in \operatorname*{argmin}_{\boldsymbol{\rho} \in \mathbb{P}_{\mathcal{T}}} \frac{1}{\tau} \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}), \qquad n \geq 1. \tag{32}$$

A further characterization of the scheme is needed for its practical implementation, but the condensed expression (32) already provides crucial informations gathered in the following theorem. Note in particular that our scheme automatically preserves mass and the positivity since the solutions $(\boldsymbol{\rho}^n)_{n \geq 1}$ belong to $\mathbb{P}_{\mathcal{T}}$.

Theorem 2.3. *For all $n \geq 1$, there exists a unique solution $\boldsymbol{\rho}^n \in \mathbb{P}_{\mathcal{T}}$ to (32). Moreover, energy is dissipated along the time steps. More precisely,*

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) \leq \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \frac{1}{\tau} \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) \leq \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^{n-1}), \qquad \forall n \geq 1. \tag{33}$$

Proof. The functional $\boldsymbol{\rho} \mapsto \frac{1}{\tau} \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}) + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho})$ l.s.c. and strictly convex on the compact set $\mathbb{P}_{\mathcal{T}}$ in view of Lemma 2.2 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 (33) is obtained by choosing $\boldsymbol{\rho} = \boldsymbol{\rho}^{n-1}$ as a competitor in (32). □

In view of (31), and after rescaling the dual variable $\boldsymbol{\phi} \leftarrow \frac{\boldsymbol{\phi}}{\tau}$, solving (32) amounts to solve the saddle point problem

$$\inf_{\boldsymbol{\rho} \geq \mathbf{0}} \sup_{\boldsymbol{\phi}} \left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{\substack{\sigma \in \Sigma \\ \sigma = K|L}} a_\sigma \rho_\sigma (\phi_K - \phi_L)^2 + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}). \tag{34}$$

which is equivalent to its dual problem

$$\sup_{\boldsymbol{\phi}} \inf_{\boldsymbol{\rho} \geq \mathbf{0}} \left\langle \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \frac{\tau}{2} \sum_{\substack{\sigma \in \Sigma \\ \sigma = K|L}} a_\sigma \rho_\sigma (\phi_K - \phi_L)^2 + \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}). \tag{35}$$

Our strategy for the practical computation of the solution to (32) is to solve the system corresponding to the optimality conditions of (35). So far, we did not take advantage of the upwind choice of the mobility (23) (we only used the linearity of $(\boldsymbol{\rho}, \boldsymbol{\phi}) \mapsto (\rho_\sigma)_{\sigma \in \Sigma}$ in the proofs of Lemmas 2.1 and 2.2, 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 2.4. *The unique solution* $(\boldsymbol{\rho}^n, \boldsymbol{\phi}^n)$ *to system*

$$\begin{cases} m_K \phi_K^n + \dfrac{\tau}{2} \displaystyle\sum_{\sigma \in \Sigma_K} a_\sigma \big((\phi_K^n - \phi_L^n)^+\big)^2 = \dfrac{\partial \mathcal{E}_\mathcal{T}}{\partial \rho_K}(\boldsymbol{\rho}^n), \\ (\rho_K^n - \rho_K^{n-1}) m_K + \tau \displaystyle\sum_{\sigma \in \Sigma_K} a_\sigma \rho_\sigma^n (\phi_K^n - \phi_L^n) = 0, \end{cases} \qquad \forall K \in \mathcal{T}, \tag{36}$$

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 (35).

System (36) is the discrete counterpart of (21), whose derivation relied on the monotonicity of the inverse of the operator $\phi \mapsto \phi + \frac{\tau}{2}|\nabla \phi|^2$. Before proving Theorem 2.4, let us show that the space discretization preserves this property at the discrete level. To this end, we introduce the functional $\boldsymbol{\mathcal{G}} = (\mathcal{G}_K)_K \in C^1(\mathbb{R}^\mathcal{T}; \mathbb{R}^\mathcal{T})$ defined by

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

Lemma 2.5. *Given* $\boldsymbol{f} \in \mathbb{R}^\mathcal{T}$*, there exists a unique solution to* $\boldsymbol{\mathcal{G}}(\boldsymbol{\phi}) = \boldsymbol{f}$*, and it satisfies*

$$\min \boldsymbol{f} \le \boldsymbol{\phi} \le \max \boldsymbol{f}. \tag{37}$$

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

$$\boldsymbol{f} \ge \widetilde{\boldsymbol{f}} \quad \Longrightarrow \quad \boldsymbol{\phi} \ge \widetilde{\boldsymbol{\phi}}. \tag{38}$$

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

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

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

$$\frac{\tau}{2m_K} \sum_{\substack{\sigma \in \Sigma_{K^*} \\ \sigma = K^*|L}} a_\sigma \left((\phi_{K^*} - \phi_L)^+\right)^2 \le \frac{\tau}{2m_K} \sum_{\substack{\sigma \in \Sigma_{K^*} \\ \sigma = K^*|L}} a_\sigma \left((\tilde{\phi}_{K^*} - \tilde{\phi}_L)^+\right)^2. \tag{39}$$

Recall $\boldsymbol{f} \ge \widetilde{\boldsymbol{f}}$ so $\mathcal{G}_{K^*}(\boldsymbol{\phi}) \ge \mathcal{G}_{K^*}(\tilde{\boldsymbol{\phi}})$ together with (39) it yields $\phi_{K^*} \ge \tilde{\phi}_{K^*}$. Finally as in K^* the difference $\phi_K - \tilde{\phi}_K$ is minimal, we obtain $\phi_K \ge \tilde{\phi}_K$ for all $K \in \mathcal{T}$. The uniqueness of the solution $\boldsymbol{\phi}$ of $\boldsymbol{\mathcal{G}}(\boldsymbol{\phi}) = \boldsymbol{f}$ follows directly. The maximum principle (37) is also a straightforward consequence of (38) as one can compare $\boldsymbol{\phi}$ to $(\min \boldsymbol{f})\mathbf{1}$ and $(\max \boldsymbol{f})\mathbf{1}$ which are fixed points of $\boldsymbol{\mathcal{G}}$. Finally, existence follows from Leray-Schauder fixed-point theorem [40] as the bounds (37) are uniform whatever $\tau \ge 0$. □

With Lemma 2.5 at hand, we can now prove Theorem 2.4.

Proof of Theorem 2.4. Uniqueness of the solution $\boldsymbol{\rho}^n$ to (32) was already proved in Theorem 2.3. Owing to (33), $\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n)$ is finite. So Lemma 2.1 ensures the existence of a discrete Kantorovitch potential $\widetilde{\boldsymbol{\phi}}^n$ satisfying (after a suitable rescaling by τ^{-1})

$$(\rho_K^n - \rho_K^{n-1}) m_K + \tau \sum_{\sigma \in \Sigma_K} a_\sigma \rho_\sigma^n (\widetilde{\phi}_K^n - \widetilde{\phi}_L^n) = 0, \qquad \forall K \in \mathcal{T}. \tag{40}$$

The above condition is the optimality condition w.r.t. $\boldsymbol{\phi}$ in (35). To compute the optimality condition w.r.t. $\boldsymbol{\rho}$ in (35) let us rewrite the objective using the definition of ρ_σ and $\boldsymbol{\mathcal{G}}$:

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

Thus (35) rewrites

$$\sup_{\boldsymbol{\phi}} \inf_{\boldsymbol{\rho} \geq \mathbf{0}} \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \langle \boldsymbol{\rho}, \boldsymbol{\mathcal{G}}(\boldsymbol{\phi}) \rangle_{\mathcal{T}} . \tag{41}$$

Denote by

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

Using (41) the optimality conditions w.r.t. $\boldsymbol{\rho}$ of (35) thus reads

$$m_K \widetilde{\phi}_K^n + \frac{\tau}{2} \sum_{\sigma \in \Sigma_{0,K}} a_\sigma \big((\widetilde{\phi}_K^n - \widetilde{\phi}_L^n)^+ \big)^2 = \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n), \qquad \forall K \in \mathcal{P}^n \tag{42}$$

and

$$m_K \widetilde{\phi}_K^n + \frac{\tau}{2} \sum_{\sigma \in \Sigma_{0,K}} a_\sigma \big((\widetilde{\phi}_K^n - \widetilde{\phi}_L^n)^+ \big)^2 \leq \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n), \qquad \forall K \in \mathcal{Z}^n. \tag{43}$$

By definition, $(\boldsymbol{\rho}^n, \widetilde{\boldsymbol{\phi}}^n)$ is a saddle point of (35), so equivalently of (41) and by strong duality is it also a saddle point of

$$\inf_{\boldsymbol{\rho} \geq \mathbf{0}} \sup_{\boldsymbol{\phi}} \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \langle \boldsymbol{\rho}, \boldsymbol{\mathcal{G}}(\boldsymbol{\phi}) \rangle_{\mathcal{T}} . \tag{44}$$

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

$$\sup_{\boldsymbol{\phi}} \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi} \right\rangle_{\mathcal{T}} - \langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{G}}(\boldsymbol{\phi}) \rangle_{\mathcal{T}} . \tag{45}$$

To prove Theorem 2.4, we have to prove that, given $\boldsymbol{\rho}^n$, we can saturate the inequality in both (42) and (43) while preserving the optimality in (45). Lemma 2.5 gives the existence of a solution $\boldsymbol{\phi}^n \in \mathbb{R}^{\mathcal{T}}$ to

$$\boldsymbol{\mathcal{G}}(\boldsymbol{\phi}^n) = \left(\frac{1}{m_K} \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}^n) \right)_{K \in \mathcal{T}}. \tag{46}$$

Note that (42) implies

$$\boldsymbol{\mathcal{G}}_K(\boldsymbol{\phi}^n) = \boldsymbol{\mathcal{G}}_K(\widetilde{\boldsymbol{\phi}}^n) \qquad \forall K \in \mathcal{P}^n$$

so

$$\langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{G}}(\boldsymbol{\phi}^n) \rangle_{\mathcal{T}} = \left\langle \boldsymbol{\rho}^n, \boldsymbol{\mathcal{G}}(\widetilde{\boldsymbol{\phi}}^n) \right\rangle_{\mathcal{T}}. \tag{47}$$

The combination of (42) and (43) is exactly $\boldsymbol{\mathcal{G}}(\boldsymbol{\phi}^n) \geq \boldsymbol{\mathcal{G}}(\widetilde{\boldsymbol{\phi}}^n)$, thus Lemma 2.5 gives $\boldsymbol{\phi}^n \geq \widetilde{\boldsymbol{\phi}}^n$. Consequently,

$$\left\langle \boldsymbol{\rho}^{n-1}, \boldsymbol{\phi}^n \right\rangle_{\mathcal{T}} \geq \left\langle \boldsymbol{\rho}^{n-1}, \widetilde{\boldsymbol{\phi}}^n \right\rangle_{\mathcal{T}} \tag{48}$$

since $\boldsymbol{\rho}^{n-1} \geq \mathbf{0}$. Incorporating (47) and (48) in (45) shows that $\boldsymbol{\phi}^n$ is a better competitor than $\widetilde{\boldsymbol{\phi}}^n$. Therefore, $(\boldsymbol{\rho}^n, \boldsymbol{\phi}^n)$ is a saddle point of (35) and satisfies (36). Finally, owing to Lemma 2.5, the solution $\boldsymbol{\phi}^n$ to (46) is unique, concluding the proof of Theorem 2.4. □

2.4. **Comparison with the classical backward Euler discretization.** The scheme (32) is based on a "first discretize then optimize" approach. We have built a discrete counterpart of $\frac{1}{2}W_2^2$ and a discrete energy $\mathcal{E}_{\mathcal{T}}$, then the discrete dynamics is chosen in an optimal way by (32). In opposition, the continuous equation (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 (1), leading to what we call a "first optimize then discretize" approach. It is classical for the semi-discretization in time of (1) to use a backward Euler scheme. If one combines this technic with upstream weighted Finite Volumes, we obtain the following fully discrete scheme:

$$(\check{\rho}_K^n - \rho_K^{n-1}) m_K + \tau \sum_{\sigma \in \Sigma_K} a_\sigma \check{\rho}_\sigma^n (\check{\phi}_K^n - \check{\phi}_L^n) = 0, \quad \text{with} \quad \check{\phi}_K^n = \frac{1}{m_K} \frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\check{\boldsymbol{\rho}}^n), \qquad \forall K \in \mathcal{T}. \tag{49}$$

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 (32): it is mass and positivity preserving as well as energy diminishing.

Proposition 2.6. *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 (49), which satisfies*

$$\mathcal{E}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n) + \frac{1}{\tau} \mathcal{A}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n; \boldsymbol{\rho}^{n-1} - \check{\boldsymbol{\rho}}^n) + \tau \mathcal{A}_{\mathcal{T}}^*(\check{\boldsymbol{\rho}}^n; \check{\boldsymbol{\phi}}^n) \leq \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^{n-1}). \tag{50}$$

Proof. Summing (49) 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 $\check{\phi}_{K^\star}^n \geq \check{\phi}_K^n$ for all $K \in \mathcal{K}^n$. Then it follows from the upwind choice of the mobility in (49) that

$$\sum_{\substack{\sigma \in \Sigma_{K^\star} \\ \sigma = K|L}} a_\sigma \check{\rho}_\sigma^n (\check{\phi}_{K^\star}^n - \check{\phi}_L^n) \leq 0,$$

so that $\check{\rho}^n_{K^\star} \geq \rho^{n-1}_{K^\star} \geq 0$, showing a contradiction. Therefore, $\mathcal{K}^n = \emptyset$ and $\check{\boldsymbol{\rho}}^n \geq \mathbf{0}$. These two *a priori* estimates (mass and positivity preservation) are uniform w.r.t. $\tau \geq 0$, thus they are sufficient to prove the existence of a solution $(\check{\boldsymbol{\rho}}^n, \check{\boldsymbol{\phi}}^n)$ to (49) thanks to a topological degree argument [40].

Let us now turn to the derivation of the energy / energy dissipation inequality (50). Multiplying (49) by $\check{\phi}^n_K$ 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 $\langle \check{\boldsymbol{\rho}}^n - \boldsymbol{\rho}^{n-1}, \check{\boldsymbol{\phi}}^n \rangle_{\mathcal{T}} \geq \mathcal{E}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n) - \mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^{n-1})$. Thus to prove (50), it remains to check that

$$\frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n; \boldsymbol{\rho}^{n-1} - \check{\boldsymbol{\rho}}^n) = \tau \mathcal{A}^*_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n; \check{\boldsymbol{\phi}}^n) = \frac{1}{\tau}\mathcal{A}^*_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n; \tau\check{\boldsymbol{\phi}}^n). \tag{51}$$

In view of (29), $\tau\check{\boldsymbol{\phi}}^n$ 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 (51) holds as a consequence of (30). □

Next proposition provides a finer energy / energy dissipation estimate than (33), which can be thought as discrete counterpart to the energy / energy dissipation inequality (EDI) which is a characterization of generalized gradient flows [2, 48].

Proposition 2.7. *Given $\boldsymbol{\rho}^{n-1} \in \mathbb{P}_{\mathcal{T}}$, let $\boldsymbol{\rho}^n$ be the unique solution to (32) and let $\check{\boldsymbol{\rho}}^n$ be a solution to (49), then*

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

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 (32), thus

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^n) + \frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) \leq \mathcal{E}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n) + \frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\check{\boldsymbol{\rho}}^n; \boldsymbol{\rho}^{n-1} - \check{\boldsymbol{\rho}}^n). \tag{52}$$

Combining this with (50) and bearing in mind that $\frac{1}{\tau}\mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) = \tau \mathcal{A}^*_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\phi}^n)$ thanks to (30), we obtain the desired inequality (52). □

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). 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 Section 2.1 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 $m \geq 1$,

$$h_K \leq \zeta d_\sigma \leq \zeta^2 h_K, \qquad \forall \sigma \in \Sigma_K,\ \forall K \in \mathcal{T}_m, \tag{53a}$$

$$\operatorname{dist}(\boldsymbol{x}_K, \overline{K}) \leq \zeta h_K, \qquad \forall K \in \mathcal{T}_m, \tag{53b}$$

and

$$\sum_{\sigma \in \sigma_K} m_{\Delta_\sigma} \leq \zeta m_K, \qquad \forall K \in \mathcal{T}_m. \tag{53c}$$

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 $V \in C^2(\overline{\Omega})$, we define the discrete counterpart of the energy (4) by

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

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

$$\frac{\partial \mathcal{E}_{\mathcal{T}}}{\partial \rho_K}(\boldsymbol{\rho}) = m_K(\log(\rho_K) + V_K) \qquad \forall K \in \mathcal{T}. \tag{54}$$

Given an initial condition $\varrho^0 \in \mathbb{P}(\Omega)$ 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 $\boldsymbol{\rho}^0 = \left(\rho_K^0\right)_{K\in\mathcal{T}}$ defined by

$$\rho_K^0 = \frac{1}{m_K}\int_K \varrho^0 \mathrm{d}\boldsymbol{x} \geq 0, \qquad \forall K \in \mathcal{T}. \tag{55}$$

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 (32) still admits a unique solution $\boldsymbol{\rho}^n$ for all $n \geq 1$ thanks to Theorem 2.3, since its proof does not use the differentiability of the energy. Thanks to the conservativity of the scheme and definition (55) of $\boldsymbol{\rho}^0$, one has

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

Let us show that $\boldsymbol{\rho}^n > \mathbf{0}$ for all $n \geq 1$. To this end, we proceed as in [55, Lemma 8.6].

Lemma 3.1. *Assume that ϱ^0 has positive mass, then the iterated solutions $(\boldsymbol{\rho}^n)_{n\geq 1}$ to scheme (32) 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 $n \geq 1$:*

$$\phi_K^n + \frac{\tau}{2m_K}\sum_{\sigma=K|L\in\Sigma_K} a_\sigma \left((\phi_K^n - \phi_L^n)^+\right)^2 = \log(\rho_K^n) + V_K, \tag{56}$$

$$(\rho_K^n - \rho_K^{n-1})m_K + \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n(\phi_K^n - \phi_L^n) = 0. \tag{57}$$

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 (32), there holds

$$\begin{aligned}\sum_{K\in\mathcal{T}} m_K\left[\rho_K^n \log \rho_K^n - \rho_{K,\epsilon}^n \log \rho_{K,\epsilon}^n\right] \leq \sum_{K\in\mathcal{T}} m_K\left(\rho_{K,\epsilon}^n - \rho_K^n\right)V_K \\ + \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}_\epsilon^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}_\epsilon^n) - \mathcal{A}_{\mathcal{T}}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n).\end{aligned} \tag{58}$$

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 (58) can be overestimated by

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

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 $\rho \mapsto \log\rho$,

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

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 (36) hold, which rewrites as (56)–(57) thanks to (54). The uniqueness of the discrete Kantorovitch potential $\boldsymbol{\phi}^n$ for all $n \ge 1$ is then provided by Theorem 2.4. □

Lemma 3.1 allows to define two functions $\rho_{\mathcal{T},\tau}$ and $\phi_{\mathcal{T},\tau}$ by setting

$$\rho_{\mathcal{T},\tau}(\boldsymbol{x},t) = \rho_K^n, \quad \phi_{\mathcal{T},\tau}(\boldsymbol{x},t) = \phi_K^n \quad \text{if } (\boldsymbol{x},t) \in K \times (t^{n-1}, t^n].$$

It follows from the conservativity of the scheme and definition (55) of $\boldsymbol{\rho}^0$ that

$$\int_\Omega \rho_{\mathcal{T},\tau}(\boldsymbol{x},t^n)\mathrm{d}\boldsymbol{x} = \langle \boldsymbol{\rho}^n, \mathbf{1}\rangle_\mathcal{T} = \langle \boldsymbol{\rho}^0, \mathbf{1}\rangle_\mathcal{T} = \int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x} > 0,$$

so that $\rho_{\mathcal{T},\tau}(\cdot,t)$ belongs to $\mathbb{P}(\Omega)$ for all $t \in (0,T)$.

The goal of this section is to prove the following theorem.

Theorem 3.2. *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 (53) 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) corresponding to the initial data ϱ^0.*

The proof is based on compactness arguments. At first in Section 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 3.2. Finally, we identify the limit value as a weak solution in Section 3.3.

Remark 3.3. *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.*

3.1. **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 3.4. *Given $\varrho^0 \in \mathbb{P}(\Omega)$ such that $\mathcal{E}(\varrho^0) < +\infty$, and let $\boldsymbol{\rho}^0$ be defined by (55), 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 (33) 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

$$\mathcal{E}_{\mathcal{T}}(\boldsymbol{\rho}^0) = T_1 + T_2 + T_3 \tag{59}$$

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 (55) of $\boldsymbol{\rho}^0$ and Jensen's inequality that

$$T_1 \le \int_\Omega [\varrho^0 \log \varrho^0 - \varrho^0] \mathrm{d}\boldsymbol{x}. \tag{60}$$

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,

$$T_3 = \int_\Omega e^{-V}\mathrm{d}\boldsymbol{x} + \sum_{K\in\mathcal{T}} m_K[e^{-V(\boldsymbol{x}_K)} - e^{-V(\widetilde{\boldsymbol{x}}_K)}] \le \int_\Omega e^{-V}\mathrm{d}\boldsymbol{x} + e^{\|V^-\|_\infty}\|\nabla V\|_\infty \mathrm{diam}(\Omega). \tag{61}$$

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,

$$T_2 = \int_\Omega \varrho^0 V \mathrm{d}\boldsymbol{x} + \sum_{K\in\mathcal{T}} m_K \rho_K^0 [V(\boldsymbol{x}_K) - V(\check{\boldsymbol{x}}_K)] \le \int_\Omega \varrho^0 V \mathrm{d}\boldsymbol{x} + \|\nabla V\|_\infty \mathrm{diam}(\Omega) \int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x}. \tag{62}$$

Combining (60)–(62) in (59) 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 3.5. *Using the convention $\log(0) = -\infty$, one has*

$$\min_{K\in\mathcal{T}} \left[\log(\rho_K^n) + V_K\right] \ge \min_{K\in\mathcal{T}} \left[\log(\rho_K^{n-1}) + V_K\right], \qquad \forall n \ge 1.$$

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 (56) 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 (57) ensures that $\rho_{K_\star}^n \ge \rho_{K_\star}^{n-1}$. On the other hand, since

$$\sum_{\sigma = K_\star | L \in \Sigma_{K_\star}} a_\sigma \left((\phi_{K_\star}^n - \phi_L^n)^+\right)^2 = 0,$$

the discrete HJ equation (56) provides that

$$\phi_{K^\star}^n = \log(\rho_{K_\star}^n) + V_{K_\star} = \min_{K\in\mathcal{T}} \left[\log(\rho_K^n) + V_K\right] \ge \log(\rho_{K_\star}^{n-1}) + V_{K_\star} \ge \min_{K\in\mathcal{T}} \left[\log(\rho_K^{n-1}) + V_K\right].$$

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 3.6. *Let $(\boldsymbol{\rho}^n, \boldsymbol{\phi}^n)$ be the iterated solution to (36), then*

$$\sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n (\phi_K^n - \phi_L^n)^2 \leq C_1. \tag{63}$$

Moreover, if $\varrho^0 \geq \rho_\star \in (0, +\infty)$, then there exists C_2 (depending on Ω, V and ϱ^0) such that

$$\sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma (\phi_K^n - \phi_L^n)^2 \leq C_2. \tag{64}$$

Proof. Since $\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) \geq 0$ for all $\boldsymbol{\rho} \in \mathbb{P}_\mathcal{T}$, summing (33) 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 (30), the left-hand side rewrites

$$\sum_{n=1}^{N} \frac{1}{\tau} \mathcal{A}_\mathcal{T}(\boldsymbol{\rho}^n; \boldsymbol{\rho}^{n-1} - \boldsymbol{\rho}^n) = \sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n (\phi_K^n - \phi_L^n)^2,$$

so that it only remains to use Lemma 3.4 to recover (63).

Finally, if ϱ^0 is bounded from below by some $\rho_\star > 0$, then Lemma 3.5 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 (64) 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

$$\rho_{\Sigma,\tau}(\boldsymbol{x}, t) = \begin{cases} \rho_\sigma^n & \text{if } (\boldsymbol{x}, t) \in \Delta_\sigma \times (t^{n-1}, t^n], \quad \sigma \in \Sigma, \\ \rho_K^n & \text{if } (\boldsymbol{x}, t) \in K \setminus \left(\bigcup_{\sigma\in\Sigma_K} \Delta_\sigma\right) \times (t^{n-1}, t^n], \quad K \in \mathcal{T}. \end{cases} \tag{65}$$

Lemma 3.7. *There exists C_3 depending only on ζ and ϱ^0 such that*

$$\int_\Omega \rho_{\Sigma,\tau}(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \leq C_3, \qquad \forall t > 0. \tag{66}$$

Moreover, there exists C_4 depending only on ζ, V and ϱ^0 such that

$$\int_\Omega \rho_{\Sigma,\tau}(\boldsymbol{x}, t) \log \rho_{\Sigma,\tau}(\boldsymbol{x}, t) \mathrm{d}\boldsymbol{x} \leq C_4, \qquad \forall t > 0. \tag{67}$$

Proof. Since $t \mapsto \rho_{\Sigma,\tau}(\cdot, t)$ is piecewise constant, it suffices to check that the above properties at each t^n, $1 \leq n \leq N$. In view of the definition of $\rho_{\Sigma,\tau}$, one has

$$\int_\Omega \rho_{\Sigma,\tau}(\boldsymbol{x}, t^n) \mathrm{d}\boldsymbol{x} \leq \sum_{K\in\mathcal{T}} \sum_{\sigma\in\Sigma_K\cap\Sigma_{\mathrm{ext}}} \rho_K^n m_K + \sum_{\sigma\in\Sigma} \rho_\sigma^n m_{\Delta_\sigma}.$$

The first term can easily be overestimated by $\int_\Omega \rho_{\mathcal{T},\tau}(\boldsymbol{x},t^n)\mathrm{d}\boldsymbol{x} = \int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x}$. 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 (53c), 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 (66) holds with $C_3 = (1+\zeta)\int_\Omega \varrho^0 \mathrm{d}\boldsymbol{x}$.

Reproducing the above calculations, one gets that

$$\begin{aligned}\int_\Omega \rho_{\Sigma,\tau}(\boldsymbol{x},t)\log\rho_{\Sigma,\tau}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x} &\le (1+\zeta)\int_\Omega \rho_{\mathcal{T},\tau}(\boldsymbol{x},t)\log\rho_{\mathcal{T},\tau}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\\ &= (1+\zeta)\left(\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}^n) + \sum_{K\in\mathcal{T}} m_K[\rho_K^n(1-V_K) - e^{-V_K}]\right).\end{aligned}$$

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 (67) holds with $C_4 = (1+\zeta)\,(C_1 + \|(1-V)^+\|_\infty)$. □

The last lemma of this section can be thought as a discrete $\left(L^\infty((0,T);W^{1,\infty}(\Omega))\right)'$ estimate on $\partial_t \rho_{\mathcal{T},\tau}$. This estimate will be used to apply a discrete nonlinear Aubin-Simon lemma [5] in the next section.

Lemma 3.8. *Let $\varphi \in C_c^\infty(Q_T)$, then define $\varphi_K^n = \frac{1}{m_K}\int_K \varphi(\boldsymbol{x},t^n)\mathrm{d}\boldsymbol{x}$ 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 (57) by φ_K^n and summing over $K\in\mathcal{T}$ and $n\in\{1,\dots,N\}$ yields

$$A := \sum_{n=1}^N \sum_{K\in\mathcal{T}} m_K(\rho_K^n - \rho_K^{n-1})\varphi_K = -\sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n(\phi_K^n - \phi_L^n)(\varphi_K^n - \varphi_L^n).$$

Applying Cauchy-Schwarz inequality on the right-hand side then provides

$$A^2 \le \left(\sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n(\phi_K^n - \phi_L^n)^2\right)\left(\sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n(\varphi_K^n - \varphi_L^n)^2\right). \tag{68}$$

The first term in the right-hand side is bounded thanks to Lemma 3.6. On the other hand, the regularity of φ ensures that there exists $\widetilde{\boldsymbol{x}}_K \in K$ such that $\varphi(\boldsymbol{x}_K,t^n) = \varphi_K^n$ for all $K\in\mathcal{T}$. Thanks to the regularity assumptions (53a)–(53b) on the mesh, there holds

$$|\varphi_K^n - \varphi_L^n| \le \|\nabla\varphi\|_\infty |\widetilde{\boldsymbol{x}}_K - \widetilde{\boldsymbol{x}}_L| \le (1 + 2\zeta(1+\zeta))\|\nabla\varphi\|_\infty d_\sigma, \qquad \sigma = K|L.$$

Hence, the second term of the right-hand side in (68) can be overestimated by

$$\sum_{n=1}^{N}\tau\sum_{\sigma=K|L\in\Sigma}a_\sigma\rho_\sigma^n(\varphi_K^n-\varphi_L^n)^2\le(1+2\zeta(1+\zeta))^2\|\nabla\varphi\|_\infty^2\sum_{n=1}^{N}\tau\sum_{\sigma=K|L\in\Sigma}m_\sigma d_\sigma\rho_\sigma^n$$
$$\le(1+2\zeta(1+\zeta))^2 d\|\nabla\varphi\|_\infty^2\iint_{Q_T}\rho_{\Sigma,\tau}\mathrm{d}\boldsymbol{x}\mathrm{d}t$$
$$\le(1+2\zeta(1+\zeta))^2 C_3 T d\|\nabla\varphi\|_\infty^2,$$

the last inequality being a consequence of Lemma 3.7. Combining all this material in (68) shows the desired estimate with $C_5=(1+2\zeta(1+\zeta))\sqrt{C_1C_3Td}$. □

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 3.4, 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 3.7, this allows to show that

$$\int_\Omega\rho_{\mathcal{T},\tau}(\boldsymbol{x},t)\log\rho_{\mathcal{T},\tau}(\boldsymbol{x},t)\mathrm{d}\boldsymbol{x}\le C_6,\qquad\forall t\in(0,T] \tag{69}$$

for some C_6 depending only on ϱ^0, ζ and V. Combining de La Vallée Poussin's theorem with Dunford-Pettis' one [58, Ch. XI, Theorem 3.6], there exists $\varrho\in L^\infty((0,T);L^1(\Omega))$ such that, up to a subsequence,

$$\rho_{\mathcal{T}_m,\tau_m}\text{ tends to }\varrho\text{ weakly in }L^1(Q_T)\text{ as }m\text{ tends to }+\infty. \tag{70}$$

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 [11, Corollary 3.9]), so that (69) yields

$$\iint_{Q_T}\varrho\log\varrho\mathrm{d}\boldsymbol{x}\mathrm{d}t\le C_6T. \tag{71}$$

Moreover, since $\rho_{\mathcal{T},\tau}\ge\alpha$ thanks to Lemma 3.5, then $\varrho\ge\alpha$ too.

Our goal is to show that ϱ is the unique weak solution to the Fokker-Planck equation (3) corresponding to the initial data ϱ^0. Even though the continuous problem is linear, (70) 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) 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 3.9. *Assume that $\varrho^0\ge\rho_\star\in(0,+\infty)$, then, up to a subsequence,*

$$\rho_{\mathcal{T}_m,\tau_m}\underset{m\to\infty}{\longrightarrow}\varrho\qquad\text{strongly in }L^1(Q_T), \tag{72}$$

$$\log\rho_{\mathcal{T}_m,\tau_m}\underset{m\to\infty}{\longrightarrow}\log\varrho\qquad\text{strongly in }L^1(Q_T), \tag{73}$$

$$\phi_{\mathcal{T}_m,\tau_m}\underset{m\to\infty}{\longrightarrow}\log\varrho+V\qquad\text{strongly in }L^1(Q_T). \tag{74}$$

Proof. Our proof of (72)–(73) relies on ideas introduced in [49] that were adapted to the discrete setting in [5]. 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 $f, g : Q_T \to \mathbb{R}$:

$$\iint_{Q_T} |fg| \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq \iint_{Q_T} \Upsilon(|f|)\mathrm{d}\boldsymbol{x}\mathrm{d}t + \iint_{Q_T} \Upsilon^*(|g|)\mathrm{d}\boldsymbol{x}\mathrm{d}t. \tag{75}$$

Now, notice that since $\rho_{\mathcal{T},\tau}$ is bounded from below thanks to Lemma 3.5 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

$$\log \rho_{\mathcal{T}_m,\tau_m} \underset{m\to\infty}{\longrightarrow} \ell \quad \text{weakly in } L^1(Q_T). \tag{76}$$

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 (71), $\Upsilon^*(\varrho)$ belongs also to $L^1(Q_T)$. Therefore, thanks to (75), 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

$$r_{\mathcal{T},\tau}(\boldsymbol{x},t) = r_K^n \quad \text{if } (\boldsymbol{x},t) \in K \times (t^{n-1}, t^n],$$

Thanks to Lemma 3.6, $\|r_{\mathcal{T},\tau}\|_{L^1(Q_T)} \leq \frac{1}{2} C_2 \tau$. 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 (56) and the triangle inequality, we obtain that for all $m \geq 1$, there holds

$$\int_0^T \int_{\Omega_{\boldsymbol{\xi}}} |\log \rho_{\mathcal{T}_m,\tau_m}(\boldsymbol{x}+\boldsymbol{\xi},t) - \log \rho_{\mathcal{T}_m,\tau_m}(\boldsymbol{x},t)| \, \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq A_{1,m}(\boldsymbol{\xi}) + A_{2,m}(\boldsymbol{\xi}) + A_{3,m}(\boldsymbol{\xi}),$$

where, denoting by $V_{\mathcal{T}}(\boldsymbol{x}) = V_K$ if $\boldsymbol{x} \in K$, we have set

$$\begin{aligned}
A_{1,m}(\boldsymbol{\xi}) &= \int_0^T \int_{\Omega_{\boldsymbol{\xi}}} |r_{\mathcal{T}_m,\tau_m}(\boldsymbol{x}+\boldsymbol{\xi},t) - r_{\mathcal{T}_m,\tau_m}(\boldsymbol{x},t)| \mathrm{d}\boldsymbol{x}\mathrm{d}t, \\
A_{2,m}(\boldsymbol{\xi}) &= \int_0^T \int_{\Omega_{\boldsymbol{\xi}}} |\phi_{\mathcal{T}_m,\tau_m}(\boldsymbol{x}+\boldsymbol{\xi},t) - \phi_{\mathcal{T}_m,\tau_m}(\boldsymbol{x},t)| \mathrm{d}\boldsymbol{x}\mathrm{d}t, \\
A_{3,m}(\boldsymbol{\xi}) &= T \int_{\Omega_{\boldsymbol{\xi}}} |V_{\mathcal{T}_m}(\boldsymbol{x}+\boldsymbol{\xi}) - V_{\mathcal{T}_m}(\boldsymbol{x})| \mathrm{d}\boldsymbol{x}.
\end{aligned}$$

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 [11, Exercise 4.34]) that there exists $\omega \in C(\mathbb{R}_+;\mathbb{R}_+)$ with $\omega(0) = 0$ such that

$$A_{1,m}(\boldsymbol{\xi}) + A_{3,m}(\boldsymbol{\xi}) \leq \omega(|\boldsymbol{\xi}|), \qquad \forall \boldsymbol{\xi} \in \mathbb{R}^d, \, \forall m \geq 0. \tag{77}$$

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:

$$\iint_{Q_T} |\nabla \phi_{\mathcal{T}_m,\tau_m}| = \sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} m_\sigma |\phi_K^n - \phi_L^n| \\ \leq \left(d|\Omega|T \sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} m_\sigma (\phi_K^n - \phi_L^n)^2 \right)^{1/2} \leq C_7.$$

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 (77) in (56) yields

$$\sup_{m\geq 1} \int_0^T \int_{\Omega_\xi} |\log \rho_{\mathcal{T}_m,\tau_m}(\boldsymbol{x}+\boldsymbol{\xi},t) - \log \rho_{\mathcal{T}_m,\tau_m}(\boldsymbol{x},t)| \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{|\boldsymbol{\xi}|\to 0}{\longrightarrow} 0. \tag{78}$$

The combination of (78) with Lemma 3.8 is exactly what one needs to reproduce the proof of [5, 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:

$$\iint_{Q_T} \rho_{\mathcal{T}_m,\tau_m} \log \rho_{\mathcal{T}_m,\tau_m} \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} \varrho \ell \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t, \qquad \forall \varphi \in C_c^\infty(Q_T). \tag{79}$$

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 (79),

$$0 \leq \iint_{Q_T} (\rho_{\mathcal{T}_m,\tau_m} - \kappa)(\log \rho_{\mathcal{T}_m,\tau_m} - \log\kappa)\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{m\to\infty}{\longrightarrow} \iint_{Q_T} (\varrho - \kappa)(\ell - \log\kappa)\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

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 (72)–(73), define

$$c_m = (\rho_{\mathcal{T}_m,\tau_m} - \varrho)(\log \rho_{\mathcal{T}_m,\tau_m} - \log\varrho) \in L^1(Q_T;\mathbb{R}_+), \qquad \forall m \geq 1.$$

Then (79) implies that

$$\iint_{Q_T} c_m \varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t \underset{m\to\infty}{\longrightarrow} 0, \qquad \forall \varphi \in C_c^\infty(Q_T),\ \varphi \geq 0.$$

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 (72)–(73) follow from Vitali's convergence theorem (see for instance [58, 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 (74) and concludes the proof of Proposition 3.9. □

Next lemma shows that $\rho_{\Sigma,\tau}$ shares the same limit ϱ as $\rho_{\mathcal{T},\tau}$.

Lemma 3.10. *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 3.7, 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 (70), 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

$$\|\log\rho_{\Sigma_m,\tau_m}-\log\rho_{\mathcal{T}_m,\tau_m}\|_{L^1(Q_T)}\underset{m\to\infty}{\longrightarrow}0, \tag{80}$$

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 3.5. Bearing in mind the definition (65) of ρ_{Σ_m,τ_m}, and one has

$$\|\log\rho_{\Sigma,\tau}-\log\rho_{\mathcal{T},\tau}\|_{L^1(Q_T)}\leq\sum_{n=1}^N\tau\sum_{\sigma=K|L\in\Sigma}m_{\Delta_\sigma}|\log\rho_K^n-\log\rho_L^n|.$$

Using (56) 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

$$R_1=\sum_{n=1}^N\tau\sum_{\sigma=K|L\in\Sigma}m_{\Delta_\sigma}|\phi_K^n-\phi_L^n|,\qquad R_2=\sum_{n=1}^N\tau\sum_{\sigma=K|L\in\Sigma}m_{\Delta_\sigma}|r_K^n-r_L^n|,$$

and

$$R_3=\sum_{\sigma=K|L\in\Sigma}m_{\Delta_\sigma}|V_K-V_L|.$$

Using again that $dm_{\Delta_\sigma}=d_\sigma m_\sigma\leq\zeta h_{\mathcal{T}}m_\sigma$ thanks to (53a), one has

$$R_1\leq\frac{\zeta}{d}h_{\mathcal{T}}\sum_{n=1}^N\tau\sum_{\sigma=K|L\in\Sigma}m_\sigma|\phi_K^n-\phi_L^n|\leq\frac{C_7\zeta}{d}h_{\mathcal{T}}\underset{m\to\infty}{\longrightarrow}0.$$

Since $|r_K^n-r_L^n|\leq r_K^n+r_L^n$, the regularity assumption (53c) 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 (53a). Therefore,

$$R_3\leq\zeta\|\nabla V\|_\infty|\Omega|h_{\mathcal{T}}\underset{m\to\infty}{\longrightarrow}0,$$

so that (80) holds, concluding the proof of Lemma 3.10. □

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). Define the vector field $\boldsymbol{F}_{\Sigma,\tau}:Q_T\to\mathbb{R}^d$ by

$$\boldsymbol{F}_{\Sigma,\tau}(\boldsymbol{x},t)=\begin{cases}d\rho_\sigma^n\frac{\phi_K^n-\phi_L^n}{d_\sigma}\boldsymbol{n}_{K\sigma} & \text{if } (\boldsymbol{x},t)\in\Delta_\sigma\times(t^{n-1},t^n],\\ 0 & \text{otherwise.}\end{cases}$$

Lemma 3.11. *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

$$\boldsymbol{G}_{\Sigma,\tau}(\boldsymbol{x},t) = \begin{cases} d\frac{\phi_L^n - \phi_K^n}{d_\sigma}\boldsymbol{n}_{K\sigma} & \text{if } (\boldsymbol{x},t) \in \Delta_\sigma \times (t^{n-1},t^n], \\ 0 & \text{otherwise,} \end{cases}$$

so that $\boldsymbol{F}_{\Sigma,\tau} = -\rho_{\Sigma,\tau}\boldsymbol{G}_{\Sigma,\tau}$. Thanks to Lemma 3.6,

$$\|\boldsymbol{G}_{\Sigma,\tau}\|^2_{L^2(Q_T)^d} = d\sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma(\phi_K^n - \phi_L^n)^2 \leq dC_2,$$

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. (74), then the weak consistency of the inflated gradient [24, 26] 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 3.6,

$$\|\boldsymbol{H}_{\Sigma,\tau}\|^2_{L^2(Q_T)^d} = d\sum_{n=1}^N \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma\rho_\sigma^n(\phi_K^n - \phi_L^n)^2 \leq dC_1,$$

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 3.7, 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 3.2, it remains to check that any limit value ϱ of the scheme is a solution to the Fokker-Planck equation (3) in the distributional sense.

Proposition 3.12. *Let ϱ be a limit value of $(\rho_{\mathcal{T}_m,\tau_m})_{m\geq 1}$ as described in Section 3.2, then for all $\varphi \in C_c^\infty(\overline{\Omega}\times[0,T))$, one has*

$$\iint_{Q_T} \varrho\partial_t\varphi \mathrm{d}\boldsymbol{x}\mathrm{d}t + \int_\Omega \varrho^0\varphi(\cdot,0)\mathrm{d}\boldsymbol{x} - \iint_{Q_T}(\varrho\nabla V + \nabla\varrho)\cdot\nabla\varphi\mathrm{d}\boldsymbol{x}\mathrm{d}t = 0. \tag{81}$$

Proof. Given $\varphi \in C_c^\infty(\overline{\Omega}\times[0,T))$, we denote by $\varphi_K^n = \varphi(\boldsymbol{x}_K,t^n)$. Then multipying (57) 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 3.9, 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 (55) 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,

$$B_2 \underset{m\to\infty}{\longrightarrow} \int_{\Omega} \varrho^0 \varphi(\cdot, 0) \mathrm{d}\boldsymbol{x}.$$

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 3.11 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 3.12, it only remains to check that

$$|B_3 - B_3'| \leq \sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} a_\sigma \rho_\sigma^n \left|\phi_K^n - \phi_L^n\right| \left| \varphi_K^{n-1} - \varphi_L^{n-1} + \frac{1}{\tau m_{\Delta_\sigma}} \int_{t^{n-1}}^{t^n} \int_{\Delta_\sigma} d_\sigma \nabla\varphi \cdot \boldsymbol{n}_{KL} \right| \mathrm{d}\boldsymbol{x}\mathrm{d}t.$$

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,

$$\left| \varphi_K^{n-1} - \varphi_L^{n-1} + \frac{1}{\tau m_{\Delta_\sigma}} \int_{t^{n-1}}^{t^n} \int_{\Delta_\sigma} d_\sigma \nabla\varphi \cdot \boldsymbol{n}_{KL} \right| \mathrm{d}\boldsymbol{x}\mathrm{d}t \leq C_\varphi d_\sigma (\tau + d_\sigma)$$

for some C_φ depending only on φ. Therefore,

$$|B_3 - B_3'| \leq C_\varphi (\tau + d_\sigma) \sum_{n=1}^{N} \tau \sum_{\sigma=K|L\in\Sigma} m_\sigma \rho_\sigma^n \left|\phi_K^n - \phi_L^n\right|.$$

Applying Cauchy-Schwarz inequality, one gets that

$$|B_3 - B_3'| \leq C_\varphi (\tau + d_\sigma) C_1 d \left\|\rho_{\Sigma,\tau}\right\|_{L^1(Q_T)} \underset{m\to\infty}{\longrightarrow} 0$$

thanks to Lemma 3.7. □

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.

4.1. **Newton method.** Due to the explicit formulation of the optimality condition of the saddle point problem (35), it appears extremely convient 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

$$\boldsymbol{J}^k \boldsymbol{d}^k = \begin{bmatrix} \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}} & \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \\ \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}} & \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}} \end{bmatrix} \begin{bmatrix} \boldsymbol{d}^k_{\boldsymbol{\phi}} \\ \boldsymbol{d}^k_{\boldsymbol{\rho}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{f}^k_{\boldsymbol{\phi}} \\ \boldsymbol{f}^k_{\boldsymbol{\rho}} \end{bmatrix}. \tag{82}$$

In the above linear system, $\boldsymbol{f}^k_{\boldsymbol{\phi}}$ and $\boldsymbol{f}^k_{\boldsymbol{\rho}}$ are the discrete HJ and continuity 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 (35) 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. To avoid this problem two possible strategies may be implemented: 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 and many more examples, the block $\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}$ is diagonal and therefore straightforward to invert. System (82) can be rewritten in term of the Schur complement and solved for $\boldsymbol{d}^k_{\boldsymbol{\phi}}$ as

$$\left[\boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\phi}} - \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}\right)^{-1} \boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\phi}}\right] \boldsymbol{d}^k_{\boldsymbol{\phi}} = \boldsymbol{f}^k_{\boldsymbol{\phi}} - \boldsymbol{J}^k_{\boldsymbol{\phi},\boldsymbol{\rho}} \left(\boldsymbol{J}^k_{\boldsymbol{\rho},\boldsymbol{\rho}}\right)^{-1} \boldsymbol{f}^k_{\boldsymbol{\rho}}, \tag{83}$$

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 4.1. *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

$$(\boldsymbol{A}^k)_{K,K} = \tau \sum_{\sigma = K|L \in \Sigma_K} a_\sigma (\phi^{n,k}_K - \phi^{n,k}_L)^+ \geq 0,$$

and, for $L \neq K$,

$$(\boldsymbol{A}^k)_{K,L} = -\tau a_\sigma (\phi^{n,k}_L - \phi^n_K)^+ \leq 0 \quad \text{if } \sigma = K|L, \qquad (\boldsymbol{A}^k)_{K,L} = 0 \quad \text{otherwise.}$$

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 [28] 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 (82). Moreover, it is possible to exploit for the solution of

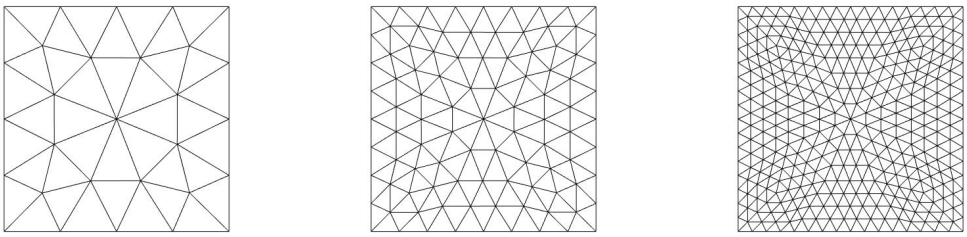


FIGURE 1. Sequence of regular triangular meshes.

system (83) solvers which are computationally more efficient, since the system is symmetric and negative definite.

4.2. **Fokker-Planck equation.** We first tackle the gradient flow of the Fokker-Planck energy, namely eq. (3). In section 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 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):

$$\varrho(x,y,t) = \exp(-\alpha t + \frac{g}{2}x)(\pi\cos(\pi x) + \frac{g}{2}sin(\pi x)) + \pi\exp(g(x-\frac{1}{2})),$$

where $\alpha = \pi^2 + \frac{g^2}{4}$. On the domain $\Omega = [0,1]^2$, the function $\varrho(x,y,t)$ is positive and satisfies the mixed boundary conditions $(\nabla\varrho + \varrho\nabla V)\cdot\boldsymbol{n}|_{\partial\Omega} = 0$. 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 1. Let us introduce the following mesh-dependent errors:

$$\begin{aligned}
\epsilon_1^n &= \sum_{K\in\mathcal{T}_m} |\rho_K^n - \varrho(\boldsymbol{x}_K, n\tau)|m_K, \quad \rightarrow \quad \text{discrete } L^1 \text{ error}\\
\epsilon_{L^\infty} &= \max_n(\epsilon_n^1), \quad \rightarrow \quad \text{discrete } L^\infty((0,T);L^1(\Omega)) \text{ error},\\
\epsilon_{L^1} &= \sum_n \tau\,\epsilon_1^n, \quad \rightarrow \quad \text{discrete } L^1((0,T);L^1(\Omega)) \text{ error},
\end{aligned}$$

where $\varrho(\boldsymbol{x}_K, n\tau_m)$ is the value in the cell center of the triangle K of the analytical solution at time $n\tau_m$, n running from 0 to the total number of time steps N_m. The upstream Finite Volume scheme with backward Euler discretization of the temporal derivative, namely scheme (49), 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 (36) 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

TABLE 1. Time-space convergence for the two schemes. Integration on the time step $[0, 0.25]$.

		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

TABLE 2. Time-space convergence for scheme (36). Integration on the time step $[0.5, 0.25]$.

		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

(49) and (36). The results are shown in Table 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 (36) exhibits the same order of convergence of scheme (49). 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 $\varrho(\boldsymbol{x}_K, 0)$ is too close to zero, and in particular equal to zero on the set $1 \times [0,1]$, and scheme (36) tends to be repulsed away from zero due to the singularity of the gradient of the first variation of the energy. In Table 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:

$$\begin{aligned}\mathcal{E}(\lim_{t\to\infty}\varrho) &= \int_\Omega \lim_{t\to\infty}(\varrho\log\varrho - \varrho g x)\mathrm{d}\boldsymbol{x} \\ &= \exp(\frac{g}{2})(\frac{\pi\log(\pi)}{g} + \frac{\pi}{2} - \frac{\pi}{g}) + \exp(-\frac{g}{2})(-\frac{\pi\log(\pi)}{g} - \frac{\pi}{2} + \frac{\pi}{g}).\end{aligned}$$

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

$$\rho_K^\infty = M\exp(-V_K),\ V_K = V(\boldsymbol{x}_K),\ \forall K \in \mathcal{T},$$

as it can be easily checked to be the unique minimizer of the discrete energy $\mathcal{E}_\mathcal{T} = \sum_{K\in\mathcal{T}} E(\rho_K)m_K$ subject to the constraint of the conservation of the mass,

$$\frac{\partial}{\partial\rho_K}\big(\mathcal{E}_\mathcal{T} + \lambda\sum_{K\in\mathcal{T}}(\rho_K - \rho_K^0)m_K\big)|_{\rho_K^\infty} = \big(\log\rho_K^\infty + 1 + V_K + \lambda\big)m_K = 0, \quad \forall K\in\mathcal{T}$$
$$\implies \quad \rho_K^\infty = \exp(-(1+\lambda) - V_K) = M\exp(-V_K), \quad \forall K\in\mathcal{T},$$

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 (49), 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 (36) instead, as it appears clear from Lemma 2.1, 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 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 2a it is noticeable that scheme (36) 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 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 (36) still dissipates faster. In the end, in Figure 2c it is remarkable 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.

4.3. **Porous medium equation.** The porous medium equation,

$$\partial_t\varrho = \Delta\varrho^m + \nabla\cdot(\varrho\nabla V),$$

has been proven in [53] to be a gradient flow in Wasserstein space with respect to the energy

$$\mathcal{E}(\rho) = \int_\Omega \frac{1}{m-1}\rho^m \mathrm{d}\boldsymbol{x} + \int_\Omega \rho V \mathrm{d}\boldsymbol{x}, \tag{84}$$

for a given m strictly greater than one. Our aim is to show that scheme (36) 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 $V(\boldsymbol{x}) = \frac{1}{2}||\boldsymbol{x} - 0.5||_2^2$. The equilibrium solution of the gradient flow should then be the Barenblatt profile $\varrho^\infty(\boldsymbol{x}) = \max((\frac{M}{2\pi})^{\frac{m-1}{m}} - \frac{m-1}{2m}||\boldsymbol{x}-0.5||_2^2, 0)^{\frac{1}{m-1}}$, with M total mass of the initial condition.

In Figure 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 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 (84), as it has been done for the Fokker-Planck energy. As expected, the solution converges towards the Barenblatt profile.

4.4. **Thin film equation.** In order to show that scheme (36) 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 particular case of a family of nonlinear fourth order equations [44]. The energy $\mathcal{E}(\rho)$ is discretized as

$$\mathcal{E}_\mathcal{T}(\boldsymbol{\rho}) = \frac{1}{2}\sum_{\sigma\in\Sigma}\Big(\frac{\rho_L - \rho_K}{d_\sigma}\Big)^2 d_\sigma m_\sigma + \sum_{K\in\mathcal{T}} \rho_K V(\boldsymbol{x}_K) m_K,$$

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 (83) is not diagonal and the Schur complement technique for the solution of the linear system (82) is not necessarily convenient anymore. In figure 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 $V(\boldsymbol{x}) = (x-1)(y-1)$.

4.5. **Salinity intrusion problem.** We want to show now that scheme (36) can be used for the solutions of systems of equations of the type of (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

$$\begin{cases} \partial_t f - \nabla\cdot(\nu f\nabla(f+g+b)) = 0 & \text{in}\,\Omega\times(0,T), \\ \partial_t g - \nabla\cdot(g\nabla(\nu f+g+b)) = 0 & \text{in}\,\Omega\times(0,T), \end{cases} \tag{85}$$

completed with the no-flux boundary conditions

$$\nabla f\cdot\boldsymbol{n} = \nabla g\cdot\boldsymbol{n} = 0 \quad \text{on}\,\partial\Omega\times(0,T),$$

and initial conditions $f(t=0) = f_0, g(t=0) = g_0$, 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 density of the fresh and salt water. Equation (85) has been proven in [38] to be a Wasserstein gradient flow with respect to the energy

$$\mathcal{E}(f,g) = \int_\Omega \Big(\frac{\nu}{2}(b+g+f)^2 + \frac{1-\nu}{2}(b+g)^2\Big)\mathrm{d}\boldsymbol{x}. \tag{86}$$

The discretization of (86) is again straightforward. In figure 6 it is represented an evolution of the two surfaces of salt and fresh water (see [1] 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 (36) works. As already said, from numerical evidences the scheme works under much more general and mild hypotheses.

Acknowledgements

CC acknowledges the support of the Labex CEMPI (ANR-11-LABX-0007-01). GT acknowledges that 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. We also thank Guillaume Carlier and Quentin Mérigot for fruitful discussions.

References

[1] A. 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.

[2] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[3] L. Ambrosio, E. Mainini, and S. Serfaty. Gradient flow of the Chapman-Rubinstein-Schatzman model for signed vortices. *Ann. Inst. H. Poincaré Anal. Non Linéaire*, 28(2):217–246, 2011.

[4] L. Ambrosio and S. Serfaty. A gradient flow approach to an evolution problem arising in superconductivity. *Comm. Pure Appl. Math.*, 61(11):1495–1539, 2008.

[5] B. Andreianov, C. Cancès, and A. Moussa. A nonlinear time compactness result and applications to discretization of degenerate parabolic–elliptic PDEs. *J. Funct. Anal.*, 273(12):3633–3670, 2017.

[6] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[7] J.-D. Benamou, G. Carlier, and M. Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. In *Gradient flows: from theory to application*, volume 54 of *ESAIM Proc. Surveys*, pages 1–17. EDP Sci., Les Ulis, 2016.

[8] M. 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.

[9] A. Blanchet. A gradient flow approach to the Keller-Segel systems. RIMS Kokyuroku's lecture notes, vol. 1837, pp. 52–73, June 2013.

[10] F. Bolley, I. Gentil, and A. Guillin. Convergence to equilibrium in Wasserstein distance for Fokker-Planck equations. *J. Funct. Anal.*, 263(8):2430–2457, 2012.

[11] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Universitext. Springer, New York, 2011.

[12] V. Calvez and T. 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.

[13] C. Cancès. Energy stable numerical methods for porous media flow type problems. *Oil & Gas Science and Technology-Rev. IFPEN*, 73:1–18, 2018.

[14] C. Cancès, T. O. Gallouët, M. Laborde, and L. 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.

[15] C. Cancès, T. O. Gallouët, and L. Monsaingeon. Incompressible immiscible multiphase flows in porous media: a variational approach. *Anal. PDE*, 10(8):1845–1876, 2017.

[16] C. Cancès and C. 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.

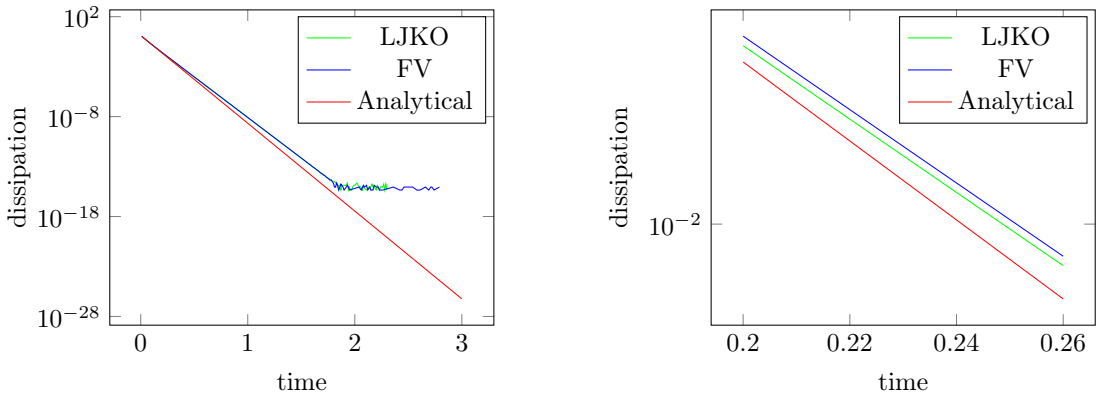
[17] C. Cancès and C. Guichard. Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Found. Comput. Math.*, 17(6):1525–1584, 2017.
[18] C. Cancès, D. Matthes, and F. Nabet. A two-phase two-fluxes degenerate Cahn-Hilliard model as constrained Wasserstein gradient flow. *Arch. Ration. Mech. Anal.*, 233(2):837–866, 2019.
[19] C. Cancès, F. Nabet, and M. Vohralík. Convergence and a posteriori error analysis for energy-stable finite element approximations of degenerate parabolic equations. HAL: hal-01894884, 2018.
[20] J. A. Carrillo, K. Craig, and F. S. Patacchini. A blob method for diffusion. *Calc. Var. Partial Differential Equations*, 58(2):53, 2019.
[21] J. A. Carrillo, K. Craig, L. Wang, and C. Wei. Primal dual methods for Wasserstein gradient flows. arXiv:1901.08081, 2019.
[22] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, and D. Slepčev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.*, 156(2):229–271, 2011.
[23] J. A. Carrillo, B. Düring, D. Matthes, and M. S. McCormick. A Lagrangian scheme for the solution of nonlinear diffusion equations using moving simplex meshes. *J. Sci. Comput.*, 73(3):1463–1499, 2018.
[24] C. Chainais-Hillairet, J.-G. Liu, and Y.-J. Peng. Finite volume scheme for multi-dimensional drift-diffusion equations and convergence analysis. *ESAIM: M2AN*, 37(2):319–338, 2003.
[25] M. Erbar and J. Maas. Gradient flow structures for discrete porous medium equations. *Discrete Contin. Dyn. Syst.*, 34(4):1355–1374, 2014.
[26] R. Eymard and T. Gallouët. H-convergence and numerical schemes for elliptic problems. *SIAM J. Numer. Anal.*, 41(2):539–562, 2003.
[27] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. Ciarlet, P. G. (ed.) et al., in Handbook of numerical analysis. North-Holland, Amsterdam, pp. 713–1020, 2000.
[28] J. Fuhrmann. Existence and uniqueness of solutions of certain systems of algebraic equations with off-diagonal nonlinearity. *Appl. Numer. Math.*, 37:359–370, 2001.
[29] K. Gärtner and L. Kamenski. Why Do We Need Voronoi Cells and Delaunay Meshes? In *Numerical Geometry, Grid Generation and Scientific Computing*, V. A. Garanzha, L. Kamenski and H. Si (ed.), *Lecture Notes in Computational Science and Engineering*, pp. 45–60, Springer International Publishing, 2019.
[30] N. Gigli and J. Maas. Gromov-Hausdorff convergence of discrete transportation metrics. *SIAM J. Math. Anal.*, 45(2):879–899, 2013.
[31] P. Gladbach, E. Kopfer, and J. Maas. Scaling limits of discrete optimal transport. arXiv:1809.01092, 2018.
[32] M. Heida. Convergences of the squareroot approximation scheme to the Fokker-Planck operator. *Math. Models Methods Appl. Sci.*, 28(13):2599–2635, 2018.
[33] M. Jacobs, I. Kim, and A. R. Mészáros. Weak solutions to the Muskat problem with surface tension via optimal transport. arXiv:1905.05370, 2019.
[34] R. Jordan, D. Kinderlehrer, and F. Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.
[35] O. Junge, D. Matthes, and H. 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.
[36] D. Kinderlehrer, L. Monsaingeon, and X. Xu. A Wasserstein gradient flow approach to Poisson-Nernst-Planck equations. *ESAIM Control Optim. Calc. Var.*, 23(1):137–164, 2017.
[37] D. Kinderlehrer and N. J. Walkington. Approximation of parabolic equations using the Wasserstein metric. *M2AN Math. Model. Numer. Anal.*, 33(4):837–852, 1999.
[38] P. Laurençot and B.-V. 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.
[39] H. Leclerc, Q. Mérigot, F. Santambrogio, and F. Stra. Lagrangian discretization of crowd motion and linear diffusion. arXiv: 1905.08507, 2019.
[40] J. Leray and J. Schauder. Topologie et équations fonctionnelles. *Ann. Sci. École Norm. Sup.*, 51((3)):45–78, 1934.
[41] W. Li, J. Lu, and L. Wang. Fisher information regularization schemes for Wasserstein gradient flows. arXiv:1907.02152.
[42] J. Maas. Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.*, 261(8):2250–2292, 2011.
[43] J. Maas and D. Matthes. Long-time behavior of a finite volume discretization for a fourth order diffusion equation. *Nonlinearity*, 29(7):1992–2023, 2016.
[44] D. Matthes, R. McCann, and G. Savar'e. A Family of Nonlinear Fourth Order Equations of Gradient Flow Type. *Communications in Partial Differential Equations*, 34(11):1352-1397, 2009

[45] D. Matthes and H. Osberger. Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation. *ESAIM Math. Model. Numer. Anal.*, 48(3):697–726, 2014.
[46] D. Matthes and H. Osberger. A convergent Lagrangian discretization for a nonlinear fourth-order equation. *Found. Comput. Math.*, 17(1):73–126, 2017.
[47] B. Maury, A. Roudneff-Chupin, and F. Santambrogio. A macroscopic crowd motion model of gradient flow type. *Math. Models Methods Appl. Sci.*, 20(10):1787–1821, 2010.
[48] A. Mielke. A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity*, 24(4):1329–1346, 2011.
[49] A. Moussa. Some variants of the classical Aubin-Lions Lemma. *J. Evol. Equ.*, 16(1):65–93, 2016.
[50] T. J. Murphy and N. J. Walkington. Control volume approximation of degenerate two-phase porous media flows. *SIAM J. Numer. Anal.*, 57(2):527–546, 2019.
[51] L. Neves de Almeida, F. Bubba, B. Perthame, and C. Pouchol. Energy and implicit discretization of the Fokker-Planck and Keller-Segel type equations. arXiv:1803.10629, 2018.
[52] F. Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: a mean-field theory. *Arch. Rational Mech. Anal.*, 141(1):63–103, 1998.
[53] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.
[54] R. Peyre. Comparison between W_2 distance and H^{-1} norm, and localization of Wasserstein distance. *ESAIM: COCV*, 24(4):1489–1501, 2018.
[55] F. Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Progress in Nonlinear Differential Equations and Their Applications 87. Birkhäuser Basel, 1 edition, 2015.
[56] Z. Sun, J. A. Carrillo, and C.-W. Shu. A discontinuous Galerkin method for nonlinear parabolic equations and gradient flow problems with interaction potentials. *J. Comput. Phys.*, 352:76–104, 2018.
[57] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
[58] A. Visintin. *Models of phase transitions*, volume 28 of *Progress in nonlinear differential equations and their applications*. Birkhäuser Boston, 1996.

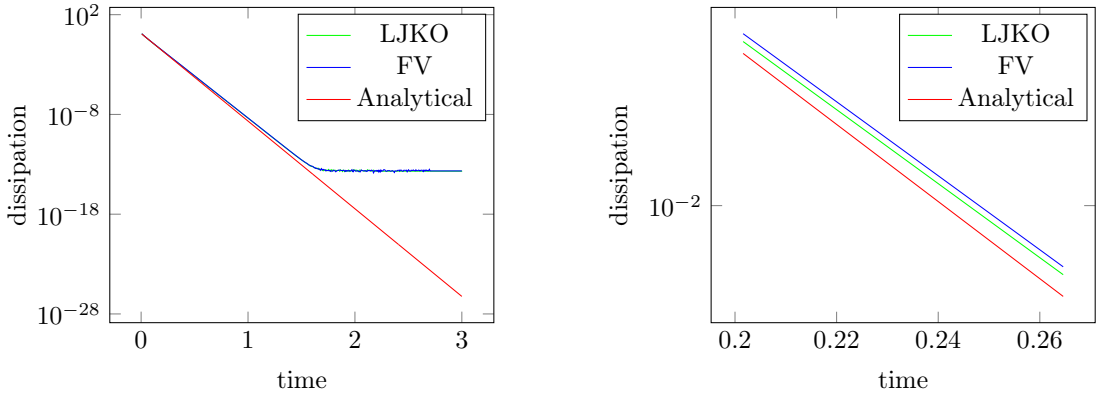
Clément Cancès (clement.cances@inria.fr): Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille

Thomas O. Gallouët (thomas.gallouet@inria.fr): Inria, Project team Mokaplan, Université Paris-Dauphine, PSL Research University, UMR CNRS 7534-Ceremade

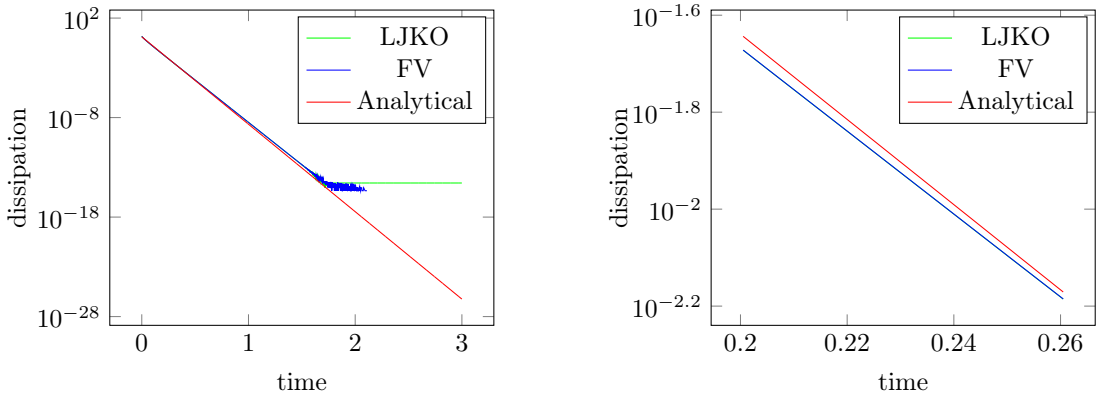
Gabriele Todeschi (gabriele.todeschi@inria.fr): Inria, Project team Mokaplan, Université Paris-Dauphine, PSL Research University, UMR CNRS 7534-Ceremade



(A) $\tau_m = 0.01,\ h_{\mathcal{T}_m} = 0.1493.$

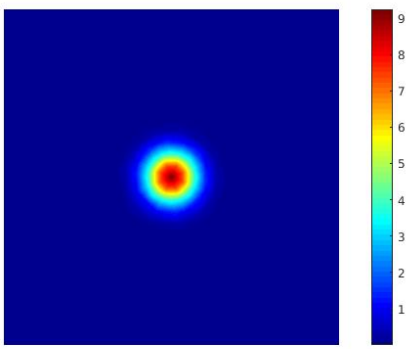


(B) $\tau_m = 0.0063,\ h_{\mathcal{T}_m} = 0.0373.$

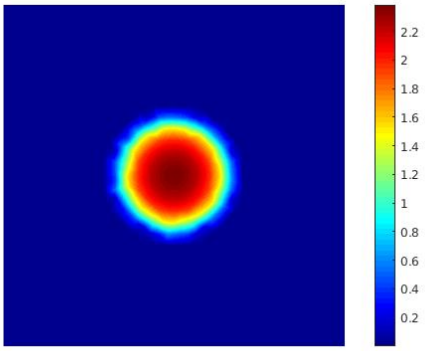


(C) $\tau_m = 0.0001,\ h_{\mathcal{T}_m} = 0.1493.$

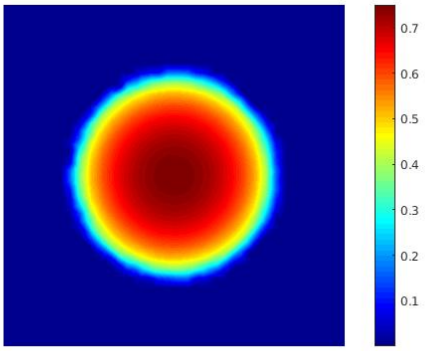
FIGURE 2. Comparison of the dissipation of the system computed with the two numerical schemes (36) (LJKO) and (49) (FV), and in the real case. Semi-logarithmic plot.



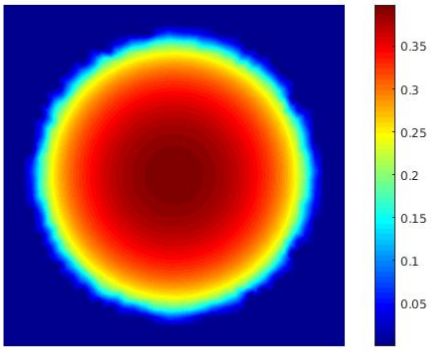
(A) t=0



(B) t=0.0001



(C) t=0.01



(D) t=1

FIGURE 3. Evolution of an initial density close to a dirac according to the porous medium equation. In each picture the scaling is different for the sake of the representation.

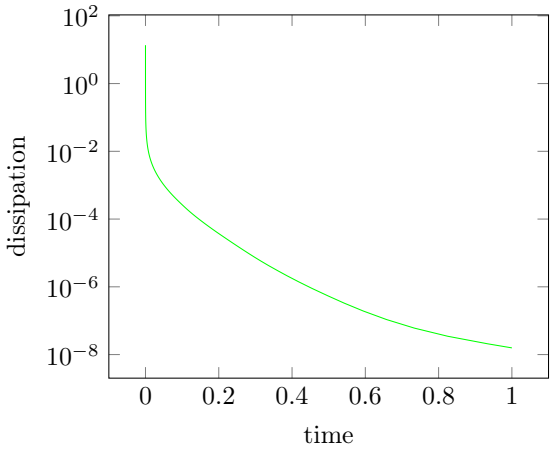
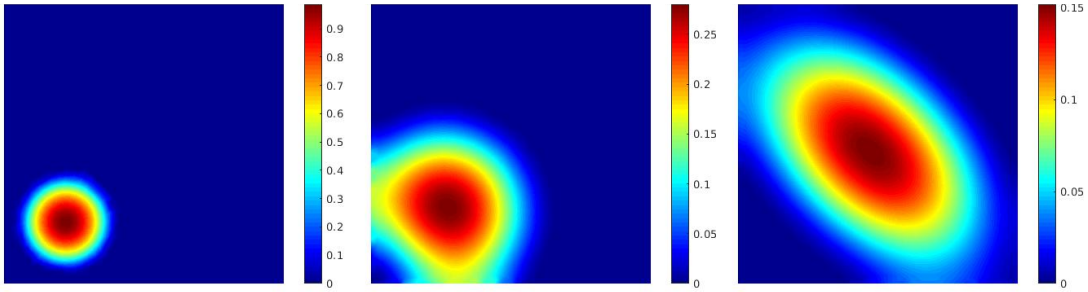
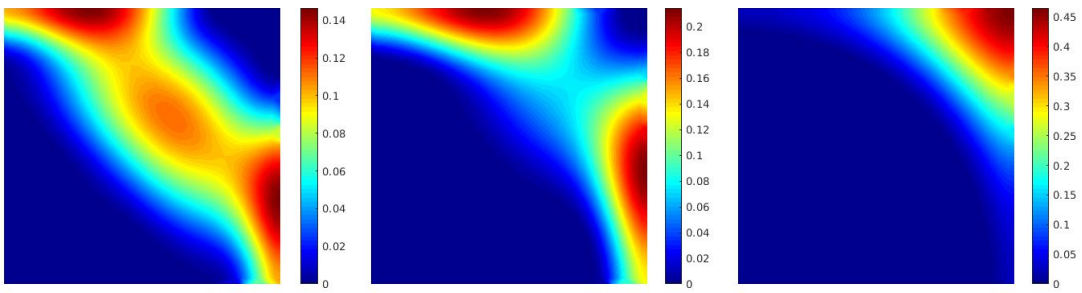


FIGURE 4. Dissipation of the energy for the porous medium equation. Semi-logarithmic plot.

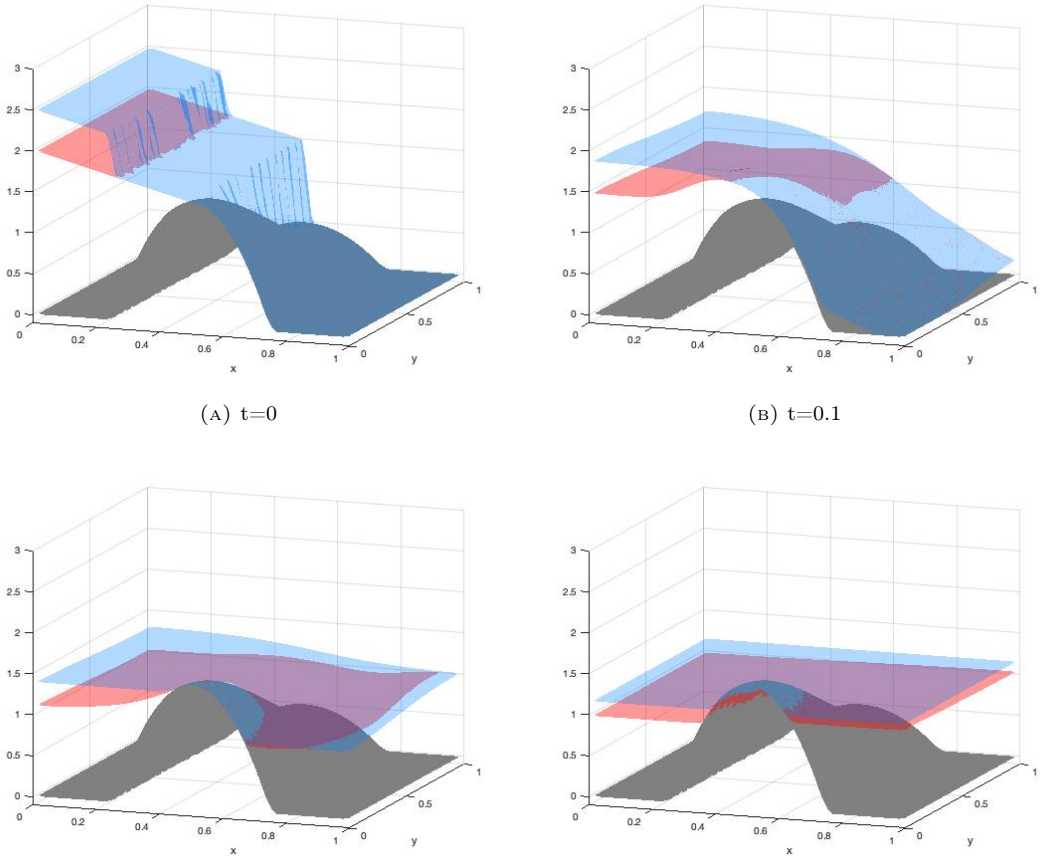


(A) t=0 (B) t=0.0015 (C) t=0.008



(D) t=0.014 (E) t=0.02 (F) t=0.3

FIGURE 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.



(A) t=0 (B) t=0.1

(C) t=0.5 (D) t=10

FIGURE 6. Evolution of the two interfaces of salt (red) and fresh (blue) water.

2.3 Euler flows

Articles:

- **A Lagrangian scheme à la Brenier for the incompressible Euler equations.** *Found Comput Math 18: 835* (2018). https://doi.org/10.1007/s10208-017-9355-y. Gallouët T.O. and Mérigot Q.
- **Convergence of a Lagrangian discretization for barotropic fluids and porous media flow.** *SIAM Journal on Mathematical Analysis* (2021) https://hal.science/hal-03234144. Gallouët T.O., Mérigot Q., Natale A.

Collaborators: The first paper is a collaboration with Q. Mérigot. It combines reinterpretation of Y. Brenier's old ideas and Q. Mérigot's new method that allows to deal numerically with semi-discrete Optimal Transport. It was done when I was a post-doc of Y. Brenier. The second paper is a collaboration with Q. Mérigot and A. Natale. At this moment A. Natale was a post-doc under our supervision.

Main contributions:

- We constructed and implemented a Lagrangian numerical scheme for the Incompressible Euler equations.
- We proved its convergence towards smooth solutions thanks to a relative entropy methods. This is not new for Lagrangian methods. Numerically we also observed a good behavior of the scheme with more rough initial conditions and wider class models based on the Incompressible Euler equations (with gravity, non homogenous fluid).
- In the second paper we constructed the same type of Lagragian scheme but for compressible Euler equations.
- We proved the convergence of the scheme towards smooth solutions thanks to a relative entropy methods.
- We proved that the same approached works also for the Wasserstein gradient flows associated to the same energy. The proof of convergence is very similar.
- We implemented the scheme in both cases: Euler flows and Wasserstein gradient flows.

Research directions: Research directions around these questions are numerous and well adapted for Phd subjects. It is one of the main axes of the new Inria ParMA team, that I will held, based at the Laboratoire Mathématiques d'Orsay (Paris Saclay University).
The main strength of these Lagrangian methods is that they are based on the physical energy and a nice geometrical structure for the PDE either Gradient flows Euler Flows or Conservative flows (where the velocity is given by the rotation of the Wasserstein gradient of the energy. So the next step is too add more physics in the model throughout the energy, deduce the numerical scheme and try to adapt the proof. let us mention some extensions.

- Fluid-structure interactions in the incompressible case. The scheme is very suited for this interaction. We have to adapt the projection step taking into account the structure. The proof of convergence works well for exemple if the motion of the structure is given. Numerically the simulations works well also with the full model (the motion of the structure is not given), a nice implementation is to approximate the whole space (fluid+structure) with particules and enforce a constraint for the points coming from the structure. This work is in progress.

- Incompressible Navier Stokes equation. The Lagrangian scheme interact perfectly with finite volume scheme such as the one presented in Section 2.2.2. Indeed the Laguerre cells makes an admissible finite volume tesselation and the quantity required to compute a finite volume approximation of the Laplacian of the velocity are all given by the Laguerre cells. Numerical simulations are very convincing see for instance the implementation done by B. Lévy in [14]. From a theoretical point of view the finite volume discretization being not consistant the proof of convergence is not straightforward but it seems to work using a nice decomposition. This work is also in progress.

- Adding some interaction terms. The next step would be to add some interactions terms in order to approximate for example Keller-Segel equations. I did not look too much in this direction yet but the recent work of D. Bresch and co-authors and S. Serfaty and co-authors [7, 2] are dealing with success with relative entropy methods and interaction terms. A natural idea would be to adapt the way they treat these interaction terms in the gronwall argument in our context. I will submit this subject for a Phd student.

- Particules approximation of the semi-geostrophic equation. The semi geostrophic equation can be recast as a conservative flow in the Wasserstein space. Numerical simulations can be done using a scheme based on semi-discrete Optimal Transport. The proof of convergence for this scheme presents some novel difficulties. This is an ongoing research conducted in collaboration with Q. Mérigot and D. Bourne.

A LAGRANGIAN SCHEME À LA BRENIER FOR THE INCOMPRESSIBLE EULER EQUATIONS

THOMAS O. GALLOUËT AND QUENTIN MÉRIGOT

Abstract. We approximate the regular solutions of the incompressible Euler equations by the solution of ODEs on finite-dimensional spaces. Our approach combines Arnold's interpretation of the solution of the Euler equations for incompressible and inviscid fluids as geodesics in the space of measure-preserving diffeomorphisms, and an extrinsic approximation of the equations of geodesics due to Brenier. Using recently developed semi-discrete optimal transport solvers, this approach yields a numerical scheme which is able to handle problems of realistic size in 2D. Our purpose in this article is to establish the convergence of this scheme towards regular solutions of the incompressible Euler equations, and to provide numerical experiments on a few simple test cases in 2D.

Contents

1. Introduction

The purpose of this article is to investigate a discretization of Euler's equation for incompressible and inviscid fluids in a domain $\Omega \subseteq \mathbb{R}^d$ with Neumann boundary conditions:

$$\begin{cases} \partial_t v(t,x) + (v(t,x)\cdot\nabla)\, v(t,x) = -\nabla p(t,x), & \text{for } t\in[0,T],\ x\in\Omega\,, \\ \operatorname{div}(v(t,x)) = 0 & \text{for } t\in[0,T],\ x\in\Omega\,, \\ v(t,x)\cdot n = 0 & \text{for } t\in[0,T],\ x\in\partial\Omega\,, \\ v(0,x) = v_0. \end{cases} \tag{1.1}$$

As noticed by Arnold [2], when expressed in Lagrangian coordinates, Euler's equations can be interpreted as the equation of geodesics in the infinite-dimensional group of measure-preserving diffeomorphisms of Ω. To see this, consider the flow map $\phi : [0,T]\times\Omega \to \Omega$

1991 *Mathematics Subject Classification.* 35Q31, 65M12, 65M50, 65Z05.

Key words and phrases. Incompressible Euler equations, Optimal Transport, Lagrangian numerical scheme, Hamiltonian.

T.O.G was supported by ANR grant ISOTACE (ANR-12-MONU-0013) and by the Fonds de la Recherche Scientifique - FNRS under Grant MIS F.4539.16.

Q.M. is supported by ANR grant MAGA (ANR-16-CE40-0014).

COMMUNICATED BY EITAN TADMOR

induced by the vector field v, that is:

$$\begin{cases} \frac{d}{dt}\phi(t,x) = v\left(t,\phi(t,x)\right) & \text{for } t\in[0,T],\ x\in\Omega\,, \\ \phi(0,\cdot) = \mathrm{id}, \\ \partial_t\phi(0,\cdot) = v_0. \end{cases} \tag{1.2}$$

Using the formula $\frac{d}{dt}\det \mathrm{D}\phi(t,x) = \mathrm{div}\left(v(t,x)\right)\det \mathrm{D}\phi(t,x)$, the incompressibility constraint $\mathrm{div}\left(v(t,x)\right) = 0$ and the initial condition $\phi(0) = \mathrm{id}$, one can check that $\phi(t,\cdot)$ belongs to the set of volume preserving maps $\mathbb{S}$, defined by

$$\mathbb{S} = \left\{ s\in \mathrm{L}^2(\Omega,\mathbb{R}^d) \mid s_{\#}\,\mathrm{Leb} = \mathrm{Leb} \right\},$$

where Leb is the restriction of the Lebesgue measure to the domain Ω and where the pushforward measure $s_{\#}\,\mathrm{Leb}$ is defined by the formula $s_{\#}\,\mathrm{Leb}(A) = \mathrm{Leb}(s^{-1}(A))$ for every measurable subset A of Ω. Euler's equations (1.1) can therefore be reformulated as

$$\begin{cases} \frac{d^2}{dt^2}\phi(t) = -\nabla p(t,\phi(t,x)) & \text{for } t\in[0,T],\ x\in\Omega\,, \\ \phi(t,\cdot)\in\mathbb{S} & \text{for } t\in[0,T], \\ \phi(0,\cdot) = \mathrm{id}, \\ \partial_t\phi(0,\cdot) = v_0. \end{cases} \tag{1.3}$$

To obtain (1.3) one simply needs to derive (1.2). This equation can be formally interpreted as the equation of geodesics in $\mathbb{S}$. In particular the pressure term in the evolution equation in (1.3) expresses that the acceleration of ϕ should be orthogonal to the tangent plane to $\mathbb{S}$ at ϕ. Indeed, note that the condition $\phi(t,\cdot)\in\mathbb{S}$ in (1.1) encodes the infinitesimal conditions $\mathrm{div}\, v(t,\cdot) = 0$ and $v(t,x)\cdot n(x) = 0$ in (1.3). This suggests that the tangent plane to $\mathbb{S}$ at a point $\phi\in\mathbb{S}$ should be the set $\{v\circ\phi \mid v\in\mathcal{H}_{\mathrm{div}}(\Omega)\}$, where $\mathcal{H}_{\mathrm{div}}(\Omega)$ denotes the set of divergence-free vector fields

$$\mathcal{H}_{\mathrm{div}}(\Omega) = \left\{ v\in \mathrm{L}^2(\Omega,\mathbb{R}^d) \mid \int_\Omega v\cdot\nabla\varphi = 0,\ \forall\varphi\in C_c^\infty(\overline{\Omega}) \right\}.$$

In addition, by the Helmoltz-Hodge decomposition, the orthogonal subspace to $\mathcal{H}_{\mathrm{div}}(\Omega)$ in $\mathrm{L}^2(\Omega,\mathbb{R}^d)$ is the space of gradients of functions. Therefore the evolution equation in (1.3) expresses that $\frac{d^2}{dt^2}\phi(t) \perp T_{\phi(t)}\mathbb{S}$, in other words that $t\mapsto\phi(t,\cdot)$ is a geodesic of $\mathbb{S}$. Note however that a solution to (1.3) does not need to be a *minimizing* geodesic between $\phi(0,\cdot)$ and $\phi(T,\cdot)$. The problem of finding a minimizing geodesic on $\mathbb{S}$ between two measure preserving maps amounts to solving equations (1.3), where the initial condition $\partial_t\phi(0,\cdot) = v_0$ is replaced by a prescribed coupling between the position of particles at initial and final times. It leads to generalized and non-deterministic solutions introduced by Brenier [6], where particles are allowed to split and cross. Shnirelman showed that this phenomenon can happen even when the measure-preserving maps $\phi(0,\cdot)$ and $\phi(T,\cdot)$ are diffeomorphisms of Ω [23].

Previous work: discretization of geodesics in $\mathbb{S}$. The first numerical experiments to recover generalized minimizing geodesics have been performed by Brenier in 1D [9]. He also proposed a scheme to compute the solutions of the Cauchy problem (1.3) in [5]. In Brenier's discretization, the measure-preserving maps are approximated by permutations of a decomposition of the domain into cubes. The numerical implementation of this idea relies on the resolution of a linear assignment problem at every timestep, whose cost is unfortunately prohibitive for domains in dimension higher than one.

The discretization we consider in this article is a variant of this approach which is more tractable computationally and leads to slightly better convergence estimates. As in [8], the measure-preserving property (or incompressibility) is enforced through a penalization term involving the squared distance to the set of measure-preserving maps $\mathbb{S}$. This squared

distance can be computed efficiently thanks to recently developed numerical solvers for optimal transport problems between probability densities and finitely-supported probability measures [3, 20, 13, 18]. This alternative discretization has already been used successfully to compute minimizing geodesics between measure-preserving maps in [21], allowing the recovery of non-deterministic solutions to Euler's equations predicted by Shnirelman and Brenier in dimension two. The object of this article is to study whether this strategy can be used to construct Lagrangian schemes for the more classical Cauchy problem for the Euler's equations (1.1), able to cope with problems of realistic size in dimension two.

Discretization in space: approximate geodesics. The construction of approximate geodesics presented here is strongly inspired by a particle scheme introduced by Brenier [8]. We first approximate the Hilbert space $\mathbb{M} = \mathrm{L}^2(\Omega, \mathbb{R}^d)$ by finite dimensional subspaces. Let N be an integer and let P_N be a tessellation of Ω into N subsets $(\omega_i)_{1\leq i\leq N}$ satisfying

$$\begin{cases} \forall i \in \{1, \dots, N\},\ \mathrm{Leb}(\omega_i) = \dfrac{1}{N}\mathrm{Leb}(\Omega) \\ h_N := \max\limits_{1\leq i\leq N} \mathrm{diam}(\omega_i) \leq \dfrac{C}{N^{1/d}} \end{cases}$$

where $C > 0$ is independent of N. We consider $\mathbb{M}_N \subseteq \mathbb{M}$ the space of functions from Ω to $\mathbb{R}^d$ which are constant on each of the subdomains (ω_i). To construct our approximate geodesics, we consider the squared distance to the set $\mathbb{S} \subseteq \mathbb{M}$ of measure-preserving maps:

$$d^2_{\mathbb{S}} : m \in \mathbb{M}_n \mapsto \min_{s\in\mathbb{S}} \|m - s\|^2_{\mathbb{M}}.$$

The approximate geodesic model is described by the differential equation in the finite-dimensional space $\mathbb{M}_N$:

$$\begin{cases} \ddot{m}(t) + \frac{\nabla \mathrm{d}^2_{\mathbb{S}}(m(t))}{2\epsilon^2} = 0, \quad \text{for } t \in [0, T]\,, \\ (m(0), \dot{m}(0)) \in \mathbb{M}^2_N \end{cases} \tag{1.4}$$

Note that the squared distance $\mathrm{d}^2_{\mathbb{S}}$ is semi-concave, so that its restriction to the finite-dimensional space $\mathbb{M}_N$ is differentiable at almost every point. This differential system is induced by the Hamiltonian $H : \mathbb{M}_N \times \mathbb{M}_N \to \mathbb{R}$

$$H(m, \dot{m}) = \frac{1}{2}\|\dot{m}\|^2_{\mathbb{M}} + \frac{d^2_{\mathbb{S}}(m)}{2\epsilon^2}. \tag{1.5}$$

We now rewrite the differential system (1.4) in terms of projection on the sets $\mathbb{S}$ and $\mathbb{M}_N$. Since the space of measure-preserving maps $\mathbb{S}$ is closed but not convex, any point in $\mathbb{M}$ admits a projection on $\mathbb{S}$, but this projection is usually not uniquely defined. To simplify the exposition we will nonetheless associate to any point $m \in \mathbb{M}$ one of its projection $P_{\mathbb{S}}(m)$, i.e. any point in $\mathbb{S}$ such that $\|P_{\mathbb{S}}(m) - m\|_{\mathbb{M}} = \mathrm{d}_{\mathbb{S}}(m)$. We also denote $P_{\mathbb{M}_N} : \mathbb{M} \to \mathbb{M}_N$ the orthogonal projection on the linear subspace $\mathbb{M}_N \subseteq \mathbb{M}$, which is a linear map. We can rewrite Eq. (1.4) in terms of these two projection operators:

$$\begin{cases} \ddot{m}(t) + \frac{m(t) - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(m(t))}{\epsilon^2} = 0, \quad \text{for } t > 0\,, \\ (m(0), \dot{m}(0)) \in \mathbb{M}^2_N \end{cases} \tag{1.6}$$

From Proposition 5.2, the double projection $P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(m)$ is uniquely defined for almost every $m \in \mathbb{M}_N$.

Remark 1.1. Equation (1.6) can be rewritten as a system of N particles in interaction, whose positions are denoted $M_1(t), \dots, M_N(t) \in \mathbb{R}^d$. Denoting $\mathbf{1}_{\omega_i}$ the indicator function of the set $\omega_i \subseteq \Omega$, we introduce

$$W : (M_1, \dots, M_N) \in (\mathbb{R}^d)^N \mapsto \mathrm{d}^2_{\mathbb{S}}(\sum_i M_i \mathbf{1}_{\omega_i}),$$

and we denote $B_i(M_1, \dots, M_N) = \nabla_{M_i} W(M_1, \dots, M_N)$. As explained in Proposition 5.2, the points $(B_i(M_1, \dots, M_N))_i$ are barycenters of a decomposition of Ω into N cells which

depend on the solution to the optimal transport problem between Leb and the empirical measure $\frac{1}{N}\sum_{1\leq i\leq N}\delta_{M_i}$. With these notations, Equation (1.6) is then equivalent to

$$\begin{cases} \ddot{M}_i(t) + \frac{1}{\epsilon^2}\left(M_i(t) - B_i(M_1(t),\dots,M_N(t))\right) = 0, \quad \text{for } t>0 \text{ and } i\in\{1,\dots,N\}, \\ (M(0),\dot{M}(0)) \in (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N \end{cases} \tag{1.7}$$

Loosely speaking, equations (1.4)–(1.6) describe a physical system where each particle $M_i(t)$ is subject to the force of a spring with stiffness $\frac{1}{\epsilon}$ attached to the point $B_i(M_1(t),\dots,M_N(t))$ which varies in time and depends on the position of all the particles. Equation (1.7) is also the Hamiltonian system associated to $\mathrm{H} : (\mathbb{R}^d)^N \times (\mathbb{R}^d)^N \to \mathbb{R}$

$$\mathrm{H}(M,\dot{M}) = \frac{1}{2}\sum_{i=1}^{N} |\dot{M}_i|^2 + \mathrm{W}(M), \tag{1.8}$$

In the case of an non-homogeneous fluid with varying volume masse, such as a mixture of oil and water, an analogue discretization would involve a system of particles with different masses ρ_i. This corresponds to replacing the Hamiltonian by

$$\mathrm{H}(M,\dot{M}) = \frac{1}{2}\sum_{i=1}^{N} \rho_i|\dot{M}_i|^2 + \mathrm{W}(M). \tag{1.9}$$

In this last formulation, it is also possible to add potential terms, such as gravitation. This will be the case for the simulation of the Rayleigh-Taylor instability in subsection 5.4.

We first prove that the system of equations (1.4) can be used to approximate regular solutions to Euler's equations (1.1). Our proof of convergence uses a modulated energy technique which is similar to that used in [8] and requires $\mathcal{C}^{1,1}$ regularity assumptions on the solution to Euler's equations. See also [10, 12] for related works.

Theorem 1.2. *Let Ω be a bounded domain of $\mathbb{R}^d$ with Lipschitz boundary. Let v, p be a strong solution of Euler's equations (1.1), let ϕ be the flow map induced by v (see (1.2)) and assume that $v, p, \partial_t v, \partial_t p, \nabla v$ and ∇p are Lipschitz on Ω, uniformly on $[0,T]$. Suppose in addition that there exists a $\mathcal{C}^1$ curve $m : [0,T] \to \mathbb{M}_N$ satisfying the initial conditions*

$$m(0) = P_{\mathbb{M}_N}(\mathrm{id}), \quad \dot{m}(0) = P_{\mathbb{M}_N}(v(0,\cdot)),$$

which is twice differentiable and satisfies the second-order equation (1.4) for all times in $[0,T]$, possibly up to a countable number of exceptions. Then,

$$\max_{t\in[0,T]} \|\dot{m} - v(t,\phi(t,\cdot))\|_{\mathbb{M}}^2 \leq C_1\frac{h_N^2}{\varepsilon^2} + C_2\varepsilon^2 + C_3 h_N \tag{1.10}$$

where the constants C_1, C_2 and C_3 only depend on Ω, on the L^∞ norm (in space) of the velocity $v(t,\cdot)$ and on the Lipschitz norms (in space) of the velocity and its first derivatives $\nabla v(t,\cdot), \partial_t v(t,\cdot)$ and of the pressure and its derivatives $p(t,\cdot)$, $\nabla p(t,\cdot)$, $\partial_t p(t,\cdot)$.

The values of C_1, C_2 and C_3 are given more precisely at the end of Section 3. Note that the hypothesis on the solution m to the differential equation (1.4) is introduced here mainly for technical reasons. Removing it is not of our main concern in this paper since we also give a proof of convergence of the fully discrete numerical scheme regardless of this assumption. It is likely that solutions to (1.4) satisfying this hypothesis can be constructed through di Perna-Lions or Bouchut-Ambrosio theory [1, 4, 19], see also [10, Appendix].

Remark 1.3. Remark that (1.10) implies the convergence of the associated flows. In particular integrating inequality (1.10) one can show that

$$\max_{t\in[0,T]} \|m(t) - \phi(t)\|_{\mathbb{M}}^2 \leq 2h_N^2 + 2T\left(C_1\frac{h_N^2}{\varepsilon^2} + C_2\varepsilon^2 + C_3 h_N\right).$$

Discretization in space and time. To obtain a numerical scheme we also need to discretize in time the Hamiltonian system (1.6). For simplicity of the analysis, we consider a simple first-order scheme called *symplectic Euler scheme* with timestep $\tau > 0$. The double projection $P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(m)$ is defined as above. The discrete solution consists of two sequences M^n, V^n in the finite-dimensional space $\mathbb{M}_N$, given by:

$$\begin{cases} (M^0, V^0) \in \mathbb{M}_N \\ V^{n+1} = V^n - \frac{\tau}{\epsilon^2}\left(M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)\right) \\ M^{n+1} = M^n + \tau V^{n+1} \end{cases} \tag{1.11}$$

Note that numerically, the piecewise-constant map $M^n : \Omega \to \mathbb{R}^d$ (resp. the piecewise-constant vector field $V^n : \Omega \to \mathbb{R}^d$) is simply encoded by an ordered list of N points (resp. N vectors), so that this scheme can be considered as describing a dynamical system involving N particles. We have the following theorem, where we denote $t^n = n\tau$.

Theorem 1.4. *Let Ω be a bounded domain of $\mathbb{R}^d$ with Lipschitz boundary, let ϵ and τ be positive numbers and let $N \in \mathbb{N}$. Let v, p be a strong solution of (1.1), let ϕ be the flow map induced by v (see (1.2)) and assume that $v, p, \partial_t v, \partial_t p, \nabla v$ and ∇p are Lipschitz on Ω, uniformly on $[0, T]$. Let $(M^n, V^n)_{n\geq 0}$ be a sequence generated by (1.11) from*

$$M^0 = P_{\mathbb{M}_N}(\mathrm{id}), \;\; V^0 = P_{\mathbb{M}_N}(v(0, \cdot)).$$

Assuming $\tau \leq \varepsilon$ and $h_N \leq \varepsilon$, we have

$$\max_{n \in \mathbb{N} \cap [0, T/\tau]} \|V^n - v(t^n, \phi(t^n, \cdot))\|^2_{\mathbb{M}} \leq C\left[\epsilon^2 + h_N + \frac{h_N^2}{\epsilon^2} + \frac{\tau}{\epsilon^2}\right],$$

where the constant C only depends on Ω, on the L^∞ norm (in space) of the velocity $v(t, \cdot)$ and on the Lipschitz norms (in space) of the velocity and its first derivatives $\nabla v(t, \cdot), \partial_t v(t, \cdot)$ and of the pressure and its derivatives $p(t, \cdot), \nabla p(t, \cdot), \partial_t p(t, \cdot)$.

In order to use the numerical scheme (1.11), one needs to be able to compute the double projection operator $P_{\mathbb{M}_N} \circ P_{\mathbb{S}}$ or equivalently the gradient of the squared distance $\mathrm{d}^2_{\mathbb{S}}$ for (almost every) m in $\mathbb{M}_N$. Brenier's polar factorization problem [7] implies that the squared distance between a map $m : \Omega \to \mathbb{R}$ and the set $\mathbb{S}$ of measure-preserving maps is equal to the squared Wasserstein distance [24] between the restriction of the Lebesgue measure to Ω, denoted Leb, and its pushforward $m_\# \,\mathrm{Leb}$ under the map m:

$$\mathrm{d}^2_{\mathbb{S}}(m) = \min_{s \in \mathbb{S}} \|m - s\|^2 = \mathrm{W}_2^2(m_\# \,\mathrm{Leb}, \mathrm{Leb}).$$

Moreover, since m is piecewise-constant over the partition $(\omega_i)_{1\leq i\leq N}$, the push-forward measure $m_\# \,\mathrm{Leb}$ if finitely supported. Denoting by $M_i \in \mathbb{R}^d$ the constant value of the map m on the subdomain ω_i we have

$$m_\# \,\mathrm{Leb} = \sum_{1\leq i\leq N} \mathrm{Leb}(\omega_i)\delta_{M_i} = \frac{1}{N}\sum_{1\leq i\leq N} \delta_{M_i}.$$

Thus, computing the projection operator $P_{\mathbb{S}}$ amounts to the numerical resolution of an optimal transport problem between the Lebesgue measure on Ω and a finitely supported measure. Thanks to recent work [3, 20, 13, 18], this problem can be solved efficiently in dimensions $d = 2, 3$. We give more details in Section 5.

Remark 1.5. The idea of using optimal transport to impose incompressibility contraints has recently been exploited as a heuristic for computational fluid dynamics simulations in computer graphics [14]. From the simulations presented in [14], it seems that the scheme behaves better numerically, and it also has the extra advantage of not depending on a penalization parameter ε. However, it comes with no mathematical convergence analysis, and even its (formal) consistence is not obvious. It would therefore be interesting to extend the convergence analysis presented in Theorem 1.4 to the scheme presented in [14]. This however probably requires new ideas, as our technique of proof relies heavily on the fact

that the space-discretization is hamiltonian, an assumption which does not seem to hold for the discretization of [14].

Remark 1.6. Our discretization (1.4) resembles (and derives from) a space-discretization of Euler's equations (1.1) introduced by Brenier in [8]. The domain is also decomposed into subdomains $(\omega_i)_{1\leq i\leq N}$, and one considers the set $\mathbb{S}_N \subseteq \mathbb{S}$, which consists of measure-preserving maps $s : \Omega \to \Omega$ that are induced by a permutation of the subdomains. Equivalently, one requires that there exists a permutation $s : \{1,\dots,N\} \to \{1,\dots,N\}$ such that $s(\omega_i) = \omega_{s(j)}$. The space-discretization considered in [8] leads to an ODE similar to (1.4), but where the squared distance to $\mathbb{S}$ is replaced by the squared distance to $\mathbb{S}_N$. This choice of discretization imposes strong contraints on the relative size of the parameters τ, h_N and ϵ, namely that $h_N = \mathrm{O}(\varepsilon^8)$ and $\tau = \mathrm{O}(\varepsilon^4)$. Such constraints still exist with the discretization that we consider here, but they are milder. In Theorem 1.4 the condition $\tau = o(\epsilon^2)$ is due to the time discretization of (1.6) and can be improved using a scheme more accurate on the conservation of the Hamiltonian (1.5). However even with an exact time discretization of the Hamiltonian, the condition $\tau = o(\epsilon)$ remains mandatory, as explained at the end of Section 4.

2. Preliminary discussion on geodesics

To illustrate the approximate geodesic scheme we focus on the very simple example of $\mathbb{R}$ seen as $\mathbb{R}\times\{0\} \subset \mathbb{R}^2$. The geodesic is given by the function $\gamma\colon [0,T] \to \mathbb{R}^2$ with

$$\begin{cases} \gamma(t) = (t,0),\ t\in[0,T], \\ \gamma(0) = (0,0), \\ \dot\gamma(0) = (1,0). \end{cases} \tag{2.1}$$

As in (1.4) we consider the solutions of the Hamiltonian system associated to:

$$H(m,\dot m) = \frac{1}{2}||\dot m||^2 + \frac{1}{2\epsilon^2} d^2_{\mathbb{R}\times\{0\}}(m). \tag{2.2}$$

That is

$$\begin{cases} \ddot m(t) = \frac{1}{\epsilon^2}\left(P_{\mathbb{R}}(m) - m\right) = \frac{1}{2\epsilon^2}\nabla d^2_{\mathbb{R}\times\{0\}}(m),\ t\in[0,T], \\ m(0) = (0,h_0), \\ \dot m(0) = (1,h_1). \end{cases} \tag{2.3}$$

where $P_{\mathbb{R}}(m)$ is the orthogonal projection from $\mathbb{R}^2$ onto $\mathbb{R}\times\{0\}$. Notice that we assumed an initial error of h_0 on the initial position and h_1 on the initial velocity. In this case the solution is explicit and reads

$$m(t) = \left(t, h_0\cos\frac{t}{\epsilon} + \epsilon h_1 \sin\frac{t}{\epsilon}\right). \tag{2.4}$$

A convenient way to quantify how far m is from being a geodesic is to use a modulated energy related to the Hamiltonian H and the solution γ. We define E_γ by

$$E_\gamma(t) = \frac{1}{2}||\dot m(t) - \dot\gamma(t)||^2 + \frac{1}{2\epsilon^2} d^2_{\mathbb{R}\times\{0\}}(m(t)). \tag{2.5}$$

A direct computation leads to

$$E_\gamma(t) = \frac{h_0^2}{\epsilon^2} + h_1^2. \tag{2.6}$$

This estimate shows that the velocity vector field $\dot m$ converges towards the geodesic velocity vector fields $\dot\gamma$ as soon as h_0 goes to 0 faster then ϵ. Our construction of approximate geodesics for the Euler equation follow this idea. Estimates (2.6) suggests that our convergence results for the incompressible Euler equation in Theorem 1.2 is sharp.

3. Convergence of the approximate geodesics model

In this section we prove Theorem 1.2.

3.1. **Strategy of the proof.** We use a modulated energy approach. Let v be a solution of (1.1) and m a solution of (1.4) and for any $t \in [0,T]$, denote $\sigma(t) = P_{\mathbb{S}}(m(t))$. In other words, $\sigma(t)$ is an arbitrary choice of a projection of $m(t)$ on $\mathbb{S}$. Equation (1.4) is the ODE associated to the Hamiltonian $H : \mathbb{M}_N \times \mathbb{M}_N \to \mathbb{R}$

$$H(m,\dot{m}) = \frac{1}{2}\|\dot{m}\|^2_{\mathbb{M}} + \frac{d^2_{\mathbb{S}}(m)}{2\epsilon^2}.$$

We therefore consider a energy involving this Hamiltonian, modulated with the exact solution v:

$$E_v(t) = \frac{1}{2}\|\dot{m}(t) - v(t,m(t))\|^2_{\mathbb{M}} + \frac{d^2_{\mathbb{S}}(m)}{2\epsilon^2}. \tag{3.1}$$

The core of the proof is to obtain a control on E_v using a Gronwall estimate. As a first step we collect some lemmas. Lemmas 3.1 and 3.2 concern the projections $\Pi_{\mathbb{M}_N}$ and $\Pi_{\mathbb{S}}$ and their orthogonality properties. Lemma 3.3 is necessary to ensure that the modulated energy introduced in (3.1) is well defined (the difficulty is that there is no reason that $m(t,\Omega) \subseteq \Omega$, and it is therefore necessary to extend v outside of Ω). Then we compute the derivative of (3.1) and modify its expression so as to identify terms of quadratic order, which are easier to control. This leads us to (3.7), which expresses the derivative of (3.1) as a sum of many terms. Each term is then estimated to obtain a Gronwall control. we keep track of the constants all along the proof.

3.2. **Preliminary lemma.** Before proving Theorem 1.2, we collect a few useful lemmas. As before, Ω is a bounded and connected domain of $\mathbb{R}^d$ with Lipschitz boundary.

Lemma 3.1 (Projection onto the measure preserving maps $\mathbb{S}$)**.** *Let $m \in \mathbb{M} = \mathrm{L}^2(\Omega,\mathbb{R}^d)$. There exists a convex function $\varphi : \Omega \to \mathbb{R}$, which is unique up to an additive constant, such that $\sigma \in \mathbb{M}$ belongs to $\Pi_{\mathbb{S}}(m)$ if and only if $m = \nabla\varphi \circ \sigma$ up to a negligible set. Moreover, $m - \sigma$ is orthogonal to the space $\mathcal{H}_{\mathrm{div}}(\Omega)\circ\sigma$, that is*

$$\forall v \in \mathcal{H}_{\mathrm{div}}(\Omega), \int_\Omega \langle m(x) - \sigma(x) | v(\sigma(x))\rangle \mathrm{d}x = 0. \tag{3.2}$$

Proof. The first part of the statement is Brenier's polar factorization theorem [7]. We first remark that

$$d^2_{\mathbb{S}}(m) = \inf_{s\in\mathbb{S}} \int \|m(x) - s(x)\|^2 dx \geq \inf_{\pi\in\Pi(m_\# \mathrm{Leb},\mathrm{Leb})} \int \|x-y\|^2 \mathrm{d}\pi(x,y) = \mathrm{W}_2^2(m_\#\mathrm{Leb},\mathrm{Leb}).$$

To prove the reverse inequality let $\nabla\varphi$ be the optimal transport map between $m_\#\mathrm{Leb} = \sum_{1\leq i\leq N}\delta_{M_i}$ and Leb. Let $L_i = \nabla\varphi^{-1}(M_i)$, by construction $\mathrm{Leb}(L_i) = \frac{1}{N}$. For any $i \in \{1...N\}$ let σ_i be a measure preserving map between ω_i and L_i, we define a measure preserving map $\sigma \in \mathbb{S}$ by $\sigma_{|\omega_i} = \sigma_i$ (anything can be done on the boundaries of the cells). By construction $m = \nabla\varphi\circ\sigma$ and $\mathrm{W}_2^2(m_\#\mathrm{Leb},\mathrm{Leb}) = \|m-\sigma\|^2$. The uniqueness of φ follows from the connectedness of the domain. Using a regularization argument we deduce the orthogonality relation

$$\int_\Omega \langle m(x)|v(\sigma(x))\rangle \mathrm{d}x = \int_\Omega \langle \nabla\varphi\circ\sigma(x)|v(\sigma(x))\rangle \mathrm{d}x = \int_\Omega \langle \nabla\varphi(x)|v(x)\rangle = -\int_\Omega \varphi \mathrm{div}\, v(x) = 0.$$

□

Lemma 3.2 (Projection onto the piecewise constant set $\mathbb{M}_N$)**.** *The projection of a function $g \in L^2(\Omega,\mathbb{R}^d)$ on $\mathbb{M}_N$ is the following piecewise constant function :*

$$\Pi_{\mathbb{M}_N}(g) = \sum_{i=1}^N G_i \mathbf{1}_{\omega_i}, \text{ with } G_i := \frac{1}{\mathrm{Leb}(\omega_i)}\int_{\omega_i} g(x)\mathrm{d}x$$

and where $\mathbf{1}_{\omega_i}$ is the indicator function of the subdomain ω_i.

Proof. It suffices to remark that for any $m \in \mathbb{M}_N$, $m = \sum_{1\leq i\leq N} M_i \mathbf{1}_{\omega_i}$,

$$\langle g|m\rangle_{\mathbb{M}} = \int_\Omega \langle m(x)|g(x)\rangle \mathrm{d}x = \sum_{1\leq i\leq N} \langle M_i| \int_{\omega_i} g(x)\mathrm{d}x\rangle = \langle m| \sum_i G_i \mathbf{1}_{\omega_i}\rangle_{\mathbb{M}} \qquad \square$$

Lemma 3.3. *Let $\Omega \subset \mathbb{R}^d$, let $(V, \|.\|)$ be a finite-dimensional normed vector space. There exists a linear map $L : \mathcal{C}^{1,1}(\Omega, V) \to \mathcal{C}^{1,1}(\mathbb{R}^d, V)$ such that for any $f \in \mathcal{C}^{1,1}(\Omega, V)$,*

(i) $Lf|_\Omega = f$,
(ii) $\|Lf\|_{\mathcal{C}^{1,1}(\mathbb{R}^d,V)} \leq C \|Lf\|_{\mathcal{C}^{1,1}(\Omega,V)}$.

Proof. This lemma is a particular case of Theorem 2 in [16]. We also refer to [11, 15] for previous results. $\square$

We are now ready to prove Theorem 1.2. In the following the dot refers to the time derivative and $\langle .|.\rangle$ to the Hilbert scalar product on $\mathbb{M}$. By abuse of notation we denote by the same name a $\mathcal{C}^{1,1}$ function defined on Ω and its (also $\mathcal{C}^{1,1}$) extension defined on the whole space $\mathbb{R}^d$ using Lemma 3.3. The space $\mathbb{R}^d$ is equipped with the canonical Euclidian norm, and the space of $d \times d$ matrices are equiped with the induced dual norm. All the Lipschitz constants that we consider are with respect to these two norms. Finally for a curve $\gamma : t \in [0, T] \mapsto \gamma(t, \cdot)$ we denote $\mathrm{Lip}_{[0,\mathrm{T}]}(\gamma) = \sup_{t\in[0,T]} \mathrm{Lip}(\gamma(t, \cdot))$.

Material derivatives. Given $(v, p) \in \mathcal{C}^1([0, T], \mathcal{C}^{1,1}(\mathbb{R}^d, \mathbb{R}^d)) \times \mathcal{C}^1([0, T], \mathcal{C}^{1,1}(\mathbb{R}^d, \mathbb{R}^d))$ and $X \in \mathbb{M}$, we define the two following functions, often called material derivatives:

$$\begin{cases} D_t v(t, X) &= \partial_t v(t, X) + (v(t, X) \cdot \nabla)\, v(t, X), \\ D_t p(t, X) &= \partial_t p(t, X) + \langle v(t, X), \nabla p(t, X))\rangle\,. \end{cases} \tag{3.3}$$

Remark that $D_t v$ and $D_t p$ are Lipschitz operators with

$$\begin{aligned} \mathrm{Lip}_{[0,\mathrm{T}]}(D_t v) &\leq \mathrm{Lip}_{[0,\mathrm{T}]}(\partial_t v) + \mathrm{Lip}_{[0,\mathrm{T}]}(v)\|\nabla v\|_{L^\infty} + \mathrm{Lip}_{[0,\mathrm{T}]}(\nabla v)\|v\|_{L^\infty} \\ &\leq \mathrm{Lip}_{[0,\mathrm{T}]}(\partial_t v) + \mathrm{Lip}_{[0,\mathrm{T}]}(v)\,\mathrm{Lip}_{[0,\mathrm{T}]}(v) + \mathrm{Lip}_{[0,\mathrm{T}]}(\nabla v)\|v\|_{L^\infty} \end{aligned} \tag{3.4}$$

$$\begin{aligned} \mathrm{Lip}_{[0,\mathrm{T}]}(D_t p) &\leq \mathrm{Lip}_{[0,\mathrm{T}]}(\partial_t p) + \mathrm{Lip}_{[0,\mathrm{T}]}(v)\|\nabla p\|_{L^\infty} + \mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p)\|v\|_{L^\infty} \\ &\leq \mathrm{Lip}_{[0,\mathrm{T}]}(\partial_t p) + \mathrm{Lip}_{[0,\mathrm{T}]}(v)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p) + \mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p)\|v\|_{L^\infty}. \end{aligned} \tag{3.5}$$

3.3. **Proof of Theorem 1.2.** We can now go to the proof of Theorem 1.2. Note that we need to use Lemma 3.3 to define the modulated energy E_v in (3.1) since the maps $m(t, \cdot) \in \mathbb{M}_N$ can send points outside of Ω when Ω is not convex.

3.3.1. *Time derivative.* We compute $\frac{d}{dt}E_v(t)$ and modify the expression in order to identify terms of quadratic order. Since the Hamiltonian $H(\dot m(t), m(t))$ is preserved, we find

$$\frac{d}{dt}E_v(t) = \underbrace{-\langle \ddot m(t), v(t, m(t))\rangle}_{I_1} \underbrace{- \langle \dot m(t) - v(t, m(t)), \partial_t v(t, m(t)) + (\dot m(t) \cdot \nabla)\, v(t, m(t))\rangle}_{I_2}. \tag{3.6}$$

Using the EDO (1.4), I_1 can be rewritten as

$$\begin{aligned} \epsilon^2 I_1 &= \langle m(t) - P_{\mathbb{M}_N}(\sigma(t)), v(t, m(t))\rangle \\ &= \langle m(t) - \sigma(t), v(t, m(t))\rangle + \langle \sigma(t) - P_{\mathbb{M}_N}(\sigma(t)), v(t, m(t))\rangle \\ &= \underbrace{\langle m(t) - \sigma(t), v(t, m(t)) - v(t, \sigma(t))\rangle}_{\epsilon^2 I_3}, \end{aligned}$$

where we have used that $\sigma(t) - P_{\mathbb{M}_N}(\sigma(t))$ is orthogonal to $\mathbb{M}_N$ and that $m(t) - \sigma(t)$ is orthogonal to $\mathcal{H}_{\mathrm{div}}(\Omega) \circ \sigma$, see Lemmas 3.2 and 3.1. To handle the term I_2 we use

the material derivatives defined by (3.3). Remark that Euler equations (1.1) implies that $D_t v(t,\sigma(t)) = -\nabla p(t,\sigma(t))$. This leads to

$$\begin{aligned} I_2 = &- \langle \dot m(t) - v(t,m(t)), \partial_t v(t,m(t)) + (v(t,m(t))\cdot\nabla)\, v(t,m(t))\rangle \\ &\underbrace{- \langle \dot m(t) - v(t,m(t)), (\dot m(t) - v(t,m(t))\cdot\nabla)\, v(t,m(t))\rangle}_{I_4} \\ = I_4 &\underbrace{- \langle \dot m(t) - v(t,m(t)), D_t v(t,m(t)) - D_t v(t,\sigma(t))\rangle}_{I_5} + \underbrace{\langle \dot m(t) - v(t,m(t)), \nabla p(t,\sigma(t))\rangle}_{I_6} \end{aligned}$$

We rewrite I_6 as

$$\begin{aligned} I_6 &= \underbrace{- \langle \dot m(t) - v(t,m(t)), \nabla p(t,m(t)) - \nabla p(t,\sigma(t))\rangle}_{I_7} + \langle \dot m(t) - v(t,m(t)), \nabla p(t,m(t))\rangle \\ &= I_7 + \frac{d}{dt}\underbrace{\int_\Omega p(t,m(t,x)))dx}_{-J(t)} - \int_\Omega \partial_t p(t,m(t,x)) - \langle v(t,m(t,x)), \nabla p(t,m(t,x))\rangle\, dx \\ &= -\frac{d}{dt}J(t) + I_7 - \underbrace{\int_\Omega D_t p(t,m(t,x))dx}_{I_8}. \end{aligned}$$

Remark 3.4. The quantity $I_5 + I_7$ would vanish if (v,p) was a solution to the Euler equations on the whole space $\mathbb{R}^d$. This is not the case in our setting, as the couple (v,p) is constructed by the extension Lemma 3.3.

Collecting the above decompositions (3.6) rewrites

$$\frac{d}{dt}E_v(t) = I_3 + I_4 + I_5 + I_7 + I_8 - \frac{d}{dt}J(t). \tag{3.7}$$

3.3.2. *Estimates.* Many of the integrals $I_3, I_4, \dots$ can be easily bounded using the energy E_v and Cauchy-Schwarz' and Young's inequalities. First,

$$\begin{aligned} I_3 &\le \left| \frac{\langle m(t) - \sigma(t), v(t,m(t)) - v(t,\sigma(t))\rangle}{\epsilon^2} \right| \\ &\le \mathrm{Lip}(v(t)) \frac{\|m(t) - \sigma(t)\|_{\mathbb{M}}^2}{\epsilon^2} \le \mathrm{Lip}_{[0,\mathrm{T}]}(v) E_v(t). \end{aligned} \tag{3.8}$$

Furthermore

$$I_4 \le \sup_{x\in\mathbb{R}^d} ||\nabla v(t,x)|| \|\dot m(t) - v(t,m(t))\|_{\mathbb{M}}^2 \le \mathrm{Lip}_{[0,\mathrm{T}]}(v) E_v(t), \tag{3.9}$$

Where C depends only on the dimension d. To estimate I_5 and later I_8 we use that $D_t v$ and $D_t p$ are Lipschitz operators with constants given by (3.4) and (3.5). For I_5 we obtain

$$\begin{aligned} I_5 &\le |\langle \dot m(t) - v(t,m(t)), D_t v(t,m(t)) - D_t v(t,\sigma(t))\rangle| \\ &\le \mathrm{Lip}_{[0,\mathrm{T}]}(D_t v) \|\dot m(t) - v(t,m(t))\|_{\mathbb{M}} \|m(t) - \sigma(t)\|_{\mathbb{M}} \\ &\le \epsilon\, \mathrm{Lip}_{[0,\mathrm{T}]}(D_t v) E_v(t), \end{aligned} \tag{3.10}$$

where we used $\mathrm{d}_{\mathbb{S}}(m(t)) = \|m(t) - \sigma(t)\|_{\mathbb{M}} \le \epsilon\sqrt{E_v(t)}$ and $\|\dot m(t) - v(t,m(t))\|_{\mathbb{M}} \le \sqrt{E_v(t)}$ to get from the second to the third line. The quantity I_7 can be bounded likewise:

$$\begin{aligned} I_7 &\le |\langle \dot m(t) - v(t,m(t)), \nabla p(t,m(t)) - \nabla p(t,\sigma(t))\rangle| \\ &\le \epsilon\, \mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p) E_v(t). \end{aligned} \tag{3.11}$$

Finally to estimate I_8 and J we can assume that $\int_\Omega p(t,x)\mathrm{d}x = 0$ since the pressure is defined up to a constant. Using that $\sigma(t)$ is measure-preserving, this gives

$$\begin{aligned}\int_\Omega D_t p(t,\sigma(t,x))dx &= \int_\Omega \partial_t p(t,\sigma(t,x)) + \langle v(t,\sigma(t,x)), \nabla p(t,\sigma(t,x)))\rangle\, dx \\ &= \int_\Omega \partial_t p(t,x))dx + \int_\Omega \langle v(t,x), \nabla p(t,x))\rangle\, dx = 0,\end{aligned}$$

Therefore, using Young's inequality,

$$\begin{aligned}I_8 &\le \left|\int_\Omega D_t p(t,m(t,x))dx - \int_\Omega D_t p(t,\sigma(t,x))dx\right| \le \mathrm{Lip}_{[0,\mathrm{T}]}(D_t p)\|m(t)-\sigma(t)\|_{L^1(\Omega)} \\ &\le \frac{1}{2}\frac{||m(t)-\sigma(t)||^2_{L^2(\Omega)}}{2\epsilon^2} + C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(D_t p)\epsilon^2 \\ &\le \frac{1}{2}E_v(t) + C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(D_t p)\epsilon^2, \qquad (3.12)\end{aligned}$$

where in this estimates and in the following estimates $C(\Omega)$ is a constant depending only on the Lebesgue measure of Ω. Similarly,

$$\begin{aligned}|J(t)| &\le \left|\int_\Omega p(t,m(t,x))) - p(t,\sigma(t,x)))dx\right| \le \mathrm{Lip}_{[0,\mathrm{T}]}(p)||m(t)-\sigma(t)||_{L^1(\Omega)} \\ &\le \frac{1}{2}E_v(t) + C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)\epsilon^2. \qquad (3.13)\end{aligned}$$

We finally remark that

$$|J(0)| \le \mathrm{Lip}_{[0,\mathrm{T}]}(p)h_N. \qquad (3.14)$$

Remark 3.5. The last two estimates show that we can add $\frac{\mathrm{d}}{\mathrm{d}t}J$ into the Gronwall argument. It is a general fact that the derivative of a controlled quantity can be added. This is a classical way of controlling the term of order one in the energy.

3.4. **Gronwall argument.** Collecting estimates (3.8), (3.9), (3.10), (3.11), (3.12), we get

$$\begin{aligned}&\frac{d}{dt}\left(E_v(t)+J(t)\right) \le I_3 + I_4 + I_5 + I_7 + I_8 \\ &\le \left[2\,\mathrm{Lip}_{[0,\mathrm{T}]}(v) + \epsilon\,\mathrm{Lip}_{[0,\mathrm{T}]}(D_t v) + \epsilon\,\mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p) + \frac{1}{2}\right]E_v(t) \\ &+ C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(D_t p)\epsilon^2\end{aligned}$$

Remark that (3.13) implies that for any $K > 0$,

$$KE_v(t) \le KE_v(t) + 2KJ(t) - 2KJ(t) \le 2KE_v(t) + 2KJ(t) + 2KC(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)\epsilon^2. \quad (3.15)$$

Therefore, setting

$$\begin{cases}\widetilde{C}_1 &= C(\Omega)\left(4\,\mathrm{Lip}_{[0,\mathrm{T}]}(v) + 2\epsilon\,\mathrm{Lip}_{[0,\mathrm{T}]}(D_t v) + 2\epsilon\,\mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p) + 1\right), \\ \widetilde{C}_2 &= C(\Omega)\left(\mathrm{Lip}_{[0,\mathrm{T}]}(D_t p) + \widetilde{C}_1\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)\right),\end{cases}$$

we obtain

$$\frac{d}{dt}\left(E_v(t)+J(t)\right) \le \widetilde{C}_1(E_v(t)+J(t)) + \widetilde{C}_2\epsilon^2.$$

We deduce from the Gronwall inequality that for any $t \in [0,T]$:

$$E_v(t) \le \left((E_v(0)+J(0)) + \widetilde{C}_2 T\epsilon^2\right)e^{\widetilde{C}_1 T} - J(t).$$

Using the estimation (3.13) one more time we obtain

$$E_v(t) \le 2\left(E_v(0) + \mathrm{Lip}_{[0,\mathrm{T}]}(p)h_N + \widetilde{C}_2 T\epsilon^2\right)e^{\widetilde{C}_1 T} + C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)\epsilon^2.$$

Finally, using that

$$E_v(0) = \frac{1}{2}\|P_{\mathbb{M}}(v_0) - v_0\|^2_{\mathbb{M}} + \frac{d^2_{\mathbb{S}}(\mathrm{Id}\,)}{2\epsilon^2} \le \frac{h_N^2}{2} + \frac{h_N^2}{2\epsilon^2}$$

we obtain

$$\begin{aligned}\|\dot m(t)-v(t,m(t))\|^2_{\mathbb{M}} &\le 2E_v(t)\\ &\le 2\left[2\left(\frac{h_N^2}{2}+\frac{h_N^2}{2\epsilon^2}+\mathrm{Lip}_{[0,\mathrm{T}]}(p)h_N+\widetilde{C}_2T\epsilon^2\right)e^{\widetilde{C}_1T}\right.\\ &\quad\left.+\,C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)\epsilon^2\right] \qquad (3.16)\\ &\le C_1'\frac{h_N^2}{\epsilon^2}+C_2'\epsilon^2+C_3'h_N \qquad (3.17)\end{aligned}$$

where

$$\begin{cases} C_1' &= 2e^{\widetilde{C}_1T}\\ C_2' &= \left(\widetilde{C}_2Te^{\widetilde{C}_1T}+C(\Omega)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)\right)\\ C_3' &= (1+\mathrm{Lip}_{[0,\mathrm{T}]}(p))e^{\widetilde{C}_1T}.\end{cases}$$

In order to estimate $\|\dot m(t)-v(t,\phi(t))\|^2_{\mathbb{M}}$ we need one additional Gronwall estimate:

$$\begin{aligned}\|\dot m(t)-v(t,\phi(t))\|^2_{\mathbb{M}} &\le 2\,\|\dot m(t)-v(t,m(t))\|^2_{\mathbb{M}}+2\,\|v(t,m(t))-v(t,\phi(t))\|^2_{\mathbb{M}}\\ &\le 2E_v(t)+2(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2\,\|m(t)-\phi(t)\|^2_{\mathbb{M}}\\ &\le 2E_v(t)+4(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2\,\|m(0)-\phi(0)\|^2_{\mathbb{M}}\\ &\quad+4(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2\left\|\int_0^t(\dot m(s)-\dot\phi(s))ds\right\|^2_{\mathbb{M}}\\ &\le 2E_v(t)+4(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2h_N^2+4T(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2\int_0^t\left\|(\dot m(s)-\dot\phi(s))\right\|^2_{\mathbb{M}}ds\\ &\le C_1'\frac{h_N^2}{\epsilon^2}+C_2'\epsilon^2+4(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2h_N^2+C_3'h_N\\ &\quad+4T(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2\int_0^t\left\|\dot m(s)-\dot\phi(s)\right\|^2_{\mathbb{M}}ds \qquad (3.18)\end{aligned}$$

where we used Jensen's inequality to obtain the second to last line. We conclude thanks to Gronwall inequality:

$$\begin{aligned}\|\dot m(t)-v(t,\phi(t))\|^2_{\mathbb{M}} &\le \left(C_1'\frac{h_N^2}{\epsilon^2}+C_2'\epsilon^2+4((\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2h_N+C_3')h_N\right)e^{4T(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2T}\\ &\le C_1\frac{h_N^2}{\epsilon^2}+C_2\epsilon^2+C_3h_N. \qquad (3.19)\end{aligned}$$

We used that ϵ and h_N are smaller than $C(\Omega)$ for (3.17) and (3.19). Observe that the right-hand side of (3.17) and (3.19) goes to zero provided that $\frac{h_N}{\epsilon}$ and ϵ go to zero. This finishes the proof of Theorem 1.2.

Remark 3.6. Using (3.4) and (3.5), the constants $\widetilde{C}_1$, $\widetilde{C}_2$ are bounded by:

$$\begin{cases} \frac{1}{C(\Omega)}\widetilde{C}_1 \le 1+4\,\mathrm{Lip}_{[0,\mathrm{T}]}(v)+2\epsilon\,\mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p)\\ \qquad\qquad +2\epsilon\left(\mathrm{Lip}_{[0,\mathrm{T}]}(\partial_t v)+(\mathrm{Lip}_{[0,\mathrm{T}]}(v))^2+\mathrm{Lip}_{[0,\mathrm{T}]}(\nabla v)\|v\|_{L^\infty}\right),\\ \frac{1}{C(\Omega)}\widetilde{C}_2 \le \mathrm{Lip}_{[0,\mathrm{T}]}(p)+\widetilde{C}_1\left[\mathrm{Lip}_{[0,\mathrm{T}]}(\partial_t p)+\mathrm{Lip}_{[0,\mathrm{T}]}(v)\,\mathrm{Lip}_{[0,\mathrm{T}]}(p)+\mathrm{Lip}_{[0,\mathrm{T}]}(\nabla p)\|v\|_{L^\infty}\right].\end{cases}$$

A close look to the explicit value of the constants $\widetilde{C}_1$, $\widetilde{C}_2$ and C_1', C_2', C_3', together with a diagonal argument shows that our scheme can be used to approximate solutions less regular than those supposed in Theorem 1.2. For example, it is possible to establish the following theorem: Let v, p be a solution of Euler's equation (1.1), where v is merely Lipschitz in space but where there exists $(v_k, p_k)_{k\in\mathbb{N}}$ a sequence of regular (in the sense of Theorem 1.2) solutions of (1.1) such that $v_k(0,\cdot)\longrightarrow v(0,\cdot)$ in $\mathbb{M}$ and $\mathrm{Lip}_T(v_k)\longrightarrow \mathrm{Lip}_T(v)$. Then there exists N_k and ϵ_k, depending polynomially on the data such that $\|\dot m_k(t)-v\left(t,m_k(t)\right)\|^2_{\mathbb{M}}$ goes to zero as k goes to infinity, where m_k is the solution of (1.6) with initial conditions

$m_k(0) = P_{\mathbb{M}_{N_k}}(\mathrm{id})$ and $\dot{m}_k(0) = P_{\mathbb{M}_{N_k}}(v_k(0))$ and with parameter $\epsilon = \epsilon_k$. If one allows an exponential dependence on the data, it is possible to approach any solution whose velocity v belongs to the L^2 closure of the regular solutions to Euler's equation.

4. Convergence of the symplectic Euler scheme

In this section we prove a statement which is slightly more general than Theorem 1.4 (see Remark 4.3), and which allows a sort of a posteriori estimates. The proof follows the proof of Theorem 1.2, but one has to deal with some additional term coming from the time discretization. It combines two Gronwall estimates. The first one is a continuous Gronwall argument on each segment $[n\tau, (n+1)\tau]$, and the second one is a discrete Gronwall estimate comparing a timestep to the next one. Both steps rely on the same modulated energy.

Theorem 4.1. *Let Ω be a bounded domain with Lipschitz boundary and let ϵ, τ positive numbers and let $N \in \mathbb{N}$. Let v, p be a strong solution of (1.1), and let ϕ be the flow map induced by v (see (1.2)). Assume that $v, p, \partial_t v, \partial_t p, \nabla v$ and ∇p are Lipschitz on Ω, uniformly on $[0,T]$. Let $(M^n, V^n)_{n\geq 0}$ be a sequence generated by (1.11) with initial conditions*

$$M^0 = P_{\mathbb{M}_N}(\mathrm{id}),\ V^0 = P_{\mathbb{M}_N}(v(0,\cdot)).$$

Finally let

$$H^n = H(M^n, V^n) = \frac{1}{2}\|V^n\|^2_{\mathbb{M}} + \frac{d^2_{\mathbb{S}}(M^n)}{2\epsilon^2}, \tag{4.1}$$

and

$$\kappa = \max_{n\in\mathbb{N}\cap[0,T/\tau]} \left(H^n - H^0\right).$$

Then, assuming $\tau \leq \epsilon$ and $h_N \leq \varepsilon$, we have

$$\max_{n\in\mathbb{N}\cap[0,T/\tau]} \|V^n - v(t^n, \phi(t^n,\cdot))\|_{\mathbb{M}} \leq C\left[\epsilon^2 + h_N + \frac{h_N^2}{\epsilon^2} + \frac{\tau}{\epsilon} + \kappa\right],$$

where the constant C only depends on Ω, on the L^∞ norm (in space) of the velocity $v(t,\cdot)$ and on the Lipschitz norms (in space) of the velocity and its first derivatives $\nabla v(t,\cdot), \partial_t v(t,\cdot)$ and of the pressure and its derivatives $p(t,\cdot), \nabla p(t,\cdot), \partial_t p(t,\cdot)$.

4.1. Preliminary lemma.

Given a solution of (1.11) and $s \in [0,1]$ and $n \in \mathbb{N}$, we denote the linear interpolates between two timesteps $n\tau$ and $(n+1)\tau$ by:

$$\begin{cases} V^{n+s} &= V^n - s\tau \frac{M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)}{\epsilon^2} \\ M^{n+s} &= M^n + s\tau V^{n+1}, \end{cases} \tag{4.2}$$

We consider the Hamiltonian H^{n+s} and modulated energy E^{n+s} defined by

$$\begin{cases} H^{n+s} &= \frac{1}{2}\|V^{n+s}\|^2_{\mathbb{M}} + \frac{\mathrm{d}^2_{\mathbb{S}}(M^{n+s})}{2\epsilon^2}, \\ E^{n+s} &= \frac{1}{2}\|V^{n+s} - v((n+s)\tau, M^{n+s})\|^2_{\mathbb{M}} + \frac{\mathrm{d}^2_{\mathbb{S}}(M^{n+s})}{2\epsilon^2}. \end{cases} \tag{4.3}$$

We start with a lemma quantifying the conservation of the Hamiltonian.

Lemma 4.2 (Conservation of the Hamiltonian)**.** *For any $s \in [0,1]$ and $n \in \mathbb{N} \cap [0, T/\tau]$,*

$$\left(1 - \frac{\tau^2}{\epsilon^2}\right) H^{n+1} \leq H^n, \tag{4.4}$$

$$H^n \leq C e^{T\tau\epsilon^{-2}}, \tag{4.5}$$

$$H^{n+s} \leq H^n + \frac{\tau^2}{\epsilon^2} H^{n+1}, \tag{4.6}$$

Proof. The proof is based on the 1-semiconcavity of $\frac{1}{2}d^2_{\mathbb{S}}$, see Proposition 5.2 for details. On the one hand the 1-semiconcavity of $\frac{1}{2}d^2_{\mathbb{S}}$ reads

$$\frac{d^2_{\mathbb{S}}(M^{n+s})}{2\epsilon^2} \leq \frac{d^2_{\mathbb{S}}(M^n)}{2\epsilon^2} + s\tau \left\langle V^{n+1}, \frac{M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)}{\epsilon^2} \right\rangle + \frac{s^2\tau^2}{2\epsilon^2}\|V^{n+1}\|^2_{\mathbb{M}},$$

where we used that $M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)$ belongs to the superdifferential of the function $\mathrm{d}^2_{\mathbb{S}}$ at M^n, and the definition of the scheme (4.2). On the other hand, (4.2) again, leads to

$$\frac{\|V^{n+s}\|^2_{\mathbb{M}}}{2} = \frac{\|V^n\|^2_{\mathbb{M}}}{2} - s\tau \left\langle V^n, \frac{M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)}{\epsilon^2} \right\rangle + s^2\tau^2 \left\| \frac{M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)}{\epsilon^2} \right\|^2_{\mathbb{M}}$$

Summing both equations and using (4.2) gives

$$H^{n+s} \leq H^n + \frac{\tau^2 s(s-1)}{\epsilon^2} \frac{\|M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)\|^2_{\mathbb{M}}}{\epsilon^2} + s^2 \frac{\tau^2}{\epsilon^2} \frac{\|V^{n+1}\|^2_{\mathbb{M}}}{2} \tag{4.7}$$

Taking $s = 1$ in (4.7) proves (4.4). The inequality (4.5) is a direct consequence of (4.4), while (4.6) follows from the combination of (4.4) and (4.7). □

Remark 4.3. Lemma 4.2 gives an upper bound for κ in Theorem 4.1 namely

$$\kappa \leq \sum_{n \in \mathbb{N} \cap [0, T/\tau]0} \left| H^{n+1} - H^n \right| \leq \frac{\tau}{\epsilon^2} T e^{T\tau\epsilon^{-2}} \left(\frac{1}{2} \|V^0\|^2_{\mathbb{M}} + \frac{h_N^2}{2\epsilon^2} \right).$$

Using this upper bound Theorem 4.1 becomes Theorem 1.4 and the condition $\kappa = o(1)$ becomes $\tau = o(\epsilon^2)$. However numerically one can expect some compensation in H^n and thus obtain a better "a posteriori bound" for κ in order to get rid of the strong assumption $\tau = o(\epsilon^2)$. Figure 5.4 illustrates the conservation of the Hamiltonian in two test cases. Notice that this estimate is not a posteriori in the usual sense since the constants in Theorem 4.1 also depend on the unknown limiting solution. The condition $\tau = o(\epsilon)$ seems mandatory for the proof techniques to work.

4.2. **The modulated energy.** Remark that with the definitions of the Hamiltonian and modulated energy, we have

$$E^{n+s} = H^{n+s} - \left\langle V^{n+s}, v\left((n+s)\tau, M^{n+s}\right) \right\rangle + \frac{1}{2} \left\| v\left((n+s)\tau, M^{n+s}\right) \right\|^2_{\mathbb{M}}, \tag{4.8}$$

so that for any $s \in [0,1]$ and any $n \in \mathbb{N}$,

$$E^{n+s} = E^n + H^{n+s} - H^n + \int_0^s d^{n+\theta} d\theta, \tag{4.9}$$

where

$$d^{n+s} = \frac{d}{ds} \left[- \left\langle V^{n+s}, v\left((n+s)\tau, M^{n+s}\right) \right\rangle + \frac{1}{2} \left\| v\left((n+s)\tau, M^{n+s}\right) \right\|^2_{\mathbb{M}} \right].$$

To evaluate d^{n+s}, we introduce $\sigma^p = P_{\mathbb{S}}(M^p)$ and we will use the compact notation

$$v^{n+s}_{M^p} = v((n+s)\tau, M^p), \qquad \partial_t v^{n+s}_{M^p} = \partial_t v((n+s)\tau, M^p), \qquad \nabla v^{n+s}_{M^p} = \nabla v((n+s)\tau, M^p),$$

$$v^{n+s}_{\sigma^p} = v((n+s)\tau, \sigma^p), \qquad \partial_t v^{n+s}_{\sigma^p} = \partial_t v((n+s)\tau, \sigma^p), \qquad \nabla v^{n+s}_{\sigma^p} = \nabla v((n+s)\tau, \sigma^p).$$

We will also use a similar notation for the material derivative of the velocity and for the pressure and its derivatives.

Remark 4.4. As before, the main idea of the following computation is to try to find terms of quadratic order in the expression. To control the remaining linear term we have to rewrite it as a derivative of a small quantity and add it in the Gronwall argument.

$$\begin{aligned} d^{n+s} &= - \left\langle \frac{d}{ds} V^{n+s}, v^{n+s}_{M^{n+s}} \right\rangle - \left\langle V^{n+s}, \frac{d}{ds} v^{n+s}_{M^{n+s}} \right\rangle + \left\langle v^{n+s}_{M^{n+s}}, \frac{d}{ds} v^{n+s}_{M^{n+s}} \right\rangle \\ &= \underbrace{\tau\epsilon^{-2} \left\langle M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n), v^{n+s}_{M^{n+s}} \right\rangle}_{I_1} \underbrace{- \left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \tau \partial_t v^{n+s}_{M^{n+s}} + \frac{d}{ds} M^{n+s} \cdot \nabla v^{n+s}_{M^{n+s}} \right\rangle}_{I_2} \end{aligned}$$

Recalling that $\sigma^n = P_{\mathbb{S}}(M^n)$, the term I_1 can be rewritten as

$$\begin{aligned} I_1 &= \tau\epsilon^{-2} \left\langle M^n - P_{\mathbb{M}_N}(\sigma^n), v^{n+s}_{M^{n+s}} \right\rangle \\ &= \tau\epsilon^{-2} \left\langle M^n - \sigma^n, v^{n+s}_{M^{n+s}} \right\rangle + \left\langle \sigma^n - P_{\mathbb{M}_N}(\sigma^n), v^{n+s}_{M^{n+s}} \right\rangle \\ &= \underbrace{\tau\epsilon^{-2} \left\langle M^n - \sigma^n, v^{n+s}_{M^{n+s}} - v^{n+s}_{\sigma^n} \right\rangle}_{I_3} \end{aligned}$$

Here we had to control the fact that, due to the double projection, the norm of the acceleration $\|M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n)\|^2_{\mathbb{M}}$ is not equal to the squared distance $d^2_{\mathbb{S}}(M^n)$. We used the orthogonality property of the double projection for that purpose. On the one hand $\sigma^n - P_{\mathbb{M}_N}(\sigma^n)$ is orthogonal to $\mathbb{M}_N$ since $\mathbb{M}_N$ is a linear subspace of $\mathbb{M}$. On the other hand $M^n - \sigma^n$ is orthogonal to the tangent space to $\mathbb{S}$ at σ^n, see Lemma 3.1.

To handle the term I_2 we use the material derivatives defined in (3.3),

$$\begin{aligned} I_2 &= -\left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \tau\partial_t v^{n+s}_{M^{n+s}} + \frac{d}{ds} M^{n+s} \cdot \nabla v^{n+s}_{M^{n+s}} \right\rangle \\ I_2 &= -\left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \tau\partial_t v^{n+s}_{M^{n+s}} + \tau v^{n+s}_{M^{n+s}} \cdot \nabla v^{n+s}_{M^{n+s}} \right\rangle \\ &\quad \underbrace{-\left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \left(\frac{d}{ds} M^{n+s} - \tau v^{n+s}_{M^{n+s}}\right) \cdot \nabla v^{n+s}_{M^{n+s}} \right\rangle}_{I_4} \\ &= I_4 \underbrace{-\tau \left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, D_t v^{n+s}_{M^{n+s}} - D_t v^{n+s}_{\sigma^{n+s}} \right\rangle}_{I_5} + \underbrace{\tau \left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \nabla p^{n+s}_{\sigma^{n+s}} \right\rangle}_{I_6}. \end{aligned}$$

We used that $D_t v^{n+s}_{\sigma^{n+s}} = -\nabla p^{n+s}_{\sigma^{n+s}}$. We now rewrite I_6 using $\frac{d}{ds} M^{n+s} = \tau V^{n+1}$:

$$\begin{aligned} I_6 &= \underbrace{\tau \left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \nabla p^{n+s}_{\sigma^{n+s}} - \nabla p^{n+s}_{M^{n+s}} \right\rangle}_{I_7} + \tau \left\langle V^{n+s} - v^{n+s}_{M^{n+s}}, \nabla p^{n+s}_{M^{n+s}} \right\rangle \\ &= I_7 + \left\langle \frac{d}{ds} M^{n+s}, \nabla p^{n+s}_{M^{n+s}} \right\rangle + \tau \left\langle V^{n+s} - V^{n+1}, \nabla p^{n+s}_{M^{n+s}} \right\rangle - \tau \left\langle v^{n+s}_{M^{n+s}}, \nabla p^{n+s}_{M^{n+s}} \right\rangle \\ &= I_7 + \frac{d}{ds} \underbrace{\int_\Omega p^{n+s}_{M^{n+s}} dx}_{-J^{n+s}} - \tau \int_\Omega (\partial_t p^{n+s}_{M^{n+s}} + \left\langle v^{n+s}_{M^{n+s}}, \nabla p^{n+s}_{M^{n+s}} \right\rangle) dx \\ &\quad + \underbrace{(1-s)\tau^2\epsilon^{-2} \left\langle M^n - P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(M^n), \nabla p^{n+s}_{M^{n+s}} \right\rangle}_{I_8} \\ &= I_7 + I_8 - \frac{d}{ds} J^{n+s} - \underbrace{\tau \int_\Omega D_t p^{n+s}_{M^{n+s}} dx}_{I_9}, \end{aligned}$$

We need to estimate all the terms in the following formula.

$$d^{n+s} = I_3 + I_4 + I_5 + I_7 + I_8 + I_9 - \frac{d}{ds} J^{n+s} \tag{4.10}$$

4.3. **Gronwall estimates on** $[n\tau, (n+1)\tau]$**.** From now and for clarity we do not track the constants anymore, and C will be a constant depending only on T, Ω, $\mathrm{Lip}_{[0,T]}(v)$, $\mathrm{Lip}_{[0,T]}(p)$, $\mathrm{Lip}_{[0,T]}(\nabla p)$, $\mathrm{Lip}_{[0,T]}(D_t v)$ and $\mathrm{Lip}_{[0,T]}(D_t p)$. The value of the constant C can

change between estimates. Using (4.2) and Young's inequality we obtain for I_3:

$$\begin{aligned}
I_3 &= \tau\epsilon^{-2}\left\langle M^n-\sigma^n, v^{n+s}_{M^{n+s}}-v^{n+s}_{\sigma^n}\right\rangle\\
&\le \tau \operatorname{Lip}_{[0,\mathrm{T}]}(v)\frac{\|M^n-\sigma^n\|_{\mathbb{M}}\|M^{n+s}-\sigma^n\|_{\mathbb{M}}}{\epsilon^2}\\
&\le \tau C\frac{\|M^n-\sigma^n\|_{\mathbb{M}}\|M^{n+s}-M^n\|_{\mathbb{M}}}{\epsilon^2}+\tau C\frac{\|M^n-\sigma^n\|_{\mathbb{M}}\|M^n-\sigma^n\|_{\mathbb{M}}}{\epsilon^2}\\
&\le \tau C\left(\frac{\|M^n-\sigma^n\|^2_{\mathbb{M}}}{\epsilon^2}+\tau\epsilon^{-1}\frac{\|M^n-\sigma^n\|_{\mathbb{M}}}{\epsilon}\|V^{n+1}\|_{\mathbb{M}}\right)\\
&\le 2\tau CE^n+C\tau^2\epsilon^{-1}H^n\\
&\le 2\tau CE^n+C\tau^2\epsilon^{-1}(H_0+\kappa).
\end{aligned}\tag{4.11}$$

Since $\frac{d}{ds}M^{n+s}=\tau V^{n+1}$, and using the definition of V^{n+1} in (1.11), I_4 can be rewritten as

$$\begin{aligned}
\tau^{-1}I_4 &= -\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},\left(V^{n+1}-v^{n+s}_{M^{n+s}}\right)\cdot\nabla v^{n+s}_{M^{n+s}}\right\rangle\\
&= -\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},\left(V^{n+s}-v^{n+s}_{M^{n+s}}\right)\cdot\nabla v^{n+s}_{M^{n+s}}\right\rangle-\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},\left(V^{n+1}-V^{n+s}\right)\cdot\nabla v^{n+s}_{M^{n+s}}\right\rangle\\
&\le \operatorname{Lip}_{[0,\mathrm{T}]}(v)\left\|V^{n+s}-v^{n+s}_{M^{n+s}}\right\|^2_{\mathbb{M}}+\tau(1-s)\epsilon^{-2}\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},(M^n-P_{\mathbb{M}}(\sigma^n))\cdot\nabla v^{n+s}_{M^{n+s}}\right\rangle\\
&\le \operatorname{Lip}_{[0,\mathrm{T}]}(v)E^{n+s}+\tau(1-s)\epsilon^{-2}\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},(M^n-\sigma^n)\cdot\nabla v^{n+s}_{M^{n+s}}\right\rangle\\
&\le C\left(E^{n+s}+\tau\epsilon^{-1}\left\|V^{n+s}-v^{n+s}_{M^{n+s}}\right\|_{\mathbb{M}}\frac{\|M^n-\sigma^n\|_{\mathbb{M}}}{\epsilon}\right)\\
&\le C\left((1+\tau\epsilon^{-1})E^{n+s}+\tau\epsilon^{-1}E^n\right)\\
&\le CE^{n+s}+CE^n
\end{aligned}\tag{4.12}$$

Note that we used that $\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},(\sigma^n-P_{\mathbb{M}_N}(\sigma^n))\cdot\nabla v^{n+s}_{M^{n+s}}\right\rangle=0$, which holds true since $\sigma^n-P_{\mathbb{M}_N}(\sigma^n)$ is orthogonal to $\mathbb{M}_N$ and since $\nabla v^{n+s}_{M^{n+s}}$ is a symmetric matrix. We also used Young's inequality to get from the second to last line. The estimates of I_5 and I_7 are similar to those in the semi-discrete case:

$$\begin{aligned}
\tau^{-1}I_5 &\le \left|\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},D_tv^{n+s}_{M^{n+s}}-D_tv^{n+s}_{\sigma^{n+s}}\right\rangle\right|\\
&\le \operatorname{Lip}_{[0,\mathrm{T}]}(D_tv)\left\|V^{n+s}-v^{n+s}_{M^{n+s}}\right\|_{\mathbb{M}}\left\|M^{n+s}-\sigma^{n+s}\right\|_{\mathbb{M}}\\
&\le C\left\|V^{n+s}-v^{n+s}_{M^{n+s}}\right\|_{\mathbb{M}}\left\|M^{n+s}-\sigma^{n+s}\right\|_{\mathbb{M}}\\
&\le \epsilon CE^{n+s}
\end{aligned}\tag{4.13}$$

The quantity I_7 is of the same kind.

$$\begin{aligned}
\tau^{-1}I_7 &\le \left|\left\langle V^{n+s}-v^{n+s}_{M^{n+s}},\nabla p^{n+s}_{\sigma^{n+s}}-\nabla p^{n+s}_{M^{n+s}}\right\rangle\right|\\
&\le \epsilon CE^{n+s}
\end{aligned}\tag{4.14}$$

For the estimation of I_8 we use $\left\langle\sigma^n-P_{\mathbb{M}_N}(\sigma^n),\nabla p^{n+s}_{M^{n+s}}\right\rangle=0$ to get

$$\begin{aligned}
I_8 &= (1-s)\tau^2\epsilon^{-2}\left\langle M^n-P_{\mathbb{M}_N}\circ P_{\mathbb{S}}(M^n),\nabla p^{n+s}_{M^{n+s}}\right\rangle\\
&\le \tau^2\epsilon^{-2}||\nabla p((n+s)\tau)||_{L^\infty(\Omega)}||M^n-\sigma^n||_{\mathbb{M}}\\
&\le \tau^2\epsilon^{-1}\operatorname{Lip}_{[0,\mathrm{T}]}(v)\frac{||M^n-\sigma^n||_{\mathbb{M}}}{\epsilon}\\
&\le \tau CE^n+\tau^2\epsilon^{-1}C.
\end{aligned}\tag{4.15}$$

To estimate J and I_9 recall that we have assumed that $\int_\Omega p(t,x)dx=0$, which implies in particular that $\int_\Omega D_tp(t,\sigma^n(t,x))dx=0$. Therefore,

$$\begin{aligned}
\tau^{-1}I_9 &\le \operatorname{Lip}_{[0,\mathrm{T}]}(D_tp)||M^{n+s}-\sigma^{n+s}||_{L^1(\Omega)}\\
&\le \frac{1}{2}E^{n+s}+C\epsilon^2.
\end{aligned}\tag{4.16}$$

Similarly

$$|J^{n+s}| = |J((n+s)\tau)| \leq \left| \int_\Omega p^{n+s}_{M^{n+s}} - p^{n+s}_{\sigma^{n+s}} dx \right| \leq \mathrm{Lip}_{[0,\mathrm{T}]}(p) ||M^{n+s} - \sigma^{n+s}||_{L^1(\Omega)}$$
$$\leq \frac{1}{2} E^{n+s} + C\epsilon^2. \tag{4.17}$$

Note also that $J^0 \leq \mathrm{Lip}_{[0,\mathrm{T}]}(p) h_N \leq C h_N$ still holds see (3.14).

4.4. **Gronwall argument on** $[n\tau, (n+1)\tau]$**.** Collecting estimates (4.11), (4.12), (4.13), (4.14), (4.15), (4.16) and (4.17) and integrating equation (4.10) from 0 to s we obtain

$$\begin{aligned} J^{n+s} + \int_0^s d^{n+\theta} d\theta &\leq J^n + 2\tau C E^n + C\tau^2 \epsilon^{-1}(H_0 + \kappa) \\ &+ \tau C \int_0^s E^{n+\theta} d\theta + \tau C E^n \\ &+ \tau \epsilon C \int_0^s E^{n+\theta} d\theta + \tau \epsilon C \int_0^s E^{n+\theta} d\theta \\ &+ \tau C E^n + \tau^2 \epsilon^{-1} C \\ &+ \frac{\tau}{2} \int_0^s E^{n+\theta} d\theta + \tau \epsilon^2 C \\ &\leq J^n + C\tau E^n + C\tau\epsilon^2 + C(H_0 + \kappa)\tau^2\epsilon^{-1} \\ &+ \tau C \int_0^s \left(E^{n+\theta} + J^{n+\theta} \right) d\theta. \end{aligned} \tag{4.18}$$

Remark that we used (3.15) to add $J^{n+\theta}$ at the last line. Remark also that we only kept the first order terms using $\epsilon \leq C$. Plugging (4.18) into (4.9) we obtain

$$E^{n+s} + J^{n+s} \leq \alpha(s) + \beta \int_0^\theta (E^{n+\theta} + J^{n+\theta}) \mathrm{d}s) \tag{4.19}$$
$$\text{where } \alpha(s) = E^n + J^n + H^{n+s} - H^n + C\tau E^n + C\tau\epsilon^2 + C(H_0 + \kappa)\tau^2\epsilon^{-1}, \ \beta = \tau C$$

so that by Gronwall lemma,

$$\begin{aligned} E^{n+1} + J^{n+1} &\leq \alpha(1) + \int_0^1 \alpha(s)\beta \exp((1-s)\beta) \mathrm{d}s \\ &\leq \left[E^n + J^n + C\tau E^n + C\tau\epsilon^2 + C(H_0 + \kappa)\tau^2\epsilon^{-1} \right] e^{C\tau} \\ &\quad + H^{n+1} - H^n + \underbrace{\int_0^1 \left(H^{n+s} - H^n \right) C\tau \exp((1-s)C\tau) \mathrm{d}s}_{R} \end{aligned}$$

Using Lemma 4.2 and in particular the upper bound (4.6) we find

$$R \leq \frac{\tau^2}{\epsilon^2} H^{n+1} \int_0^s C\tau e^{C\tau(1-\theta)} \leq C \frac{\tau^2}{\epsilon^2} (H_0 + \kappa) \left[e^{C\tau} - 1 \right]$$

so that

$$\begin{aligned} E^{n+1} + J^{n+1} &\leq \left[(1 + C\tau) \left(E^n + J^n \right) \right. \\ &\quad + C\tau\epsilon^2 + C(H_0 + \kappa)\tau^2\epsilon^{-1} \\ &\quad \left. + H^{n+s} - H^n + \tau^2\epsilon^{-2}(H_0 + \kappa) \left[e^{C\tau} - 1 \right] \right] e^{C\tau}. \end{aligned} \tag{4.20}$$

4.5. **Discrete Gronwall step.** From (4.20) and the discrete Gronwall inequality we deduce that for any $n \in \mathbb{N} \cap [0, T/\tau]$,

$$\begin{aligned} E^n + J^n &\leq \Big[E^0 + J^0 + CT\epsilon^2 + CT(H_0+\kappa)\tau\epsilon^{-1} + H^n - H^0 \\ &\quad + \tau^2\epsilon^{-2}(H_0+\kappa)\frac{T}{\tau}\left[e^{C\tau} - 1\right]\Big](1+C\tau)^n e^{CT} \\ &\leq C\left[E^0 + J^0 + \epsilon^2 + (H_0+\kappa)\tau\epsilon^{-1} + \kappa + (H_0+\kappa)\tau^2\epsilon^{-2}e^{CT}\right]e^{CT} \\ &\leq C\left[E^0 + J^0 + \epsilon^2 + (H_0+\kappa)\tau\epsilon^{-1} + \kappa\right]e^{CT}. \end{aligned}$$

We used the mean value theorem to obtain the second to last line. Using (4.17) one last time and $H_0 \leq C$ leads us to

$$\begin{aligned} E^n &\leq C\left[E^0 + J^0 + \epsilon^2 + \tau\epsilon^{-1} + \kappa\right] + C\epsilon^2 \\ &\leq C\left[\epsilon^2 + h_N + \frac{h_N^2}{\epsilon^2} + \kappa + \frac{\tau}{\epsilon}\right]. \end{aligned}$$

where the second line incorporates the initial error. It leads

$$\max_{n\in\mathbb{N}\cap[0,T/\tau]} \|V^n - v(t^n, M^n)\|_{\mathbb{M}}^2 \leq C\left[\epsilon^2 + h_N + \frac{h_N^2}{\epsilon^2} + \kappa + \frac{\tau}{\epsilon}\right].$$

A third Gronwall estimate, similar to the one done to obtain (3.19), concludes the proof:

$$\max_{n\in\mathbb{N}\cap[0,T/\tau]} \|V^n - v(t^n, \phi(t^n,\cdot))\|_{\mathbb{M}}^2 \leq C\left[\epsilon^2 + h_N + \frac{h_N^2}{\epsilon^2} + \kappa + \frac{\tau}{\epsilon}\right]..$$

Remark 4.5. A close look at the constant leads to a similar result as the one given in Remark 3.6: namely the convergence of the numerical scheme towards less regular solutions of the Euler's equations.

Remark 4.6. The method of the proof is robust and could easily be adapted to other numerical scheme. Any improvement to the estimate given in Lemma 4.2 (conservation of the Hamiltonian) will lead to improved convergence estimates for the numerical scheme.

5. Numerical implementation and experiments

5.1. **Numerical implementation.** We discuss here the implementation of the numerical scheme (1.11) and in particular the computation of the double projection $P_{\mathbb{M}_N} \circ P_{\mathbb{S}}(m)$ for a piecewise constant function $m \in \mathbb{M}_N$. Using Brenier's polar factorisation theorem, the projection of m on $\mathbb{S}$ amounts to the resolution of an optimal transport problem between Leb and the finitely supported measure $m_{\#}\,\mathrm{Leb}$. Such optimal transport problems can be solved numerically using the notion of Laguerre diagram from computational geometry.

Definition 5.1 (Laguerre diagram)**.** Let $M = (M_1, \dots, M_N) \in (\mathbb{R}^d)^N$ and let $\psi_1, \dots, \psi_N \in \mathbb{R}$. The Laguerre diagram is a decomposition of $\mathbb{R}^d$ into convex polyhedra defined by

$$\mathrm{Lag}_i(M,\psi) = \left\{x \in \mathbb{R}^d \mid \forall j \in \{1,\dots,N\},\ \|x - M_i\|^2 + \psi_i \leq \|x - M_j\|^2 + \psi_j\right\}.$$

In the following proposition, we denote $\Pi_{\mathbb{S}}(m) = \{s \in \mathbb{S} \mid \|m - s\| = \mathrm{d}_{\mathbb{S}}(m)\}$, and for a bounded subset $A \subseteq \mathbb{R}^d$ with positive measure we set $\mathrm{bary}(A) := \frac{1}{\mathrm{Leb}(A)}\int_A x\mathrm{d}x$.

Proposition 5.2. *Let $m \in \mathbb{M}_N \setminus \mathbb{D}_N$ and define $M_i = m(\omega_i) \in \mathbb{R}^d$. Assume that Ω is a bounded and connected domain of $\mathbb{R}^d$ with Lipschitz boundary. Then, there exist scalars $(\psi_i)_{1\leq i\leq N}$, which are unique up to an additive constant, such that*

$$\forall i \in \{1,\dots,N\}, \qquad \mathrm{Leb}(\mathrm{Lag}_i(M,\psi)) = \frac{1}{N}\mathrm{Leb}(\Omega). \tag{5.1}$$

We denote $L_i := \mathrm{Lag}_i(M,\psi)$. Then, a function $s \in \mathbb{S}$ is a projection of m on $\mathbb{S}$ if and only if it maps the subdomain ω_i to the Laguerre cell L_i up to a negligible set, that is:

$$\Pi_{\mathbb{S}}(m) = \{s \in \mathbb{S} \mid \forall i \in \{1,\dots,N\},\ \mathrm{Leb}(s(\omega_i)\Delta L_i) = 0\}, \tag{5.2}$$

where $A\Delta B$ denotes the symmetric difference between sets A and B. Moreover, the squared distance $\mathrm{d}^2_{\mathbb{S}}$ is differentiable at m and, setting $B_i = \frac{1}{\mathrm{Leb}(L_i)}\int_{L_i} x\mathrm{d}x$, one has

$$\begin{aligned} \mathrm{d}^2_{\mathbb{S}}(m) &= \sum_{1\leq i\leq N}\int_{L_i}\|x-M_i\|^2\,\mathrm{d}x, \\ \nabla\mathrm{d}^2_{\mathbb{S}}(m) &= 2(m - P_{\mathbb{M}_N}\circ P_{\mathbb{S}}(m)) \text{ with } P_{\mathbb{M}_N}\circ P_{\mathbb{S}}(m) = \sum_{1\leq i\leq N} B_i\mathbf{1}_{\omega_i}. \end{aligned} \tag{5.3}$$

Proof. The existence of a vector $(\psi_i)_{1\leq i\leq N}$ satisfying Equation (5.1) follows from optimal transport theory (see Section 5 in [3] for a short proof), and its uniqueness follows from the connectedness of the domain Ω. In addition, the map $T:\Omega\to\{M_1,\dots,M_N\}$ defined by $T(L_i)=M_i$ (up to a negligible set) is the gradient of a convex function and therefore a quadratic optimal transport between Leb and the measure $\frac{1}{N}\mathrm{Leb}(\Omega)\sum_i\delta_{M_i}$. By Brenier's polar factorization theorem, summarized in Lemma 3.1,

$$\begin{aligned} s\in\Pi_{\mathbb{S}}(m) \iff m = T\circ s \text{ a.e.} &\iff \forall i\in\{1,\dots,N\},\ \mathrm{Leb}(\omega_i\Delta(T\circ s)^{-1}(\{M_i\})) = 0 \\ &\iff \forall i\in\{1,\dots,N\},\ \mathrm{Leb}(s(\omega_i)\Delta L_i) = 0, \end{aligned}$$

where the last equality holds because s is measure preserving. To prove the statement on the differentiability of $\mathrm{d}^2_{\mathbb{S}}$, we first note that the function $\mathrm{d}^2_{\mathbb{S}}$ is 1-semi-concave, since

$$D(m) := \|m\|^2 - \mathrm{d}^2_{\mathbb{S}}(m) = \|m\|^2 - \min_{s\in\mathbb{S}}\|m-s\|^2 = \max_{s\in\mathbb{S}} 2\langle m|s\rangle - \|s\|^2$$

is convex. The subdifferential of D at m is given by $\partial D(m) = \{P_{\mathbb{M}_N}(s)\mid s\in\Pi_{\mathbb{S}}(m)\}$, so that D (and hence $\mathrm{d}^2_{\mathbb{S}}$) is differentiable at m if and only if $P_{\mathbb{M}_N}(\Pi_{\mathbb{S}}(m))$ is a singleton. Now, note from Lemma 3.2 that for $s\in\Pi_{\mathbb{S}}(m)$

$$P_{\mathbb{M}_N}(s) = \sum_{1\leq i\leq N}\mathrm{bary}(s(\omega_i))\mathbf{1}_{\omega_i} = \sum_{1\leq i\leq N}\mathrm{bary}(L_i)\mathbf{1}_{\omega_i}.$$

This shows that $P_{\mathbb{M}_N}(\Pi_{\mathbb{S}}(m))$ is a singleton, and therefore establishes the differentiability of $\mathrm{d}^2_{\mathbb{S}}$ at m, together with the desired formula for the gradient. □

The main difficulty to implement the numerical scheme (1.11) is the resolution of the discrete optimal transport problem (5.1), a non-linear system of equations which must be solved at every iteration. We resort to the damped Newton's algorithm presented in [17] (see also [22]) and more precisely on its implementation in the PyMongeAmpere library[1].

5.1.1. *Construction of the fixed tessellation of the domain.* The fixed tessellation $(\omega_i)_{1\leq i\leq N}$ of the domain Ω is a collection of Laguerre cells that are computed through a simple fixed-point algorithm similar to the one presented in [13]. We start from a random sampling $(C_i^0)_{1\leq i\leq N}$ of Ω. At a given step $k\geq 0$, we compute $(\psi_i)_{1\leq i\leq N}\in\mathbb{R}^N$ such that

$$\forall i\in\{1,\dots,N\},\ \mathrm{Leb}(\mathrm{Lag}_i(C,\psi)) = \frac{1}{N}\mathrm{Leb}(\Omega),$$

and we then update the new position of the centers (C_i^{k+1}) by setting $C_i^{k+1} := \mathrm{bary}(\mathrm{Lag}_i(C^k,\psi))$. After a few iterations, a fixed-point is reached and we set $\omega_i := \mathrm{Lag}_i(C^k,\psi)$.

5.1.2. *Iterations.* To implement the symplectic Euler scheme for (1.6), we start with $M_i^0 := \mathrm{bary}(\omega_i)$ and $V_i^0 := v_0(M_i^0)$. Then, at every iteration $k\geq 0$, we use Algorithm 1 in [17] to compute a solution $(\psi_i^k)_{1\leq i\leq N}\in\mathbb{R}^N$ to Equation (5.1) with $M = M^k$, i.e. such that

$$\forall i\in\{1,\dots,N\},\ \mathrm{Leb}(\mathrm{Lag}_i(M^k,\psi^k)) = \frac{1}{N}\mathrm{Leb}(\Omega).$$

[1] https://github.com/mrgt/PyMongeAmpere

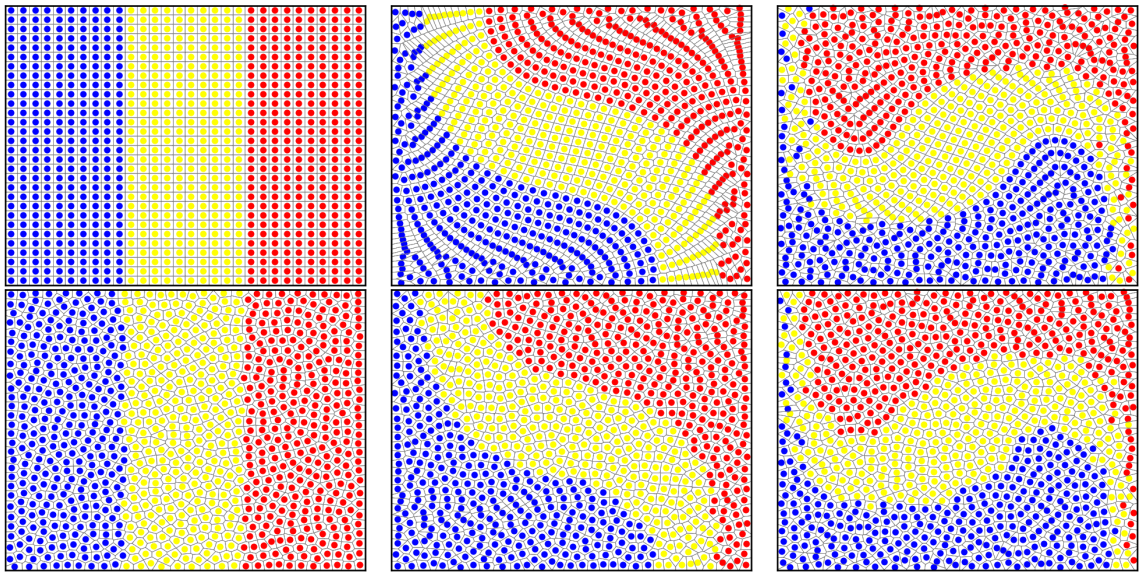


FIGURE 1. (Top row) Beltrami flow in the square, with $N = 900$ particles, $\tau = 1/50$ and $\varepsilon = .1$. The particles are colored depending on their initial position in the square. From left to right, we display the Laguerre cells and their barycenters at timesteps $k = 0, 24$ and 49. The partition $(\omega_i)_{1\leq i\leq N}$ is induced by a regular grid. (Bottom row) Same experiment, but where the partition $(\omega_i)_{1\leq i\leq N}$ is optimized using the algorithm described in §5.1.1.

Finally, we update the positions $(M_i^{k+1})_{1\leq i\leq N}$ and the speeds $(V_i^{k+1})_{1\leq i\leq N}$ by setting

$$\begin{cases} V_i^{k+1} = V_i^k + \dfrac{\tau}{\varepsilon^2}(\mathrm{bary}(\mathrm{Lag}_i(M^k, \psi^k)) - M_i^k) \\ M_i^{k+1} = M_i^k + \tau V_i^{k+1} \end{cases} \tag{5.4}$$

5.2. **Beltrami flow in the square.** Our first test case is constructed from a stationary solution to Euler's equation in 2D. On the unit square $\Omega = [-\frac{1}{2}, \frac{1}{2}]^2$, we consider the Beltrami flow constructed from the time-independent pressure and speed:

$$\begin{cases} p_0(x_1, x_2) = \dfrac{1}{2}(\sin(\pi x_1)^2 + \sin(\pi x_2)^2) \\ v_0(x_1, x_2) = (-\cos(\pi x_1)\sin(\pi x_2), \sin(\pi x_1)\cos(\pi x_2)) \end{cases}$$

In Figure 1, we display the computed numerical solution using a low number of particles ($N = 900$) in order to show the shape of the Laguerre cells associated to the solution.

5.3. **Kelvin-Helmoltz instability.** For this second test case, the domain is the rectangle $\Omega = [0, 2]\times[-.5, .5]$ periodized in the first coordinate by making the identification $(4, x_2) \sim (0, x_2)$ for $x_2 \in [-.5, .5]$. The initial speed v_0 is discontinuous at $x_2 = 0$: the upper part of the domain has zero speed, and the bottom part has unit speed:

$$v_0(x_1, x_2) = \begin{cases} 0.5 \text{ if } x_2 \geq 0 \\ 1 \text{ if } x_2 < 0 \end{cases}$$

This speed profile corresponds to a stationnary but unstable solution to Euler's equation. If the subdomains $(\omega_i)_{1\leq i\leq N}$ are computed following §5.1.1, the perfect symmetry under horizontal translations is lost, and in Figure 2 we observe the formation of vortices whose radius increases with time. This experiment involves $N = 200\,000$ particles, with parameters $\tau = 0.002$ and $\varepsilon = 0.005$, and $2\,000$ timesteps. As displayed in Figure 2, the hamiltonian of the system is very well preserved despite the roughness of the solution. This behaviour shows that the estimate of Lemma 4.2 might be overly pessimistic, and requires further investigation.

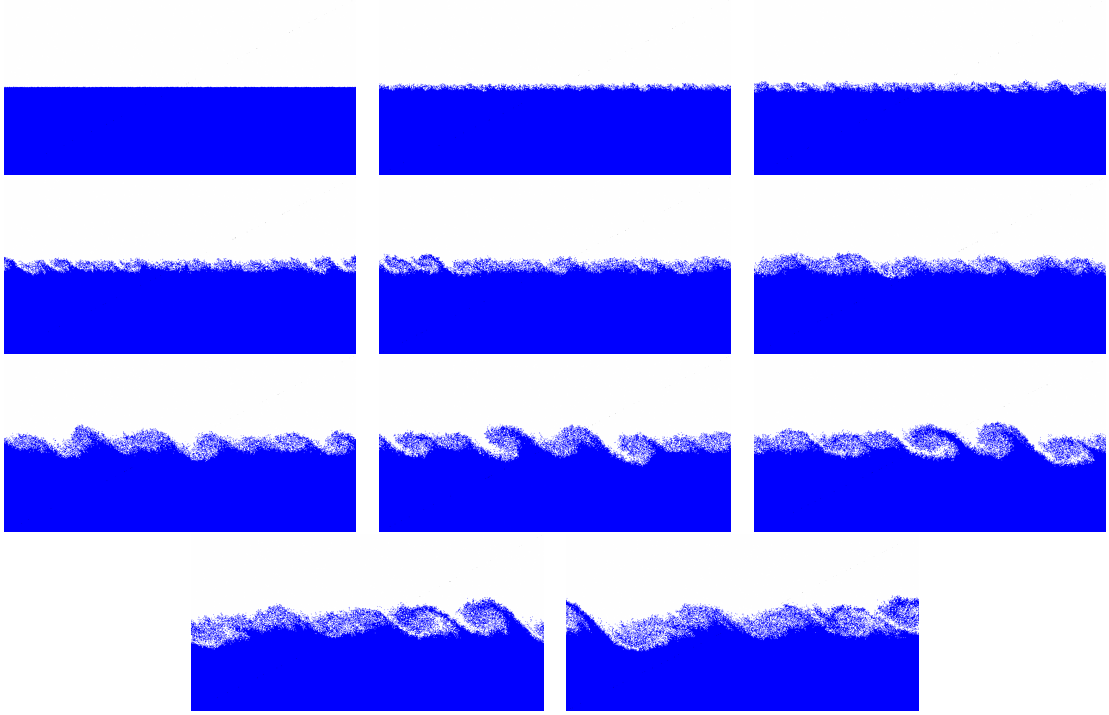


FIGURE 2. Numerical illustration of the Kelvin-Helmotz instability on a rectangle with periodic conditions (in the horizontal coordinate) involving a discontinuous initial speed. The parameters are given in §5.4.

5.4. **Rayleigh-Taylor instability.** For this last test case, the particles are assigned a density ρ_i, and are subject to the force of the gravity $\rho_i G$, where $G = (0, -10)$. This changes the numerical scheme to

$$\begin{cases} \rho_i V_i^{k+1} = \rho_i V_i^k + \tau \left(\dfrac{1}{\varepsilon^2} (\mathrm{bary}(\mathrm{Lag}_i(M^k, \psi^k) - M_i^k) + \rho_i G \right) \\ M_i^{k+1} = M_i^k + \tau V_i^{k+1} \end{cases} \tag{5.5}$$

The computational domain is the rectangle $\Omega = [-1, 1] \times [-3, 3]$, and the initial distribution of particles is given by $C_i = \mathrm{bary}(\omega_i)$, where the partition $(\omega_i)_{1 \le i \le N}$ is constructed according to §5.1.1. The fluid is composed of two phases, the heavy phase being on top of the light phase:

$$\rho_i = \begin{cases} 3 & \text{if } C_{i2} > \eta \cos(\pi C_{i1}) \\ 1 & \text{if } C_{i2} \le \eta \cos(\pi C_{i1}) \end{cases},$$

where $\eta = 0.2$ in the experiment and where we denoted C_{i1} and C_{i2} the first and second coordinates of the point C_i. Finally, we have set $N = 50\,000$, $\varepsilon = 0.002$ and $\tau = 0.001$ and we have run 2000 timesteps. The computation takes less than six hours on a single core of a regular laptop. Note that it does not seem straighforward to adapt the techniques used in the proofs of convergence presented here to this setting, where the force depends on the density of the particle. Our purpose with this test case is merely to show that the numerical scheme behaves reasonably well in more complex situations.

Software. The software developed for generating the results presented in this article is publicly available at https://github.com/mrgt/EulerLagrangianOT

ACKNOWLEDGEMENTS

We would like to thank Yann Brenier who pointed out to us the reference [8] on which this article elaborates, and for several interesting discussions at various stages of this work. We also thank Pierre Bousquet who indicated the reference [16] to us.

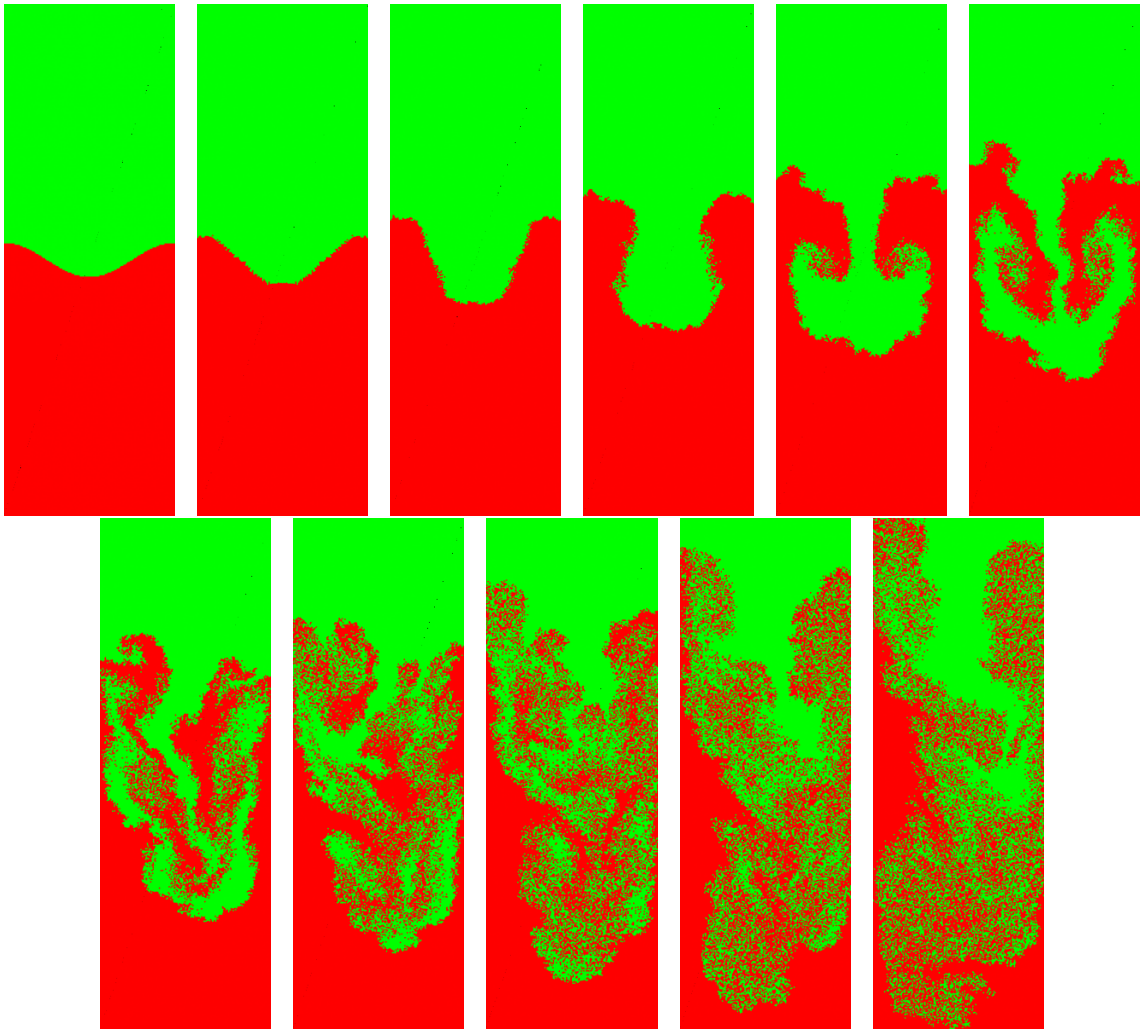


FIGURE 3. Numerical illustration of the Rayleigh-Taylor instability occuring when a heavy fluid (in green) is placed over a lighter fluid (in red) at timesteps $n = 0, 200, 400, \ldots, 2000$. The parameters are given in §5.4.

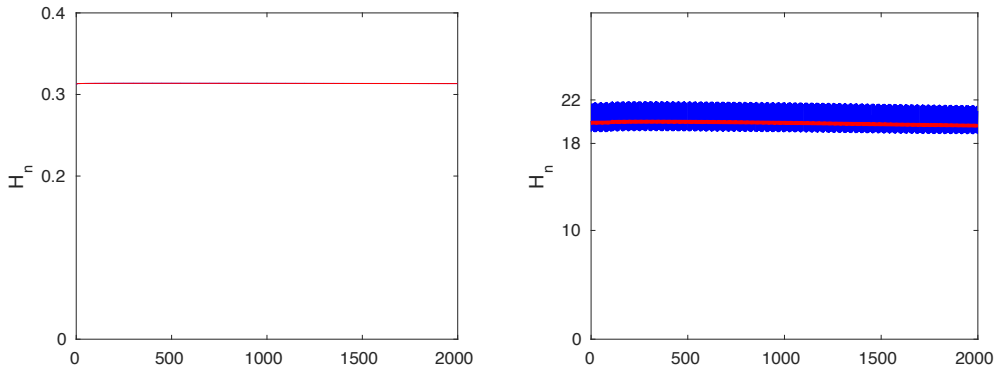


FIGURE 4. (Left) Value of the Hamiltonian during iterations of the algorithm, for the Kelvin-Helmoltz instability presented in §5.3 and using the symplectic Euler integrator. (Right) Same figure but for the Rayleigh-Taylor instability presented in §5.4, using the symplectic Euler integrator (in blue) and using the velocity Verlet integrator (in red).

References

[1] L. Ambrosio. Transport equation and cauchy problem for BV vector fields. *Inventiones mathematicae*, 158(2):227–260, 2004.

[2] V. Arnold. Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Annales de l'institut Fourier*, 16(1):319–361, 1966.

[3] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.

[4] F. Bouchut. Renormalized solutions to the vlasov equation with coefficients of bounded variation. *Archive for rational mechanics and analysis*, 157(1):75–90, 2001.

[5] Y. Brenier. A combinatorial algorithm for the Euler equations of incompressible flows. In *Proceedings of the Eighth International Conference on Computing Methods in Applied Sciences and Engineering (Versailles, 1987)*, 1989.

[6] Y. Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 1989.

[7] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[8] Y. Brenier. Derivation of the Euler equations from a caricature of Coulomb interaction. *Communications in Mathematical Physics*, 212(1):93–104, 2000.

[9] Y. Brenier. Generalized solutions and hydrostatic approximation of the Euler equations. *Physica D. Nonlinear Phenomena*, 2008.

[10] Y. Brenier and G. Loeper. A geometric approximation to the euler equations: the vlasov–monge–ampere system. *Geometric And Functional Analysis*, 14(6):1182–1218, 2004.

[11] Y. Brudnyi and P. Shvartsman. Whitney's extension problem for multivariate $c^{\{1,\omega\}}$-functions. *Transactions of the American Mathematical Society*, 353(6):2487–2512, 2001.

[12] M. Cullen, W. Gangbo, and G. Pisante. The semigeostrophic equations discretized in reference and dual variables. *Archive for rational mechanics and analysis*, 185(2):341–363, 2007.

[13] F. de Goes, K. Breeden, V. Ostromoukhov, and M. Desbrun. Blue noise through optimal transport. *ACM Transactions on Graphics (TOG)*, 31(6):171, 2012.

[14] F. de Goes, C. Wallez, J. Huang, D. Pavlov, and M. Desbrun. Power particles: an incompressible fluid solver based on power diagrams. *ACM Transactions on Graphics (TOG)*, 34(4):50, 2015.

[15] C. Fefferman. Whitney's extension problem for c^m. *Annals of mathematics*, pages 313–359, 2006.

[16] C. Fefferman et al. Extension of $c^{\{m,\Omega\}}$-smooth functions by linear operators. *Revista Matematica Iberoamericana*, 25(1):1–48, 2009.

[17] J. Kitagawa, Q. Mérigot, and B. Thibert. Convergence of a newton algorithm for semi-discrete optimal transport. *arXiv preprint arXiv:1603.05579*, 2016.

[18] B. Lévy. A numerical algorithm for L^2 semi-discrete optimal transport in 3d. *ESAIM M2AN*, 49(6), 2015.

[19] P.-L. Lions. Sur les équations différentielles ordinaires et les équations de transport. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 326(7):833–838, 1998.

[20] Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.

[21] Q. Mérigot and J.-M. Mirebeau. Minimal geodesics along volume preserving maps, through semi-discrete optimal transport. *arXiv preprint arXiv:1505.03306*, 2015.

[22] J.-M. Mirebeau. Discretization of the 3d monge-ampere operator, between wide stencils and power diagrams. *arXiv preprint arXiv:1503.00947*, 2015.

[23] A. I. Shnirelman. Generalized fluid flows, their approximation and applications. *Geometric and Functional Analysis*, 1994.

[24] C. Villani. *Optimal transport: old and new*. Springer Verlag, 2009.

Département de mathématiques, Université de Liège, Allée de la découverte 12, B-4000 Liège, Belgique.

E-mail address: thomas.gallouet@ulg.ac.be

Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France.

E-mail address: quentin.merigot@math.u-psud.fr

CONVERGENCE OF A LAGRANGIAN DISCRETIZATION FOR BAROTROPIC FLUIDS AND POUROUS MEDIA FLOW

THOMAS O. GALLOUËT, QUENTIN MÉRIGOT, AND ANDREA NATALE

Abstract. When expressed in Lagrangian variables, the equations of motion for compressible (barotropic) fluids have the structure of a classical Hamiltonian system in which the potential energy is given by the internal energy of the fluid. The dissipative counterpart of such a system coincides with the porous medium equation, which can be cast in the form of a gradient flow for the same internal energy. Motivated by these related variational structures, we propose a particle method for both problems in which the internal energy is replaced by its Moreau-Yosida regularization in the L^2 sense, which can be efficiently computed as a semi-discrete optimal transport problem. Using a modulated energy argument which exploits the convexity of the problem in Eulerian variables, we prove quantitative convergence estimates towards smooth solutions. We verify such estimates by means of several numerical tests.

1. Introduction

The Euler equations describing the evolution of a barotropic fluid in a compact domain $M \subset \mathbb{R}^d$ with Lipschitz boundary and on a time interval $[0, T]$ are given by the following system of equations:

$$\begin{cases} \partial_t(\rho u) + \nabla \cdot (\rho u \otimes u) + \nabla P(\rho) = 0\,, \\ \partial_t \rho + \operatorname{div}(\rho u) = 0\,, \end{cases} \tag{1.1}$$

where $\rho(t,x) \geq 0$ is the fluid density, $u(t,x) \in \mathbb{R}^d$ is the Eulerian velocity and the function $P : [0,\infty) \to \mathbb{R}$ defines the pressure as a function of the density. The first equation in (1.1) is generally referred to as the momentum equation, whereas the second is the continuity equation and describes local mass conservation in the fluid. The system is supplemented by the initial and boundary conditions:

$$\rho(0,\cdot) = \rho_0\,, \quad u(0,\cdot) = u_0\,, \quad u \cdot n_{\partial M} = 0 \text{ on } [0,T] \times \partial M\,,$$

where $n_{\partial M}$ is the outward normal to the boundary ∂M. Smooth solutions conserve the total energy

$$\int_M \frac{1}{2}|u|^2 \rho \mathrm{d}x + \int_M U(\rho)\,\mathrm{d}x\,, \tag{1.2}$$

where $U : [0,\infty) \to \mathbb{R}$ is a smooth strictly convex function, superlinear at infinity, defining the internal energy of the fluid. This is related to the pressure by the thermodynamic relations

$$P(r) = rU'(r) - U(r)\,, \quad P'(r) = rU''(r). \tag{1.3}$$

Date: September 26, 2023.

Different choices of the internal energy U lead to different models. The two most classical examples are:

(1) polytropic fluids, which correspond to $U(r) = r^m/(m-1)$ with $m > 1$, and $P(r) = r^m$ (these include isentropic fluids, and the Saint-Venant system modelling gravity driven shallow water flows for $m = 2$);
(2) isothermal fluids, which correspond to $U(r) = r\log(r) - r$ and $P(r) = r$.

Adding a friction term $-\zeta\rho u$ on the right-hand side of the momentum equation, i.e. the first equation in the system (1.1), and considering the high friction limit $\zeta \to \infty$, one formally obtains $u = -\nabla U'(\rho)$, which substituted into the continuity equation yields

$$\partial_t \rho - \Delta P(\rho) = 0\,. \tag{1.4}$$

In particular, the choice $U(r) = r^m/(m-1)$ with $m > 1$ and $P(r) = r^m$, which is associated with polytropic fluids, yields the porous medium equation. Similarly, the choice $U(r) = r\log r - r$ and $P(r) = r$ corresponding to isothermal fluids, yields the heat equation.

1.1. **Lagrangian formulation.** For both the compressible Euler system (1.1) and its high friction limit (1.4), the density evolves according to the continuity equation with respect to a time-dependent vector field u. Let $S_0 \subseteq M$ be the support of the initial density ρ_0 and $X : [0,T]\times S_0 \to M$ be the flow associated with u, i.e. the time-dependent map satisfying the flow equation

$$\dot{X}_t = u(t, X_t) \tag{1.5}$$

with initial condition $X_0 = \mathrm{Id}|_{S_0}$, where Id is the identity map on $\mathbb{R}^d$. If ρ_0 and u are sufficiently regular, then the flow equation (1.5) and the continuity equation have both a unique strong solution, and the density is the pushforward of ρ_0 by the flow, i.e. $\rho(t,\cdot) = X_{t\#}\rho_0$, where the pushforward is defined by the condition

$$(X_{t\#}\rho_0)[B] = \rho_0[X_t^{-1}(B)] \quad \text{for any } B \subset M. \tag{1.6}$$

In general, equation (1.6) defines $X_{t\#}\rho_0$ only as a measure on M. However, if X_t is a smooth invertible map, $X_{t\#}\rho_0$ is absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$, and we identify it with its smooth density.

Using equation (1.5) and (1.6), the total energy of the fluid (1.2) can then be written in terms of X only as follows:

$$\int_M \frac{1}{2}|\dot{X}_t|^2 \rho_0 \mathrm{d}x + \int_M U(X_{t\#}\rho_0)\,\mathrm{d}x\,. \tag{1.7}$$

Let $\mathbb{X} \coloneqq L^2_{\rho_0}(S_0;\mathbb{R}^d)$. In the smooth setting, we can interpret the energy (1.7) as a functional on curves of smooth invertible maps in $C^\infty(S_0;M)$, viewed as a manifold in $\mathbb{X}$ with the induced metric. The associated Euler-Lagrange equations coincide with Newton's second law:

$$\ddot{X}_t = -\nabla_{\mathbb{X}}\mathcal{F}(X_t)\,, \quad \mathcal{F}(\sigma) \coloneqq \int_M U(\sigma_{\#}\rho_0)\,\mathrm{d}x\,, \tag{1.8}$$

where we identify the gradient $\nabla_{\mathbb{X}}\mathcal{F}(X_t)$ with an element of $\mathbb{X}$ (see Remark 2.3 for a formal computation of $\nabla_{\mathbb{X}}\mathcal{F}(X_t)$). Equation (1.8) is the Lagrangian equivalent to the

momentum equation in (1.1), and in particular from its solutions one can retrive the solutions to the Euler system (1.1) using the flow equation (1.5) and the definition of pushforward (1.6).

In the case of the high friction limit (1.4), the flow evolves according to a gradient flow dynamics, which correspond to the equation:

$$\dot{X}_t = -\nabla_{\mathbb{X}} \mathcal{F}(X_t)\,. \tag{1.9}$$

Here, equation (1.9) is equivalent to the condition $u = -\nabla U'(\rho)$, and from its solutions one can retrive the solutions to (1.4) by pushforward of the initial density as in (1.6).

The point of view described above for the compressible Euler system is one of the possible generalizations of the approach developed by Arnold for the incompressible Euler equations (see, e.g., Proposition 2.7 in [18]), which he intrepreted as the geodesic equation on the group of volume-preserving diffeomorphisms with the L^2 metric [1]. On the other hand, the gradient flow structure in (1.9) is the Lagrangian counterpart of the Wasserstein gradient flow interpretation of equation (1.4), developed in the celebrated works of Otto [27] and Jordan, Kinderlehrer, and Otto [17].

In this paper, we will construct discrete versions of the systems (1.8) and (1.9) in which the flow is approximated by a curve of (non-smooth and non-injective) maps belonging to a finite-dimensional subpace of $\mathbb{X}$. As a consequence of this extrinsic point of view, we will regard the internal energy $\mathcal{F}$ in equation (1.8) as a real-valued functional on the whole space $\mathbb{X}$, which we set to $+\infty$ when $\sigma_{\#}\rho_0$ is not absolutely continuous with respect to the Lebesgue measure $\mathrm{d}x$ restricted to M.

1.2. **Space discretization.** We now turn to the design of the Lagrangian scheme, i.e. an evolutive system for a finite number of particles, to approximate both Euler and gradient flows. In order to define the evolution of the particles we introduce a discrete equivalent of the Lagrangian variational structure highlighted in the previous section. This also allows us to preserve at the discrete level the link between the two models described above.

Let $N \in \mathbb{N}^*$ and consider a partition $\mathcal{P}_N := (P_i)_{1\le i\le N}$ of the initial support $S_0 \subseteq M$ in N regions with $h_N := \max_i \operatorname{diam}(P_i) \lesssim N^{-d}$. We define $\mathbb{X}_N \subset \mathbb{X}$ as the space of functions that are constant on each subdomain P_i, i.e.

$$\mathbb{X}_N := \{X_N \in \mathbb{X} \mid X_N(\omega) = X_N^i \in \mathbb{R}^d \ \text{ for a.e. } \omega \in P_i,\ 1 \le i \le N\}\,.$$

Then, we discretize the flow X by a curve $X_N : [0,T] \to \mathbb{X}_N$, and for any $t \in [0,T]$ we identify $X_N(t)$ with the vector of the position of the particles $(X_N^i(t))_i \in \mathbb{R}^{dN}$ where $X_N^i(t) \in \mathbb{R}^d$ is the image of any point in P_i by the map $X_N(t)$ and therefore carries a mass $\rho_0[P_i]$. As in the continuous case the density of the fluid is given by the pushforward $\rho_N(t) = X_N(t)_{\#}\rho_0$, or more explicitly by the sum of all the particles weighted by their respective masses:

$$\rho_N(t) = \sum_{i=1}^{N} \rho_0[P_i]\delta_{X_N^i(t)}\,. \tag{1.10}$$

Since $\rho_N(t)$ is not absolutely continuous, the internal energy $\mathcal{F}$ is identically $+\infty$ on all of $\mathbb{X}_N$, and in order to define our numerical approximation, we need to replace it by a regularized version. In this paper we consider the Moreau-Yosida regularization of $\mathcal{F}$, which is given by

$$\mathcal{F}_\varepsilon(X) \coloneqq \inf_{\sigma\in\mathbb{X}} \frac{\|X-\sigma\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(\sigma)\,, \tag{1.11}$$

for any $X \in \mathbb{X}$ and for a fixed $\varepsilon > 0$. Note that problem (1.11) always admits minimizers when $X \in \mathbb{X}_N$, but these are in general not unique.

In order to mimic the continuous case, the discrete dynamics is thus given by the Euler (resp. gradient) flow of $\mathcal{F}_\varepsilon$ in $\left(\mathbb{X}_N, L^2_{\rho_0}\right)$. More precisely, the space discretization of the Euler system (1.1) reads as follows:

$$\ddot{X}_N(t) = -P_{\mathbb{X}_N}\nabla_{\mathbb{X}}\mathcal{F}_\varepsilon(X_N(t))\,, \quad X_N(0) = \mathrm{Id}_N\,, \quad \dot{X}_N(0) = u_0\circ \mathrm{Id}_N\,, \tag{1.12}$$

where $P_{\mathbb{X}_N}$ is the $L^2_{\rho_0}$ projection onto $\mathbb{X}_N$, and we set $\mathrm{Id}_N \coloneqq P_{\mathbb{X}_N}\mathrm{Id}|_{S_0}$. Note that the left-hand side of equation (1.12) can be identified with the vector collecting the acceleration of the particles $(\ddot{X}^i_N(t))_i \in \mathbb{R}^{dN}$. The right-hand side is just the gradient of $\mathcal{F}_\varepsilon$ viewed as a function on $\mathbb{X}_N$, and it is uniquely defined for almost every point in $\mathbb{X}_N$ (see Proposition 5.2 for a precise statement). In particular, we have

$$P_{\mathbb{X}_N}\nabla_{\mathbb{X}}\mathcal{F}_\varepsilon(X_N) = \frac{X_N - P_{\mathbb{X}_N}X^\varepsilon_N}{\varepsilon}\,, \quad X^\varepsilon_N \in \operatorname*{argmin}_{\sigma\in\mathbb{X}} \frac{\|X_N-\sigma\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(\sigma)\,, \tag{1.13}$$

for almost any $X_N \in \mathbb{X}_N$. As in the continuous setting, the total energy of the system at time t is given by the sum of the kinetic and internal energy, where we replace now the internal energy by its regularized version:

$$\mathcal{E}_\varepsilon(t, X_N) \coloneqq \sum_{i=1}^N \frac{1}{2}|\dot{X}^i_N(t)|^2\rho_0[P_i] + \mathcal{F}_\varepsilon(X_N(t))\,, \tag{1.14}$$

and this is conserved by smooth solutions of (1.12).

Similarly, the discrete version of the gradient flow (1.9) is given by

$$\dot{X}_N(t) = -P_{\mathbb{X}_N}\nabla_{\mathbb{X}}\mathcal{F}_\varepsilon(X_N(t))\,, \quad X_N(0) = \mathrm{Id}_N\,. \tag{1.15}$$

Here, the total energy at time t is simply given by the internal energy $\mathcal{F}_\varepsilon(X_N(t))$, and it is dissipated by smooth solutions of (1.15).

1.3. **Time discretization.** The variational structure of the space-discrete systems described so far can be exploited to design a stable time discretization. The method we describe here consists in considering different approximations of the energy in each time step, and is modelled on the strategy proposed by Brenier in [3].

Let $\tau > 0$ a fixed time step, $N_T \in \mathbb{N}^*$ be the number of time steps with $T = \tau N_T$, and $t_n \coloneqq n\tau$ for any $0 \le n \le N_T$. We define a discrete-time approximation of system (1.12), by considering the C^1 curves $X_N : [0,T] \mapsto \mathbb{X}_N$ satisfying in each time interval $[t_n, t_{n+1})$ the equation

$$\ddot{X}_N(t) = -\frac{X_N(t) - P_{\mathbb{X}_N}X^\varepsilon_N(t_n)}{\varepsilon}\,, \tag{1.16}$$

where

$$X_N^\varepsilon(t_n) \in \operatorname*{argmin}_{\sigma\in\mathbb{X}} \frac{\|X_N(t_n)-\sigma\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(\sigma)\,, \tag{1.17}$$

and with the same initial condition as in (1.12). This system is conservative in each interval $[t_n, t_{n+1})$ for the energy

$$\mathcal{E}_\varepsilon^n(t, X_N) := \sum_{i=1}^N \frac{1}{2}|\dot X_N^i(t)|^2\rho_0[P_i] + \frac{\|X_N(t)-X_N^\varepsilon(t_n)\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(X_N^\varepsilon(t_n))\,. \tag{1.18}$$

The total energy $\mathcal{E}_\varepsilon(t, X_N)$ defined in (1.14) is however dissipated in general since, by definition of the regularized energy $\mathcal{F}_\varepsilon$, we have

$$\mathcal{E}_\varepsilon(t_{n+1}, X_N) \le \mathcal{E}_{\varepsilon,\tau}^n(t_{n+1}, X_N) = \mathcal{E}_{\varepsilon,\tau}^n(t_n, X_N) = \mathcal{E}_\varepsilon(t_n, X_N)\,. \tag{1.19}$$

The discrete-time approximation of the gradient flow (1.15) is given by a continuous curve $X_N : [0,T] \mapsto \mathbb{X}_N$ which on each interval $[t_n, t_{n+1})$ is the gradient flow on $\mathbb{X}_N$ for the energy:

$$\frac{\|X_N(t)-X_N^\varepsilon(t_n)\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(X_N^\varepsilon(t_n))\,. \tag{1.20}$$

More explicitly, a discrete solution is any C^0 curve $X_N : [0,T] \mapsto \mathbb{X}_N$ which satisfies in each time interval $[t_n, t_{n+1})$,

$$\dot X_N(t) = -\frac{X_N(t) - P_{\mathbb{X}_N} X_N^\varepsilon(t_n)}{\varepsilon}\,, \tag{1.21}$$

with $X_N^\varepsilon(t_n)$ defined as in (1.17), and the same initial condition as in (1.15). Also in this case the internal energy $\mathcal{F}_\varepsilon(X_N(t))$ is dissipated along the evolution, since we have

$$\mathcal{F}_\varepsilon(X_N(t_{n+1})) \le \frac{\|X_N(t_{n+1})-X_N^\varepsilon(t_n)\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(X_N^\varepsilon(t_n)) \le \mathcal{F}_\varepsilon(X_N(t_n))\,. \tag{1.22}$$

1.4. **Relation with previous works and convergence results.** Using a Lagrangian formulation for the discretization of problems (1.1) and (1.4) enables us to reproduce the conservative and gradient flow structure of the corresponding models. In turn, this allows us to construct stable numerical methods as in (1.16) and (1.21) to discretize their solutions. Similar strategies were already explored in the 1990s, during the emergence of particle methods, for example in the context of the discretization of the incompressible Euler equations in the works of Buttke [4] and Russo [28]. Such methods can be seen as instances of the more general Smoothed Particle Hydrodynamics (SPH) discretizations, where the interaction forces amongst the particles are computed by reconstructing the fluid density through convolution with a fixed kernel (see, e.g., the review articles [24, 26] and references therein), and which have been widely used in the context of the discretization of fluid models.

Recent SPH methods explicitely exploit the variational structure of the models for the construction of the method itself as in [10]. In the same article, the Authors also established a general (non-quantitative) convergence result towards measure-valued solutions of problem (1.1) for its discretization in space only. In another recent work [12],

the Authors proved quantitative convergence estimates with modulated energy techniques but limited to the case $P(r) = r^2$ and for the discretization in space only. This last work also highlights how the choice of the kernel is crucial to obtain convergence.

The discretization strategy we use in this paper is closely related to the one developed by Brenier [3], who proposed a discretization of incompressible Euler which replaces the incompressibility constraint by a potential term given by the L^2 distance from the set of measure-preserving maps, discretized as permutations of a fixed regular grid. The potential term used by Brenier can be reinterpreted as a Moreau-Yosida regularization (as in (1.11)) of an energy given by the convex indicator function of the Lebesgue measure. Gallouët and Mérigot [13] later used a similar approach, but rephrased as a particle method, which allowed them to employ efficient semi-discrete optimal transport techniques to compute the discrete solution, and at the same time improved the convergence estimates of [3] using a modulated energy approach. Note that the use of semi-discrete optimal transport techniques to simulate fluids was first launched by the work of Mérigot and Mirebeau [25] to solve the geodesic problem associated with the incompressible Euler equations.

Our convergence results generalize the one in [13] to the compressible and gradient flow setting. Differently from SPH methods, here the density is reconstructed via a Moreau-Yosida regularization (i.e. as the push-forward of ρ_0 by the regularized flow X_N^ε), which eliminates the problem of selecting a kernel, the reconstruction being deeply linked with the energy itself (see Proposition 5.2). On the other hand, the kernel length-scale parameter of SPH methods is replaced here by the parameter ε in the regularized functional (1.11).

The main results of this paper are contained in Theorem 1.1 and 1.2 below. The central issue of the proofs is the construction of an appropriate modulated energy to measure the discrepancy error between the discrete and continuous solution. In this work we construct a modulated energy exploiting the convexity of the energy in the Eulerian setting, which is lost in the Lagrangian formulation, and the particular structure of the Moreau-Yosida regularization. It should be noted that for convex energies, modulated energy estimates of the type we use here are classical tools for the study of problems (1.1) and (1.4) (see, e.g., Chapter 5 in [8]): namely, to prove weak-strong stability and uniqueness results, and to establish convergence in the high friction limit from entropy weak solutions of the Euler equations (1.1) with friction to porous media flow (1.4) [21]. Note also that such tecnhiques are not limited to the cases we consider in this article, and can be generalized to treat also less regular energies (see, e.g., [15, 22], for a framework covering the Euler-Korteweg and Euler-Poisson theory).

Another important point is related to the time discretization. The method we use in this work, described in Section 1.3, directly derives from that used by Brenier in [3] for the incompressible Euler equations. It is specially adapted to the structure of the Moreau-Yosida regularization, and consists in devicing a quadratic approximation of the energy (see equation (1.20)) which dominates the regularized energy over each time-step. This naturally implies the stability of the discrete solutions (see equations (1.19) and (1.22)), which is an essential element for the convergence results below. Note that symplectic integrators [16] could be another natural choice for the discretization of the

Hamiltonian system (1.12). This choice was explored in [13] for incompressible Euler, but it is more difficult to analyze due to the lack of an explicit control of the continuous energy of the system. Another approach which we do not explore in this paper is the time discretization developed in [14, 7] (see also its numerical implementation in [31]) which is better adapted to the non-smooth setting since it is designed to overcome the non-uniqueness issues related to the notion of entropy solutions.

The convergence estimate we obtain for the discretization of (1.1) is the following:

Theorem 1.1. *Suppose that $(\rho, u) : [0,T]\times M \to [0,\infty)\times\mathbb{R}^d$ is a strong solution to (1.1) such that $u\cdot n_{\partial M} = 0$ on $[0,T]\times\partial M$, with $U : [0,\infty) \to \mathbb{R}$ being a smooth strictly convex and superlinear function such that (3.8) holds. Suppose that $u \in C^1([0,T], C^{2,1}(M,\mathbb{R}^d))$, $\rho_0 \in C^{1,1}(M)$, and that either $\rho_0 \geq \rho_{min} > 0$ or that U admits a right third derivative at 0, i.e. $|U'''_+(0)| < \infty$. Suppose in addition that $X_N : [0,T] \to \mathbb{X}_N$ is a C^1 curve which satisfies (1.16) for all times in $[0,T]$, with initial conditions $X_N(0) = \mathrm{Id}_N$ and $\dot X_N(0) = u(0, \mathrm{Id}_N(\cdot))$. Then, denoting by X the flow associated with u satisfying $X(0) = \mathrm{Id}|_{S_0}$,*

$$\sup_{t\in[0,T]} \|\dot X_N(t) - u(t, X_N(t))\|^2_{\mathbb{X}} + \|X_N(t) - X(t)\|^2_{\mathbb{X}} \leq C\Big(\frac{h_N^2}{\varepsilon} + h_N + \varepsilon + \frac{\tau}{\varepsilon}\Big), \tag{1.23}$$

where $C > 0$ depends only on $\sup_{t\in[0,T]}(\|u(t)\|_{C^{2,1}} + \|\partial_t u(t)\|_{C^{2,1}})$, $\|\rho_0\|_{C^{1,1}}$, and on U, T and d.

For what concerns the discretization of dissipative problems of the type (1.4), several Lagrangian discretizations based on their gradient flow structure (1.9) have already been developed (see, e.g., the method in [6] which is close to SPH methods, or in general the review [5] and references therein). The discretization we consider here has been studied in [23] (in the time-continuous setting), where the Authors considered more general energies than those we treat here, modelling for example congestion phenomena, and proved the convergence of the discrete measures (1.10) to solutions of the associated PDE in dimesion one. The result requires an a priori estimate on the regularized flow X^ε_N which is not proven in higher dimensions. Here we circumvent this issue using the same arguments as in Theorem 1.1, and in particular by a careful choice of a modulated energy and by exploiting the smoothness of the continuous solutions. The convergence estimate we obtain for the discretization of problem (1.4) is the following:

Theorem 1.2. *Suppose that $\rho : [0,T]\times M \to [0,\infty)$ is a strong solution to (1.9) such that $\nabla U'(\rho)\cdot n_{\partial M} = 0$ on $[0,T]\times\partial M$, with $U : [0,\infty) \to \mathbb{R}$ being a smooth strictly convex and superlinear function such that (3.8) holds. Suppose that $u := -\nabla U'(\rho)$ is of class $C^{2,1}$ in space, uniformly in time, $\rho_0 \in C^{1,1}(M)$, and that either $\rho_0 \geq \rho_{min} > 0$ or that U admits a right third derivative at 0, i.e. $|U'''_+(0)| < \infty$. Suppose in addition that $X_N : [0,T] \to \mathbb{X}_N$ is a C^0 curve which satisfies (1.21) for all times in $[0,T]$ with initial conditions $X_N(0) = \mathrm{Id}_N$. Then, denoting by X the flow associated with u satisfying $X(0) = \mathrm{Id}|_{S_0}$,*

$$\sup_{t\in[0,T]} \int_0^t \|\dot X_N(s) - u(s, X_N(s))\|^2_{\mathbb{X}}\,\mathrm{d}s + \|X_N(t) - X(t)\|^2_{\mathbb{X}} \leq C\Big(\frac{h_N^2}{\varepsilon} + h_N + \varepsilon + \frac{\tau}{\varepsilon}\Big), \tag{1.24}$$

where $C > 0$ depends only on $\sup_{t\in[0,T]} \|\nabla U'(\rho(t))\|_{C^{2,1}}$, $\|\rho_0\|_{C^{1,1}}$, and on U, T and d.

Remark 1.3. *The modulated energy we use to prove the estimates above has an additional term, if one compares it to the left-hand sides of equations* (1.23) *and* (1.24)*, which is associated with the internal energy* $\mathcal{F}$ *and which is omitted in order to simplify the statements. This term is discussed in detail in Section 3 and actually implies a stronger control on the reconstructed density associated with the regularized flow* X_N^ε*.*

2. Moreau-Yosida regularization

In this section we collect some properties of the regularized energy in (1.11). We provide an equivalent Eulerian formulation of such an energy using the L^2-Wasserstein distance on the space of positive measures of fixed mass, and we also give a characterization of its gradient in terms of the pressure, which will be useful to prove our convergence results.

We start by introducing the Eulerian counterpart to the internal energy functional in (1.8), which we obtain by regarding this as a function of the density rather than the Lagrangian flow map. More precisely, denoting by $\mathcal{M}_+(\mathbb{R}^d)$ the set of positive finite measures on $\mathbb{R}^d$, we define $\mathcal{U} : \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R}$ as follows:

$$\mathcal{U}(\rho) := \begin{cases} \int_M U(\rho)\,\mathrm{d}x & \text{if } \rho \ll \mathrm{d}x \llcorner M, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.1}$$

Then, the functional $\mathcal{F} : \mathbb{X} \to \mathbb{R}$ in (1.8) can be equivalently defined by

$$\mathcal{F}(X) := \mathcal{U}(X_\# \rho_0).$$

We define $\mathcal{U}_\varepsilon(\rho) : \mathcal{M}_+(\mathbb{R}^d) \to \mathbb{R}$ as the Moreau-Yosida regularization of $\mathcal{U}$ with respect to the L^2-Wasserstein distance, i.e.

$$\mathcal{U}_\varepsilon(\rho) := \min_{\mu \in \mathcal{M}_+(\mathbb{R}^d)} \frac{W_2^2(\rho,\mu)}{2\varepsilon} + \mathcal{U}(\mu)\,. \tag{2.2}$$

The quantity $W_2(\rho,\mu)$ is the L^2-Wasserstein distance between ρ and μ (see, e.g., Chapter 5 in [29]), and it can be defined via the following minimization problem:

$$W_2^2(\rho,\mu) := \min_{\gamma \in \Pi(\rho,\mu)} \int |x-y|^2 \,\mathrm{d}\gamma(x,y)\,,$$

where $\Pi(\rho,\mu)$ is the set of positive measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals ρ and μ, and we set $W_2^2(\rho,\mu) = +\infty$ whenever ρ and μ have different total mass. Since U is strictly convex and superlinear, for any $\rho \in \mathcal{M}_+(\mathbb{R}^d)$ (with finite second moment) the function minimized in problem (2.2) is lower semi-continuous with respect to the Wasserstein metric (see, e.g., Proposition 7.7 in [29]) and therefore it admits a unique minimizer which we denote ρ^ε. The link between the Eulerian (2.2) and Lagrangian form (1.11) of the regularized energy is established in the following lemma.

Lemma 2.1. *Let* $X_N \in \mathbb{X}_N$ *and* $\rho_N = (X_N)_\# \rho_0$*, with* $\rho_0 \in \mathcal{M}_+(\mathbb{R}^d)$ *such that* $\rho_0 \ll \mathrm{d}x \llcorner M$*. Then,* $\mathcal{F}_\varepsilon(X_N) = \mathcal{U}_\varepsilon(\rho_N)$*. In particular, there exists a convex function* ψ :

$\mathbb{R}^d \to \mathbb{R}$, *whose gradient is uniquely defined, such that* X_N^ε *is a minimizer associated with* X_N *in problem* (1.11), *i.e.*

$$X_N^\varepsilon \in \operatorname*{argmin}_{\sigma\in\mathbb{X}} \frac{\|X_N-\sigma\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(\sigma)\,,$$

if and only if $X_N = \nabla\psi \circ X_N^\varepsilon$ *up to a negligible set. Moreover, let*

$$\rho_N^\varepsilon \coloneqq \operatorname*{argmin}_{\mu\in\mathcal{M}_+(\mathbb{R}^d)} \frac{W_2^2(\rho_N,\mu)}{2\varepsilon} + \mathcal{U}(\mu)\,.$$

Then, $\rho_N^\varepsilon = (X_N^\varepsilon)_\# \rho_0$.

Proof. Let $\Pi(\rho_N,\mu)$ the set of positive measures on $\mathbb{R}^d\times\mathbb{R}^d$ with marginals $\rho_N = (X_N)_\#\rho_0$ and μ. Since ρ_0 is a.c., for any $\mu\in\mathcal{M}_+(\mathbb{R}^d)$ with the same total mass of ρ_0, there exists a $\sigma\in\mathbb{X}$ such that $\sigma_\#\rho_0 = \mu$, and we can construct a measure $(X_N,\sigma)_\#\rho_0 \in \Pi(\rho_N,\mu)$. This implies that

$$\min_{\gamma\in\Pi(\rho_N,\mu)} \int \frac{|x-y|^2}{2\varepsilon}\,\mathrm{d}\gamma(x,y) + \mathcal{U}(\mu) \le \frac{\|X_N-\sigma\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{U}(\sigma_\#\rho_0)\,. \tag{2.3}$$

Therefore, taking the infimum over σ on both sides of (2.3) yields $\mathcal{U}_\varepsilon(\rho_N) \le \mathcal{F}_\varepsilon(X_N)$.

To prove the reverse inequality, consider again $\rho_N = (X_N)_\#\rho_0 = \sum_i \rho_0[P_i]\delta_{X_N^i}$ and let ρ_N^ε the associated minimizer of problem (2.2). By Brenier's theorem [2], there exists a unique transport map given by the gradient of a convex function ψ such that $(\nabla\psi)_\#\rho_N^\varepsilon = \rho_N$ and $W_2^2(\rho_N,\rho_N^\varepsilon) = \int_M |\nabla\psi - \mathrm{Id}|^2 \mathrm{d}\rho_N^\varepsilon$. This coincides with the optimal transport map from ρ_N^ε to ρ_N. For any $1\le i\le N$, denote $L_i \coloneqq (\nabla\psi)^{-1}(X_N^i)$ so that $\rho_N^\varepsilon[L_i] = \rho_0[P_i]$, and let $\sigma_i : P_i \to L_i$ be any map such that $(\sigma_i)_\#\rho_0|_{P_i} = \rho_N^\varepsilon|_{L_i}$. Then we can take $X_N^\varepsilon\in\mathbb{X}$ to be the map defined by $X_N^\varepsilon|_{P_i} = \sigma_i$. Clearly, $X_N = \nabla\psi\circ X_N^\varepsilon$ by construction and

$$\mathcal{F}_\varepsilon(X_N) \le \frac{\|X_N - X_N^\varepsilon\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{U}((X_N^\varepsilon)_\#\rho_0) = \int_M \frac{|\nabla\psi-\mathrm{Id}|^2}{2\varepsilon}\mathrm{d}\rho_N^\varepsilon + \mathcal{U}(\rho_N^\varepsilon) = \mathcal{U}_\varepsilon(\rho_N)\,.$$

Therefore, we have the equality $\mathcal{U}_\varepsilon(\rho_N) = \mathcal{F}_\varepsilon(X_N)$. Finally, using again equation (2.3) we deduce that if X_N^ε is any minimizer $\rho_N^\varepsilon = (X_N^\varepsilon)_\#\rho_0$. □

Using the optimality conditions of the minimization problem (2.2), one can actually provide an explicit expression for the minimizer ρ_N^ε corresponding to an empirical measure ρ_N. Such a characterization is proven in Proposition 11 in [30], but we recall the precise statement in Proposition 5.2 below. In particular, this shows that ρ_N^ε has a continuous bounded density on M. In turn, this allows us to prove the following statement which is a slight adaptation of Lemma 6.1 in [9].

Lemma 2.2. *Let* $X_N\in\mathbb{X}_N$ *and define* X_N^ε *and* ρ_N^ε *as in Lemma 2.1. For any* $v\in C^1(M,\mathbb{R}^d)$ *with* $v\cdot n_{\partial M} = 0$ *on* ∂M, *we have*

$$\int_{S_0} \frac{X_N - X_N^\varepsilon}{\varepsilon}\cdot v\circ X_N^\varepsilon\,\rho_0\mathrm{d}x = -\int_M P(\rho_N^\varepsilon)\mathrm{div}\,v\,\mathrm{d}x\,. \tag{2.4}$$

Proof. We follow the proof of Lemma 6.1 in [9] and introduce first the flow of v, i.e. for $\delta > 0$ we define $Y : (-\delta, \delta) \times M \to M$ as the solution to the flow equation $\dot{Y}_s = v \circ Y_s$ for $s \in (-\delta, \delta)$ and with $Y_0 = \mathrm{Id}$, the identity map on M. Note that $Y_s : M \to M$ is a C^1 diffeomorphism, since v is C^1 and it is tangent to the boundary, and we have

$$\partial_s \det \nabla Y_s = (\operatorname{div} v \circ Y_s) \det \nabla Y_s \,. \tag{2.5}$$

Then we define $\rho_s := (Y_s)_{\#}\rho_N^\varepsilon$, and identifying ρ_N^ε with its density with respect to $\mathrm{d}x \llcorner M$ we have

$$\rho_s = \frac{\rho_N^\varepsilon}{\det \nabla Y_s} \circ Y_s^{-1} \,, \tag{2.6}$$

which can be directly deduced via a change variables in the integral formulation of the definition of the pushforward (1.6). Moreover, the function

$$g : s \in (-\delta, \delta) \to \frac{W_2^2(\rho_N, \rho_s)}{2\varepsilon} + \mathcal{U}(\rho_s) \in \mathbb{R}$$

has a minimum at $s = 0$. Since ρ_N^ε is bounded, using equation (2.6), (2.5), and the definition of P in (1.3) we obtain

$$\frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} \mathcal{U}(\rho_s) = \frac{\mathrm{d}}{\mathrm{d}s}\bigg|_{s=0} \int_M U\left(\frac{\rho_N^\varepsilon}{\det \nabla Y_s}\right) \det \nabla Y_s \,\mathrm{d}x = -\int_M P(\rho_N^\varepsilon) \operatorname{div} v \,\mathrm{d}x \,. \tag{2.7}$$

We now introduce $\gamma_s = (\nabla\psi, Y_s)_{\#}\rho_N^\varepsilon$, so that $W_2^2(\rho_N, \rho_s) \leq \int |x - y|^2 \mathrm{d}\gamma_s(x, y)$, which implies

$$\begin{aligned} W_2^2(\rho_N, \rho_s) - W_2^2(\rho_N, \rho_N^\varepsilon) &\leq \int_M |\nabla\psi - Y_s|^2 \mathrm{d}\rho_N^\varepsilon - \int_M |\nabla\psi - \mathrm{Id}|^2 \mathrm{d}\rho_N^\varepsilon \\ &= \int_M (Y_s - \mathrm{Id}) \cdot (\mathrm{Id} + Y_s - 2\nabla\psi) \mathrm{d}\rho_N^\varepsilon \,. \end{aligned}$$

Therefore,

$$0 \leq g(s) - g(0) \leq \frac{1}{2\varepsilon} \int_M (Y_s - \mathrm{Id}) \cdot (\mathrm{Id} + Y_s - 2\nabla\psi) \mathrm{d}\rho_N^\varepsilon + \mathcal{U}(\rho_s) - \mathcal{U}(\rho_N^\varepsilon) \,.$$

Dividing by s, taking the limit for $s \to 0$ and using equation (2.7) gives

$$\int_M \frac{\nabla\psi - \mathrm{Id}}{\varepsilon} \cdot v \,\mathrm{d}\rho_N^\varepsilon \leq -\int_M P(\rho_N^\varepsilon) \operatorname{div} v \,\mathrm{d}x \,. \tag{2.8}$$

Since the same also holds replacing v by $-v$, equality holds and we obtain equation (2.4) by a change of variables on the left-hand side of (2.8).

□

Remark 2.3. *Note that using the same computation of equation (2.7), and performing a change of variables on its right-hand side, one can formally identify $\nabla_{\mathbb{X}} \mathcal{F}(X_t) = \nabla U'(\rho_t) \circ X_t$ in equation (1.8) and (1.9).*

3. Modulated energy

In this section we introduce the two main quantities that we will need to measure the distance between continuous and discrete solutions of problems (1.1) and (1.4). These are constructed as discrete versions of the classical relative kinetic and internal energy of the system expressed in Eulerian variables. Here we adapt these definitions to our discrete Lagrangian setting and to the regularized energy (1.11).

The relative kinetic energy in the discrete setting is defined as follows:

Definition 3.1 (Relative kinetic energy)**.** Given a curve $u : [0,T] \to C^0(\mathbb{R}^d;\mathbb{R}^d)$, the relative kinetic energy of a discrete flow $X_N : [0,T] \to \mathbb{X}_N$ with respect to u at time t is given by

$$\begin{aligned} \mathcal{K}_u(t, X_N) &:= \frac{1}{2}\|\dot{X}_N(t,\cdot) - u(t, X_N(t,\cdot))\|^2_{\mathbb{X}} \\ &= \frac{1}{2}\sum_{i=1}^N |\dot{X}^i_N(t) - u(t, X^i_N(t))|^2 \rho_0[P_i]\,. \end{aligned} \tag{3.1}$$

Remark 3.2. *The choice of the relative kinetic energy in definition 3.1 can be motivated as follows. The kinetic energy can be viewed as a convex function of the density ρ and the momentum $m = \rho u$ given by*

$$\int_M \frac{|m|^2}{2\rho}\,. \tag{3.2}$$

Then, it is natural to measure the distance between two states (ρ, m) and $(\tilde\rho, \tilde m)$, with $\tilde m = \tilde\rho\tilde u$, by considering the difference between the value of the functional (3.2) at (ρ, m) and the linear part of its Taylor expansion at $(\tilde\rho, \tilde m)$ in the direction $(\rho - \tilde\rho, m - \tilde m)$. The resulting quantity is given by

$$\int_M \frac{1}{2}|u - \tilde u|^2 \rho \,\mathrm{d}x\,, \tag{3.3}$$

which is precisely the Eulerian counterpart to equation (3.1).

In the order to define the relative internal energy in the discrete setting, for any $\rho, \tilde\rho \in C^0(M, (0,\infty))$ we first define

$$\mathcal{U}(\rho|\tilde\rho) := \int_M U(\rho|\tilde\rho)\,\mathrm{d}x\,, \tag{3.4}$$

where

$$U(r|s) := U(r) - U(s) - U'(s)(r - s)\,. \tag{3.5}$$

If $|U'_+(0)| < +\infty$, equation (3.4) defines $\mathcal{U}(\rho|\tilde\rho)$ for any $\rho, \tilde\rho \in C^0(M, [0,\infty))$. Since we assume U to be strictly convex, $\mathcal{U}(\rho|\tilde\rho) \geq 0$ and it vanishes if and only if $\rho = \tilde\rho$.

The relative internal energy in the discrete setting is defined in order to fit the solutions of the numerical schemes detailed in Section 1.3, and in particular the corresponding time discretization, which we recall in the definition below.

Definition 3.3 (Discrete relative internal energy)**.** Let $\tau > 0$ a fixed time step, $N_T \in \mathbb{N}^*$ be the number of time steps with $T = \tau N_T$, and $t_n := n\tau$ for any $0 \leq n \leq N_T$. Given a curve $\rho : [0,T] \to C^0(\mathbb{R}^d; [0,\infty))$, the discrete relative internal energy of a discrete flow $X_N : [0,T] \to \mathbb{X}_N$ with respect to ρ at time $t \in [t_n, t_{n+1})$ is given by

$$\mathcal{F}_{\varepsilon,\rho}(t, X_N) := \frac{\|X_N(t) - X_N^\varepsilon(t_n)\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{U}(\rho_N^\varepsilon(t_n)|\rho(t))\,, \tag{3.6}$$

where $X_N^\varepsilon(t_n)$ is any fixed minimizer of problem (1.11), i.e.

$$X_N^\varepsilon(t_n) \in \operatorname*{argmin}_{\sigma \in \mathbb{X}} \frac{\|X_N(t_n) - \sigma\|_{\mathbb{X}}^2}{2\varepsilon} + \mathcal{F}(\sigma)\,,$$

and $\rho_N^\varepsilon(t_n) := (X_N^\varepsilon(t_n))_\# \rho_0$.

Remark 3.4. *In the smooth Eulerian setting the relative internal energy would be given just by the functional in equation (3.4). Importantly, even if the potential energy of the discrete system is a convex functional on $\mathbb{X}$, the discrete relative internal energy in (3.6) does not correspond to this point of view and should rather be regarded as an approximation of (3.4). The same holds for the definition of the relative kinetic energy above, which does not coincide with the one obtained interpreting the kinetic energy as convex functional on $\mathbb{X}$. This time however there is no approximation since if we replaced X_N by a smooth injective flow we could recover (3.3) from (3.1) by a simple change of variables.*

The convergence proof in Section 4 will rely on a Grönwall argument based on the discrete relative energies (3.1) and (3.6). It will require us to control the time variation of the total discrete relative energy by itself. The advantage of adopting an Eulerian rather than Lagrangian point of view in the definitions above is that, in the Eulerian case, such a control can be enforced by exploiting simple algebraic properties of the functions P and U. More precisely, we will need to control the relative pressure

$$P(r|s) := P(r) - P(s) - P'(s)(r - s) \tag{3.7}$$

by $U(r|s)$. To this end, we will make the following assumption: there exists a constant $A > 0$ such that

$$|P''(r)| \leq A\, U''(r) \quad \forall\, r > 0\,. \tag{3.8}$$

This assumption is verified in the classical cases of interest of power laws and of the entropy. It implies the following lemma, which is an extract of Lemma 3.3 in [15].

Lemma 3.5. *Let U and P be smooth functions on $[0,\infty)$ verifying (1.3) and (3.8). Then*

$$|P(r|s)| \leq AU(r|s) \quad \forall\, r, s > 0\,. \tag{3.9}$$

Proof. We have $P(r|s) = (r-s)^2 \int_0^1 (1-\theta) P''((1-\theta)s + \theta r)\, \mathrm{d}\theta$ and similarly for $U(r|s)$. Hence, using equation (3.8),

$$|P(r|s)| \leq (r-s)^2 \int_0^1 (1-\theta)|P''((1-\theta)s + \theta r)|\, \mathrm{d}\theta \leq A\, U(r|s)\,.$$

□

Remark 3.6. *In the following, in order to treat the case of the convergence towards solutions with vanishing density we will need to add the hypothesis that U admits a right third derivative at 0, i.e. $|U'''_+(0)| < \infty$. Note that in this setting, if equation (3.9) holds for $r, s > 0$, then it holds by continuity for $r, s \geq 0$.*

4. Convergence of the fully discrete scheme

In this section we use the discrete relative energies introduced in Section 3 to prove our convergence results for the space-time discretization of problems (1.1) and (1.4) defined in Section 1.3.

Since the image of the discrete solution $X_N(t)$ (i.e. the particles' positions) may not be contained in the domain M, an essential ingredient of the proof is the possibility to extend the exact solution of the continuous models outside the domain. Importantly, besides keeping the same regularity, the extended density and veloctiy will need to satisfy the continuity equation also outside the domain. We construct such extended variables in the following lemma, by exploting the properties of the continuity equation and using an extension theorem due to Fefferman [11].

Lemma 4.1. *Let $u : [0,T] \times M \to \mathbb{R}^d$ be such that $u \cdot n_{\partial M} = 0$ on $[0,T] \times \partial M$, and $\rho_0 : M \to [0,\infty)$. If u is of class $C^{2,1}$ in space, uniformly in time, and ρ_0 is of class $C^{1,1}$, then there exist $\tilde{u} : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ and $\tilde{\rho} : [0,T] \times \mathbb{R}^d \to \mathbb{R}$ such that:*

(1) *$\tilde{u}$ is an extension of u, i.e. $\tilde{u}(t)|_M = u(t)$ for all $t \in [0,T]$, and there exists a constant $C > 0$ only depending on d such that*

$$\sup_{t\in[0,T]} \|\tilde{u}(t)\|_{C^{2,1}} \leq C \sup_{t\in[0,T]} \|u(t)\|_{C^{2,1}} ; \tag{4.1}$$

moreover, if $u \in C^1([0,T], C^{2,1}(M,\mathbb{R}^d))$ then

$$\sup_{t\in[0,T]} \|\partial_t\tilde{u}(t)\|_{C^{2,1}} \leq C \sup_{t\in[0,T]} \|\partial_t u(t)\|_{C^{2,1}} ; \tag{4.2}$$

(2) *the couple $(\tilde{\rho}, \tilde{u})$ solves the continuity equation:*

$$\partial_t\tilde{\rho} + \operatorname{div}(\tilde{\rho}\tilde{u}) = 0 \quad \text{on } [0,T] \times \mathbb{R}^d,$$

and in particular the curve $\rho : t \in [0,T] \to \tilde{\rho}(t)|_M$ is the unique solution to the continuity equation on $[0,T] \times M$ associated with u and initial conditions $\rho(0) = \rho_0$; if $\rho_0 \geq \rho_{min} > 0$, then $\tilde{\rho} \geq \tilde{\rho}_{min} > 0$, where $\tilde{\rho}_{min}$ only depend on ρ_{min}, $\sup_{t\in[0,T]} \|u(t)\|_{C^{2,1}}$, T and d; moreover, $\sup_{t\in[0,T]} \|\tilde{\rho}(t)\|_{C^{1,1}}$ only depends on $\|\rho_0\|_{C^{1,1}}$, $\sup_{t\in[0,T]} \|u(t)\|_{C^{2,1}}$, T, d (and on ρ_{min} in the case $\rho_0 \geq \rho_{min} > 0$).

Proof. The first part is just an application of the construction proposed by Fefferman in [11] to extend Hölder continuous functions. In particular, by theorem 2 in [11], for any $k \geq 0$ there exists a linear bounded operator $L_k : C^{k,1}(M) \to C^{k,1}(\mathbb{R}^d)$ such that the norm of L_k is bounded by a constant depending only on d and k, and for any $f \in C^{k,1}(\mathbb{R}^d)$ one has $L_k f|_M = f$. Then, setting $\tilde{u}(t) = L_2\, u(t)$ (applied componentwise) for all $t \in [0,T]$ for a given extension operator L_2, we obtain the estimate (4.1) by the boundedness of L_2. In the case where $u \in C^1([0,T], C^{2,1}(M,\mathbb{R}^d))$, by linearity of L_2 we have $\partial_t\tilde{u} = L_2\partial_t u$, from which we deduce (4.2).

For the second part, we first introduce $X : [0,T] \times \mathbb{R}^d \to \mathbb{R}^d$ the flow of $\tilde{u}$, i.e. the solution to the flow equation $\dot{X}_t = \tilde{u}(t, X_t)$ with initial conditions $X_0 = \mathrm{Id}$. For all times $t \in [0,T]$, X_t is a $C^{2,1}$ diffeomorphism of $\mathbb{R}^d$ and by construction the $C^{2,1}$ norm of X_t and X_t^{-1} only depend on that of u and on T. Note, in particular, that the Jacobian determinant solves

$$\partial_t \det \nabla X_t = \operatorname{div} \tilde{u}(t, X_t) \det \nabla X_t \,,$$

which implies that for all $(t,x) \in [0,T] \times \mathbb{R}^d$,

$$\max\left\{ \det\nabla X_t(x), \frac{1}{\det\nabla X_t(x)} \right\} \leq \exp\left(\int_0^t \|\operatorname{div} \tilde{u}(t)\|_\infty \, \mathrm{d}t \right) . \tag{4.3}$$

Now, if ρ_0 is not strictly-positive, we define an extension $\tilde{\rho}_0 : \mathbb{R}^d \to \mathbb{R}$ of ρ_0 on the whole space by $\tilde{\rho}_0 \coloneqq L_1 \rho_0$ (and note that $\tilde{\rho}_0$ may be negative). Then, we define for all $t \in [0,T]$

$$\tilde{\rho}(t) = \frac{\tilde{\rho}_0}{\det \nabla X_t} \circ X_t^{-1} \,, \tag{4.4}$$

and therefore the regularity of $\tilde{\rho}$ in space derives from that of $\tilde{\rho}_0$, X_t^{-1} and $\det \nabla X_t$, and from the bound (4.3). Moreover, by direct computation one can check that $\tilde{\rho}$ solves the continuity equation with velocity $\tilde{u}$. On the other hand, if $\rho_0 \geq \rho_{min} > 0$, we define $\tilde{\rho}_0 \coloneqq \exp(L_1 \log(\rho_0))$ and $\tilde{\rho}$ as above. Then, the lower bound on $\tilde{\rho}$ can be deduced from equations (4.4) and (4.3). □

In the following, we finally prove Theorem 1.1 and 1.2, which establish a bound on the rate of convergence for our space-time discretizations of problems (1.1) and (1.4), respectively.

Proof of Theorem 1.1. Throughout the proof we will denote by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ the inner product and norm on $\mathbb{X}$, respectively, i.e. the L^2 inner product and norm weighted by ρ_0. Moreover, for any function $f : [0,T] \to C^{0,1}(E)$ with $E \subseteq \mathbb{R}^d$, we will denote by $\mathrm{Lip}_T(f) \coloneqq \sup_{t \in [0,T]} \mathrm{Lip}(f(t))$ and we will use the same notation for vector-valued functions.

We denote by $\tilde{u}$ and $\tilde{\rho}$ the extensions of u and ρ, respectively, constructed via Lemma 4.1. Note that if ρ is not strictly-positive, $\tilde{\rho}$ may be negative. However, since in the case we suppose that $|U_+'''(0)| < +\infty$, we replace U by a C^3 extension defined on $\mathbb{R}$ (which we still denote by U with an abuse of notation), e.g., by setting $U(r) = \sum_{n=0}^3 U_+^{(n)}(0) r^n / n!$ for $r < 0$. Then, $U^{(n)}(\tilde{\rho})$ is Lipschitz in space, uniformly in time, for $n = 0, 1, 2$.

We define the relative energy as follows:

$$\mathcal{E}_{\rho,u}(t, X_N) \coloneqq \mathcal{K}_{\tilde{u}}(t, X_N) + \mathcal{F}_{\varepsilon,\rho}(t, X_N) + \frac{1}{2} \|X_N(t) - X(t)\|^2 \,. \tag{4.5}$$

Note that besides the relative kinetic and internal energy, we also included an additional term in (4.5) given by the squared L^2 distance between the flows and which will help us deal with the fact that the image of $X_N(t)$ may not be included in M. Note also that while the relative kinetic energy needs to be computed using the extended velocity field $\tilde{u}$, for the relative internal energy we can use indifferently either ρ or $\tilde{\rho}$ since it is defined via an integral over the (fixed) domain M.

The strategy of the proof is the following. First of all, we compute separately the time derivative of the three terms in (4.5) for $t \in [t_n, t_{n+1})$. We then apply Grönwall's inequality on the same time interval to obtain a first estimate. Finally, we use a discrete Grönwall's inequality for $0 \leq n \leq N_T$ to prove the result.

Step 1: Time derivative of the relative kinetic energy. We introduce the material derivative

$$D_t\tilde{u}(t) \coloneqq \partial_t\tilde{u}(t) + \tilde{u}(t)\cdot\nabla\tilde{u}(t)\,.$$

Then, using equation (1.12), we have

(4.6)
$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}_{\tilde{u}}(t,X_N) &= \langle \ddot{X}_N(t) - \partial_t\tilde{u}(t,X_N(t)) - \dot{X}_N(t)\cdot\nabla\tilde{u}(t,X_N(t)), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle \\ &= -\langle (\dot{X}_N(t) - \tilde{u}(t,X_N(t)))\cdot\nabla\tilde{u}(t,X_N(t)), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle \\ &\quad - \langle \varepsilon^{-1}(X_N(t) - X^\varepsilon_N(t_n)) + D_t\tilde{u}(t,X_N(t))), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle\,,\end{aligned}$$

where we replaced $\ddot{X}_N(t)$ using (1.12), and we removed the projection onto $\mathbb{X}_N$, since $\dot{X}_N(t) - \tilde{u}(t,X_N(t)) \in \mathbb{X}_N$. Observe that the system (1.1) implies

$$\rho_0 D_t\tilde{u}(t,X(t)) = -\rho_0\nabla U'(\tilde{\rho}(t,X(t))\,.$$

Then, adding and subtracting $\nabla U'(\tilde{\rho}(t,X(t)))$ and $\nabla U'(\tilde{\rho}(t,X_N(t)))$ in the last inner product in (4.6), we obtain

(4.7)
$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{K}_{\tilde{u}}(t,X_N) = &-\langle (\dot{X}_N(t) - \tilde{u}(t,X_N(t)))\cdot\nabla\tilde{u}(t,X_N(t)), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle \\ &- \langle D_t\tilde{u}(t,X_N(t))) - D_t\tilde{u}(t,X(t))), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle \\ &+ \langle \nabla U'(\tilde{\rho}(t,X(t))) - \nabla U'(\tilde{\rho}(t,X_N(t))), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle \\ &- \langle \varepsilon^{-1}(X_N(t) - X^\varepsilon_N(t_n)) - \nabla U'(\tilde{\rho}(t,X_N(t))), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle\,.\end{aligned}$$

Step 2: Time derivative of the relative internal energy. First of all, we define the following quantity which will be useful for the computations below and also later in the Grönwall argument (see also Remark 4.2 below):

$$H^n(t) \coloneqq \int_{\mathbb{R}^d} U'(\tilde{\rho}(t))\mathrm{d}(\rho^\varepsilon_N(t_n) - \rho_N(t))\,. \tag{4.8}$$

We now compute the time derivatives of the different terms in $\mathcal{F}_{\varepsilon,\rho}(t,X_N)$ (defined by equation (3.6)) for $t \in [t_n, t_{n+1})$. By the same computations as in (2.7), we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{U}(\rho(t)) = -\int_M P(\rho(t))\mathrm{div}\,u(t)\,\mathrm{d}x\,. \tag{4.9}$$

For the time derivative of the discrete energy we can arrange the terms in order to obtain a similar quantity. In particular, we have

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\|X_N(t)-X_N^\varepsilon(t_n)\|^2}{2\varepsilon}+\mathcal{U}(\rho_N^\varepsilon(t_n))\right)\\
&\qquad=\varepsilon^{-1}\langle X_N(t)-X_N^\varepsilon(t_n),\dot X_N(t)\rangle\\
&\qquad=\varepsilon^{-1}\langle X_N(t)-X_N^\varepsilon(t_n),\dot X_N(t)-\tilde u(t,X_N(t))\rangle\\
&\qquad\quad+\varepsilon^{-1}\langle X_N(t)-X_N^\varepsilon(t_n),\tilde u(t,X_N(t))-\tilde u(t,X_N^\varepsilon(t_n))\rangle\\
&\qquad\quad+\varepsilon^{-1}\langle X_N(t)-X_N(t_n),u(t,X_N^\varepsilon(t_n))\rangle\\
&\qquad\quad+\varepsilon^{-1}\langle X_N(t_n)-X_N^\varepsilon(t_n),u(t,X_N^\varepsilon(t_n))\rangle\,,
\end{aligned}\tag{4.10}$$

and note that by Lemma 2.2, the last term in in equation (4.10) can be written as follows

$$\varepsilon^{-1}\langle X_N(t_n)-X_N^\varepsilon(t_n),u(t,X_N^\varepsilon(t_n))\rangle=-\int_M P(\rho_N^\varepsilon(t_n))\mathrm{div}\,u(t)\,\mathrm{d}x\,.\tag{4.11}$$

We write the time derivative of the remaining term in $\mathcal{F}_{\varepsilon,\rho}(t,X_N)$ as follows:

$$\frac{\mathrm{d}}{\mathrm{d}t}\int_M U'(\rho(t))(\rho_N^\epsilon(t_n)-\rho(t))\,\mathrm{d}x=\frac{\mathrm{d}}{\mathrm{d}t}H^n(t)+\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}U'(\tilde\rho(t))\mathrm{d}(\rho_N(t)-\rho(t)).\tag{4.12}$$

Note that here we identify $\rho(t)$ with a measure on $\mathbb{R}^d$ extending it by zero, and we will use the same convention also in the following. Then, we compute

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}U'(\tilde\rho(t))\,\mathrm{d}(\rho_N(t)-\rho(t))=&\langle\nabla U'(\tilde\rho(t))\circ X_N(t),\dot X_N(t)\rangle\\
&-\int_M u(t)\cdot\nabla U'(\rho(t))\rho(t)\,\mathrm{d}x\\
&-\int_{\mathbb{R}^d}U''(\tilde\rho(t))\mathrm{div}(\tilde\rho(t)\tilde u(t))\,\mathrm{d}(\rho_N(t)-\rho(t)).
\end{aligned}\tag{4.13}$$

Remark that here we used the fact that the continuity equation holds also for the extended functions $(\tilde\rho,\tilde u)$, which is due to the construction described in Lemma 4.1. Using $\mathrm{div}(\tilde\rho\tilde u)=\tilde\rho\,\mathrm{div}\tilde u+\nabla\tilde\rho\cdot\tilde u$ and then using $P'(r)=rU''(r)$, we get

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_{\mathbb{R}^d}U'(\tilde\rho(t))\,\mathrm{d}(\rho_N(t)-\rho(t))=&\langle\nabla U'(\tilde\rho(t))\circ X_N(t),\dot X_N(t)-\tilde u(t,X_N(t))\rangle\\
&-\int_{\mathbb{R}^d}P'(\tilde\rho(t))\mathrm{div}\tilde u(t)\mathrm{d}(\rho_N(t)-\rho(t))\,.
\end{aligned}\tag{4.14}$$

Putting this back into equation (4.12), we find

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\int_M U'(\rho(t))(\rho_N^\epsilon(t_n)-\rho(t))\,\mathrm{d}x=&\frac{\mathrm{d}}{\mathrm{d}t}H^n(t)+\langle\nabla U'(\tilde\rho(t))\circ X_N(t),\dot X_N(t)-\tilde u(t,X_N(t))\rangle\\
&-\int_{\mathbb{R}^d}P'(\tilde\rho(t))\mathrm{div}\tilde u(t)\mathrm{d}(\rho_N(t)-\rho_N^\epsilon(t_n))\\
&-\int_M P'(\rho(t))\mathrm{div}\,u(t)(\rho_N^\epsilon(t_n)-\rho(t))\,\mathrm{d}x\,.
\end{aligned}\tag{4.15}$$

Note that we have added and subtracted $\rho_N^\epsilon(t_n)$ in the last integral, which allows us to retrieve $P(\rho_N^\varepsilon(t_n)|\rho)$ when combining all terms. In fact, replacing (4.11) into (4.10), and subtracting the contributions from (4.9) and (4.15), we obtain

(4.16)
$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{F}_{\varepsilon,\rho}(t,X_N) &= \langle \varepsilon^{-1}(X_N(t)-X_N^\varepsilon(t_n)) - \nabla U'(\tilde\rho(t,X_N(t))), \dot X_N(t)-\tilde u(t,X_N(t))\rangle\\ &\quad+\varepsilon^{-1}\langle X_N(t)-X_N^\varepsilon(t_n), \tilde u(t,X_N(t))-\tilde u(t,X_N^\varepsilon(t_n))\rangle\\ &\quad-\int_M P(\rho_N^\varepsilon(t_n)|\rho(t))\mathrm{div}\, u(t)\,\mathrm{d}x\\ &\quad+\int_{\mathbb{R}^d} P'(\tilde\rho(t))\mathrm{div}(\tilde u(t))\mathrm{d}(\rho_N(t)-\rho_N^\epsilon(t_n))\\ &\quad+\varepsilon^{-1}\langle X_N(t)-X_N(t_n), \tilde u(t,X_N^\varepsilon(t_n))\rangle - \frac{\mathrm{d}}{\mathrm{d}t}H^n(t)\,.\end{aligned}$$

We finally observe that the first term on the right-hand side of equation (4.16) coincides with the opposite of the last term in (4.7). Therefore the two terms cancel out when adding the two equations. The decomposition of the time derivative in (4.10) is designed to exploit this feature, which is a consequence of energy conservation.

Step 3: Grönwall's argument on $[t_n, t_{n+1})$. Combining

$$\frac{\mathrm{d}}{\mathrm{d}t}\frac{1}{2}\|X_N(t)-X(t)\|^2 = \langle \dot X_N(t)-\dot X(t), X_N(t)-X(t)\rangle$$

with equations (4.7) and (4.16), we obtain

(4.17)
$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{E}_{\rho,u}(t,X_N) &= -\langle D_t\tilde u(t,X_N(t))) - D_t\tilde u(t,X(t))), \dot X_N(t)-\tilde u(t,X_N(t))\rangle\\ &\quad+\langle \nabla U'(\tilde\rho(t,X(t))) - \nabla U'(\tilde\rho(t,X_N(t))), \dot X_N(t)-\tilde u(t,X_N(t))\rangle\\ &\quad-\langle (\dot X_N(t)-\tilde u(t,X_N(t)))\cdot\nabla\tilde u(t,X_N(t)), \dot X_N(t)-\tilde u(t,X_N(t))\rangle\\ &\quad+\langle \dot X_N(t)-\dot X(t), X_N(t)-X(t)\rangle\\ &\quad+\varepsilon^{-1}\langle X_N(t)-X_N^\varepsilon(t_n), \tilde u(t,X_N(t))-\tilde u(t,X_N^\varepsilon(t_n))\rangle\\ &\quad-\int_M P(\rho_N^\varepsilon(t_n)|\rho(t))\mathrm{div}\, u(t)\,\mathrm{d}x\\ &\quad+\int_{\mathbb{R}^d} P'(\tilde\rho(t))\mathrm{div}(\tilde u(t))\mathrm{d}(\rho_N(t)-\rho_N^\epsilon(t_n))\\ &\quad+\varepsilon^{-1}\langle X_N(t)-X_N(t_n), u(t,X_N^\varepsilon(t_n))\rangle - \frac{\mathrm{d}}{\mathrm{d}t}H^n(t)\\ &=: J_1+J_2+J_3+J_4+J_5+J_6+J_7+J_8-\frac{\mathrm{d}}{\mathrm{d}t}H^n(t)\,.\end{aligned}$$

Applying Cauchy-Schwarz and then Young's inequality to the first two terms we obtain

$$J_1+J_2 \le 2(\mathrm{Lip}_T(D_t\tilde u)+\mathrm{Lip}_T(\nabla U'(\tilde\rho)))\left(\mathcal{K}_{\tilde u}(t,X_N)+\frac{1}{2}\|X_N(t)-X(t)\|^2\right), \tag{4.18}$$

where $D_t\tilde u$ and $\nabla U'(\tilde\rho)$ are interpreted as functions on $[0,T]\times\mathbb{R}^d$.

For J_4, we have

$$\begin{aligned} J_4 &= \langle \dot{X}_N(t) - \tilde{u}(t, X_N(t)), X_N(t) - X(t)\rangle \\ &\quad + \langle \tilde{u}(t, X_N(t)) - \tilde{u}(t, X(t)), X_N(t) - X(t)\rangle \\ &\leq \mathcal{K}_{\tilde{u}}(t, X_N) + (1 + 2\mathrm{Lip}_T(\tilde{u}))\,\frac{1}{2}\|X_N(t) - X(t)\|^2\,, \end{aligned} \tag{4.19}$$

Using the estimate (4.19), we find

$$\begin{aligned} \sum_{i=3}^{6} J_i &\leq (1 + 2\mathrm{Lip}_T(\tilde{u}))\left(\mathcal{K}_{\tilde{u}}(t, X_N) + \frac{\|X_N(t) - X(t)\|^2}{2}\right) \\ &\quad + \mathrm{Lip}_T(\tilde{u})\frac{\|X_N(t) - X_N^\varepsilon(t_n)\|^2}{\varepsilon} + A\mathrm{Lip}_T(\tilde{u})\mathcal{U}(\rho_N^\varepsilon(t_n)|\rho(t)) \\ &\leq (1 + A'\mathrm{Lip}_T(\tilde{u}))\mathcal{E}_{\rho,u}(t, X_N)\,, \end{aligned} \tag{4.20}$$

where $A' \coloneqq \max(A, 2)$, and we used for J_6 the inequality given in Lemma 3.5 (see also Remark 3.6). Hence, combining (4.18) and (4.20) we obtain

$$\sum_{i=1}^{6} J_i \leq C_1\mathcal{E}_{\rho,u}(t, X_N)\,, \tag{4.21}$$

where $C_1 \coloneqq 2\mathrm{Lip}_T(D_t\tilde{u}) + 2\mathrm{Lip}_T(\nabla U'(\tilde{\rho})) + A'\,\mathrm{Lip}_T(\tilde{u}) + 1$. For J_7 we have

$$\begin{aligned} J_7 &\leq \mathrm{Lip}_T(P'(\tilde{\rho})\,\mathrm{div}\,\tilde{u})W_1(\rho_N(t), \rho_N^\varepsilon(t_n)) \\ &\leq C_2\left(\frac{\varepsilon}{2} + \frac{W_2^2(\rho_N(t), \rho_N^\varepsilon(t_n))}{2\varepsilon}\right) \\ &\leq C_2\left(\frac{\varepsilon}{2} + \frac{\|X_N(t) - X_N^\varepsilon(t_n)\|^2}{2\varepsilon}\right)\,, \end{aligned} \tag{4.22}$$

where $W_1(\cdot,\cdot)$ denotes the L^1-Wasserstein distance and we have used the inequality $W_1(\rho_N(t), \rho_N^\varepsilon(t_n)) \leq W_2(\rho_N(t), \rho_N^\varepsilon(t_n))$ (see Chapter 5 in [29]), and where $C_2 \coloneqq \mathrm{Lip}_T(P'(\tilde{\rho})\,\mathrm{div}\,\tilde{u})$.

For J_8 we have

$$\begin{aligned} J_8 &= \frac{1}{\varepsilon}\int_{t_n}^{t} \langle \dot{X}_N(t'), u(t, X_N^\varepsilon(t_n))\rangle\,\mathrm{d}t' \\ &\leq \frac{1}{\varepsilon}\int_{t_n}^{t} \|\dot{X}_N(t')\|\|u(t, X_N^\varepsilon(t_n))\|\,\mathrm{d}t' \\ &\leq \frac{\tau}{\varepsilon}\left(\mathcal{E}_\varepsilon(t_n, X_N) - \min\mathcal{U} + \frac{1}{2}\|u\|^2_{L^\infty([0,T]\times M)}\right) \\ &\leq \frac{\tau}{\varepsilon}\left(\mathcal{E}_\varepsilon(0, X_N) + C_3\right)\,, \end{aligned} \tag{4.23}$$

where we used the conservation-dissipation of the energy $\mathcal{E}_\varepsilon$ (1.19) for the last two inequalities, and where $C_3 \coloneqq \|u\|^2_{L^\infty([0,T]\times M)}/2 - \min\mathcal{U}$.

Using the same argument as for J_7, we obtain

$$
\begin{aligned}
|H^n(t)| \leq & |\mathrm{Lip}_T(U'(\tilde{\rho}))|^2 \varepsilon + \frac{W_2^2(\rho_N(t), \rho_N^\varepsilon(t_n))}{4\varepsilon} \\
\leq & C_4 \varepsilon + \frac{1}{2} \mathcal{E}_{\rho,u}(t, X_N)\,,
\end{aligned} \tag{4.24}
$$

where $C_4 \coloneqq |\mathrm{Lip}_T(U'(\tilde{\rho}))|^2$.

This last inequality allows us to include $H^n(t)$ in the Grönwall argument. In particular, let $E^n(t) \coloneqq \mathcal{E}_{\rho,u}(t, X_N) + H^n(t)$. Combining the estimates (4.21), (4.22), (4.23), into (4.17), we find

$$
\frac{\mathrm{d}}{\mathrm{d}t} E^n(t) \leq (C_1 + C_2) \mathcal{E}_{\rho,u}(t, X_N) + \frac{C_2}{2} \varepsilon + (C_3 + \mathcal{E}_\varepsilon(0, X_N)) \frac{\tau}{\varepsilon}\,.
$$

Adding and subtracting $2(C_1+C_2)H^n(t)$, using the bound (4.24) and rearranging terms, this implies

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} E^n(t) \leq & 2(C_1 + C_2) E^n(t) + (\frac{C_2}{2} + 2(C_1 + C_2) C_4) \varepsilon + (C_3 + \mathcal{E}_\varepsilon(0, X_N)) \frac{\tau}{\varepsilon} \\
=: & C_5 E^n(t) + C_6 \varepsilon + (C_3 + \mathcal{E}_\varepsilon(0, X_N)) \frac{\tau}{\varepsilon}\,.
\end{aligned}
$$

Applying Grönwall inequality over the interval $[t_n, s]$ with $t_n < s < t_{n+1}$, we obtain

$$
E^n(t_{n+1}^-) \coloneqq \lim_{s \to t_{n+1}^-} E^n(s) \leq (E^n(t_n) + C_6 \varepsilon \tau + (C_3 + \mathcal{E}_\varepsilon(0, X_N)) \frac{\tau^2}{\varepsilon}) \exp(C_5 \tau)\,.
$$

In order to apply a discrete Grönwall inequality, we need to replace the left-hand side with $E^{n+1}(t_{n+1}) \coloneqq \mathcal{E}_{\rho,u}(t_{n+1}, X_N) + H^{n+1}(t_{n+1})$. This is indeed possible, since by definition of $X_N^\varepsilon(t_{n+1})$ and continuity of ρ, ρ_N and X_N^ε we have

$$
\begin{aligned}
\mathcal{F}_{\varepsilon,\rho}(t_{n+1}, X_N) + H^{n+1}(t_{n+1}) = & \frac{\|X_N(t_{n+1}) - X_N^\varepsilon(t_{n+1})\|^2}{2\varepsilon} + \mathcal{U}(\rho_N^\varepsilon(t_{n+1})) - \mathcal{U}(\rho(t_{n+1})) \\
& - \int_{\mathbb{R}^d} U'(\tilde{\rho}(t_{n+1})) \mathrm{d}(\rho_N(t_{n+1}) - \mathbf{1}_M \rho(t_{n+1})) \\
\leq & \frac{\|X_N(t_{n+1}) - X_N^\varepsilon(t_n)\|^2}{2\varepsilon} + \mathcal{U}(\rho_N^\varepsilon(t_n)) - \mathcal{U}(\rho(t_{n+1})) \\
& - \int_{\mathbb{R}^d} U'(\tilde{\rho}(t_{n+1})) \mathrm{d}(\rho_N(t_{n+1}) - \mathbf{1}_M \rho(t_{n+1})) \\
= & \mathcal{F}_{\varepsilon,\rho}(t_{n+1}^-, X_N) + H^n(t_{n+1}^-)\,.
\end{aligned} \tag{4.25}
$$

Hence we get

$$
E^{n+1}(t_{n+1}) \leq (E^n(t_n) + C_6 \varepsilon \tau + (C_3 + \mathcal{E}_\varepsilon(0, X_N)) \frac{\tau^2}{\varepsilon}) \exp(C_5 \tau)\,.
$$

Remark 4.2. *Note that the quantity*

$$
\mathcal{F}_{\varepsilon,\rho}(t, X_N) + H^n(t)\,, \quad \text{for } t \in [t_n, t_{n+1})\,, \tag{4.26}
$$

can be regarded as a different approximation of the relative internal energy of the continuous setting (3.4). Using this quantity instead of simply $\mathcal{F}_{\varepsilon,\rho}(t, X_N)$ allows us to relate the estimates across different time steps as in equation (4.25) wihtout having to deal

with the discontinuities in time of ρ_N^ε. Nonetheless, the sum in (4.26) is not positive in general, which is why we define the relative internal energy by $\mathcal{F}_{\varepsilon,\rho}(t, X_N)$ only.

Step 4: Discrete Grönwall's argument. Since $t_{N_T} = \tau N_T = T$, we obtain

$$E^{N_T}(T) \leq E^0(0)\exp(C_5 T) + (C_6\varepsilon + (C_3 + \mathcal{E}_\varepsilon(0, X_N))\frac{\tau}{\varepsilon})\frac{\exp(C_5(T+\tau)) - 1}{C_5}\,.$$

Using once again equation (4.24), this implies

$$\begin{aligned}\mathcal{E}_{\rho,u}(T, X_N) \leq &(\mathcal{E}_{\rho,u}(0, X_N) + H^0(0))\exp(C_5 T)\\ &+ (C_6\varepsilon + (C_3 + \mathcal{E}_\varepsilon(0, X_N))\frac{\tau}{\varepsilon})\frac{\exp(C_5(T+\tau)) - 1}{C_5}\\ &+ \frac{1}{2}\mathcal{E}_{\rho,u}(T, X_N) + C_4\varepsilon\,.\end{aligned}$$

Hence, we get

$$\begin{aligned}\mathcal{E}_{\rho,u}(T, X_N) \leq &2(\mathcal{E}_{\rho,u}(0, X_N) + H^0(0))\exp(C_5 T)\\ &+ 2(C_6\varepsilon + (C_3 + \mathcal{E}_\varepsilon(0, X_N))\frac{\tau}{\varepsilon})\frac{\exp(C_5(T+\tau)) - 1}{C_5} + 2C_4\varepsilon\,.\end{aligned} \tag{4.27}$$

In order to conclude the proof we need to estimate the initial energy $\mathcal{E}_\varepsilon(0, X_N)$ and the quantity $\mathcal{E}_{\rho,u}(0, X_N) + H^0(0)$. Note that, due to the initial conditions (1.12)

$$\begin{aligned}\mathcal{E}_\varepsilon(0, X_N) &= \sum_{i=1}^N \frac{1}{2}|\dot{X}_N^i(0)|^2\rho_0[P_i] + \mathcal{F}_\varepsilon(X_N(0))\\ &\leq \frac{1}{2}\|u(0)\|^2_{L^\infty(M)} + \mathcal{U}(\rho(0)) + \frac{W_2^2(\rho(0), \rho_N(0))}{2\varepsilon}\\ &\leq C_3 + \mathcal{U}(\rho(0)) + \frac{\delta_N^2}{2\varepsilon}\,,\end{aligned}$$

where δ_N is the error in the initial conditions in the Wasserstein distance, i.e.

$$\delta_N \coloneqq W_2(\rho_N(0), \rho(0))\,. \tag{4.28}$$

In order to bound the quantity $\mathcal{E}_{\rho,u}(0, X_N) + H^0(0)$, we first estimate the term

$$\begin{aligned}\mathcal{F}_\rho(0, X_N) + H^0(0) = &\frac{W_2^2(\rho_N(0), \rho_N^\varepsilon(0))}{2\varepsilon}\\ &+ \mathcal{U}(\rho_N^\varepsilon(0)) - \mathcal{U}(\rho(0)) - \int_{\mathbb{R}^d} U'(\tilde{\rho}(0))\mathrm{d}(\rho_N(0) - \mathbf{1}_M\rho(0))\,.\end{aligned}$$

By definition of $\rho_N^\varepsilon(0)$ we find

$$\frac{W_2^2(\rho_N(0), \rho_N^\varepsilon(0))}{2\varepsilon} + \mathcal{U}(\rho_N^\varepsilon(0)) - \mathcal{U}(\rho(0)) \leq \frac{W_2^2(\rho_N(0), \rho(0))}{2\varepsilon}\,.$$

Moreover,

$$\left|\int_{\mathbb{R}^d} U'(\tilde{\rho}(0))\mathrm{d}(\rho_N(0) - \mathbf{1}_M\rho(0))\right| \leq \mathrm{Lip}(U'(\rho(0)))W_1(\rho_N(0),\rho(0)) \leq \frac{C_0^2}{2}\varepsilon + \frac{W_2^2(\rho_N,\rho_N(0))}{2\varepsilon},$$

where $C_0 \coloneqq \mathrm{Lip}(U'(\rho(0)))$. Combining the two estimates and recalling (4.28) we get

$$\mathcal{F}_\rho(0,X_N) + H^0(0) \leq \frac{C_0^2}{2}\varepsilon + \frac{W_2^2(\rho_N(0),\rho(0))}{\varepsilon} = \frac{C_0^2}{2}\varepsilon + \frac{\delta_N^2}{\varepsilon}\,. \tag{4.29}$$

The remaining terms in the relative energy $\mathcal{E}_{\rho,u}(0,X_N)$ can be estimated by the fact that $\mathcal{K}(0,X_N) = 0$ (due to the initial conditions (1.12)) and the bound

$$\delta_N = W_2(\rho_N(0),\rho(0)) = \|P_{\mathbb{X}_N}\mathrm{Id} - \mathrm{Id}\| \leq \sqrt{\rho_0[M]}h_N\,, \tag{4.30}$$

which follows from definition of h_N.

We conclude by replacing the estimates above into equation (4.27) and estimating the constants using Lemma 4.1. □

We now turn to the proof of Theorem 1.2. We will focus only on the differences with the proof of Theorem 1.1. In particular the kinetic energy will not be taken into account in the definition of the energy.

Proof of Theorem 1.2. The proof follows the same line as the one of Theorem 1.1. We denote by $\tilde{\rho}$ and $\tilde{u}$ the extensions of ρ and $u \coloneqq -\nabla U'(\rho)$ constructed via Lemma 4.1. In particular, note that $\tilde{u} \neq -\nabla U'(\tilde{\rho})$ outside the domain. In the case where $|U'''_+(0)| < \infty$, we also extend U as a C^3 function on $\mathbb{R}$ as in the proof of Theorem 1.1.

Then we take as relative energy

$$\mathcal{Z}_\rho(t,X_N) \coloneqq \mathcal{F}_{\varepsilon,\rho}(t,X_N) + \frac{1}{2}\|X_N(t) - X(t)\|^2\,. \tag{4.31}$$

By equation (4.16), the time derivative of $\mathcal{Z}_{\rho,u}(t,X_N)$ satisfies

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{Z}_\rho(t,X_N) + \frac{\mathrm{d}}{\mathrm{d}t}H^n(t) = \sum_{i=4}^{8} J_i - \langle \dot{X}_N(t) + \nabla U'(\tilde{\rho}(t,X_N(t))), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle\,, \tag{4.32}$$

where the terms $H^n(t)$ and J_i are defined as in equation (4.8) and (4.17), respectively. Adding and subtracting $\tilde{u}(t,X_N(t))$ and $\nabla U(\tilde{\rho}(t,X(t))$ in the last term we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{Z}_\rho(t,X_N) + \frac{\mathrm{d}}{\mathrm{d}t}H^n(t) + 2\mathcal{K}_{\tilde{u}}(t,X_N) = \sum_{i=4}^{10} J_i\,, \tag{4.33}$$

where

$$J_9 \coloneqq \langle \tilde{u}(t,X(t)) - \tilde{u}(t,X_N(t)), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle\,,$$

$$J_{10} \coloneqq \langle \nabla U(\tilde{\rho}(t,X(t))) - \nabla U(\tilde{\rho}(t,X_N(t))), \dot{X}_N(t) - \tilde{u}(t,X_N(t))\rangle\,.$$

The estimates for the terms J_i are analogous to those in the proof of Theorem 1.1. In particular, we obtain

$$\sum_{i=4}^{6} J_i \leq \mathcal{K}_{\varepsilon,\tilde{u}}(t, X_N) + C_1 \mathcal{Z}_\rho(t, X_N)\,,$$

where now $C_1 \coloneqq 1 + A' \mathrm{Lip}_T(\tilde{u})$, and as in the previous proof $A' \coloneqq \max(2, A)$. The terms J_7 and H^n are estimated as in equations (4.22) and (4.24), respectively, with the same constants C_2 and C_4. For J_8, proceeding as in (4.23), we obtain

$$\begin{aligned} J_8 &\leq \frac{1}{\varepsilon}\int_{t_n}^{t} \left(\frac{1}{2}\|\dot{X}_N(t')\|^2 + \frac{1}{2}\|u(t, X_N^\varepsilon(t_n))\|^2\right) \mathrm{d}t' \\ &\leq \frac{1}{2\varepsilon}\Big(\frac{\|X_N(t_n) - X_N^\varepsilon(t_n)\|^2}{2\varepsilon} - \frac{\|X_N(t_{n+1}) - X_N^\varepsilon(t_n)\|^2}{2\varepsilon}\Big) + \frac{\tau}{2\varepsilon}\|u\|^2_{L^\infty([0,T]\times M)} \\ &\leq \frac{1}{2\varepsilon}\left(\mathcal{F}_\varepsilon(X_N(t_n)) - \mathcal{F}_\varepsilon(X_N(t_{n+1}))\right) + \frac{\tau}{2\varepsilon}\|u\|^2_{L^\infty([0,T]\times M)} \\ &=: \frac{\Delta^n}{2\varepsilon} + C_3\frac{\tau}{\varepsilon}\,, \end{aligned} \tag{4.34}$$

where we used the equation

$$\|\dot{X}_N(t)\|^2 = -\frac{\mathrm{d}}{\mathrm{d}t}\frac{\|X_N(t) - X_N^\varepsilon(t_n)\|^2}{2\varepsilon}$$

to pass from the first to the second line, and the inequality

$$\frac{\|X_N(t_{n+1}) - X_N^\varepsilon(t_n)\|^2}{2\varepsilon} + \mathcal{U}(\rho_N^\varepsilon(t_n)) \geq \mathcal{F}_\varepsilon(X_N(t_{n+1}))$$

to pass from the second to the third line. Finally, the last two terms are estimated as follows

$$\begin{aligned} J_9 + J_{10} &\leq \frac{1}{2}\mathcal{K}_{\varepsilon,\tilde{u}}(t, X_N(t)) + 2(\mathrm{Lip}_T(\tilde{u}) + \mathrm{Lip}_T(\nabla U'(\tilde{\rho}))\|X_N(t) - X(t)\|^2 \\ &=: \frac{1}{2}\mathcal{K}_{\varepsilon,\tilde{u}}(t, X_N(t)) + C_5\mathcal{Z}_\rho(t, X_N)\,. \end{aligned}$$

Introducing $Z^n(t) \coloneqq \mathcal{Z}_\rho(t, X_N) + H^n(t)$, and proceding as in the previous proof, we obtain

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t}Z^n(t) + \frac{1}{2}\mathcal{K}_{\varepsilon,\tilde{u}}(t, X_N) &\leq 2(C_1 + C_2 + C_5)Z^n(t) + \Big(\frac{C_2}{2} + 2(C_1 + C_2 + C_5)C_4\Big)\varepsilon \\ &\quad + \frac{\Delta^n}{2\varepsilon} + C_3\frac{\tau}{\varepsilon} \\ &=: C_6 Z^n(t) + C_7\varepsilon + C_3\frac{\tau}{\varepsilon} + \frac{\Delta^n}{2\varepsilon}\,. \end{aligned}$$

Therefore, by the same reasoning as above

$$\begin{aligned} \mathcal{Z}_\rho(T, X) + \frac{1}{2}\int_0^T \mathcal{K}_{\varepsilon,\tilde{u}}(t, X_N) \leq& 2\Big(\mathcal{Z}_{\rho,u}(0, X_N) + H^0(0) + C_7\varepsilon + C_3\frac{\tau}{\varepsilon}\Big)\exp(C_6 T) \\ &+ \frac{\tau}{\varepsilon}\left(\mathcal{F}_\varepsilon(X_N(0)) - \mathcal{F}_\varepsilon(X_N(T))\right)\exp(C_6 T) + 2C_4\varepsilon\,. \end{aligned}$$

However, note that

$$\mathcal{F}_\varepsilon(X_N(0)) - \mathcal{F}_\varepsilon(X_N(T)) \leq \mathcal{F}_\varepsilon(X_N(0)) - \min \mathcal{U} \leq \mathcal{U}(\rho(0)) - \min \mathcal{U} + \frac{W^2(\rho(0), \rho_N(0))}{2\varepsilon}.$$

We conclude the proof using (4.29) and (4.30) to bound this latter term as well as $\mathcal{Z}_\rho(0, X_N) + H^0(0)$, and using Lemma 4.1 to bound the constants in the final estimate. □

Remark 4.3. *Note that the dependency of our estimates on h_N is only due to the bound (4.30). In particular, the error estimates in Theorem 1.1 and 1.2 hold also replacing h_N with δ_N.*

Remark 4.4. *We observe that we can obtain similar convergence estimates also on the Lagrangian velocity as it can be easily verified with the following triangular inequality*

$$\begin{aligned} \|\dot{X}_N(t) - \dot{X}(t)\|_{\mathbb{X}} &\leq \|\dot{X}_N(t) - \tilde{u}(t, X_N(t))\|_{\mathbb{X}} + \|\tilde{u}(t, X_N(t)) - \tilde{u}(t, X(t))\|_{\mathbb{X}} \\ &\leq \sqrt{2\mathcal{K}_{\tilde{u}}(t, X_N)} + \mathrm{Lip}_T(\tilde{u})\|X_N(t) - X(t)\|_{\mathbb{X}}. \end{aligned} \tag{4.35}$$

Remark 4.5. *The regularity of the exact solutions required in Theorem 1.1 and 1.2 is chosen in order to apply the extension Lemma 4.1. However, examining the constants appearing in the estimates above, one can see that this is stronger than what is actually required from the extended variables themselves. For example, one can check that the proof still holds if u is of class $C^{1,1}$ on $[0,T]\times M$ with $C^{1,1}$ divergence in space, uniformly in time, and admits an extension $\tilde{u}$ with the same regularity. If M is sufficiently regular, say simply connected with a smooth boundary, such an extension can be constructed using Fefferman's extension theorem [11] as in Lemma 4.1 but applied to the potentials obtained via the Helmholtz decomposition of u.*

5. Implementation

5.1. **Computation of the Moreau-Yosida regularization.** In this section we describe how the schemes (1.16) and (1.21) can be implemented. In particular, we show that computing the gradient vector field driving the dynamics amounts to solving a semi-discrete optimal transport problem at each time step.

Definition 5.1 (Laguerre diagram)**.** The Laguerre diagram of $(x_1, \dots, x_N) \in (\mathbb{R}^d)^N$ with weights $(w_1, \dots, w_N) \in \mathbb{R}^N$ is a decomposition of M into N subsets $(L_i)_i$ defined by

$$L_i := \{x \in M \mid \forall j \in \{1, \dots, N\}, |x - x_i|^2 + w_i \leq |x - x_j|^2 + w_j\}.$$

In the following we will identify $\mathbb{X}_N$ with $(\mathbb{R}^d)^N$, i.e. we regard an element $X_N \in \mathbb{X}_N$ as the collection of the particle positions $(X_N^i)_i \in (\mathbb{R}^d)^N$. With this identification, the functional $\mathcal{F}_\varepsilon$ can be interepreted as a function on $(\mathbb{R}^d)^N$, and its gradient at a given point as a vector in $(\mathbb{R}^d)^N$. Let us introduce the set

$$\mathcal{D}_N := \{(x_1, \dots, x_N) \in (\mathbb{R}^d)^N \mid x_i = x_j \text{ for some } i \neq j\}.$$

In the following proposition we collect the results of Proposition 11 and 13 in [30] adapted to our setting. It gives the explicit expression of the regularized density and

the gradient of the regularized energy appearing in the time-continuous schemes given by (1.12) and (1.15).

Proposition 5.2. *Let $X = (x_1, \dots, x_N) \in (\mathbb{R}^d)^N \setminus \mathcal{D}_N$, and set $\rho_N = \sum_i \rho_0[P_i]\delta_{x_i}$. Then the unique minimizer ρ_N^ε of problem satisfies*

$$\rho_N^\varepsilon(x) = (2\varepsilon U')^{-1}((w_i - |x - x_i|^2) \vee U'(0)), \quad \forall\, x \in L_i\,,$$

where $(L_i)_i$ is the Laguerre diagram associated with the positions $(x_1, \dots, x_N)$ and the weights $(w_1, \dots, w_N)$, which are uniquely defined up to an additive constant by the condition $\rho_N^\varepsilon[L_i] = \rho_0[P_i]$. Moreover, $\mathcal{F}_\varepsilon$ interpreted as a function on $(\mathbb{R}^d)^N$ is continuously differentiable on $(\mathbb{R}^d)^N \setminus \mathcal{D}_N$ and

$$\nabla_{x_i}\mathcal{F}_\varepsilon(X) = \rho_0[P_i]\frac{x_i - b_i(X)}{\varepsilon}\,, \quad b_i(X) := \frac{1}{\rho_0[P_i]}\int_{L_i} x\rho_N^\varepsilon \mathrm{d}x\,.$$

Remark 5.3 (Power energies)**.** *If the energy is defined by the power function*

$$U(r) = \frac{r^m}{m-1}\,,$$

for $m > 1$, then the minimizer ρ_N^ε has the following form:

$$\rho_N^\varepsilon(x) = \left[\left(\frac{m-1}{m}\right)\frac{(w_i - |x - x_i|^2)_+}{2\varepsilon}\right]^{\frac{1}{m-1}} \quad \forall\, x \in L_i\,.$$

Actually, in order to compute the solutions of the fully-discrete scheme, we do not need the expression for the gradient in Proposition 5.2, but we just need to identify $P_{\mathbb{X}_N} X_N^\varepsilon(t_n)$ in (1.16) and (1.21). For this, assume that $(X_N^i(t_n))_i \in (\mathbb{R}^d)^N \setminus \mathcal{D}_N$ and let $(L_i)_i$ be the Laguerre diagram associated with $\rho_N^\varepsilon(t_n)$. Then, for any $Y_N \in \mathbb{X}_N$, we have

$$\int_{S_0} X_N^\varepsilon(t_n)\cdot Y_N \rho_0 \,\mathrm{d}x = \sum_i Y_N^i \cdot \int_{P_i} X_N^\varepsilon(t_n)\rho_0 \,\mathrm{d}x = \sum_i Y_N^i \cdot \int_{L_i} x\rho_N^\varepsilon(t_n)\,\mathrm{d}x\,.$$

Therefore,

$$P_{\mathbb{X}_N} X_N^\varepsilon(t_n)(\omega) = b_i(X_N(t_n)) \quad \forall\, \omega \in P_i\,,$$

where $b_i(X_N(t_n)) \in \mathbb{R}^d$ is the barycenter of $\rho_N^\varepsilon(t_n)$ restricted on L_i.

Remark 5.4 (Initialization by optimal quantization)**.** *The partition $\mathcal{P}_N$ of the support $S_0 \subseteq M$ of ρ_0 which is required to define the space $\mathbb{X}_N$ (see Section 1.2) can be itself defined as the interesection of a Laguerre diagram $(L_i)_i$ with S_0. For instance, assuming the masses to be equal, i.e. $m_i = \rho_0[M]/N$ for $i = 1, \dots, N$, one can select the vector of positions $(x_1, \dots, x_N) \in (\mathbb{R}^d)^N$ defining the diagram to belong to the* argmin *of*

$$(y_1, \dots, y_N) \in (\mathbb{R}^d)^N \mapsto W_2\left(\sum_i \frac{\rho_0[M]}{N}\delta_{y_i}, \rho_0\right),$$

so that there exists a vector of weights $(w_1, \dots, w_N) \in \mathbb{R}^N$ such that $\rho_0[L_i] = \rho_0[M]/N$. Then one can define the initial conditions $X_N(0)$ by $X_N(0)|_{L_i} = x_i$, and therefore $\rho_N(0) = \sum_i \frac{\rho_0[M]}{N}\delta_{x_i}$. With this choice $\delta_N = W_2(\rho_N(0), \rho_0) \lesssim N^{-1/d}$ (see, e.g., [20]). In view of Remark 4.3, this ensures the convergence of the schemes independently of the size of the partion h_N.

5.2. **Time integration and linear potentials.** The schemes (1.12) and (1.15) can be easily generalized to the case when the energy of the system contains an additional linear term of the form

$$\int_M V \,\mathrm{d}\rho\,,$$

where $V \in C^{1,1}(M)$ is a given function. At the discrete level, it is more convenient to treat this term independently of the Moreau-Yosida regularization, i.e. by adding to the discrete energy the term

$$\int_{\mathbb{R}^d} \tilde{V} \,\mathrm{d}\rho_N = \int_M \tilde{V} \circ X_N \,\mathrm{d}\rho_0\,, \tag{5.1}$$

where $\tilde{V}$ is a $C^{1,1}$ extension of V, e.g., constructed using Fefferman's extension theorem [11]. Then, in view of Proposition 5.2 the discrete scheme (1.12) would be replaced by

$$\ddot{X}^i_N(t) = -\frac{X^i_N(t) - b_i(X_N(t_n))}{\varepsilon} - \nabla\tilde{V}(X^i_N(t))\,, \tag{5.2}$$

for all $i \in \{1,\dots,N\}$ and $t \in [t_n, t_{n+1})$, where b_i is defined as in Proposition 5.2. Therefore for each time-step one needs to:

(1) find the optimal density $\rho^\varepsilon_N(t_n)$ and the associated barycenters $b_i(X_N(t_n))$: as in [19], this is done by applying a damped Newton's method to solve the system of optimality conditions $\rho^\varepsilon_N(t_n)[L_i] = \rho_0[P_i]$ from Proposition 5.2;
(2) solve N decoupled systems of ODEs in (5.2), which can be done explicitly for particular choices of $\tilde{V}$.

The same holds for the scheme (1.15) upon replacing $\ddot{X}^i_N(t)$ by $\dot{X}^i_N(t)$.

Finally, note that even with the additional term (5.1), the proofs of convergence above still apply without modifying the relative energies and with only minor modifications. In particular, the constant in Theorem 1.1 and 1.2 would additionally depend on $\mathrm{Lip}(\nabla\tilde{V})$.

6. Numerical tests

In this section we demonstrate numerically the behavior of the scheme in terms of convergence with mesh and time-step refinement. The tests presented hereafter correspond to the internal energy/pressure function

$$U(r) = P(r) = r^2\,, \tag{6.1}$$

for which the Euler equations (1.1) yield the shallow water equations without rotation and the gradient flow (1.4) yields the porous medium equation with a quadratic nonlinearity. Note, however, that in tests below the vector field $\nabla U'(\rho)$ is not Lipschitz (in fact, it is discontinuous at the boundary of the support of ρ), so they are outside the limits of applicability of our theorems. For all the tests the discrete initial condition are determined by optimal quantization with respect to the Wasserstein distance as in Remark 5.4.

For the computation of the Moreau-Yosida regularization, we used the open-source library `sd-ot`, which is available at https://github.com/sd-ot.

$1/\sqrt{N}$	ΔX	rate	$\Delta\mathcal{U}$	rate
1.25e-01	4.71e-02	-	1.66e-02	-
6.25e-02	2.78e-02	7.62e-01	9.39e-03	8.21e-01
3.12e-02	1.55e-02	8.44e-01	5.11e-03	8.77e-01
1.56e-02	8.24e-03	9.08e-01	2.72e-03	9.12e-01

TABLE 1. Errors and convegence rates for the Barenblatt solution of the porous medium equation, with $\varepsilon = \sqrt{\tau} = 1/\sqrt{N}$.

6.1. **Convergence: porous medium equation.** The porous medium equation (1.4) associated with the energy (6.1) admits the following exact solution

$$\rho(t,x) = \frac{1}{\sqrt{t}}\left(C^2 - \frac{1}{16\sqrt{t}}|x|^2\right)_+ \tag{6.2}$$

on any time interval $[t_0, T]$, with $t_0 > 0$. Initial conditions are given by optimal quantization of the Barenblatt profile at given time. Equation (6.2) describes the evolution of the so-called Barenblatt profile. The internal energy decays according to

$$\mathcal{U}(t) = \frac{16\pi C^6}{3\sqrt{t}},$$

whereas the Lagrangian flow is given by

$$X(t,x) = x\left(\frac{t}{t_0}\right)^{1/4}. \tag{6.3}$$

Here we take $t_0 = 1/16$, $T = 1$ and $C = 1/3$, and we monitor the following quantities:

$$\Delta X \coloneqq \|X_N(T) - X(T, X_N(0))\|_{\mathbb{X}}, \quad \Delta\mathcal{U} \coloneqq |\mathcal{U}_\varepsilon(\rho_N(T)) - \mathcal{U}(T)|. \tag{6.4}$$

Note that ΔX is an order one approximation of the L^2 distance between the discrete and continuous flows appearing in the convergence estimates. For a given number of particles N, we take $\varepsilon = \sqrt{\tau} = 1/\sqrt{N}$, which implies a rate of convergence of $1/2$ according to Theorem 1.2 (see also Remark 5.4). In Figure 1 we show the density ρ_N^ε for fixed N and at different times and the associated Laguerre diagram. Table 1 collects the errors and the associated convergence rates which confirm our estimate. Figure 2 shows the time evolution of the internal energy, which decreases monotonically in accordance with the stability estimate (1.22).

6.2. **Convergence: Euler equation.** We perform two different convergence tests for the Euler model (1.1). For the first we construct an exact solution of the equation by a time rescaling of the Barenblatt solution above, i.e. we take

$$\rho(t,x) = \frac{4}{1+2t+5t^2}\left(C^2 - \frac{1}{4(1+2t+5t^2)}|x|^2\right)_+. \tag{6.5}$$

This is an exact solution of the model associated with the Lagrangian flow

$$X(t,x) = x\sqrt{5t^2+2t+1} \tag{6.6}$$

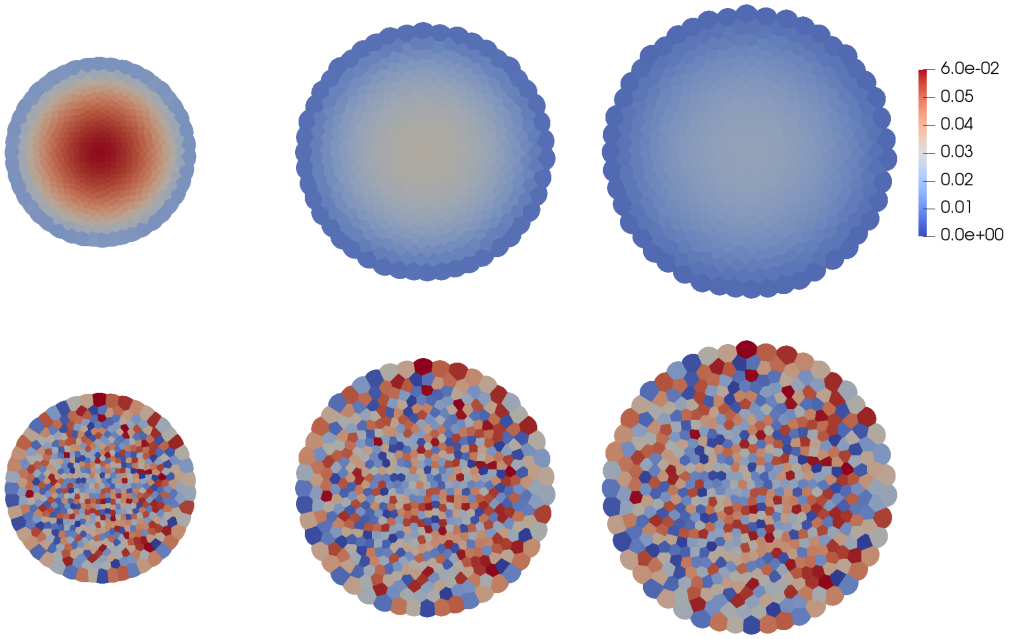


FIGURE 1. Evolution of the density ρ_N^ε for the Barenblatt solution of the porous medium equation for $N = 576$, and $\varepsilon = \sqrt{\tau} = 1/\sqrt{N}$. Upper row: weights evolution; lower row: Laguerre diagram evolution.

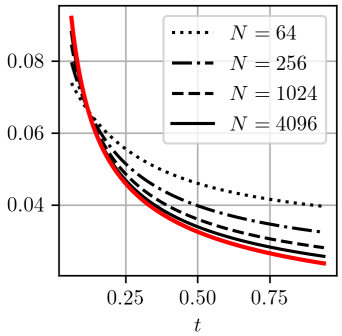


FIGURE 2. Time evolution of the discrete internal energy $\mathcal{U}_\varepsilon(\rho_N(t))$ for the Barenblatt solution of the porous medium equation (the red line corresponds to the exact energy evolution).

and the initial conditions $\dot{X}(0,x) = x$. In this case the exact kinetic and internal energy evolutions are given by

$$\mathcal{K}(t) = \frac{4\pi C^6 (10t+2)^2}{3(5t^2+2t+1)}\,, \quad \mathcal{U}(t) = \frac{64\pi C^6}{3(1+2t+5t^2)}\,. \tag{6.7}$$

$1/\sqrt{N}$	ΔX	rate	$\Delta\mathcal{E}$	rate
1.25e-01	4.36e-02	-	2.46e-02	-
6.25e-02	2.77e-02	6.53e-01	1.68e-02	5.52e-01
3.12e-02	1.61e-02	7.83e-01	1.02e-02	7.13e-01
1.56e-02	8.80e-03	8.71e-01	5.71e-03	8.44e-01

TABLE 2. Errors and convegence rates for the Barenblatt solution of the Euler equations, with $\varepsilon = \sqrt{\tau} = 1/\sqrt{N}$.

For this test, we take $t_0 = 0$, $T = 0.6$ and $C = 1/3$, and we monitor the flow error ΔX defined in equation (6.4) and the total energy error

$$\Delta\mathcal{E} \coloneqq |\mathcal{E}_\varepsilon(T, X_N) - \mathcal{E}(T)|\,, \tag{6.8}$$

where $\mathcal{E}(T) = \mathcal{K}(T) + \mathcal{U}(T)$.

For the second test, we add to the system a linear confinement potential

$$V(x) = \frac{5}{8}|x|^2\,. \tag{6.9}$$

Then, we consider the exact solutions associated with the steady density

$$\rho(x) = \left(C^2 - \frac{1}{16}|x|^2\right)_+ \tag{6.10}$$

and the rigid rotation given by the flow

$$X(t,x) = R(t)x\,, \quad R(t) = \begin{pmatrix} \cos(t) & \sin(t) \\ -\sin(t) & \cos(t) \end{pmatrix}. \tag{6.11}$$

Both the kinetic and internal energy are constant during the evolution and they are given by

$$\mathcal{K}(t) = \frac{64\pi C^6}{3}\,, \quad \mathcal{U}(t) = \frac{16\pi C^6}{3}\,. \tag{6.12}$$

For this test we take $t_0 = 0$, $T = 1$ and $C = 1/3$, and we monitor the same quantities as above.

As before, for a given number of particles N, we take $\varepsilon = \sqrt{\tau} = 1/\sqrt{N}$, which implies a rate of convergence of $1/2$ according to Theorem 1.1 (see also Remark 5.4). Tables 2 and 3 collect the errors and the associated convergence rates for the two tests and confirm our error estimate. Figures 3 and 4 show the time evolution of the total, kinetic and internal energy; note that the discrete total energy decreases monotonically in accordance with the stability estimate (1.19).

ACKNOWLEDGEMENTS

This work was supported by a public grant as part of the Investissement d'avenir project, reference ANR-11-LABX-0056-LMH, LabEx LMH, and by a grant from the French ANR (MAGA, ANR-16-CE40-0014).

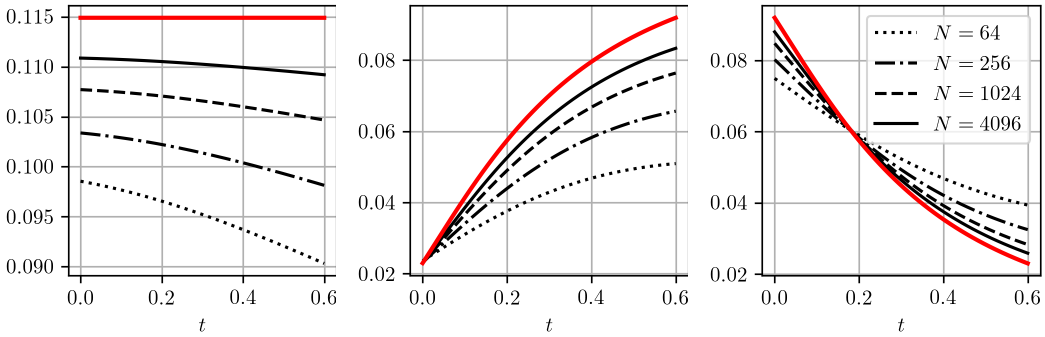


FIGURE 3. Time evolution of the discrete total energy $\mathcal{E}_\varepsilon(t, X_N)$ (left), kinetic energy (center), and internal energy $\mathcal{U}_\varepsilon(\rho_N(t))$ (right) for the Barenblatt solution of the Euler equations (the red line corresponds to the exact energy evolution).

$1/\sqrt{N}$	ΔX	rate	$\Delta\mathcal{E}$	rate
1.25e-01	7.28e-02	-	3.00e-02	-
6.25e-02	3.76e-02	9.55e-01	1.59e-02	9.18e-01
3.12e-02	1.92e-02	9.71e-01	8.16e-03	9.61e-01
1.56e-02	9.84e-03	9.61e-01	4.28e-03	9.29e-01

TABLE 3. Errors and convegence rates for the rigid rotation solution of the Equation equation, with $\varepsilon = \sqrt{\tau} = 1/\sqrt{N}$.

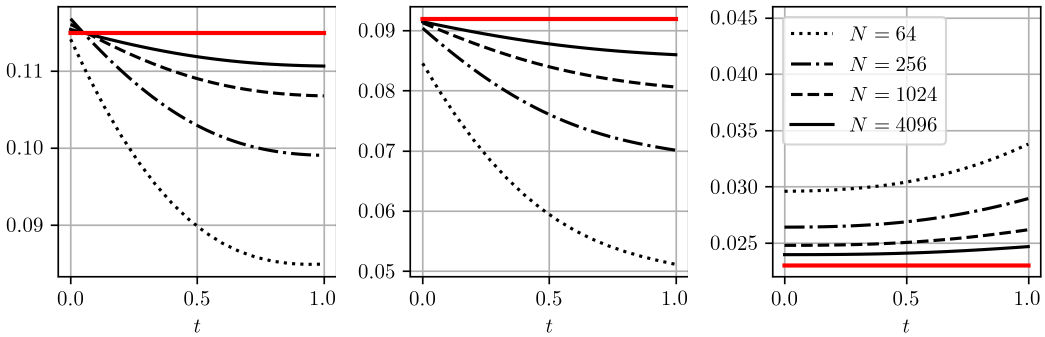


FIGURE 4. Time evolution of the discrete total energy $\mathcal{E}_\varepsilon(t, X_N)$ (left), kinetic energy (center), and internal energy $\mathcal{U}_\varepsilon(\rho_N(t))$ (right) for the rigid rotation solution of the Euler equations (the red line corresponds to the exact energy evolution).

References

[1] Vladimir Arnold. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. In *Annales de l'institut Fourier*, volume 16.

[2] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[3] Yann Brenier. Derivation of the Euler Equations from a Caricature of Coulomb Interaction. *Communications in Mathematical Physics*, 212(1):93–104, 2000.

[4] Tomas F Buttke. Velicity methods: Lagrangian numerical methods which preserve the Hamiltonian structure of incompressible fluid flow. In *Vortex flows and related numerical methods*, pages 39–57. Springer, 1993.

[5] Jose A. Carrillo, Daniel Matthes, and Marie-Therese Wolfram. Chapter 4 - Lagrangian schemes for Wasserstein gradient flows. In Andrea Bonito and Ricardo H. Nochetto, editors, *Geometric Partial Differential Equations - Part II*, volume 22 of *Handbook of Numerical Analysis*, pages 271–311. Elsevier, 2021.

[6] José Antonio Carrillo, Katy Craig, and Francesco S Patacchini. A blob method for diffusion. *Calculus of Variations and Partial Differential Equations*, 58(2):1–53, 2019.

[7] Fabio Cavalletti, Marc Sedjro, and Michael Westdickenberg. A variational time discretization for the compressible Euler equations. *Trans. Amer. Math. Soc.*, 371:5083–5155, 2019.

[8] Constantine M Dafermos, Constantine M Dafermos, Constantine M Dafermos, Grèce Mathématicien, Constantine M Dafermos, and Greece Mathematician. *Hyperbolic conservation laws in continuum physics*, volume 3. Springer, 2005.

[9] Lawrence C Evans. Partial differential equations and Monge-Kantorovich mass transfer. *Current developments in mathematics*, 1997(1):65–126, 1997.

[10] Joep HM Evers, Iason A Zisis, Bas J van der Linden, and Manh Hong Duong. From continuum mechanics to sph particle systems and back: Systematic derivation and convergence. *ZAMM-Journal of Applied Mathematics and Mechanics/Zeitschrift für Angewandte Mathematik und Mechanik*, 98(1):106–133, 2018.

[11] Charles Fefferman. Extension of $C^{m,\omega}$-smooth functions by linear operators. *Revista Matematica Iberoamericana*, 25(1):1–48, 2009.

[12] Tino Franz and Holger Wendland. Convergence of the smoothed particle hydrodynamics method for a specific barotropic fluid flow: constructive kernel theory. *SIAM Journal on Mathematical Analysis*, 50(5):4752–4784, 2018.

[13] Thomas O Gallouët and Quentin Mérigot. A Lagrangian scheme à la Brenier for the incompressible Euler equations. *Foundations of Computational Mathematics*, 18(4):835–865, 2018.

[14] Wilfrid Gangbo and Michael Westdickenberg. Optimal transport for the system of isentropic Euler equations. *Communications in Partial Differential Equations*, 34(9):1041–1073, 2009.

[15] Jan Giesselmann, Corrado Lattanzio, and Athanasios E Tzavaras. Relative energy for the Korteweg theory and related Hamiltonian flows in gas dynamics. *Archive for Rational Mechanics and Analysis*, 223(3):1427–1484, 2017.

[16] Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric numerical integration: structure-preserving algorithms for ordinary differential equations*, volume 31. Springer Science & Business Media, 2006.

[17] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

[18] Boris Khesin, Gerard Misiolek, and Klas Modin. Geometric hydrodynamics via Madelung transform. *Proceedings of the National Academy of Sciences*, 115(24):6165–6170, 2018.

[19] Jun Kitagawa, Quentin Mérigot, and Boris Thibert. Convergence of a newton algorithm for semi-discrete optimal transport. *Journal of the European Mathematical Society*, 21(9):2603–2651, 2019.

[20] Benoit Kloeckner. Approximation by finitely supported measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 18(2):343–359, 2012.

[21] Corrado Lattanzio and Athanasios E Tzavaras. Relative entropy in diffusive relaxation. *SIAM Journal on Mathematical Analysis*, 45(3):1563–1584, 2013.

[22] Corrado Lattanzio and Athanasios E Tzavaras. From gas dynamics with large friction to gradient flows describing diffusion theories. *Communications in Partial Differential Equations*, 42(2):261–290, 2017.
[23] Hugo Leclerc, Quentin Mérigot, Filippo Santambrogio, and Federico Stra. Lagrangian discretization of crowd motion and linear diffusion. *SIAM Journal on Numerical Analysis*, 58(4):2093–2118, 2020.
[24] Steven J Lind, Benedict D Rogers, and Peter K Stansby. Review of smoothed particle hydrodynamics: towards converged Lagrangian flow modelling. *Proceedings of the Royal Society A*, 476(2241):20190801, 2020.
[25] Quentin Mérigot and Jean-Marie Mirebeau. Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport. *SIAM Journal on Numerical Analysis*, 54(6):3465–3492, 2016.
[26] Joe J Monaghan. Smoothed particle hydrodynamics. *Reports on progress in physics*, 68(8):1703, 2005.
[27] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[28] Giovanni Russo. On the impulse formulation of the Euler Equations. In *Proceedings of the IX International Conference on Waves and Stability in Continuous Media, Rendiconti del Circolo Matematico di Palermo, Serie II, Suppl*, volume 57, pages 447–542, 1998.
[29] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.
[30] Clément Sarrazin. Lagrangian discretization of variational mean field games. *arXiv preprint arXiv:2010.11519*, 2020.
[31] Michael Westdickenberg and Jon Wilkening. Variational particle schemes for the porous medium equation and for the system of isentropic Euler equations. *ESAIM: Mathematical Modelling and Numerical Analysis*, 44(1):133–166, 2010.

Thomas O. Gallouët (thomas.gallouet@inria.fr), Team Mokaplan, Inria Paris 75012 Paris, CEREMADE, CNRS, UMR 7534, Université Paris-Dauphine, PSL University, 75016 Paris, France

Quentin Mérigot (quentin.merigot@universite-paris-saclay.fr), Université Paris-Saclay, CNRS, Laboratoire de mathématiques d'Orsay, 91405, Orsay, France and Institut universitaire de France (IUF)

Andrea Natale (andrea.natale@inria.fr), Inria, Univ. Lille, CNRS, UMR 8524 - Laboratoire Paul Painlevé, F-59000 Lille, France

Chapter 3

Unbalanced Optimal transport, geometry and PDE

3.1 Regularity theory and geometry of unbalanced optimal transport.

Articles:

- **Regularity theory and geometry of unbalanced optimal transport.** *Submitted 2023* Gallouët T.O., Ghezzi R. et Vialard F.X. https://hal.science/hal-03498098v1.

Collaborators: Roberta Ghezzi and Francois Xavier Vialard

Main contributions:

- We investigate the regularity of optimal transport maps for Unbalanced optimal Transport, making the link with regularity of a classical Optimal Transport problem.
- We provide an equivalent of the Brenier Polar Factorization Theorem in the UOT case.
- We explicit the link between c-convexity and a cone-distance convexity linked to UOT.
- We also explicite the link between c-convex functions and cone-distance convex functions.
- It allows us to show for instance that the MTW condition on the Cone with the cone distance implies the MTW condition for the cost associated cost c on the base space.

Research directions: A natural follow-up of this paper is to understand how to use this new polar factorization theorem in order to compute numerical approximations of some PDE, in the spirit of what we have done in Section 2.3 with Lagrangian numerical scheme. We also want to pursue this investigation onto the link between the geometry of the underlying space and the one on the cone space. Our main motivation is to provide more efficient numerical methods for some problems related to UOT.

REGULARITY THEORY AND GEOMETRY OF UNBALANCED OPTIMAL TRANSPORT

THOMAS GALLOUËT, ROBERTA GHEZZI, AND FRANÇOIS-XAVIER VIALARD

Abstract. Using the dual formulation only, we show that regularity of unbalanced optimal transport also called entropy-transport inherits from regularity of standard optimal transport. We then provide detailed examples of Riemannian manifolds and costs for which unbalanced optimal transport is regular. Among all entropy-transport formulations, Wasserstein-Fisher-Rao metric, also called Hellinger-Kantorovich, stands out since it admits a dynamic formulation, which extends the Benamou-Brenier formulation of optimal transport. After demonstrating the equivalence between dynamic and static formulations on a closed Riemannian manifold, we prove a polar factorization theorem, similar to the one due to Brenier and Mc-Cann. As a byproduct, we formulate the Monge-Ampère equation associated with Wasserstein-Fisher-Rao (WFR) metric, which also holds for more general costs. Last, we study the link between c-convex functions for the cost induced by the WFR metric and the cost on the cone; the main result is that weak Ma-Trudinger-Wang condition on the cone implies it for the cost induced by WFR.

Contents

1. Introduction

In the past few years, optimal transport has seen an impressive development mainly driven by applied fields in which real data require robust and largely applicable models. In many applications, data are modeled by probability distributions. To compare two such distributions, optimal transport (OT) provides a distance which is geometrically meaningful. Indeed, OT lifts a distance on the base space to the space of probability measures. In OT, the underlying idea consists in explaining variation of mass between measures via displacement, thereby having a global constraint of equal total mass for the two measures. The last constraint can easily be alleviated with global renormalization but

The second author is supported by project "ConDiTransPDE", Control, diffusion and transport problems in PDEs and applications, project number E83C22001720005, funded by Università degli Studi di Roma "Tor Vergata", Rome Italy.

the obtained model will not be able to account for possible local change of mass. Considering this shortcoming [20, 4], it was natural to enrich the model using local change of mass as proposed by the last author and co-authors and independently by others in [8, 9, 25, 30].

When looking for a generalization of optimal transport to unnormalized measures, there are at least two possible directions. The first one consists in extending the Kantorovich formulation of optimal transport, which is static in contrast to the Benamou-Brenier formulation. This idea amounts to relax the marginal constraints using some divergence such as the relative entropy (Kullback-Leibler). By doing so, it is not trivial to know whether the resulting functional gives a proper distance between positive densities. The second one is to start by the dynamic formulation of Benamou and Brenier [3], which is of interest since it uncovers the Riemannian-like structure of the Wasserstein metric for the L^2 cost. A natural Riemannian tensor on the space of densities which is one-homogeneous with respect to rescaling of mass is the Hessian of the entropy, known as the Fisher-Rao metric when restricted to the set of probability densities.

The latter idea was the starting point of the concurrent works [8, 9, 25, 30] that introduced what is now called *unbalanced optimal transport* and which has seen several applications in data sciences. Arguably, the most significant result on this model is the equivalence between the static formulation and the dynamic formulation [9, 30]. Importantly, the article [30] gives another characterization of unbalanced optimal transport as a standard optimal transport problem on the cone over the base manifold with second order moment constraints. This formulation was exploited in [19, 41] to reformulate the Camassa-Holm equation as a standard incompressible Euler equation on an extension of the cone. Then, generalized flows à la Brenier were studied in [18] for the Camassa-Holm equation and its higher-dimensional extension. Other interesting extensions and related works of the unbalanced framework include the projection of this distance to the set of probability measures using homogeneity property [28] and gradient flows that retain more convexity than standard Wasserstein gradient flows [27, 26]. The dynamic formulation of unbalanced optimal transport has also drawn some interest [5, 2], also for defining new metrics between metric measure spaces [34, 39]. Applications of unbalanced optimal transport are numerous [45, 36, 38, 39, 15], in particular in data science and computer vision, since this model is more robust in some sense than standard optimal transport and computationally feasible using entropic regularization [10].

An open question in this unbalanced framework is the issue of regularity. In the context of standard optimal transport, regularity appeared after Brenier stated the existence of an optimal transport map under mild conditions in Euclidean space. Since then, an "implicit" regularity of optimal transport was discovered in [7] and following works, see [13] for a recent overview. Regularity does not hold in general but it is observed when the underlying densities are regular and have convex support in Euclidean space. These results are based on Monge-Ampère equation and they have triggered a number of works concerned with extensions to Riemannian manifolds [32].

Contributions and structure of the article. In this paper, we address the question of regularity of unbalanced optimal transport. We focus on two important instances of the problem which give rise to a metric on the space of positive Radon measures, namely the Wasserstein-Fisher-Rao (or Hellinger-Kantorovich) and the Gaussian-Hellinger distances. Obviously, there is not just a single map as in standard optimal transport. However, the objects of interest are still encoded via optimal potentials, on which regularity can be studied. Alternatively, regularity can also be tackled from the primal formulation. Indeed, a plan which minimizes the primal formulation of unbalanced optimal transport is an optimal transport plan between its marginals.

From the above remarks, it is expected that regularity of the potentials is inherited from regularity theory for optimal transport. This fact is proven in Section 2 in Theorem 4 by studying the dual formulation and in particular its first-order optimality condition which encodes optimal transport between the optimal marginals of the primal formulation. Starting from the general formulation of [30], our regularity theorem requires Lipschitz regularity of the optimal potentials. Existence of Lipschitz potentials is proven in Section 2.2, under geometric conditions on the measures. Under these conditions, we obtain our results for Gaussian-Hellinger and Wasserstein-Fisher-Rao in Section 2.3. In particular, Gaussian-Hellinger is regular on the sphere and the Euclidean space, whereas

Wasserstein-Fisher-Rao is regular only on the sphere but not on the Euclidean space. We then focus in Section 3 on the Wasserstein-Fisher-Rao metric for which we show the equivalence between static and dynamic formulations on a closed Riemannian manifold. To derive our main contribution in this section, we take advantage of a geometric point of view to show a polar factorization [6, 32] theorem on a semi-direct product of groups, which is the natural extension of the diffeomorphism group to the unbalanced setting. Such a decomposition inherits the regularity results of unbalanced optimal transport. Last, we study c-convex function for the cost on the cone and the cost induced by the Wasserstein-Fisher-Rao metric. Our main result is to show that the so-called weak MTW condition on the cone implies the same condition for the cost induced by Wassertein-Fisher-Rao.

2. Regularity of unbalanced optimal transport

2.1. **From optimal transport regularity to unbalanced optimal transport regularity.** In what follows, we use the notation X, Y for two spaces that are either Euclidean spaces, bounded convex sets of Euclidean spaces, or Riemannian manifolds. In fact, results in this section apply to the more general setting of [30] but since we are interested in regularity theory, we choose to focus on the aforementioned cases.

We consider the general case of an entropy function, that replaces the relative entropy.

Definition 1. An *entropy function* $F : \mathbb{R} \to [0, +\infty]$ is a convex, lower semi-continuous, nonnegative function such that $F(1) = 0$ and $F(x) = +\infty$ if $x < 0$. Its recession constant is $F'_\infty = \lim_{r\to+\infty} \frac{F(r)}{r}$.

Proposition 1. *The Legendre-Fenchel transform of F, denoted by F^*, has a domain of definition* $\mathrm{dom}(F^*) = (-\infty, F'_\infty]$ *and it satisfies*

$$\partial F^*(\mathrm{dom}(F^*)) \subset \mathbb{R}_{\geq 0}\,. \tag{2.1}$$

Moreover, if $\partial F(0) = +\infty$, then $\partial F^*(\mathrm{dom}(F^*)) \subset \mathbb{R}_{>0}$.

Remark 1. *The hypothesis $\partial F(0) = +\infty$ is satisfied, for instance, by the choice $F(x) = x\log(x) - x + 1$, arguably the most important and most frequent entropy function used in unbalanced optimal transport. In this case, the Legendre-Fenchel transform is $F^*(x) = e^x - 1$.*

Definition 2. Let F be an entropy function and μ, ν be Radon measures on a Riemannian manifold M. The *Csiszàr divergence* associated with F is

$$D_F(\mu, \nu) = \int_M F\left(\frac{\mathrm{d}\mu(x)}{\mathrm{d}\nu(x)}\right) \mathrm{d}\nu(x) + F'_\infty \int_M \mathrm{d}\mu^\perp\,, \tag{2.2}$$

where $\mu^\perp$ is the orthogonal part of the Lebesgue decomposition of μ with respect to ν.

For $F(x) = x\log(x) - x + 1$, D_F is also known as *Kullback-Leibler divergence* or *relative entropy*, and it reads

$$\mathrm{KL}(\mu, \nu) = \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right) \mathrm{d}\nu + |\nu| - |\mu|\,. \tag{2.3}$$

Given F, the resulting divergence D_F is jointly convex and lower semi-continuous on the space of pairs of finite and positive Radon measures, see [30, Corollary 2.9]. We can now define the primal formulation of unbalanced optimal transport, which is similar to the Kantorovich formulation of optimal transport. We denote by $\mathcal{M}_+(X)$ the space of finite and positive Radon measures on X. As is standard in optimal transportation, we need a cost function, which can be unbounded in our setting.

Definition 3. A function $c : M \times M \to \mathbb{R} \cup \{+\infty\}$ is a *cost function* if it is bounded below.

Definition 4 (Kantorovich UOT)**.** Let $(\rho_0, \rho_1) \in \mathcal{M}_+(X) \times \mathcal{M}_+(Y)$ and F_0, F_1 be entropy functions. The *unbalanced optimal transport problem* is defined as

$$\mathrm{UOT}(\rho_0, \rho_1) = \inf_{\gamma \in \mathcal{M}_+(X\times Y)} D_{F_0}(\gamma_0, \rho_0) + D_{F_1}(\gamma_1, \rho_1) + \int_{X\times Y} c(x, y)\,\mathrm{d}\gamma(x, y)\,, \tag{2.4}$$

where γ_0, γ_1 are marginals of γ, and $c : X \times Y \to \mathbb{R} \cup \{+\infty\}$ is a cost function.

The distance between two Dirac masses can be computed explicitly: let $\rho_0 = r\delta_x$, $\rho_1 = s\delta_y$, in order to compute $\mathrm{UOT}(\rho_0, \rho_1)$ one has to compute the local quantity

$$\mathcal{D}((x,r),(y,s)) \coloneqq \inf_{z\in\mathbb{R}_{>0}} rF_0(z/r) + sF_1(z/s) + c(x,y)z\,. \tag{2.5}$$

This quantity can be computed explicitly for the Kullback-Leibler divergence for both F_0 and F_1 and is equal to $\mathcal{D}((x,r),(y,s)) = r + s - 2\sqrt{rs}e^{-c(x,y)/2}$; it will be useful for example to derive the Monge formulation of UOT.

The UOT problem has many equivalent reformulations, in this section, we rely on the dual formulation of (2.4) given by the Fenchel-Rockafellar theorem.

Proposition 2 (Dual UOT)**.** *The dual formulation of* (2.4) *is*

$$\mathcal{S}(z_0,z_1) := \sup_{(z_0,z_1)\in C_b(X)\times C_b(Y)} -\int_X F_0^*(-z_0(x))\,\mathrm{d}\rho_0(x) - \int_Y F_1^*(-z_1(y))\,\mathrm{d}\rho_1(y) \tag{2.6}$$

under the constraint

$$z_0(x) + z_1(y) \le c(x,y)\,. \tag{2.7}$$

For a proof in the general case, see for instance [30, Proposition 4.3].

Our goal is to show that regularity of unbalanced optimal transport follows from regularity of standard optimal transport for the cost c. This result can be expected since once the optimal marginals γ_0, γ_1 are fixed in (2.4), optimizing on the plan γ (with fixed marginals) is indeed a standard optimal transport problem between γ_0 and γ_1 for the cost c.

Lemma 3 (Linearized UOT)**.** *Assume that the entropy functions F_i are differentiable on their domain. Let $(z_0^\star, z_1^\star) \in C_b(X)\times C_b(Y)$ be a pair of optimal potentials for the dual problem* (2.6) *satisfying* $\mathrm{range}(-z_i^\star) \subset \mathrm{dom}(F_i^*)$*. Then $(z_0^\star, z_1^\star)$ is a solution of the standard optimal transport problem*

$$\sup_{(z_0,z_1)\in C_b(X)\times C_b(Y)} \int_X z_0(x)\,\mathrm{d}\tilde{\rho}_0(x) + \int_Y z_1(y)\,\mathrm{d}\tilde{\rho}_1(y) \tag{2.8}$$

under the constraint $z_0(x) + z_1(y) \le c(x,y)$ where $\tilde{\rho}_i = F_i^{*\prime}(-z_i^\star)\rho_i$ for $i = 0, 1$.

Proof. Let $(\delta z_0, \delta z_1) \in C_b(X)\times C_b(Y)$ denotes the first order admissible variations of z_0, z_1 satisfying the inequality constraint $z_0(x) + \delta z_0(x) + z_1(y) + \delta z_1(y) \le c(x,y)$. Given some potentials $(z_0, z_1) \in C_b(X)\times C_b(Y)$, one can differentiate the dual functional (2.6) to get

$$\int_X \delta z_0(x) F_0^{*\prime}(-z_0(x))\,\mathrm{d}\rho_0(x) + \int_Y \delta z_1(y) F_1^{*\prime}(-z_1(y))\,\mathrm{d}\rho_1(y)\,,$$

At $(z_0^\star, z_1^\star)$ the optimality implies for all admissible $(\delta z_0^\star, \delta z_1^\star)$

$$\int_X \delta z_0^\star(x) F_0^{*\prime}(-z_0^\star)\,\mathrm{d}\rho_0(x) + \int_Y \delta z_1^\star(y) F_1^{*\prime}(-z_1^\star(y))\,\mathrm{d}\rho_1(y) \le 0,$$

or equivalently by linearity

$$\int_X (z_0^\star + \delta z_0^\star(x))\, F_0^{*\prime}(-z_0^\star)\,\mathrm{d}\rho_0(x) + \int_Y (z_1^\star + \delta z_1^\star(y))\, F_1^{*\prime}(-z_1^\star(y))\,\mathrm{d}\rho_1(y) \le \int_X z_0^\star(x) F_0^{*\prime}(-z_0^\star)\,\mathrm{d}\rho_0(x) + \int_Y z_1^\star(y) F_1^{*\prime}(-z_1^\star(y))\,\mathrm{d}\rho_1(y),$$

for all $(\bar{z}_0, \bar{z}_1) = (z_0^\star + \delta z_0^\star, z_1^\star + \delta z_1^\star)$ satisfying $\bar{z}_0(x) + \bar{z}_1(y) \le c(x,y)$. It exactly says that $(z_0^\star, z_1^\star)$ is optimal in the constraint problem (2.8).

□

Remark 2. *An immediate consequence of this proof is that the corresponding Radon measures $\tilde{\rho}_i$ have the same total mass. Indeed, given a pair of potentials (z_0, z_1) satisfying* (2.7)*, for every $\lambda \in \mathbb{R}$ the pair $(z_0 + \lambda, z_1 - \lambda)$ still satisfies* (2.7)*. However, the linearized objective functional differs with*

the term $\lambda(|\tilde{\rho}_0| - |\tilde{\rho}_1|)$ where $|\cdot|$ denotes total mass. This term can be made arbitrarily large unless $|\tilde{\rho}_0| = |\tilde{\rho}_1|$, thus contradicting the fact that the linearization is bounded.

Remark 3. *Following Lemma 3 and Brenier's work* [6, Section 1.4]*, a potential z_0 solution of Proposition 2 can be taken as a definition of variational solutions for a* UOT*-Monge-Ampère equation given by :*

$$\det\left[-\nabla^2 z_0(x) + (\nabla^2_{xx} c)(x, \varphi(x))\right] = \left|\det\left[(\nabla_{x,y} c)(x, \varphi(x))\right]\right| \frac{F_0^{*\prime}(-z_0^\star)\rho_0(x)}{F_1^{*\prime}(-z_1^\star(y))\rho_1 \circ \varphi(x)}\,. \tag{2.9}$$

See Proposition 16 for more detailed computation in particular cases like Gaussian-Hellinger or Hellinger-Kantorovich problems.

Regularity results for such optimal potentials are therefore regularity results for these PDE. The following definition is useful to state the main result of this section.

Definition 5. Let $(\rho_0, \rho_1) \in \mathcal{M}_+(X) \times \mathcal{M}_+(Y)$ be two measures which are absolutely continuous with respect to a reference volume with densities $(\rho_0, \rho_1) \in C^{k,\alpha}(X) \times C^{k,\alpha}(Y)$ for a given non-negative integer k, $\alpha \in (0,1)$. We say that (ρ_0, ρ_1) is a *k-regular pair of measures* if, for every $0 \leq l \leq k$ and every pair $(\lambda_0, \lambda_1) \in C^{l,\alpha}(X) \times C^{l,\alpha}(Y)$ of positive functions bounded away from zero and infinity, the optimal potentials, for the classical balanced optimal transport problem between the pair $\tilde{\rho}_0 = \lambda_0\rho_0/|\lambda_0\rho_0|$ and $\tilde{\rho}_1 = \lambda_1\rho_1/|\lambda_1\rho_1|$, are of class $C^{l+2,\alpha}$.

This definition/assumption encapsulates the regularity of balanced optimal transport needed for its extension to the unbalanced setting. This condition is realized in [12, Theorem 3.3] for C^k positive densities whose support is a convex domain and which are bounded away from zero and infinity. More generally, this definition fits well with the regularity theory developed for Monge-Ampère equation. Indeed, there is often geometric assumptions on the support of the measures, for instance convexity in the Euclidean case, which are left unchanged under pointwise multiplication with a positive function.

We now state the main result of this section which says that unbalanced optimal transport inherits the regularity of standard optimal transport associated with the cost c.

Theorem 4 (Reduction to standard optimal transport)**.** *Assume that*

(1) the Fenchel-Legendre transform of the entropy functions have domain $[0, +\infty)$, are C^{k+1} on $(0, \infty)$ and $\partial F_i(0) = +\infty$, $i = 0, 1$;

(2) the pair of measures (ρ_0, ρ_1) is k-regular;

(3) the optimal potentials for unbalanced optimal transport $(z_0^\star, z_1^\star)$ are Lipschitz continuous. Then, the optimal pair $(z_0^\star, z_1^\star)$ is of class $C^{k+2,\alpha}(X) \times C^{k+2,\alpha}(Y)$.

Assumption (1) ensures that the resulting marginals are sufficiently smooth and with unchanged support, i.e., the multiplicative term $F_i^{*\prime}(-z_i^\star)$ does not vanish. Existence of Lipschitz potentials is in general a consequence of Lipschitz continuity of the cost. However, for unbounded costs, it requires more assumptions, as detailed in the next section for the Wasserstein-Fisher-Rao metric.

Assumption (2) says that a theory of regularity for a class of optimal maps in the case of classical optimal transport is available. This is true for example under conditions on the Ma-Trudinger-Wang tensor see [43, Chapter 12] for instance. Some links between the MTW tensor on the underlying space X and the MTW tensor on the cone over X is discussed in Section 4.

Proof. The proof is a straightforward bootstrap argument based on the combination of Lemma 3 and Hypothesis (2). Since the optimal potentials are Lipschitz, Lemma 3 gives that these potentials are optimal for a classical balanced optimal transport problem between a new pair of densities which inherits smoothness from the potentials and the initial densities, namely $\tilde{\rho}_i = F_i^{*\prime}(-z_i^\star)\rho_i$. Hypothesis (1) gives that $F_i^{*\prime}(-z_i^\star)$ is C^l if $z_i \in C^l$ for $l \leq k$. It implies that the regularity of $\tilde{\rho}_i$ is given by that of z_i. At the initialization step of the bootstrap, they are only Lipschitz, then applying Lemma and Hypothesis (2), the optimal potentials gain in regularity to be $C^{3,1}$. Then, in turn, we obtain that the marginals $\tilde{\rho}_i$ are $C^{\min(k,3)}$. Iterating this bootstrap argument gives the result, the optimal potentials are $C^{k+2,\alpha}$ and the optimal marginals $\tilde{\rho}_i$ are $C^{k,\alpha}$. □

2.2. **Existence of Lipschitz potentials for unbounded costs.** The choice of cost c in formulation (2.4) may vary. For instance, in usual applications outside mathematics, the Euclidean squared distance is often used. From the mathematical point of view, the case of

$$c(x,y) = -\log\left(\cos^2\left(d(x,y)\wedge\frac{\pi}{2}\right)\right) \tag{2.10}$$

stands out since it appears in the static formulation of the Wasserstein-Fisher-Rao metric. Importantly, this cost is unbounded as well as its gradients, since it blows up when $d(x,y)$ is close to $\pi/2$. In this section we prove existence of Lipschitz potential for the maximization problem in (2.6), (2.7) for unbounded costs under an admissibility assumption on the source and target measure. Such condition may be interpreted by saying that pure creation/destruction of mass is forbidden or, in other words, mass transport must be performed between the source and target measure on the whole supports.

For simplicity, we consider the case where M is either a compact Riemannian manifold or a convex and compact domain in Euclidean space. Let us recall the notion of conjugate function. Let $c : M\times M\to\mathbb{R}\cup\{+\infty\}$ be a cost function. The c-conjugate of a function $z : M\to\mathbb{R}$ is defined by

$$\hat{z}(x) = \inf_{y\in M} c(x,y) - z(y)\,.$$

We now define a class of functions that will be considered in this section as costs. In particular, such costs can be unbounded.

Definition 6. A function $c : M\times M\to\mathbb{R}\cup\{+\infty\}$ is a locally Lipschitz *cost function* if it is bounded below and if, for every $L\in\mathbb{R}$, the restriction of c on the sub-level $c^{-1}((-\infty,L])$ is Lipschitz.

Obviously, the Lipschitz constant on a sub-level may depend on the chosen L.

Definition 7 (Admissible measures)**.** A pair of Radon positive measures (ρ_1,ρ_2) is *admissible* if, denoting $K_i = \mathrm{Supp}(\rho_i)$, $K_i\neq\emptyset$ $i=0,1$, and there holds

$$\max\left(\sup_{x\in K_1}\inf_{y\in K_2} c(x,y), \sup_{y\in K_2}\inf_{x\in K_1} c(x,y)\right) < \infty\,. \tag{2.11}$$

We denote this finite number by $c_H(\rho_1,\rho_2)$.

When considering the distance as cost function, being admissible simply means that the supports of the source and target measure have finite Hausdorff distance.

Proposition 5. *Let F_0,F_1 be entropy functions that have finite value at 0. Let $(\rho_0,\rho_1)\in\mathcal{M}_+(M)^2$ be a pair of admissible measures. Then there exists an optimal pair $(z_0,z_1)\in C(M)^2$ for the maximization problem in (2.6). Moreover, z_i is locally Lipschitz on K_i, $i=0,1$ and $z_1=\hat{z}_0$.*

Let us first prove an auxiliary technical lemma.

Lemma 6. *Let (ρ_0,ρ_1) be an admissible pair of measures. Then, there exist $x_1,\dots,x_k\in M$ and $r_1,\dots,r_k>0$ such that $\rho_0(B(x_i,r_i))>0$ and for any $y\in K_1$, there exists $\bar{i}\in\{1,\dots,k\}$ such that $\sup_{x\in B(x_{\bar{i}},r_{\bar{i}})} c(x,y) < c_H(\rho_0,\rho_1)+1$.*

Proof. Recall that K_i, $i=0,1$, is the support of ρ_i. Since the pair (ρ_0,ρ_1) is admissible, for every $y\in K_1$, there exists $B(x_y,r_y)$ and $B(y,\delta_y)$ small enough such that $\sup_{x_1\in B(x_y,r_y),y_1\in B(y,\delta_y)} c(x_1,y_1) < c_H(\rho_0,\rho_1)+1$ and $\rho_0(B(x_y,r_y))>0$. As K_1 is compact, there exists a finite number of points $(y_i)_{i=1,\dots,k}$ such that $K_1\subset\cup_{i=1}^k B(x_y,r_y)$. Therefore with $x_i=x_{y_i}$ and $r_i=r_{y_i}$, for $i=1,\dots,k$, the announced result is satisfied. □

Proof of Proposition 5. Recall that $\mathcal{S}(z_0,z_1)$ denotes the functional in the maximization problem (2.6). Remark that $\mathcal{S}(0,0)=0$, hence the supremum in (2.6) is nonnegative. Moreover, taking the c-conjugate of z_0 improves the value of $\mathcal{S}$, i.e., $\mathcal{S}(z_0,\hat{z}_0)\geq\mathcal{S}(z_0,z_1)$. Iterating this alternate optimization enables to restrict the optimization set to pairs of potentials that satisfy $z_1=\hat{z}_0$ and $z_0=\hat{z}_1$ (indeed, the c-conjugate is an involution on its range). We prove that the set

$$\mathcal{E} = \{(z_0,z_1)\in C(M)^2 \mid (2.7) \text{ is satisfied, } \mathcal{S}(z_0,z_1)\geq 0 \text{ and } z_1=\hat{z}_0,\ z_0=\hat{z}_1\} \tag{2.12}$$

is equibounded and equi-Lipschitz, i.e., there exists a constant $L > 0$ such that for every pair $(z_0, z_1) \in \mathcal{E}$, $z_0|_{\mathrm{supp}(\rho_0)}$ and $z_1|_{\mathrm{supp}(\rho_1)}$ are locally L-Lipschitz.

Let us start by equiboundedness of $\mathcal{E}$. Consider $B(x_i, r_i)$ for $i = 1, \dots, k$ given by Lemma 6 for the measure ρ_0 such that

$$\inf_{y \in \mathrm{Supp}(\rho_1)} \min_{i=1,\dots,k} c(y, x_i) \leq c_H + 1 \,.$$

Since $F_0^*(x) \geq \langle x, 0\rangle - F_0(0) = -F_0(0)$, for every $i \in 1, \dots, k$, there holds

$$0 \leq \mathcal{S}(z_0, \hat{z}_0) \leq -\rho_0(B(x_i, r_i))F_0^*(-\tilde{z}) + F_0(0)\rho_0(M) + F_1(0)\rho_1(M)$$

where $\tilde{z} = \max(\sup_{x \in B(x_i, r_i)} z_0(x), 0)$. As a consequence, denoting $\delta > 0$ the minimum of $\rho_0(B(x_i, r_i))$ for $i = 1, \dots, k$, one has, since $F_0^*(-\tilde{z}) \geq 0$,

$$-F_0(0)\rho_0(M) - F_1(0)\rho_1(M) \leq -\delta F_0^*(-\tilde{z}) \,. \tag{2.13}$$

Moreover, since $F_0^*(x) \geq \langle x, 1\rangle - F_0(1) = x$, the following lower bound

$$\tilde{z} \geq \frac{-F_0(0)\rho_0(M) - F_1(0)\rho_1(M)}{\delta}$$

holds. Set $\kappa = (-F_0(0)\rho_0(M) - F_1(0)\rho_1(M))/\delta$. Denote by $\alpha \overset{\text{def.}}{=} \inf_{x \in M} \min_i c(x, x_i)$, then

$$\begin{aligned} \hat{z}_0(y) &\leq \inf_x c(x, y) - z_0(x) \\ &\leq \alpha - \kappa \,. \end{aligned}$$

where x_i is chosen such that $c(x_i, y) < c_H(\rho_0, \rho_1) + 1$. Hence $\hat{z}_0$ is bounded above. As a direct consequence, z_0 is bounded below. By symmetry of the hypothesis on ρ_0, ρ_1, we obtain that there exists A, B, depending only on ρ_0, ρ_1, F_0^*, F_1^* and $c_H(\rho_0, \rho_1)$ such that $B \leq z_0 \leq A$ and $B \leq \hat{z}_0 \leq A$, for every $(z_0, \hat{z}_0) \in \mathcal{E}$.

We now prove that there exists a uniform constant L such that for every pair $(z_0, z_1) \in \mathcal{E}$, z_i is Lipschitz continuous with constant L. Let $(z_0, z_1) \in \mathcal{E}$. By definition of $\mathcal{E}$, $z_0 = \hat{z}_1$. Since z_1 is bounded above by A, the infimum is attained at a point $y(x)$ such that $c(x, y(x)) \leq B - A$,

$$\hat{z}_0(x) = c(x, y(x)) - z_1(y(x))$$

and moreover, for every $x' \in M$,

$$\hat{z}_0(x') \leq c(x', y(x)) - z_1(y(x)) \,.$$

Subtracting the two previous formulas gives

$$\hat{z}_0(x') - \hat{z}_0(x) \leq c(x', y(x)) - c(x, y(x)) \,.$$

Let L be the Lipschitz constant of c on the sublevel $c^{-1}((-\infty, B - A])$, then

$$|\hat{z}_0(x') - \hat{z}_0(x)| \leq L d(x, x') \,.$$

Therefore $\mathcal{E}$ is not empty, equibounded and equi-Lipschitz. As a consequence, existence of an optimal pair (z_0, z_1) for (2.6) with the required properties is obtained with a standard argument based on Ascoli–Arzelà theorem for compactness and dominated convergence theorem for the convergence of the functional $\mathcal{S}$. □

As concerns uniqueness, an obvious sufficient condition is given by the following statement.

Proposition 7. *If F_0^* and F_1^* are strictly convex, the optimal pair (z_0, z_1) is unique ρ_0 and ρ_1 a.e.*

Proof. The maximization problem (2.6) is strictly convex. □

Collecting the previous results leads to existence and uniqueness of optimal Lipschitz potentials for (2.4).

Corollary 8. *Let* $F_0(x) = F_1(x) = x\log(x) - x + 1$ *and*

$$c(x,y) = \frac{1}{2}d(x,y)^2, \text{ or } c(x,y) = -\log\left(\cos^2\left(d(x,y) \wedge \delta\pi/2\right)\right) \tag{2.14}$$

for some $\delta > 0$*. Then, for every pair of admissible measures, there exists a unique pair of Lipschitz continuous optimal potentials for the dual formulation* (2.6)*.*

Note that any pair of measures is admissible for the quadratic cost.

Combining Theorem 4 and Corollary 8, regularity results for the costs in (2.14) can be inferred in different ways depending on the choice of the ambient space M. When $M = \mathbb{R}^d$, the quadratic cost supports regularity theorems for optimal transport. For the second cost in (2.14), regularity results also hold for $M = S^d$ the unit sphere of dimension d and for the sphere of radius $1/2$ (see Section 2.3). In [30], such cases are named after Gaussian-Hellinger for the quadratic case, and Hellinger-Kantorovich for the other cost. The latter is also known as Wesserstein-Fisher-Rao distance (see for instance [8, 9]).

2.3. **Two important costs for regularity of unbalanced optimal transport.** We discuss the case of two important costs in unbalanced optimal transport. The first one is the most commonly used in practical applications, the Euclidean squared cost. The second one arises naturally from the dynamic formulation which was originally proposed to introduce this model.

Gaussian-Hellinger distance: Euclidean space and spheres. Regularity in these two cases is an immediate consequence of Theorem 4 and the regularity of optimal transport, for which sufficient conditions ensuring assumption (2) in Theorem 4 are well-known. We simply detail the case of the Euclidean space, for which the following statement holds true, as a consequence of [12, Theorem 3.3].

Corollary 9. *Let* X, Y *be convex sets in* $\mathbb{R}^d$ *and let* $(\mu,\nu) \in \mathcal{M}_+(X) \times \mathcal{M}_+(Y)$ *be a pair of measures which are absolutely continuous with respect to the Lebesgue measure, with densities* (f,g) *bounded away from zero and infinity. Assume the entropy functions* F_0, F_1 *have strictly convex and differentiable Fenchel-Legendre transforms with infinite slope at* 0*.*

If $(f,g) \in C^{k,\alpha}(\overline{X}) \times C^{k,\alpha}(\overline{Y})$ *for some positive integer* k *and* $\alpha \in (0,1)$*, then, the pair of optimal potentials* (z_0, z_1) *in the dual formulation* (2.6) *for the quadratic cost* $\frac{1}{2}\|x-y\|^2$ *belongs to* $C^{k+2,\alpha}(X) \times C^{k+2,\alpha}(Y)$ *and* ∇z_0 *is a* $C^{k+1,\alpha}$*-diffeomorphism between* $\overline{X}$ *and* $\overline{Y}$*.*

Wasserstein-Fisher-Rao distance. We consider the case of a d-dimensional Riemannian manifold M having constant sectional curvature, i.e., M may be the Euclidean space, a d-sphere, or the hyperbolic space and

$$c(x,y) = -\log\left(\cos\left(d(x,y) \wedge \frac{\pi}{2}\right)^2\right). \tag{2.15}$$

Here we provide sufficient conditions to ensure assumption (2) in Theorem 4 based on the study of Ma-Trudinger-Wang tensor for the cost (2.15) on such manifolds.

Since [31], the study of the so-called Ma-Trudinger-Wang (MTW) tensor allows to provide sufficient conditions to imply regularity of potential functions in optimal transport, see [43, Chapter 12].

In particular: MTW weak condition states that MTW tensor must be nonnegative for every pair of points and every pair of c-orthogonal vectors; MTW strong condition states that MTW weak condition holds true and the tensor vanishes only at vanishing vectors. MTW tensor for costs of the type $c(x,y) = l(d(x,y))$ was analysed in [29] for even smooth functions $l : \mathbb{R} \to [0,+\infty)$ having invertible derivative. In particular, authors characterize MTW weak and strong conditions on manifolds with constant sectional curvature in terms of some computable explicit functions, see [29, Theorem 5.3].

Proposition 10. *Let* M *be a Riemannian manifold with constant sectional curvature and let* $c : M \times M \to \mathbb{R} \cup \{+\infty\}$ *be as in* (2.15)*.*

Then

(i) *MTW weak condition for c fails if M is either the Euclidean space $\mathbb{R}^d$, either the hyperbolic space $\mathbb{H}^d$ or the d-sphere of radius $R > 1$ with the induced metric;*
(ii) *MTW weak condition holds for c if M is the d-sphere of radius 1 with the induced metric;*
(iii) *MTW strong condition holds for c if M is the d-sphere of radius $R = 1/2$ and $|v| = \sqrt{g(v,v)}$ denotes the norm with respect to the metric tensor on M with the induced metric.*

Let us make simple comments on these results. Since the cone construction is curvature decreasing, one cannot expect the MTW weak condition to be satisfied when the Riemannian manifold has nonpositive curvature, such as the Euclidean space or the hyperbolic space. However, in nonnegative curvature, there is a better chance to observe regularity for the cost (2.15). We further generalize this connection in Section 4.

Proof. We start by recalling the main results in [29]. Consider a cost function $J(x,y) = l(d(x,y))$, where $l : \mathbb{R} \to [0, +\infty[\to \mathbb{R}$ is a smooth, even function such that $l''(s) > 0$. Set $h(s) = (l')^{-1}(s)$. Then the J-exponential map can be computed as

$$J\text{-}\exp_x(v) = \exp_x\left(\frac{h(|v|)}{|v|}v\right),$$

where $\exp_x$ denotes the Riemannian exponential on M and $|v| = \sqrt{g_x(v,v)}$ denotes the norm with respect to the metric tensor on M. By definition, the MTW tensor is

$$MTW_x(u,v,w) = -\frac{3}{2}\partial_s^2\partial_t^2|_{s=t=0} J(\exp_x(tu), J\text{-}\exp_x(v+sw)),$$

where $x \in M$, and u, v, w are tangent vectors at x. Define $A(s) = \frac{1}{h(s)}$, and

$$B(s) = \begin{cases} s\coth(h(s)), & \text{if } M = \mathbb{R}^d, \\ \frac{s}{h(s)}, & \text{if } M = \mathbb{H}^d, \\ s\cot(h(s)), & \text{if } M \text{ is the unit sphere.} \end{cases}$$

By [29, Proposition 5.1], whenever u and w are J-orthogonal, the MTW tensor can be simplified to

$$MTW_x(u,v,w) = -\frac{3}{2}\left(\alpha(|v|)|u_0|^2|w_0|^2 + \beta(|v|)|u_0|^2|w_1|^2 + \gamma(|v|)|u_1|^2|w_0|^2 + \delta(|v|)|u_1|^2|w_1|^2\right),$$

where $u = u_0 + u_1$, $w = w_0 + w_1$, $u_0, w_0 \in \mathrm{span}\{v\}, u_1, w_1 \in (\mathrm{span}\{v\})^\perp$ and coefficients are given by

$$\alpha(s) = \frac{s^2A''(s) + 6(A(s) - B(s)) - 4s(A'(s) - B'(s))}{s^2}, \tag{2.16}$$

$$\beta(s) = \frac{sA'(s) - 2(A(s) - B(s))}{s^2}, \tag{2.17}$$

$$\gamma(s) = B''(s), \tag{2.18}$$

$$\delta(s) = \frac{B'(s)}{s}, \tag{2.19}$$

in terms of functions A, B defined above. By Theorem 5.3 in [29], the MTW tensor satisfies MTW weak condition if and only if, for every $s \in [0, |l'(D)|]$, with D the diameter of M, four inequalities hold

$$\beta(s) \le 0,\ \gamma(s) \le 0,\ \delta(s) \le 0,\ \alpha(s) + \delta(s) \le 2\sqrt{\beta(s)\gamma(s)}. \tag{2.20}$$

Moreover, MTW strong condition holds if and only if the four inequalities are strict for every $s \in (0, |l'(D)|]$.

Note that cost c in (2.15) is of the type $l(d(x,y))$, for $l(s) = -\log(\cos^2(s))$. We compute explicitly functions A, B for the hyperbolic space and for the Euclidean space. In both cases, $\beta(0) > 0$, whence MTW weak condition fails.

When M is the d-sphere of radius $R \in (0, +\infty)$, we interpret the cost c in (2.15) as $c(x,y) = l_R(d(x,y))$ where $l_R(x,y) = -\log(\cos^2(Rs))$. Hence we set $B(s) = s\cot(h_R(s))$, with $h_R = (l_R')^{-1}$ and apply [29, Proposition 5.1] to compute the MTW tensor on the d-sphere of radius R by means

of the MTW tensor on the unit d-sphere with rescaled distance. Note that MTW conditions (weak or strong) must hold for $s \in [0, |l_R'(D)|]$, where $D = \pi$ is the diameter of the unit sphere.

Computing explicitly, $\alpha(0) = \beta(0) = \gamma(0) = \delta(0) = \frac{1}{3}\left(1 - \frac{1}{R^2}\right)$. Therefore we conclude that when $R > 1$ MTW weak condition fails. On the other hand, an explicit computation gives

$$\begin{aligned} &\text{for } R = 1, \qquad \alpha(s) = \beta(s) = \gamma(s) = \delta(s) \equiv 0, \\ &\text{for } R = \frac{1}{2}, \qquad \alpha(s) = \beta(s) = \gamma(s) = \delta(s) \equiv -1\,. \end{aligned}$$

Hence for $R = 1$ MTW weak condition holds and MTW vanishes on c-orthogonal vectors, whereas for $R = 1/2$ MTW strong condition holds. □

We end this section with remarks concerning MTW conditions on the d-sphere of radius $R \in (0,1) \setminus \{1/2\}$ with the induced metric. Using (2.16), (2.17), (2.18), (2.19), an easy computation gives

$$\begin{aligned} \alpha_R(s) &= \frac{12R^2}{s^2} - \frac{2}{s}\cot\left(\frac{1}{R}\arctan(s/(2R))\right) - \frac{8}{s^2+4R^2}\csc^2\left(\frac{1}{R}\arctan(s/(2R))\right), \\ \beta_R(s) &= \frac{2}{s^2}\left(s\cot\left(\frac{1}{R}\arctan(s/(2R))\right) - 2R^2\right), \\ \gamma_R(s) &= \frac{8}{(s^2+4R^2)^2}\csc^2\left(\frac{1}{R}\arctan(s/(2R))\right)\left(s\cot\left(\frac{1}{R}\arctan(s/(2R))\right) - 2R^2\right), \\ \delta_R(s) &= \frac{1}{s}\cot\left(\frac{1}{R}\arctan(s/(2R))\right) - \frac{2}{s^2+4R^2}\csc^2\left(\frac{1}{R}\arctan(s/(2R))\right). \end{aligned}$$

A simple computation allows to prove that for $R \in (0, 1/2)$ the functions β_R and γ_R are non positive for every $s \in (0, 2R\tan(\pi R))$. To see this, consider the auxiliary function

$$\xi(s) = s\cot\left(\frac{1}{R}\arctan(s/(2R))\right) - 2R^2.$$

Then $\beta_R(s) = \frac{2}{s^2}\xi(s)$ and $\gamma_R(s) = \frac{8\xi(s)}{(s^2+4R^2)^2}\csc^2(\frac{\arctan(s/(2R))}{R}))^2)$. We are going to show that, for every $R \in (0,1/2)$ and every $s \in (0, 2R\tan(\pi R))$, $\xi(s) < 0$. Note that for $R \in (0,1/2)$, $\frac{\arctan(s/(2R))}{R}) \in (0,\pi)$. Hence $\xi(s) < 0$ is equivalent to $s\cot(\frac{\arctan(s/(2R))}{R}) < 2R^2$ which in turn is equivalent to $\frac{\arctan(s/(2R))}{R} > \operatorname{arccot}(\frac{2R^2}{s})$. Using $\arctan x = \operatorname{arccot}(1/x)$, the last inequality is equivalent to $\operatorname{arccot}(2R/s) > R\operatorname{arccot}(\frac{2R^2}{s})$. Set $v = 2R/s$ and define $k(v) = \operatorname{arccot}(v) - R\operatorname{arccot}(Rv)$. To show that $\xi(s) < 0$ it is sufficient to prove that $k(v) > 0$ on $(0, +\infty)$. This is an easy consequence of the fact that $k(0) = \pi/2(1-T) > 0$, $\lim_{v\to+\infty} k(v) = 0$ and

$$k'(v) = \frac{R^2}{1+R^2v^2} - \frac{1}{1+v^2} < 0, v \in (0, +\infty).$$

To test the last two conditions in (**??**), let us plot the 0-level sets of the functions $\delta_R(\cdot), (\alpha_R + \delta_R - 2\sqrt{\beta_R\gamma_R})(\cdot)$ in the region $(R,s) \in (0,1)\times(0,25)$. We plot also the function $w(R) = |l_R'(\pi)| = |2R\tan(\pi R)|$. Recall that the MTW strong condition holds if the four functions $\beta_R(\cdot), \gamma_R(\cdot), \delta_R(\cdot), (\alpha_R + \delta_R - 2\sqrt{\beta_R\gamma_R})(\cdot)$ are strictly negative for every $s \in (0, |2R\tan(\pi R)|]$.

3. The Wasserstein-Fisher-Rao metric

In this section, we detail the case of the Wasserstein-Fisher-Rao (WFR) metric on a smooth compact Riemannian manifold M, which is the cornerstone of unbalanced optimal transport as introduced in [25, 8, 30]. Recall that the Wassertein-Fisher-Rao corresponds to the cost function given in 2.15 and to the Kullback-Leibler divergence for the marginal penalization (i.e., both entropy functions are given by $F(x) = x\log(x) - x + 1$. First we prove the equivalence of several definitions of this metric. In particular we introduce an equivalent of the Monge formulation of standard OT to this

unbalanced setting. Using this formulation we prove the existence of unbalanced optimal transport maps and an unbalanced version of Brenier polar factorization Theorem on the automorphism group of the cone $\mathcal{C}(M)$ see Theorem 18. A regularity theory for such maps is obtained in section 2 and it is linked to an unbalanced Monge-Ampère equation, see section 3.4.

3.1. **Equivalent formulations of WFR metric.** As in classical optimal transport, the Wasserstein-Fisher-Rao metric can be defined in many ways. Here we detail five of them, namely: Monge, Kantorovich, semi-couplings, dual and dynamical formulation. The Kantorovich formulation is the one introduced in Definition 4 and the dual formulation is given in Proposition 2. For the sake of clarity we instantiate them hereafter. The starting point of all these formulations is certainly the dynamical formulation of the WFR metric which appears as a generalization of Benamou-Brenier formula by introducing a source term in the continuity equation. This is the formulation we first present below.

In the sequel, let (M, g) be a compact Riemannian manifold, let vol denote the Riemannian volume on M and let div denote the divergence of a vector field with vol.

3.1.1. *Dynamical formulation of.* Given $\rho_0, \rho_1 \in \mathcal{M}_+(M)$ and $a, b > 0$, we start by the following optimization problem

$$\inf_{\rho,v,\alpha} \frac{1}{2}\int_0^1 \left(\int_\Omega a^2 g_x(v(x),v(x)) + b^2\alpha^2(x)\,\mathrm{d}\rho_t(x)\right)\,\mathrm{d}t$$

under the constraints of the generalized continuity equation, with time boundary conditions

$$\partial_t\rho + \operatorname{div}(\rho v) = \alpha\rho\,, \rho(0,\cdot) = \rho_0, \rho(1,\cdot) = \rho_1\,.$$

Here the control variables are α, the growth rate (also called *Malthusian parameter*) and v, a vector field, both depending on time t and position $x \in M$.

Remark 4. *For $\alpha \equiv 0$, the dynamic formulation above is the well-known Benamou-Brenier formulation of the optimal transport problem* [3].

We now give the definition, relying on convexity, which allows to account for every positive Radon measure and not only those with density with respect to the reference volume measure.

Definition 8 (Dynamical formulation of WFR metric)**.** Let $\rho_0, \rho_1 \in \mathcal{M}_+(M)$, the WFR metric is defined by

$$\mathrm{WFR}^2(\rho_0,\rho_1) = \inf_{\rho,\mathsf{m},\mu} \mathcal{J}(\rho,\mathsf{m},\mu)\,,$$

where

$$\mathcal{J}(\rho,\mathsf{m},\mu) = a^2\int_0^1\int_M \frac{g_x^{-1}(\tilde{\mathsf{m}}(t,x),\tilde{\mathsf{m}}(t,x))}{\tilde\rho(t,x)}\,\mathrm{d}\nu(t,x) + b^2\int_0^1\int_M \frac{\tilde\mu(t,x)^2}{\tilde\rho(t,x)}\,\mathrm{d}\nu(t,x) \tag{3.1}$$

over the set (ρ,m,μ) satisfying $\rho \in \mathcal{M}_+([0,1]\times M)$, $\mathsf{m} \in (\Gamma^0_M([0,1]\times M, TM))^*$ which denotes the dual of time dependent continuous vector fields on M (time dependent sections of the tangent bundle), $\mu \in \mathcal{M}([0,1]\times M)$ subject to the constraint

$$\int_{[0,1]\times M} \partial_t f\,\mathrm{d}\rho + \int_{[0,1]\times M} (\mathsf{m}(\nabla_x f) - f\,\mathrm{d}\mu) = \int_M f(1,\cdot)\,\mathrm{d}\rho_1 - \int_M f(0,\cdot)\,\mathrm{d}\rho_0 \tag{3.2}$$

satisfied for every test function $f \in C^1([0,1]\times M, \mathbb{R})$. Moreover, $\nu \in \mathcal{M}_+([0,1]\times M)$ is chosen such that ρ, m, μ are absolutely continuous with respect to ν and $\tilde\rho, \tilde{\mathsf{m}}, \tilde\mu$ denote their Radon-Nikodym derivative with respect to ν.

Note that due to the one-homogeneity of the formulas with respect to $(\tilde\rho, \tilde{\mathsf{m}}, \tilde\mu)$, the functional $\mathcal{J}$ is well-defined, i.e., it does not depend on the choice of the dominating measure ν. Moreover, the divergence is defined by duality on the space $C^1(M)$. Formula (3.1) in Definition 8 is called dynamic since the time variable is involved and only length-space structures can be defined in this way. It is of interest to show that the variational problem admits a so-called static formulation that does not involve the time variable.

3.1.2. *Semi-couplings formulation.* The semi-couplings formulation already appears in [9] and in another form in [30]. In both references, equivalence between semi-couplings and dynamical formulation is proved in the Euclidean case. We now extend those results to a Riemannian setting.

Given $\rho_0, \rho_1 \in \mathcal{M}_+(M)$, set

$$\Gamma(\rho_0,\rho_1) \stackrel{\text{def.}}{=} \left\{(\gamma_0,\gamma_1) \in \left(\mathcal{M}_+(M^2)\right)^2 : p^1_*\hat{\gamma_0} = \rho_0,\, p^2_*\hat{\gamma_1} = \rho_1\right\},$$

where p^1 and p^2 denote the projection on the first and second factors of the product M^2. Moreover, consider the cone

$$\mathscr{C}(M) = \{(x,r) \mid x \in M, r > 0\},$$

endowed with the Riemannian metric

$$h_{(x,r)} = a^2 r^2 g_x + 4b^2\, \mathrm{d}r^2,$$

where g is the Riemannian metric on M, and a, b appear in the definition of WFR metric. Finally, denote by $d_{\mathscr{C}(M)}$ the distance on $\mathscr{C}(M)$ associated with the Riemannian metric h.

Theorem 11 (Semi-couplings formulation of WFR metric)**.** *The WFR distance satisfies*

$$\mathrm{WFR}^2(\rho_0,\rho_1) = \min_{(\hat{\gamma_0},\hat{\gamma_1})\in\Gamma(\rho_0,\rho_1)} \int_{M^2} d^2_{\mathscr{C}(M)}\left((x,\sqrt{\frac{\mathrm{d}\hat{\gamma_0}}{\mathrm{d}\hat{\gamma}}}),(y,\sqrt{\frac{\mathrm{d}\hat{\gamma_1}}{\mathrm{d}\hat{\gamma}}})\right)\mathrm{d}\hat{\gamma}(x,y)\,, \tag{3.3}$$

where γ is any measure that dominates γ_0, γ_1.

The functional

$$\mathcal{S}(\hat{\gamma_0},\hat{\gamma_1}) \stackrel{\text{def.}}{=} \int_{M^2} d^2_{\mathscr{C}(M)}\left((x,\sqrt{\frac{\mathrm{d}\hat{\gamma_0}}{\mathrm{d}\hat{\gamma}}}),(y,\sqrt{\frac{\mathrm{d}\hat{\gamma_1}}{\mathrm{d}\hat{\gamma}}})\right)\mathrm{d}\hat{\gamma}(x,y)$$

is well-defined, i.e., it does not depend on the choice of the measure $\hat{\gamma}$. Indeed, the square distance function $d^2_{\mathscr{C}(M)}$ is two-homogeneous with respect to dilation of the mass variables, since $h_{(x,\lambda r)} = a^2(\lambda r)^2 g_x + 4b^2\lambda^2\,\mathrm{d}r^2$. As a consequence of Rockafellar's theorem [35, Theorem 5], $\mathcal{S}$ is convex and lower-semicontinuous on the space of Radon measures as the Legendre-Fenchel transform of a convex functional on the space of continuous functions.

Our proof of Theorem 11 is an adaptation to the Riemannian case of the one in [9, Theorem 4.3], to which we refer the reader for technical details. The same reasoning, based on a simple regularization argument which is intrinsic on Riemannian manifolds, applies under minor adaptations to the standard Wasserstein distance W_2 on Riemannian manifolds, see for instance the comments in [44, Remark 8.3]. A different proof of the equivalence between dynamical and semi-coupling formulation for the Wasserstein distance W_2 in the Riemannian setting is given in [1] which uses the Nash isometric embedding theorem.

Proof of Theorem 11. First of all, the set Γ is weak* closed, the functional S is weakly continuous and lower semicontinuous. Therefore, the fact that the minimum for $\mathcal{S}$ is attained follows by application of the direct method of calculus of variations. In the following, we denote by $\mathcal{S}^2(\rho_0,\rho_1)$ the right hand side of (3.3).

Since $d_{\mathscr{C}(M)}$ is a distance on $\mathscr{C}(M)$, one can prove that $\mathcal{S}$ is a distance on $\mathcal{M}_+(M)$ and $\mathcal{S}$ is continuous w.r.t. the weak* topology, as done in [9].

We claim that, for every pair of measures (ρ_0,ρ_1) that are finite linear combination of Dirac masses, the inequality

$$\mathcal{S}^2(\rho_0,\rho_1) \geq \mathrm{WFR}^2(\rho_0,\rho_1),$$

holds. To see this, note that for $\rho_0 = \sum_i a_i\delta_{x_i}$ and $\rho_1 = \sum_j b_j\delta_{y_j}$, for finite sets of points $\{x_i, y_j\}_{i,j} \subset M$, the minimization problem (3.3) can be reduced to a linear optimization problem in finite dimension. Indeed, the optimal semi-couplings can be proved to have support on the

product of the support of ρ_0 and ρ_1. As a consequence, the optimal semi-couplings can be written as $\gamma^k = \sum_{i,j} m^k_{i,j} \delta_{(x_i,y_j)}$ for $k = 0, 1$. Then, one has

$$\begin{aligned}\mathcal{S}^2(\rho_0,\rho_1) &= \sum_{i,j} d^2_{\mathscr{C}(M)}\left((x_i, m^0_{i,j}), (y_j, m^1_{i,j})\right)\\ &\geq \sum_{i,j} \mathrm{WFR}^2(m^0_{i,j}\delta_{x_i}, m^1_{i,j}\delta_{y_j}) \geq \mathrm{WFR}^2(\rho_0,\rho_1)\,,\end{aligned}$$

where the first inequality comes from the fact that the distance on the cone (with mass coordinates) for a geodesic $(x(t), m(t))$ is given by the evaluation of WFR on the path $m(t)\delta_{x(t)}$. The second inequality is given by subadditivity of WFR^2. Since linear By density of finite linear combination of Dirac masses and weak* continuity of both WFR and $\mathcal{S}$, the inequality $\mathcal{S}^2(\rho_0,\rho_1) \geq \mathrm{WFR}^2(\rho_0,\rho_1)$ holds on $(\mathcal{M}_+(M))^2$.

We now prove the reverse inequality which follows using the convexity of $(\rho_0,\rho_1) \mapsto \mathrm{WFR}^2(\rho_0,\rho_1)$. By subadditivity of WFR^2, one has, for every $\rho_2 \in \mathcal{M}_+(M)$

$$\mathrm{WFR}^2(\rho_0+\rho_2,\rho_1+\rho_2) \leq \mathrm{WFR}^2(\rho_0,\rho_1)\,. \tag{3.4}$$

Using the triangular inequality and the fact that the WFR metric is bounded above (up to a multiplicative constant) by the Hellinger distance, we also have, for $\varepsilon_1 > 0$

$$\mathrm{WFR}(\rho_0,\rho_1) \leq \mathrm{WFR}(\rho_0+\varepsilon_1\,\mathrm{vol},\rho_1+\varepsilon_1\,\mathrm{vol}) + 2\,\mathrm{cst}\,\sqrt{\varepsilon_1}\,. \tag{3.5}$$

Let us be more precise on the previous inequality: Consider now a path ρ, m, μ which is a solution to the continuity equation (3.2), then so is the path $\rho + \varepsilon_1\,\mathrm{vol}, \mathsf{m}, \mu$ satisfying the boundary conditions $\rho(0) = \rho_0$, $\rho(1) = \rho_1$. Note that $\varepsilon_1\,\mathrm{vol}$ is constant in time and space. In addition, it is obvious that

$$\mathcal{J}(\rho+\varepsilon_1\,\mathrm{vol},\mathsf{m},\mu) \leq \mathcal{J}(\rho,\mathsf{m},\mu)\,.$$

To prove the final result, it suffices to prove that $\mathcal{S}(\rho_0+\varepsilon_1\,\mathrm{vol},\rho_1+\varepsilon_1\,\mathrm{vol}) \leq \mathcal{J}(\rho+\varepsilon\,\mathrm{vol},\mathsf{m},\mu)+\varepsilon_0$ for any $\varepsilon_0 > 0$. This will be done via a smoothing argument which is standard in the Euclidean case using convolution but has never been adapted, to the best of our knowledge, to work on Riemannian manifolds (see [44, Remarks 8.3]).

Our goal is to prove that there exists a path of smooth quantities $(\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon)$ for which $\mathcal{J}(\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon)$ is close to $\mathcal{J}(\rho,\mathsf{m},\mu)$ and ρ_ε is strictly positive and the time endpoints of the path are close in the weak-* topology. The conclusion can then be obtained by integrating the flow defined by the vector field $(\mathsf{m}_\varepsilon/\rho_\varepsilon, \mu_\varepsilon/\rho_\varepsilon)$. It gives that $\mathcal{S}(\rho_\varepsilon(0),\rho_\varepsilon(1)) \leq \mathcal{J}(\rho_\varepsilon,\mathsf{m}_\varepsilon,\mu_\varepsilon)$ and the conclusion is similar to the Euclidean case [9, Theorem 5].

By compactness of M, it is sufficient to locally smooth the path on M by iteration of this smoothing. Therefore, we will work on a chart U around a point $x_0 \in M$. By Moser's lemma, it is possible to choose the chart such that the volume form is the Lebesgue measure.

Averaging over perturbations of identity: We construct perturbations (of compact support) of the identity which will be local translations around x_0 and which will play the role of the translations in the standard convolution formula. We consider a ball $B(x_0,r_0)$ and a function u whose support is contained in $B(x_0,r_0)$ and is constant equal to 1 on $B(x_0,r_1)$ for $0 < r_1 < r_0$. For a given vector $v \in \mathbb{R}^d$, we consider the map $\Phi_v(x) = x + u(x)v$ which is a smooth diffeomorphism. We extend Φ to the whole manifold M by defining it as identity outside of U.

Let $k : \mathbb{R}^{d+1} \to \mathbb{R}_+$ be a smooth symmetric function whose support is contained in the unit ball and such that $\int k(y)\,\mathrm{d}y = 1$ and define for $\varepsilon > 0$, $k_\varepsilon(x) = k(x/\varepsilon)/\varepsilon^{d+1}$ whose support is thus contained in the ball of radius ε. We define the mollifier $\mathsf{k}_\varepsilon \star$ acting on $f \in C([0,1]\times U, \mathbb{R})$ by

$$(\mathsf{k}_\varepsilon \star f)(s,x) = \int_{\mathbb{R}}\int_U k_\varepsilon(s,v) f(t+s, \Phi_v^{-1}(x))\,\mathrm{d}v\,\mathrm{d}s\,, \tag{3.6}$$

which is well defined for ε small enough, extending the function outside the time interval $[0,1]$ as a constant. Moreover, for ε sufficiently small, it coincides with the usual convolution on a

neighborhood of x_0. By duality, it is well defined on Radon measures and extends trivially to vector valued measures as follows:

$$(\mathsf{k}_\varepsilon \star \rho)(s,x) = \int_{\mathbb{R}}\int_U k_\varepsilon(s,v)(\Phi_v)_*(\rho(t+s))\,\mathrm{d}v\,\mathrm{d}s\,, \tag{3.7}$$

$$(\mathsf{k}_\varepsilon \star m)(s,x) = \int_{\mathbb{R}}\int_U k_\varepsilon(s,v)\,\mathrm{Ad}^*_{\Phi_v^{-1}}(\mathsf{m}(t+s))\,\mathrm{d}v\,\mathrm{d}s\,. \tag{3.8}$$

We consider the path $(\Phi_v)_*(\rho)$ which satisfies the continuity equation for the triple of measures $\Big((\Phi_v)_*(\rho), \mathrm{Ad}^*_{\Phi_v^{-1}}(\mathsf{m}), (\Phi_v)_*(\mu)\Big)$ and average over v to consider

$$(\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon) = (\mathsf{k}_\varepsilon \star \rho, \mathsf{k}_\varepsilon \star \mathsf{m}, \mathsf{k}_\varepsilon \star \mu)\ . \tag{3.9}$$

As a convex combination, this path satisfies the continuity equation and the boundary conditions are close in the weak-* topology when ε tends to 0. An important remark is that, for ε small enough, $\mathsf{k}_\varepsilon \star \mathrm{Ad}^*_{\Phi_v^{-1}}(\mathsf{m})$ reduces to the standard convolution on m in a small neighborhood of x_0 since $D\Phi_v = \mathrm{Id}$ in a neighborhood of x_0 since $u \equiv 1$ on $B(x_0, r_1)$.

Use of convexity of $\mathcal{J}$: For notation convenience, we denote by f the integrand of $\mathcal{J}$ and we make the abuse of notation to use ρ, m, μ instead of their corresponding densities w.r.t. ν a dominating measure.

Under the change of variables $y = \Phi_v^{-1}(x)$ (we use one homogeneity hereafter) leads to

$$\begin{aligned}\mathcal{J}(\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon) = \int_{[0,1]\times M} f\left(x, (\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon)\right)\,\mathrm{d}\nu(x) \leq \\ \int_{\mathbb{R}}\int_U\int_{[0,1]\times M} k_\varepsilon(s,v) f(\Phi_v(y), (\rho(t+s), D\Phi_v(t,y)\mathsf{m}(t+s), \mu(t+s)))\,\mathrm{d}\nu(t,y)\,\mathrm{d}t\,\mathrm{d}s\,\mathrm{d}v\,.\end{aligned} \tag{3.10}$$

Moreover, since the metric g on M is smooth and in particular uniformly continuous on M and since $\|D\Phi_v - \mathrm{Id}\,\| \leq \mathrm{cst}\|v\|$ for a constant that only depends on u, we thus have, for any $\varepsilon_2 > 0$, the existence of $\delta > 0$ such that if $\|v\| \leq \delta$ then,

$$|g(x)(w,w) - g(\Phi_v(x))(D\Phi_v(x)w, D\Phi_v(x)w)| \leq \varepsilon_2\, g(x)(w,w)\,, \tag{3.11}$$

for every $w \in T_xM$. Therefore, a direct estimation leads to

$$\begin{aligned}\left|\int_{\mathbb{R}\times M} k_\varepsilon(s,v) f(\Phi_v(x), (\rho(t+s), \mathsf{m}(t+s), \mu(t+s)))\,\mathrm{d}\nu(t,x) - \int_{[0,1]\times M} f(x, (\rho(t), \mathsf{m}(t), \mu(t)))\,\mathrm{d}\nu(t,x)\right| \\ \leq \varepsilon_2 \mathcal{J}(\rho, \mathsf{m}, \mu)\,,\end{aligned} \tag{3.12}$$

and as a consequence the desired result,

$$\mathcal{J}(\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon) \leq \mathcal{J}(\rho, \mathsf{m}, \mu) + \varepsilon_2 \mathcal{J}(\rho, \mathsf{m}, \mu)\,. \tag{3.13}$$

Since this averaging reduces to standard convolution in the coordinate chart U in a small neighborhood of x_0, it implies that $(\rho_\varepsilon, \mathsf{m}_\varepsilon, \mu_\varepsilon)$ is smooth in a neighborhood of x_0 and $\rho_\varepsilon \geq \varepsilon_1$ vol. By compactness of M, iterating a finite number of times this argument gives the desired path. □

Next, we prove the equivalence of these two formulations with a particular UOT problem introduced in Section 2.

3.1.3. *Kantorovich formulation and dual formulation.* As in [9] the application of Fenchel-Rockafellar duality Theorem gives the dual formulation of WFR. This is summarized in the following proposition.

Proposition 12 (Dual formulation of WFR)**.** *On (M,g), it holds*

$$\mathrm{WFR}^2(\rho_0,\rho_1) = \sup_{(\phi,\psi)\in C(M)^2} \int_M \phi(x)\,\mathrm{d}\rho_0(x) + \int_M \psi(y)\,\mathrm{d}\rho_1(y) \tag{3.14}$$

subject to $\forall (x,y) \in M^2$,

$$\begin{cases} \phi(x) \leq 1\,, \quad \psi(y) \leq 1\,, \\ (1-\phi(x))(1-\psi(y)) \geq \cos^2\left(d(x,y) \wedge (\pi/2)\right)\,. \end{cases} \tag{3.15}$$

A reformulation of this linear optimization problem is

$$\mathrm{WFR}^2(\rho_0,\rho_1) = \sup_{(z_0,z_1)\in C(M)^2} \int_M 1 - e^{-z_0(x)}\,\mathrm{d}\rho_0(x) + \int_M 1 - e^{-z_1(y)}\,\mathrm{d}\rho_1(y) \tag{3.16}$$

subject to $\forall (x,y) \in M^2$,

$$z_0(x) + z_1(y) \leq -\log\left(\cos^2\left(d(x,y)\wedge(\pi/2)\right)\right)\,. \tag{3.17}$$

Interestingly this last formulation is exactly the dual formulation of UOT defined in Proposition 2 with the cost $c(x,y) = -\log\left(\cos^2\left(d(x,y)\wedge(\pi/2)\right)\right)$ and dual entropy functions $F_0^*(x) = F_1^*(x) = F^*(x) = e^x - 1$. As noticed in Remark 1 the associated entropy function is therefore $F(x) = x\log(x) - x + 1$ leading to the Kullback-Leibler divergence, which reads

$$\mathrm{KL}(\mu,\nu) = \int \frac{\mathrm{d}\mu}{\mathrm{d}\nu} \log\left(\frac{\mathrm{d}\mu}{\mathrm{d}\nu}\right)\,\mathrm{d}\nu + |\nu| - |\mu|\,. \tag{3.18}$$

Existence of Lipschitz solutions to the dual problem has been proved under admissibility condition on the measures in Section 2.2. Without these assumptions, existence of potentials can be proved in a less regular space of functions in [30, Section 6.2].

Proposition 13 (Kantorovich formulation of WFR)**.** *With the same notations as above it holds*

$$\mathrm{WFR}^2(\rho_0,\rho_1) = \inf_{\gamma\in\mathcal{M}_+(M^2)} \mathrm{KL}(p^1_*\gamma,\rho_0) + \mathrm{KL}(p^2_*\gamma,\rho_1) - \int_{M^2} \log(\cos^2(d(x,y)\wedge(\pi/2)))\,\mathrm{d}\gamma(x,y)\,. \tag{3.19}$$

3.2. **A Monge formulation.** OT supports an interesting geometric framework. Indeed, the push-forward action of the diffeomorphisms group on the space of densities is a (formal) Riemannian submersion to the space of densities endowed with the Wasserstein metric, see [21, 14] for more details. This structure also exists in the case of UOT, as already explained in [19]. We briefly recall it hereafter.

3.2.1. *The formal Riemannian submersion and Monge formulation of* WFR. Recall that a Riemannian submersion is a submersion π between two Riemannian manifolds M and N, such that $\mathrm{d}\pi$ is an isometry between the orthogonal of its kernel and its range. An important property of Riemannian submersion is that every geodesic on N can be lifted (called horizontal lift) to a unique geodesic on M (having the same length), up to the choice of a basepoint in M. In the following, the roles of M and N are taken by $\mathrm{Diff}(M)$, the group of diffeomorphisms of M and $\mathrm{Dens}_p(M)$ the space of probability densities on M. We choose the reference volume form ρ_0 on M and define

$$\begin{aligned} &\pi_0 : \mathrm{Diff}(M) \to \mathrm{Dens}_p(M) \\ &\pi(\varphi) = \varphi_*\rho_0 \end{aligned}$$

which is a (formal) Riemannian submersion of the metric $L^2(M,\rho_0)$ on $\mathrm{Diff}(M)$ to the Wasserstein W_2 metric on $\mathrm{Dens}_p(M)$. Using the horizontal lift property of geodesics mentioned above, the Benamou and Brenier dynamic formulation [3] can be rewritten on the group $\mathrm{Diff}(M)$ as the Monge problem,

$$W_2(\rho_0,\rho_1)^2 = \inf_{\varphi\in\mathrm{Diff}(M)} \left\{ \int_\Omega d_M^2(\varphi(x),x)\,\rho_0(x)\,\mathrm{dvol}(x) \;:\; \varphi_*\rho_0 = \rho_1 \right\}\,. \tag{3.20}$$

In the unbalanced case, the group $\mathrm{Diff}(M)$ is replaced with the semidirect product of groups between $\mathrm{Diff}(M)$ and the space of positive functions on M which is a group under pointwise multiplication. It is not a direct product but a semidirect one, where the composition law is defined such

that the map π given by

$$\pi_1 : (\mathrm{Diff}(M) \ltimes C(M, \mathbb{R}_{>0})) \times \mathrm{Dens}(M) \mapsto \mathrm{Dens}(M)$$
$$\pi_1\left((\varphi, \lambda), \rho\right) \stackrel{\text{def.}}{=} \varphi_*(\lambda\rho)$$

is a left-action of the group $\mathrm{Diff}(M) \ltimes C(M, \mathbb{R}_{>0})$ on the space of densities. Similarly to the optimal transport case, this action is actually a Riemannian submersion between $L^2(M, M \times \mathbb{R}_{>0})$ and $\mathrm{Dens}(M)$ endowed with the WFR metric. Note that the L^2 metric is defined by a density (the initial density) on M and a metric on $M \times \mathbb{R}_{>0}$ (see [16] for more details) and this Riemannian metric is completely specified by the unbalanced optimal transport model, namely

$$g_{(x,m)}(\mathrm{d}x, \mathrm{d}m) = a^2 m\, \mathrm{d}x^2 + b^2 \frac{\mathrm{d}m^2}{m}\cdot \tag{3.21}$$

Up to the change of variable $m = r^2$, we find that the metric can be rewritten as

$$g_{(x,r)}(\mathrm{d}x, \mathrm{d}r) = a^2 r^2\, \mathrm{d}x^2 + 4b^2\, \mathrm{d}r^2\,, \tag{3.22}$$

which is called a cone metric[1]. Since it is a classical formulation of this metric, we adopt this change of variable in the rest of the paper. In particular, the action is changed into

$$\pi : (\mathrm{Diff}(M) \ltimes C(M, \mathbb{R}_{>0})) \times \mathrm{Dens}(M) \mapsto \mathrm{Dens}(M)$$
$$\pi\left((\varphi, \lambda), \rho\right) \stackrel{\text{def.}}{=} \varphi_*(\lambda^2\rho)\,,$$

and the metric on $M \times \mathbb{R}_{>0}$ is the cone metric (3.22). We now adopt the notation $\mathscr{C}(M)$ for the $M \times \mathbb{R}_{>0}$ equipped with the cone metric. In fact, as done in [19] we can identify this semidirect product of groups with the automorphism group of the cone $\mathscr{C}(M)$ (since it has a multiplicative group structure in the $\mathbb{R}_{>0}$ component). Thus, to shorten the notations, we use $\mathrm{Aut}(\mathscr{C}(M))$ instead of $\mathrm{Diff}(M) \ltimes C(M, \mathbb{R}_{>0})$. We now state the (formal) Riemannian submersion result obtained in [19].

Proposition 14. *Let $\rho_0 \in \mathrm{Dens}(M)$ be a positive density and π be the map*

$$\pi : \mathrm{Aut}(\mathscr{C}(M)) \mapsto \mathrm{Dens}(M)$$
$$\pi(\varphi, \lambda) = \varphi_*(\lambda^2\rho_0)\,.$$

Then, π is a Riemannian submersion between $\mathrm{Aut}(\mathscr{C}(M))$ endowed with the metric $L^2(M, \rho_0, \mathscr{C}(M))$ and $\mathrm{Dens}(M)$ with the WFR *metric.*

For details about the proof, we refer the reader to [19]. This proposition can be used to deduce a static or Monge formulation of the variational problem.

Definition 9. Let $(\rho_0, \rho_1) \in \mathcal{M}_+(M^2)$. The Monge formulation of WFR is given by

$$\begin{aligned} \text{M-WFR}^2(\rho_0, \rho_1) &= \inf_{(\varphi,\lambda)} \left\{ \int_M d^2_{\mathscr{C}(M)}\left[(x,1), (\varphi(x), \lambda(x))\right] d\rho_0(x) \,:\, \varphi_*(\lambda^2\rho_0) = \rho_1 \right\}, \\ &= \inf_{(\varphi,\lambda)} \left\{ d^2_{\mathrm{Aut}(\mathscr{C}(M))}\left[(\mathrm{Id}, 1), (\varphi, \lambda)\right] \,:\, \varphi_*(\lambda^2\rho_0) = \rho_1 \right\} \end{aligned} \tag{3.23}$$

where the infimum is taken over $(\varphi, \lambda) \in \mathrm{Diff}(M) \ltimes C(M, \mathbb{R}_{>0})$ and $(\mathrm{Id}, 1)$ denotes the identity in $\mathrm{Aut}(\mathscr{C}(M))$.

This Monge formulation extends to more general divergences and costs. Indeed, one can formulate

$$\text{M-UOT}^2(\rho_0, \rho_1) = \inf_{(\varphi,\lambda)} \left\{ \int_M \mathcal{D}_{\mathscr{C}(M)}\left[(x,1), (\varphi(x), \lambda(x))\right]^2 d\rho_0(x) \,:\, \varphi_*(\lambda^2\rho_0) = \rho_1 \right\},$$

where

$$\mathcal{D}_{\mathscr{C}(M)}((x,r),(y,s))^2 = \inf_{z \in \mathbb{R}_{>0}} r^2 F_0(z/r^2) + s^2 F_1(z/s^2) + c(x,y)z\,. \tag{3.24}$$

[1]It is interesting to check that other Riemannian metrics on the cone can be chosen provided they are two-homogeneous in the radial variable. Some of the results of this article carry over such cases.

Importantly, the quantity[2] $\mathcal{D}_{\mathscr{C}(M)}$ is not necessarily a power of a distance on the cone but it is the case in the three following situations. When $F_0 = F_1$ is the relative entropy, two cases are known, $c(x,y) = -\log(\cos(\min(d(x,y), \frac{\pi}{2}))^2)$ for which $\mathcal{D}_{\mathscr{C}(M)}((x,r),(y,s))$ is almost the distance on the cone but not exactly[3] since $\mathcal{D}_{\mathscr{C}(M)}((x,r),(y,s))^2 = r^2 + s^2 - 2rs\cos(\min(d(x,y), \frac{\pi}{2}))$. The equality between the two seemingly different Monge formulations actually holds.

The *Gaussian-Hellinger* case is recovered for $c(x,y) = d^2(x,y)$ which gives $\mathcal{D}_{\mathscr{C}(M)}((x,r),(y,s))^2 = r^2 + s^2 - 2rse^{-d(x,y)^2/2}$. The last known case is for *partial optimal transport* where the divergences are taken as the total variation of measures given by the entropy function $F(x) = |x-1|$ and the $c = d^q$. Then, for $q \geq 1$, $\mathcal{D}_{\mathscr{C}(M)}((x,r),(y,s))^q = r + s - (\min(r,s))\min(0, 2 - d(x,y)^q)$ gives a distance.

A consequence of the semi-couplings formulation is the relaxation inequality $\text{M-WFR}^2(\rho_0,\rho_1) \geq \text{WFR}^2(\rho_0,\rho_1)$: for any ϕ consider $\gamma(x,y) = (x,\phi(x))\rho_0$, $\gamma_0(x,y) = \gamma(x,y)$ and $\gamma_1(x,y) = \lambda^2(x)\gamma(x,y)$. The converse inequality does not hold in general since in the case of unbalanced transport not only the particles can split but also they can reach the apex of the cone. However under our admissibility condition on (ρ_0,ρ_1) we prove that $\text{M-WFR}^2(\rho_0,\rho_1) = \text{WFR}^2(\rho_0,\rho_1)$ in Proposition 17.

3.3. **Kantorovich relaxation: the conic formulation.** This yet another but important formulation was introduced in [30] and can be interpreted as a natural Kantorovich relaxation of the Monge formulation. Indeed, instead of making the map φ stochastic, one makes both the map and the rescaling stochastics. From a cost on the cone defined by minimization in (3.24), one defines the *conic formulation*

$$\text{C-OT}(\rho_0,\rho_1) = \inf_{\gamma\in\Gamma_{\mathcal{C}}} \int_{\mathcal{C}(M)\times\mathcal{C}(M)} \mathcal{D}_{\mathscr{C}(M)}((x,r),(y,s))^2 d\gamma((x,r),(y,s))\,, \tag{3.25}$$

where $\Gamma_{\mathcal{C}}$ denotes the set of positive Radon measures γ on the product of cones such that

$$\begin{cases} \rho_0(x) = \int_{\mathbb{R}} r^2 [p^1_*\gamma](x,dr)\,, \\ \rho_1(y) = \int_{\mathbb{R}} s^2 [p^1_*\gamma](y,ds)\,. \end{cases} \tag{3.26}$$

These constraints are moment constraints instead of marginal constraints in standard OT. Moreover, this formulation does not require the plan to be a probability measure on the product space although it can also be restricted to the set of probability measure by action with dilations, see [30, 18]. In fact, formula (3.24) is 2-homogeneous so that the mass can always be rescaled pointwisely. Last, the moment constraint is the natural relaxation of the action by pushforward and rescaling $\varphi_*(\lambda^2\rho_0)$. Note that from the numerical point of view, introducing this additional radial variable is costly, yet it is amenable to entropic regularization, see [37]. The proof of equivalence with the formulations introduced above can be found in [30]. For our purpose and to prepare the discussion of c-convex functions in Section 4, we simply note that the dual solutions of this problem are also dual solutions of an OT problem; the optimal potentials take the form $r^2p(x)$ and $s^2q(y)$ for functions p, q defined on M. These potentials are necessarily 2-homogeneous functions in the radial variable.

3.4. **Monge solution and polar factorization on the automorphism group.** The geometric structure used to show Brenier's polar factorization theorem [6] in standard optimal transport relies on the Riemannian submersion and solution of Monge problem. Thanks to results given in Section 3.2.1 and after finding a solution to the Monge problem M-WFR we generalise in this section polar factorization to the unbalanced framework.

[2]Note that with respect to the first section we made the slight change of variable with the square root to remain consistent with the definition of the group action.

[3]For the cone distance, the minimum is taken with π rather than $\pi/2$, this difference is explained by the fact that at the level of the measures, the transformation can occur simultaneously for both Dirac masses.

3.4.1. *Monge solution of* WFR. To show the existence of a solution to Monge problem (3.23) we start by solving $\mathrm{WFR}(\rho_0,\rho_1)$, in the dual form (3.16), (3.17) and we provide geometric properties of such solution (see Proposition 16). To prove Proposition 16 there are two different arguments: one is based on results in Section 2 and the existence of Lipschitz potentials; the other one mimics the standard case of optimal transport with minor adaptions due to the cost. This latter approach leads to a pair of approximately differentiable potentials. For completeness we give both proofs.

Lemma 15 (sub-differentiability)**.** *Let $y\in M$, the function g defined on M by $g(x)=\cos^2\left(d(x,y)\right)$ is sub-differentiable.*

Proof. The function $d^2(\cdot,y)$ is super-differentiable see [32, Proposition 6] for instance. Therefore $d^2_{\pi/2}(\cdot,y)=(d(x,y)\wedge(\pi/2))$ is also super-differentiable and the function g is sub-differentiable as the combination of a decreasing C^1 function and the super-differentiable function $d^2_{\pi/2}(\cdot,y)$, see [32, Lemma 5]. □

Proposition 16 (Brenier's variational solution of WFR-Monge-Ampère)**.** *Let $(\rho_0,\,\rho_1)\in\mathcal{M}_+(M^2)$ and let (z_0,z_1) be the generalized optimal potentials for $\mathrm{WFR}^2(\rho_0,\rho_1)$. Suppose that (ρ_0,ρ_1) is admissible and $\rho_0\ll\mathrm{vol}$, then z_0 is ρ_0 a.e. unique and approximate differentiable on $Supp(\rho_0)$. The optimal plan γ in the formulation (13) is unique, with marginals $\gamma_0=e^{-z_0}\rho_0$, $\gamma_1=e^{-z_1}\rho_1$ and concentrated on the graph of*

$$x\mapsto\varphi(x)=\exp_x^M\left(-\arctan\left(\frac{\|\tilde\nabla z_0(x)\|}{2}\right)\frac{\tilde\nabla z_0(x)}{\|\tilde\nabla z_0(x)\|}\right)=\text{c-exp}(-\nabla z_0(x))\,, \tag{3.27}$$

that is $\varphi_*\gamma_0=\gamma_1$ and $\gamma=(\mathrm{Id}\times\varphi)_*\gamma_0$. Finally

$$\mathrm{WFR}^2(\rho_0,\rho_1)=\int_M 1-e^{-z_0(x)}\,\mathrm{d}\rho_0(x)+\int_M 1-e^{-z_1(y)}\,\mathrm{d}\rho_1(y)\,. \tag{3.28}$$

Note that (z_0,z_1) may not be continuous as needed in (3.16) but (3.28) still holds true. The approximate differentiable proof of this proposition (being more technical) is given in Appendix A, we prefer to discuss the corresponding formulation of the Monge-Ampère equation hereafter and a simple sketch of proof following the results in Section 2.

Direct proof. Corollary 8 gives a pair of Lipschitz potentials (z_0,z_1) solution of $\mathrm{WFR}^2(\rho_0,\rho_1)$. Lemma 3 proves that this pair is also solution of a classical Optimal Transport problem between $\gamma_0=e^{-z_0}\rho_0$, $\gamma_1=e^{-z_1}\rho_1$ for the cost $c(x,y)=-\log\left(\cos^2\left(d(x,y)\wedge(\pi/2)\right)\right)$. The hypothesis on ρ_0 and Classical optimal transport theory arguments gives the existence of a map $\varphi(x)=\text{c-exp}(-\nabla z_0(x))$ solution of this OT problem. In particular $\varphi_*\gamma_0=\gamma_1$. □

Remark 5. *Note that the map $\varphi(x)=\text{c-exp}(-\nabla z_0(x))$ is a solution to a standard OT problem from $\gamma_0=e^{-z_0}\rho_0$ to $\gamma_1=e^{-z_1}\rho_1$ for the cost $c(x,y)=-\log\left(\cos^2\left(d(x,y)\wedge(\pi/2)\right)\right)$. Therefore, OT regularity theory applies to z_0 with fixed marginals γ_0,γ_1. In particular, higher regularity of z_0 increases regularity of γ_0 and γ_1 and, in turn, a bootstrap argument improves regularity of z_0 (see also the strategy in the proof of Theorem 4).*

As a consequence of the underlying classical OT structure, the potential found in Proposition 16, denoted by z, is a solution of a Monge-Ampère equation with a right-hand side that also depends on the potential. We recall how to derive the equation supposing that z is C^2. Remember that $c(x,y)=-\log(\cos^2(d_{\pi/2}(x,y)))$ and $\varphi(x)=\exp_x^M\left(-\arctan\left(\frac12\|\nabla z(x)\|\right)\frac{\nabla z(x)}{\|\nabla z(x)\|}\right)$, therefore

$$2\sqrt2\tan(d_{\pi/2}(x,\varphi(x)))\frac{\sqrt2}{2d_{\pi/2}(x,\varphi(x))}\nabla\left(\frac12 d^2_{\pi/2}(x,\varphi(x))\right)=(\nabla_x c)(x,\varphi(x))$$

and the sub-differentiable equality (A.4) reads

$$\nabla z(x)-(\nabla_x c)(x,\varphi(x))=0\,. \tag{3.29}$$

Observe that by definition of $\text{c-exp}_x(v) = [(-\nabla_x c)(x,\cdot)]^{-1}(v)$ (3.29) is exactly

$$\varphi(x) = \text{c-exp}(-\nabla z(x)).$$

Differentiating (3.29) and taking the determinant yields

$$\det\left[-\nabla^2 z(x) + (\nabla^2_{xx} c)(x,\varphi(x))\right] = \left|\det\left[(\nabla_{x,y} c)(x,\varphi(x))\right]\right| \left|\det(\nabla\varphi)\right| . \tag{3.30}$$

Notice that the c-concavity property of z implies that $-\nabla^2 z + (\nabla^2_{xx} c)(x,\varphi(x))$ is a nonnegative symmetric matrix. To obtain the equation on z, observe that $\varphi_*\left((1+\frac{1}{4}\|\nabla z\|^2)e^{-2z}\rho_0\right) = \rho_1$ (see the proof of Proposition 17 below for details) or equivalently

$$|\det(\nabla\varphi)| = e^{-2z}\left(1+\frac{1}{4}\|\nabla z\|^2\right)\frac{f}{g\circ\varphi},$$

for smooth z and smooth positive measures ρ_0 and ρ_1 with densities f and g with respect to the volume measure vol. Together with (3.30), we obtain the WFR-Monge-Ampère equation defined by

$$\det\left[-\nabla^2 z(x) + (\nabla^2_{xx} c)(x,\varphi(x))\right] = \left|\det\left[(\nabla_{x,y} c)(x,\varphi(x))\right]\right| e^{-2z(x)}\left(1+\frac{1}{4}\|\nabla z(x)\|^2\right)\frac{f(x)}{g\circ\varphi(x)}, \tag{3.31}$$

where φ is given by (3.32) and satisfies the second boundary value problem: φ maps the support of ρ_0 towards the support of ρ_1.

Remark 6. *Another possibility is to write directly the Monge-Ampère equation satisfied by φ as an optimal map pushing γ_0 to γ_1 that is*

$$\det\left[-\nabla^2 z(x) + (\nabla^2_{xx} c)(x,\varphi(x))\right] = \left|\det\left[(\nabla_{x,y} c)(x,\varphi(x))\right]\right| \frac{e^{-z_0(x)}\rho_0(x)}{e^{-z_1(\varphi(x))}\rho_1\circ\varphi(x)}.$$

Using $z_0(x)+z_1(\varphi(x)) = c(x,\varphi(x))$ and $1+\frac{1}{4}\|\nabla z_0(x)\|^2 = e^{c(x,\varphi(x))}$ one recovers the WFR*-Monge-Ampère equation (3.31).*

Remark 7. *Following Brenier* [6, Section 1.4] *Proposition 16 can be taken as a definition of variational solutions for the* WFR*-Monge-Ampère equation (3.31) with second boundary value problem. The question of regularity of such a solution of a* WFR*-Monge-Ampère equation is a consequence of the results proved in Section 2. In particular as we saw it depends on the regularity of classical OT and therefore on the study of the Ma-Trudinger-Wang tensor associated to c see* [11], [42, Section 12].

Thanks to Proposition 16 we are now able to prove the existence, under some assumptions on the initial density, of a solution to the Monge problem M-WFR.

Proposition 17 (Solution to the Monge problem M-WFR and equivalence to WFR)**.** *Let ρ_0,ρ_1 be admissible and such that ρ_0 has density w.r.t. the volume measure on M. Then, there exists a ρ_0 a.e. unique c-convex function on M, z, approximatively differentiable ρ_0-a.e., such that the associated unbalanced transport couple (φ,λ) defined by*

$$\varphi(x) = \exp^M_x\left(-\arctan\left(\frac{1}{2}\|\tilde\nabla z(x)\|\right)\frac{\tilde\nabla z(x)}{\|\tilde\nabla z(x)\|}\right) \tag{3.32}$$

and

$$\lambda(x) = e^{-z(x)}\sqrt{1+\frac{1}{4}\|\tilde\nabla z(x)\|^2} \tag{3.33}$$

is a solution of the Monge problem (3.23) and satisfies

$$\pi[(\varphi,\lambda),\rho_0] = \varphi_*\left(\lambda^2\rho_0\right) = \varphi_*\left((1+\frac{1}{4}\|\tilde\nabla z\|^2)e^{-2z}\rho_0\right) = \rho_1 . \tag{3.34}$$

Moreover, (φ,λ) is the unique ρ_0 a.e. unbalanced transport couple associated to a c-concave potential, also unique, such that $\pi[(\varphi,\lambda),\rho_0] = \rho_1$. The potential z is characterized by

$$\text{M-WFR}^2(\rho_0,\rho_1) = \text{WFR}^2(\rho_0,\rho_1) = \int_M 1-e^{-z(x)}\,\mathrm{d}\rho_0(x) + \int_M 1-e^{-z^c(y)}\,\mathrm{d}\rho_1(y), \tag{3.35}$$

Proof. Existence: Let (z_0, z_1) be the optimal potentials for $\mathrm{WFR}^2(\rho_0, \rho_1)$. From Proposition 16, we know that $x \mapsto \varphi(x) = \exp_x^M \left(-\arctan\left(\frac{\|\tilde\nabla z_0(x)\|}{2}\right) \frac{\tilde\nabla z_0(x)}{\|\tilde\nabla z_0(x)\|}\right)$ is well defined ρ_0 a.e. and $\varphi_*(\gamma_0) = \gamma_1$ where $\gamma_i = \sigma_i \rho_i = e^{-z_i}\rho_i$, $i = 0, 1$. Therefore

$$\begin{aligned}
\rho_1 &= \sigma_1^{-1}\gamma_1 = \sigma_1^{-1}\varphi_*(\gamma_0) = \sigma_1^{-1}\varphi_*\left(\sigma_0\rho_0\right)\\
&= \varphi_*\left(e^{-z_0}\sigma_1^{-1}\circ\varphi\rho_0\right) = \varphi_*\left(e^{-z_0}e^{z_1\circ\varphi}\rho_0\right) = \varphi_*\left(e^{-z_0}e^{c(\cdot,\varphi(\cdot))}e^{-z_0}\rho_0\right)\\
&= \varphi_*\left(e^{-2z_0}\left(1+\frac14\|\tilde\nabla z_0\|^2\right)\rho_0\right) = \varphi_*\left(\left(e^{-z_0}\sqrt{1+\frac14\|\tilde\nabla z_0\|^2}\right)^2\rho_0\right)\\
&= \pi\left[\left(\varphi, e^{-z_0}\sqrt{1+\frac14\|\tilde\nabla z_0\|^2}\right), \rho_0\right].
\end{aligned}$$

We used that ρ_0 a.e. $z_0(x) + z_1(\varphi(x)) = c(x, \varphi(x))$, $1 + \tan^2(x) = 1/\cos^2(x)$ and thus $1 + \frac14\|\tilde\nabla z_0(x)\|^2 = e^{c(x,\varphi(x))}$. Equation (3.28) is exactly (3.35).

To prove uniqueness, consider z to be a c-concave function, such that (φ, λ) are well defined through (3.32), (3.33) and $\pi[(\varphi, \lambda), \rho_0] = \rho_1$. Then, we claim that $\gamma = [\mathrm{Id} \times \varphi]_*(e^{-z}\rho_0)$ is an optimal plan for $\mathrm{WFR}^2(\rho_0, \rho_1)$ in (13). Indeed, let us check that γ satisfies the optimality conditions of [30, Theorem 6.3(b)].

- γ is concentrated on the set of equality for a pair (z, z^c) of c-concave functions. By definition of φ, it holds ρ_0 a.e. and therefore $\gamma_0 = e^{-z}\rho_0$ a.e.

$$z(x) + z^c(\varphi(x)) = c(x, \varphi(x))\,. \tag{3.36}$$

  Thus, (z, z^c) satisfies for all $(x, y) \in M \times M$, $z(x) + z^c(y) \le c(x, y)$ with equality γ a.e.
- The marginals are absolutely continuous with respect to ρ_0 and ρ_1. It holds true for $\gamma_0 = e^{-z}\rho_0$. Note then that ρ_0 a.e.

$$\lambda^2(x) = e^{-2z(x)}(1 + \frac14\|\tilde\nabla z(x)\|^2) = e^{-z(x)}e^{z^c(\varphi(x))}\,.$$

  It yields

$$\rho_1 = \varphi_*(\lambda^2\rho_0) = \varphi_*(e^{z^c(\varphi(x))}e^{-z(x)}\rho_0) = e^{z^c}\varphi_*(\gamma_0) = e^{z^c}\gamma_1\,,$$

  thus $\gamma_1 = e^{-z^c}\rho_1$ and γ is optimal for $\mathrm{WFR}^2(\rho_0, \rho_1)$.

In particular it implies $\text{M-WFR}^2(\rho_0, \rho_1) = \mathrm{WFR}^2(\rho_0, \rho_1)$. The computation (A.5) yields (3.35) and uniqueness of the generalized optimal potentials for $\mathrm{WFR}^2(\rho_0, \rho_1)$ in Proposition (16) implies uniqueness of (z, φ, λ).

□

3.4.2. *Polar factorization.* We are left with proving a polar factorization theorem for the automorphism group of the cone $\mathrm{Aut}(\mathscr{C}(M))$.

Definition 10. The *generalized automorphism semigroup* of $\mathscr{C}(M)$ is the set of measurable maps (denoted by $\mathcal{M}es$ below) (φ, λ) from M to $\mathscr{C}(M)$

$$\overline{\mathrm{Aut}}(\mathscr{C}(M)) = \{(\varphi, \lambda) \in \mathcal{M}es(M, M) \ltimes \mathcal{M}es(M, \mathbb{R}_{>0})\}\,, \tag{3.37}$$

endowed with the semigroup law

$$(\varphi_1, \lambda_1)\cdot(\varphi_2, \lambda_2) = (\varphi_1\circ\varphi_2, (\lambda_1\circ\varphi_2)\lambda_2)\,.$$

We also consider the stabilizer of the volume measure in the automorphisms of $\mathscr{C}(M)$. It is defined by

$$\overline{\mathrm{Aut}}_{\mathrm{vol}}(\mathscr{C}(M)) = \left\{(s, \lambda) \in \overline{\mathrm{Aut}}(\mathscr{C}(M)) \,:\, \pi\left((s, \lambda), \mathrm{vol}\right) = \mathrm{vol}\right\}\,. \tag{3.38}$$

By abuse of notation, any $(s,\lambda) \in \overline{\mathrm{Aut}}_{\mathrm{vol}}(\mathscr{C}(M))$ will be denoted $\left(s, \sqrt{\mathrm{Jac}(s)}\right)$ meaning that for every continuous function $f \in C(M,\mathbb{R})$

$$\int_M f(s(x))\sqrt{\mathrm{Jac}(s)}^2 \,\mathrm{d}\,\mathrm{vol}(x) = \int_M f(x)\,\mathrm{d}\,\mathrm{vol}(x)\,. \tag{3.39}$$

Theorem 18 (Polar factorization)**.** *Let $(\phi,\lambda) \in \overline{\mathrm{Aut}}(\mathscr{C}(M))$ be an element of the generalized automorphism group of the half-densities bundle such that $\rho_1 = \pi_0\left[(\phi,\lambda),\mathrm{vol}\right]$ is an absolute continuous admissible measure. Then, there exists a unique minimizer, characterized by a c-concave function z_0, to the Monge formulation (3.23) between* vol *and ρ_1 and there exists a unique measure preserving generalized automorphism $(s,\sqrt{\mathrm{Jac}(s)}) \in \overline{\mathrm{Aut}}_{\mathrm{vol}}(\mathscr{C}(M))$ such that* vol *a.e.*

$$(\phi,\lambda) = \exp^{\mathscr{C}(M)}\left(-\frac{1}{2}\tilde{\nabla} p_{z_0}, -p_{z_0}\right) \circ (s,\sqrt{\mathrm{Jac}(s)}) \tag{3.40}$$

or equivalently

$$(\phi,\lambda) = \left(\varphi, e^{-z_0}\sqrt{1+\|\tilde{\nabla} z_0\|^2}\right)\cdot (s,\sqrt{\mathrm{Jac}(s)})\,, \tag{3.41}$$

where $p_{z_0} = e^{z_0} - 1$ and

$$\varphi(x) = \exp_x^M\left(-\arctan\left(\frac{1}{2}\|\tilde{\nabla} z_0(x)\|\right)\frac{\tilde{\nabla} z_0(x)}{\|\tilde{\nabla} z_0(x)\|}\right)\,. \tag{3.42}$$

Moreover $(s,\sqrt{\mathrm{Jac}(s)})$ is the unique $L^2(M,\mathscr{C}(M))$ projection of (ϕ,λ) onto $\overline{\mathrm{Aut}}_{\mathrm{vol}}(\mathscr{C}(M))$.

Proof of Theorem 18. We denote $\rho_0 = \mathrm{vol}$ and $\rho_1 = \pi_0\left[(\phi,\lambda),\rho_0\right]$. Let (z_0,z_1) be a solution of $\mathrm{WFR}^2(\rho_0,\rho_1)$ and γ be an optimal unbalanced transport plan. By symmetry, (z_1,z_0) is a solution of $\mathrm{WFR}^2(\rho_1,\rho_0)$ and γ^t is an optimal unbalanced transport plan. Let finally (φ_0,λ_0) and (φ_1,λ_1) be the two transport couples given by application of Proposition 16 to (ρ_0,ρ_1) and (ρ_1,ρ_0). We divide the proof into four small steps. We also denote $\mathrm{dom}(f)$ the domain of definition of the function f.

Step 1: φ_0 and φ_1 are inverse maps. On $U = \varphi_0^{-1}(\mathrm{dom}\tilde{\nabla} z_1) \cap \mathrm{dom}(\tilde{\nabla} z_0)$ which has full γ_0 and therefore ρ_0 measure (we use here the admissible condition to say that γ_0 and ρ_0 have the same support), we have

$$z_0(x) + z_1(\varphi_0(x)) = c(x,\varphi_0(x))$$

and thus $\varphi_1(\varphi_0(x)) = x$. Similarly, it holds $\varphi_0(\varphi_1(y)) = y$ on $V = \varphi_1^{-1}(\mathrm{dom}\,\tilde{\nabla} z_0)\cap\mathrm{dom}(\tilde{\nabla} z_1)$ which has full ρ_1 measure.

Step 2: (φ_0,λ_0) and (φ_1,λ_1) are inverse in $\overline{\mathrm{Aut}}$. From Step 1, ρ_1 a.e. it holds $\varphi_0(\varphi_1(y)) = y$. Thus, ρ_1 a.e.

$$(\varphi_0,\lambda_0)\cdot(\varphi_1,\lambda_1) = (\varphi_0\circ\varphi_1, \lambda_0\circ\varphi_1\lambda_1) = (\mathrm{Id},(\lambda_0\circ\varphi_1)\lambda_1)\,.$$

Moreover by (3.34) of Proposition 17 applied twice

$$\pi\left[(\varphi_0,\lambda_0)\cdot(\varphi_1,\lambda_1),\rho_1\right] = \pi\left[(\varphi_0,\lambda_0),\pi\left[(\varphi_1,\lambda_1),\rho_1\right]\right] = \pi\left[(\varphi_0,\lambda_0),\rho_0\right] = \rho_1\,.$$

It implies that

$$\pi\left[(\mathrm{Id},(\lambda_0\circ\varphi_1)\lambda_1),\rho_1\right] = \pi\left[(\varphi_0,\lambda_0)\cdot(\varphi_1,\lambda_1),\rho_1\right] = \rho_1\,.$$

In other words, we have ρ_1 a.e. $(\lambda_0\circ\varphi_1)\lambda_1 = 1$ and ρ_1 a.e.

$$(\varphi_0,\lambda_0)\cdot(\varphi_1,\lambda_1) = (\mathrm{Id},1)\,.$$

Step 3: polar factorization. Let $(s,\lambda_s) = (\varphi_1,\lambda_1)\cdot(\phi,\lambda) = (\varphi_1\circ\phi, \lambda_1\circ\phi\lambda)$. By construction, one has

$$\pi\left[(s,\lambda_s),\rho_0\right] = \pi\left[(\varphi_1,\lambda_1)\cdot(\phi,\lambda),\rho_0\right] = \pi\left[(\varphi_1,\lambda_1),\pi\left[(\phi,\lambda),\rho_0\right]\right] = \pi\left[(\varphi_1,\lambda_1),\rho_1\right] = \rho_0\,.$$

Therefore, (s,λ_s) belongs to $\overline{\mathrm{Aut}}_{\mathrm{vol}}$ and $\lambda_s = \sqrt{\mathrm{Jac}(s)}$ holds in the weak sense (3.39). Thus

$$(\phi,\lambda) = (\mathrm{Id},1)\cdot(\phi,\lambda) = (\varphi_0,\lambda_0)\cdot(\varphi_1,\lambda_1)\cdot(\phi,\lambda) = (\varphi_0,\lambda_0)\cdot(s,\sqrt{\mathrm{Jac}(s)})\,.$$

It proves the polar factorization.

Step 4: Uniqueness. The pair of c-concave potentials (z_0, z_1) is optimal for $\mathrm{WFR}(\rho_0, [(\varphi_0, \lambda_0), \rho_0]) = \mathrm{WFR}(\rho_0, \rho_1)$ and therefore by Proposition 17, z_i are unique ρ_i a.e.. We deduce that the projection $(s, \sqrt{\mathrm{Jac}(s)}) = (\varphi_1, \lambda_1) \cdot (\phi, \lambda)$ is also unique ρ_0 a.e.. Indeed the positivity of λ implies that $\mathrm{Supp}(\lambda^2 \rho_0) = \mathrm{Supp}(\rho_0)$, thus ϕ maps $\mathrm{Supp}(\rho_0)$ onto $\mathrm{Supp}(\rho_1)$ and the uniqueness of φ_1 and λ_1, ρ_1 a.e., implies the uniqueness of s and $\sqrt{\mathrm{Jac}(s)}$, ρ_0 a.e.. To prove that $(s, \sqrt{\mathrm{Jac}(s)})$ is the $L^2(M, \mathscr{C}(M))$ projection of (ϕ, λ) onto $\overline{\mathrm{Aut}}_{\mathrm{vol}}(\mathscr{C}(M))$, we observe

$$\begin{aligned}
\inf_{(\sigma, \sqrt{\mathrm{Jac}(\sigma)}) \in \overline{\mathrm{Aut}}_{\mathrm{vol}}(\mathscr{C}(M))} \int_M d^2_{\mathscr{C}(M)} \left((\phi, \lambda), (\sigma, \sqrt{\mathrm{Jac}(\sigma)}) \right) \rho_0 &\geq \mathrm{WFR}^2(\rho_0, \rho_1) \\
&= \int_M d^2_{\mathscr{C}(M)} \big((\varphi_0, \lambda_0), (\mathrm{Id}, 1) \big) \rho_0 \\
&= \int_M d^2_{\mathscr{C}(M)} \left((\varphi_0, \lambda_0) \cdot (s, \sqrt{\mathrm{Jac}(s)}), (s, \sqrt{\mathrm{Jac}(s)}) \right) \rho_0 \\
&= \int_M d^2_{\mathscr{C}(M)} \left((\phi, \lambda), (s, \sqrt{\mathrm{Jac}(s)}) \right) \rho_0 \,,
\end{aligned}$$

which gives the result. □

As in OT, Theorem 18 could be extended, for example, to any admissible ρ_1 without the absolute continuity assumption. In such a case, one looses uniqueness of the measure preserving generalized automorphism $(s, \sqrt{\mathrm{Jac}(s)})$. An other extension is to project on the subset of $\overline{\mathrm{Aut}}(\mathscr{C}(M))$:

$$\overline{\mathrm{Aut}}_{\rho_0, \mu_0}(\mathscr{C}(M)) = \left\{ (s, \lambda) \in \overline{\mathrm{Aut}}(\mathscr{C}(M)) \,\middle|\, \pi\left((s, \lambda), \rho_0\right) = \mu_0 \right\} ,$$

in the spirit of [44, Theorem 3.15]. The proof is similar to the one given above. Last, linearization of this polar factorization leads to an Helmholtz decomposition for velocity vector fields. As explain previously this last three results are not limited to the case of WFR. A similar analysis for the Gaussian-Hellinger case is even easier to compute. For instance for Gaussian-Hellinger in $\mathbb{R}^d$ the optimal potential z would be semi-concave, thus φ a gradient of a convex function:

$$\varphi(x) = x - \nabla z(x), \tag{3.43}$$

and

$$\lambda(x) = e^{-z(x) + \frac{1}{4}\|\nabla z(x)\|^2}. \tag{3.44}$$

This formulation can be particularly adapted for statistical or numerical applications. We leave these for future works.

4. The Ma-Trudinger-Wang tensor in the WFR case. Some relations between c-convex functions and $\mathrm{d}_{\mathscr{C}}$-convex functions

In this section we investigate the link between c-convex functions on the base space M and $\mathrm{d}^2_{\mathscr{C}(M)}$-convex functions on $\mathscr{C}(M)$. As a consequence, we provide a relation between the MTW-tensor on M for the cost c and the MTW-tensor on $\mathscr{C}(M)$ for the cost $\mathrm{d}^2_{\mathscr{C}(M)}$. Since for instance the connexity of the c-subdifferential is a synthetic formulation of $MTW_c(0)$. For simplicity, we denote by $\mathrm{d}_{\mathscr{C}}$ the distance on $\mathscr{C}(M)$.

We prove two fundamental facts. Lemma 19 states that a function is c-convex on M if and only if its (suitably defined) lift is $\mathrm{d}^2_{\mathscr{C}}$-convex on $\mathscr{C}(M)$. Lemma 20 is concerned with explicit computations along c-segments.

Let us recall the definition of cost-convex functions.

Definition 11. [43, Definition 5.2] Let $X \times Y \subset M \times M$ be a subset and c be a cost function on $X \times Y$. A function $f : X \to \mathbb{R} \cup \{+\infty\}$ is c-convex if it is not identically $+\infty$ and if there exists a function $g : Y \to \mathbb{R} \cup \{\pm\infty\}$ such that, for every $x \in X$,

$$f(x) = \sup_{y \in Y} g(y) - c(x, y) \,.$$

The c-subdifferential of f at point $\bar{x}$, denoted by $\partial_c f(\bar{x})$, is the set of $y \in Y$ such that, for every $x \in X$,

$$f(x) \geq f(\bar{x}) + c(\bar{x}, y) - c(x, y).$$

In the sequel we set $\cos_+(x,y) := \cos\left(\min(d(x,y), \frac{\pi}{2})\right)$ and we consider the cost $c(x,y) = -\log(\cos_+^2(x,y))$. The corresponding distance on the cone is given by

$$\mathrm{d}^2_{\mathscr{C}}((x,r),(y,s)) = r^2 + t^2 - 2rt\cos_+(x,y).$$

Definition 12. Given a function $f : M \to \mathbb{R}$ we define the *lift* of f to $\mathscr{C}(M)$ as the function $F_f : \mathcal{C}(M) \mapsto \mathbb{R}$ as

$$F_f(x,r) = r^2(e^{f(x)} - 1).$$

This definition is natural with formulation 3.16 of WFR in mind seen as a dual formulation on the cone.

Lemma 19. *Let $X \times Y \subset M \times M$ and $f : M \to \mathbb{R}$. Then f is c-convex on $X \times Y$ if and only if F_f is $\mathrm{d}^2_{\mathscr{C}}$-convex on $(X \times \mathbb{R}_+) \times (Y \times \mathbb{R}_+)$. In particular, given $(\bar{x}, \bar{r}) \in X \times \mathbb{R}_+$, $y \in \partial_c f(\bar{x})$ if and only if $(y,s) \in \partial_{\mathrm{d}^2_{\mathscr{C}}} F_f(\bar{x}, \bar{r})$ where $s = \bar{r}\frac{e^{f(\bar{x})}}{\cos_+(\bar{x},y)}$. Finally $(F_f)^{\mathrm{d}^2_{\mathscr{C}}} = F_{f^c}$.*

Proof. By Definition 11, it is sufficient to prove the second statement.

The function f is c-convex on $X \times Y$, if and only if for every $\bar{x} \in X$ the c-subdifferential of f at $\bar{x}$ is not empty. In particular, for every $\bar{x} \in X$ there exists $y \in Y$ such that, for every $x \in X$,

$$\begin{aligned} f(x) &\geq f(\bar{x}) + c(\bar{x}, y) - c(x, y) \\ &= f(\bar{x}) - \log(\cos_+^2(\bar{x}, y)) + \log(\cos_+^2(x, y)), \end{aligned}$$

or, equivalently, for every $x \in X$,

$$e^{f(x) - f(\bar{x})} \frac{\cos_+^2(\bar{x}, y)}{\cos_+^2(x, y)} \geq 1. \tag{4.1}$$

Let now $\bar{r} \in \mathbb{R}_+$. Then $(y,s) \in \partial_{\mathrm{d}^2_{\mathscr{C}}} F_f(\bar{x}, \bar{r})$ if and only if, for every $(x,r) \in X \times \mathbb{R}_+$, the following inequality holds true

$$r^2(e^{f(x)} - 1) \geq \bar{r}^2(e^{f(\bar{x})} - 1) + \mathrm{d}^2_{\mathscr{C}}\left((\bar{x}, \bar{r}), (y, s)\right)) - \mathrm{d}^2_{\mathscr{C}}\left((x, r), (y, s)\right)). \tag{4.2}$$

Using the definition of $\mathrm{d}_{\mathscr{C}}$, (4.2) is equivalent to

$$r^2 e^{f(x)} \geq \bar{r}^2 e^{f(\bar{x})} - 2s\bar{r}\cos_+(d(\bar{x}, y)) + 2sr\cos_+(d(x, y)). \tag{4.3}$$

Adding $s^2\cos_+^2(x,y)e^{-f(x)} + s^2\cos_+^2(\bar{x},y)e^{-f(\bar{x})}$ to both sides of (4.3), the inequality becomes

$$\begin{aligned} &e^{f(x)}\left(r - s\cos_+(x,y)e^{-f(x)}\right)^2 - e^{f(\bar{x})}\left(\bar{r} - s\cos_+(\bar{x},y)e^{-f(\bar{x})}\right)^2 \\ &+ s^2\cos_+^2(x,y)e^{-f(x)}\left(\frac{\cos_+^2(\bar{x},y)}{\cos_+^2(x,y)}e^{f(x)-f(\bar{x})} - 1\right) \geq 0 \end{aligned} \tag{4.4}$$

For F_f to be $\mathrm{d}^2_{\mathscr{C}}$-convex, (4.4) must be satisfied for every $(x,r) \in X \times \mathbb{R}_+$. When this is the case, evaluating (4.4) at $x = \bar{x}$ implies that, for every $r \in \mathbb{R}_+$,

$$\left(r - s\cos_+(\bar{x},y)e^{-f(\bar{x})}\right)^2 - \left(\bar{r} - s\cos_+(\bar{x},y)e^{-f(\bar{x})}\right)^2 \geq 0. \tag{4.5}$$

For a given $\bar{r} \in \mathbb{R}_+$ (4.5) holds for every $r \in \mathbb{R}_+$ if and only if

$$s = \bar{r}\frac{e^{f(\bar{x})}}{\cos_+(\bar{x},y)}.$$

Thus, the (unique) value of s has been identified and we now evaluate (4.4) at this value. Inequality (4.4) holds for every $(x,r) \in X \times \mathbb{R}_+$ if and only if

$$e^{f(x)}\left(r - s\cos_+(x,y)e^{-f(x)}\right)^2 + s^2\cos_+^2(x,y)e^{-f(x)}\left(\frac{\cos_+^2(\bar{x},y)}{\cos_+^2(x,y)}e^{f(x)-f(\bar{x})} - 1\right) \geq 0. \tag{4.6}$$

If (4.6) holds true for every $(x,r) \in X \times \mathbb{R}_+$, then evaluating at $r = s\cos_+(x,y)e^{-f(x)}$ we infer that

$$\frac{\cos_+^2(\bar{x},y)}{\cos_+^2(x,y)} e^{f(x)-f(\bar{x})} - 1 \geq 0$$

must be satisfied for every $x \in X$, that is to say (4.1), i.e., $y \in \partial_c f(\bar{x})$.

The other direction is obvious since in Formula (4.6) the first term is a square and the second term is nonnegative due to (4.1). The proof of $(F_f)^{\mathrm{d}^2_{\mathscr{C}}} = F_{f^c}$ is done similarly or can be seen as a consequence of the identification of the subdiffentials. □

The next lemma makes a link between the notions of c-segment on M and $\mathrm{d}^2_{\mathscr{C}}$-segment on $\mathscr{C}(M)$. Let us recall the definition of cost-segments on a manifold.

Definition 13. [43, Definition 12.10] Let $c : M \times M \to \mathbb{R}$ be a cost, $\bar{x} \in M$, and consider the parameterized segment between $q_0, q_1 \in T_{\bar{x}}M$ given by $[0,1] \ni t \mapsto q_t = (1-t)q_0 + tq_1$. The c-segment, whenever it is defined, is given by the parameterized curve

$$[0,1] \ni t \mapsto y_t := -(\nabla_x c(\bar{x},\cdot))^{-1} q_t.$$

In this case, we refer to $\bar{x}$ as the *base point* of the c-segment. Recalling that, by definition, $\text{c-exp}_{\bar{x}}(v) = y$ if and only if $-\nabla_x c(\bar{x},y) = v$, c-segments coincide with the image under c-exponential map of segments in the tangent space. In the sequel we also use the notation $[y_0,y_1]^c_{\bar{x}}(t)$ for the c-segment given by $\text{c-exp}_{\bar{x}}(q_t)$, where $y_i = \text{c-exp}_{\bar{x}}(q_i), i = 0,1$.

Lemma 20 (Link between cost-convex segment)**.** *Let $y_t = [y_0,y_1]^c_{\bar{x}}(t) = \text{c-exp}_{\bar{x}}(q_t)$ be a c-segment on M. For every $\bar{r} \in \mathbb{R}_+^*$, $a_0 > -2\bar{r}$ there exist $s_0, s_1 \in \mathbb{R}_+$ such that $[(y_0,s_0),(y_1,s_1)]^{\mathrm{d}^2_{\mathscr{C}}}_{\bar{x},\bar{r}}(t) = \mathrm{d}^2_{\mathscr{C}}\text{-exp}_{(\bar{x},\bar{r})}(p_t,a_0)$ is a $\mathrm{d}^2_{\mathscr{C}}$-segment of $\mathscr{C}(M)$. Moreover s_t, p_t are given by*

$$s_t = \frac{2\bar{r}+a_0}{2\cos_+(\bar{x},y_t)}, \quad p_t = \left(\bar{r}^2 + \frac{a_0}{2}\bar{r}\right) q_t. \tag{4.7}$$

Conversely, let $[(y_0,s_0),(y_1,s_1)]^{\mathrm{d}^2_{\mathscr{C}}}_{\bar{x},\bar{r}}(t) = \mathrm{d}^2_{\mathscr{C}}\text{-exp}_{(\bar{x},\bar{r})}(p_t,a_0)$ be a $d_{\mathscr{C}(M)}$-segment of $\mathscr{C}(M)$ with $a_0 > -2\bar{r}$. Then $[y_0,y_1]^c_{\bar{x}}(t) = \text{c-exp}_{\bar{x}}(q_t)$ is a c-segment of M with the choice $q_t = \frac{2p_t}{2\bar{r}^2+\bar{r}a_0}$.

We can state a longer but more exhaustive statement. Let $t \mapsto y_t \in M$ be the c-segment on M with endpoints y_0, y_1, base point $\bar{x}$, given by the image under $\text{c-exp}_{\bar{x}}$ of the segment $q_t = (1-t)q_0 + tq_1 \in T_{\bar{x}}M$. For every $\bar{r} \in \mathbb{R}_+^*$ and for every $a_0 > -2\bar{r}$ there exist $s_0, s_1 \in \mathbb{R}_+$ such that the curve $t \mapsto (y_t, s_t) \in \mathscr{C}(M)$, with

$$s_t = \frac{2\bar{r}+a_0}{2\cos_+(\bar{x},y_t)},$$

is the $\mathrm{d}^2_{\mathscr{C}}$-segment with endpoints $(y_0,s_0),(y_1,s_1)$ and base point $(\bar{x},\bar{r})$, given by the image under $\mathrm{d}^2_{\mathscr{C}}\text{-exp}_{(\bar{x},\bar{r})}$ of the segment

$$p_t = \left(\bar{r}^2 + \frac{a_0}{2}\bar{r}\right) q_t, \quad a_t \equiv a_0.$$

Conversely, let $t \mapsto (y_t,s_t) \in \mathscr{C}(M)$ be the $\mathrm{d}^2_{\mathscr{C}}$-segment with endpoints $(y_0,s_0),(y_1,s_1)$, base point $(\bar{x},\bar{r})$, given by the image under $\mathrm{d}^2_{\mathscr{C}}\text{-exp}_{(\bar{x},\bar{r})}$ of the segment (p_t,a_t), with $p_t = (1-t)p_0 + tp_1 \in T_{\bar{x}}M$ and $a_t \equiv a_0 > -2\bar{r}$. Then $t \mapsto y_t \in M$ is the c-segment of M with endpoints y_0, y_1, base point $\bar{x}$, given by the image under $\text{c-exp}_{\bar{x}}$ of the segment $q_t = \frac{2p_t}{2\bar{r}^2+\bar{r}a_0}$.

Proof. Recall that $c(x,y) = -\log(\cos_+^2(x,y))$ and $\mathrm{d}^2_{\mathscr{C}}((x,r),(y,t)) = r^2 + t^2 - 2rt\cos_+(x,y)$. Thus,

$$-\nabla_x c(\bar{x},z) = \partial_x[\log(\cos_+^2(\bar{x},z))] = 2\frac{\partial_x[\cos_+(\bar{x},z)]}{\cos_+(\bar{x},z)},$$
$$-\nabla_{(x,r)}\ \mathrm{d}^2_{\mathscr{C}}((\bar{x},\bar{r}),(z,s)) = 2\left(\bar{r}s\partial_x[\cos_+(\bar{x},z)], -\bar{r} + s\cos_+(\bar{x},z)\right).$$

Therefore, a curve $t \mapsto y_t \in M$ is the c-segment $[y_0, y_1]^c_{\bar{x}}(t)$ if and only if there exist $q_0, q_1 \in T_{\bar{x}}M$ for which y_t satisfies

$$(1-t)q_0 + tq_1 = -\nabla_x c(\bar{x}, y_t) = 2\frac{\partial_x[\cos_+(\bar{x}, y_t)]}{\cos_+(\bar{x}, y_t)}, \tag{4.8}$$

where $y_i = \text{c-exp}_{\bar{x}}(q_i), i = 0, 1$ (for simplicity set $q_t = (1-t)q_0 + tq_1$). Similarly, a curve $t \mapsto (y_t, s_t) \in \mathscr{C}(M)$ is a $\mathrm{d}^2_{\mathscr{C}}$-segment if and only if there exist $a_0, a_1 > 0$ and $p_0, p_1 \in T_{\bar{x}}M$ for which (y_t, s_t) satisfies

$$\begin{cases} -\partial_x \ \mathrm{d}^2_{\mathscr{C}}((\bar{x}, \bar{r}), (y_t, s_t)) = 2\bar{r}s_t\partial_x[\cos_+(\bar{x}, y_t)] = (1-t)p_0 + tp_1 \\ -\partial_r \ \mathrm{d}^2_{\mathscr{C}}((\bar{x}, \bar{r}), (y_t, s_t)) = -2\bar{r} + 2s_t\cos_+(\bar{x}, y_t) = (1-t)a_0 + ta_1\,. \end{cases} \tag{4.9}$$

Let $t \mapsto y_t = [y_0, y_1]^c_{\bar{x}}(t) = \text{c-exp}_{\bar{x}}(q_t)$, with $q_t = (1-t)q_0 + tq_1 \in T_{\bar{x}}M$. For simplicity, we look for solutions of (4.9) where $a_0 = a_1$. If $t \mapsto \cos_+(\bar{x}, y_t)$, the second equation gives

$$s_t = \frac{a_0 + 2\bar{r}}{2\cos_+(\bar{x}, y_t)}, \tag{4.10}$$

which is strictly positive if $a_0 > -2\bar{r}$. Plugging such choice of s_t in the first equation of system (4.9), we look for $p_0, p_1 \in T_{\bar{x}}M$ satisfying

$$\bar{r}\frac{a_0 + 2\bar{r}}{\cos_+(\bar{x}, y_t)}\partial_x[\cos_+(\bar{x}, y_t)] = (1-t)p_0 + tp_1.$$

Using (4.8), the identity above reads

$$\frac{\bar{r}}{2}(a_0 + 2\bar{r})q_t = (1-t)p_0 + tp_1,$$

which is satisfied by the choice $p_i = \frac{\bar{r}}{2}(a_0 + 2\bar{r})q_i$, $i = 0, 1$.

Conversely, assume a $\mathrm{d}^2_{\mathscr{C}}$-segment is given by

$$t \mapsto (y_t, s_t) = [(y_0, s_0), (y_1, s_1)]^{\mathrm{d}^2_{\mathscr{C}}}_{(\bar{x}, \bar{r})}(t) = \mathrm{d}^2_{\mathscr{C}}\text{-exp}_{(\bar{x}, \bar{r})}(p_t, a_0)\,,$$

where $p_t = (1-t)p_0 + tp_1$ and $a_0 > -2\bar{r}$. Then the pair (y_t, s_t) satisfies (4.9). Define $q_t = \frac{2p_t}{a_0\bar{r} + 2\bar{r}^2}$. Since $t \mapsto p_t$ is affine, so is $t \mapsto q_t$. Moreover by (4.9), q_t satisfies

$$q_t = \partial_x[\log(\cos^2_+(d(\bar{x}, y_t)))] = -\nabla_x c(\bar{x}, y_t).$$

Therefore $t \mapsto y_t$ is a c-segment between endpoints y_0, y_1 and with base point $\bar{x}$. □

A direct consequence of the correspondence between c-segments and $\mathrm{d}^2_{\mathscr{C}}$-segments is the following.

Corollary 21 (Link between cost convexity domains)**.** *Let $Y \times \mathbb{R}_{>0} \subset \mathscr{C}(M)$ be a $\mathrm{d}^2_{\mathscr{C}}$-convex set with respect to $(\bar{x}, \bar{r}) \in \mathscr{C}(M)$. Then $Y \subset M$ is a c-convex set with respect to $\bar{x} \in M$.*

Proof. By definition see [43, Definition 12.11], $Y \times \mathbb{R}_+ \subset \mathscr{C}(M)$ is $\mathrm{d}^2_{\mathscr{C}}$-convex set with respect to $(\bar{x}, \bar{r})$ if every pair of points in $Y \times \mathbb{R}_+$ can be joined by a $\mathrm{d}^2_{\mathscr{C}}$-segment with base point $(\bar{x}, \bar{r})$. Take $y_0, y_1 \in Y$ such that there exists $q_0, q_1 \in T_{\bar{x}}M$ with the property $y_i = \text{c-exp}_{\bar{x}}(q_i), i = 0, 1$. Let $a_0 > -2\bar{r}$ and define

$$\begin{aligned} p_i &= \left(\bar{r}^2 + \frac{a_0}{2}\bar{r}\right)q_i, \ i = 0, 1, \\ s_i &= \frac{2r + a_0}{2\cos_+(\bar{x}, y_i)}, \ i = 0, 1. \end{aligned}$$

By construction, the $\mathrm{d}^2_{\mathscr{C}}$-segment

$$t \mapsto (y_t, s_t) := \mathrm{d}^2_{\mathscr{C}}\text{-exp}_{(\bar{x}, \bar{r})}((1-t)p_0 + tp_1, a_0)$$

is contained in $Y \times \mathbb{R}_+$ and has endpoints $(y_0, s_0), (y_1, s_1)$. By Lemma 20, the curve $t \mapsto y_t$ coincides with the c-segment $\text{c-exp}_{\bar{x}}(q_t)$, where $q_t = (1-t)q_0 + tq_1$. □

A synthetic formulation for the sign of the MTW_{cost} tensor is also given by the quasi or plain convexity of a particular functional, the so-called *support function (see Lemma below)* along a cost-segment see for instance [43, Theorem 12.36, Proposition 12.25(i)] [23, Theorem 2.7][**?**, Section 1.5.b,c,d]. We turn now to the second crucial lemma of this section which makes the link between this support function defined on the base space and the one defined on the cone.

Lemma 22. *Assume $t \mapsto y_t = [y_0, y_1]^c_{\bar{x}}(t) \in M$ is a c-segment with base point $\bar{x}$ and let $h_x : [0,1] \to \mathbb{R}$ denote the support function on y_t, namely*

$$h_x(t) = c(\bar{x}, y_t) - c(x, y_t)\,.$$

Let $t \mapsto (y_t, s_t) = [(y_0, s_0), (y_1, s_1)]^{\mathrm{d}^2_{\mathscr{C}}}_{(\bar{x},\bar{r})}(t)$ be any $\mathrm{d}^2_{\mathscr{C}}$-segment associated to $[y_0, y_1]^c_{\bar{x}}(t)$ throughout Lemma 20 and denote by $H_{(x,r)} : [0,1] \to \mathbb{R}$ the corresponding support function,

$$H_{(x,r)}(t) = \mathrm{d}^2_{\mathscr{C}}((\bar{x},\bar{r}),(y_t,s_t) - \mathrm{d}^2_{\mathscr{C}}((x,r),(y_t,s_t)).$$

Then h_x and $H_{(x,r)}$ satisfy the following identity

$$h_x(t) = 2\log\left(\frac{H_{(x,r)}(t) - \bar{r}^2 + r^2}{a_0\bar{r} + 2\bar{r}^2} + 1\right).$$

Remark that for $h_x, H_{(x,r)}$ to be well defined the cost c must satisfies some smoothness condition such that q_t is in the definition domain of c-exp.

Proof. By definition

$$\begin{aligned} h_x(t) = c(\bar{x}, y_t) - c(x, y_t) &= -\log(\cos^2_+(\bar{x}, y_t)) + \log(\cos^2_+(x, y_t)) \\ &= 2\log\left(\frac{\cos_+(x, y_t)}{\cos_+(\bar{x}, y_t)}\right). \end{aligned} \tag{4.11}$$

The support function on $\mathscr{C}(M)$ is given by

$$\begin{aligned} H_{(x,r)}(t) &= \mathrm{d}^2_{\mathscr{C}}((\bar{x},\bar{r}),(y_t,s_t)) - \mathrm{d}^2_{\mathscr{C}}((x,r),(y_t,s_t)) \\ &= \bar{r}^2 - r^2 + 2rs_t\cos_+(x, y_t) - 2\bar{r}s_t\cos_+(\bar{x}, y_t). \end{aligned}$$

Since (y_t, s_t) is a $\mathrm{d}^2_{\mathscr{C}}$-segment, it satisfies (4.7), whence

$$2\bar{r}s_t = \frac{a_0\bar{r} + 2\bar{r}^2}{\cos_+(\bar{x}, y_t)} = \frac{\bar{a}}{\cos_+(\bar{x}, y_t)},$$

where $\bar{a} = a_0\bar{r} + 2\bar{r}^2 > 0$. Thus

$$H_{x,r}(t) - \bar{r}^2 + r^2 = \bar{a}\left(\frac{r\cos_+(d(x, y_t))}{\bar{r}\cos_+(d(\bar{x}, y_t))} - 1\right).$$

Finally

$$\log\left(H_{x,r}(t) - \bar{r}^2 + r^2 + \bar{a}\right) - \log\bar{a} = \log\left(\frac{r\cos_+(d(x, y_t))}{\bar{r}\cos_+(d(\bar{x}, y_t))}\right),$$

which provides the statement thanks to (4.11).

□

Thanks to the link we made between c-convex functions/c-segment on the cone and on the base manifold, we are able to provide an example of answer to the question raised in [22, Example 3.9] which is

> It remains interesting to find more general sufficient conditions on a Riemannian manifold (M, g) and function f "..." for $f(d(x, y))$ to be strictly or weakly regular i.e $MTW_{f(d)}(0)$ holds.

We prove hereafter the following sufficient condition: if the cost on the cone satisfies $MTW_{\mathrm{d}^2_{\mathscr{C}}}(0)$, then it $MTW_c(0)$ holds where $c(x,y) = -\log(\cos^2_+(d(x,y)))$. Recall that this cost is associated with the Wasserstein-Fisher-Rao metric, see Corollary 8. Importantly, this proof holds for any cost $d(x,y)$ on the base manifold as long as $\nabla_x d(x,\cdot)$ is injective and continuous with inverse continuous on a small neighborhood of all $y_0 \in M$. We have two proofs of this result based on two different synthetic formulations of $MTW_c(0)$. One is based on the quasi-convexity of the c-segment the other one on the assumption (C) that we now recall.

Definition 14. [43, p.288] A cost c on $M \times M$ satisfies Assumption (C) if for every c-convex function f and for every $x \in M$ in its domain, the c-subdifferential $\partial_c f(x)$ is connected.

Lemma 23. *If $\mathrm{d}^2_{\mathscr{C}}$ satisfies assumption (C) on $\mathscr{C}(M)$ then c satisfies assumption (C) on M.*

Proof. To prove assumption (C) for c, let $f : M \to \mathbb{R}$ be a c-convex function. (Note that on both M and $\mathscr{C}(M)$ connectedness is equivalent to path-connectedness.) Take $y_1, y_2 \in \partial_c f(\bar{x})$. Then, by Lemma 19, $(y_i, s_i) \in \partial_{\mathrm{d}^2_{\mathscr{C}}} F_f(\bar{x}, \bar{r})$, where $s_i = \frac{\bar{r}e^{f(\bar{x})}}{\cos_+(\bar{x}, y_i)}$. By assumption (C) on $\mathrm{d}^2_{\mathscr{C}}$, $\partial_{\mathrm{d}^2_{\mathscr{C}}} F_f(\bar{x}, \bar{r})$ is connected, hence there exists a continuous path $t \mapsto (y_t, s_t) \in \partial_{\mathrm{d}^2_{\mathscr{C}}} F_f(\bar{x}, \bar{r})$, with endpoints $(y_0, s_0), (y_1, s_1)$. Again, by Lemma 19, $(y_t, s_t) \in \partial_{\mathrm{d}^2_{\mathscr{C}}} F_f(\bar{x}, \bar{r})$ if and only if

$$y_t \in \partial_c f(\bar{x}), \quad s_t = \frac{\bar{r}e^{f(\bar{x})}}{\cos_+(\bar{x}, y_t)}.$$

In particular, $t \mapsto y_t$ is a continuous path in $\partial_c f(\bar{x})$ between endpoints y_0, y_1, whence $\partial_c f(\bar{x})$ is connected. □

We can now state and prove the main Theorem of this section. Recall that a cost c satisfies the MTW weak condition if and only if, for every pair of points the MTW tensor associated with c computed at any pair of c-orthogonal vectors is nonnegative (see also Section 2.3).

Theorem 24. *If $\mathrm{d}^2_{\mathscr{C}}$ on $\mathscr{C}(M)$ satisfies the MTW weak condition, then the cost c on M satisfies the MTW weak condition.*

We give two proofs of Theorem 24.

Proof 1. Recall that, under some convexity assumptions, [43, Theorem 12.42] states that assumption (C) is equivalent to MTW weak condition. Both costs $\mathrm{d}^2_{\mathscr{C}}$ on $\mathscr{C}(M)$ and c on M satisfy the requirements in [43, Theorem 12.42]. Therefore, applying the result to $\mathrm{d}^2_{\mathscr{C}}$ we deduce that $\mathrm{d}^2_{\mathscr{C}}$ satisfies assumption (C). By Lemma 23 also c satisfies assumption (C) on M. Applying [43, Theorem 12.42] to c we conclude that c satisfies MTW weak condition. □

Proof 2. By the results [43, Proposition 12.15 (i), Theorem 12.42], under the same convexity assumptions, MTW weak condition for a cost is equivalent to the quasi-convexity of the support function along any cost-segment, see [23, Theorem 2.7],[?, Section 1.5.b,c,d]. [43, Theorem 12.36, Proposition 12.25(i)] Assume $\mathrm{d}^2_{\mathscr{C}}$ satisfies MTW weak condition on $\mathscr{C}(M)$. Then, the support $H_{(x,r)}(t) = \mathrm{d}^2_{\mathscr{C}}((\bar{x}, \bar{r}), (y_t, s_t)) - \mathrm{d}^2_{\mathscr{C}}((x, r), (y_t, s_t))$ function along any $\mathrm{d}^2_{\mathscr{C}}$-segment $t \mapsto (y_t, s_t)$ is quasi-convex, i.e.,

$$H_{(x,r)}(t) \leq \max\left(H_{(x,r)}(0), H_{(x,r)}(1)\right).$$

Let $t \mapsto y_t \in M$ be a c-segment, $x \in M$. By Lemma 20, y_t is the projection on M of a $\mathrm{d}^2_{\mathscr{C}}$-segment $t \mapsto (y_t, s_t)$. Moreover, by Lemma 22, the support function $t \mapsto h_x(t)$ along y_t and $t \mapsto H_{(x,r)}(t)$ are related by

$$h_x(t) = 2\log\left(\frac{H_{(x,r)}(t) - \bar{r}^2 + r^2}{a_0\bar{r} + 2\bar{r}^2} + 1\right).$$

By hypothesis, $H_{(x,r)}(t)$ is quasi-convex. Since log is an increasing function, $\max\left(H_{(x,r)}(0), H_{(x,r)}(1)\right) = H_{(x,r)}(j)$ is equivalent to $\max(h_x(0), h_x(1)) = h_x(j)$. Since $a_0\bar{r} + 2\bar{r}^2 > 0$, quasi-convexity of $t \mapsto H_{(x,r)}(t)$ implies quasi-convexity of $t \mapsto h_x(t)$. Finally, we apply [43, Proposition 12.15 (i), Theorem 12.42] to the cost c and we deduce that c satisfies MTW weak condition. □

Note that this theorem can be checked by direct computations[4] however the above proof uses a *synthetic* strategy as illustrated in [43, Chapter 26].

Remark 8. *With Theorem* [43, 12.42] *in mind a summary of this section could be the following, which give some weaker results: Lemma 22 states an equivalence for c to be regular on M and for $\mathrm{d}^2_{\mathscr{C}}$ to be regular on a specific set of $\mathrm{d}^2_{\mathscr{C}}$-segments of $\mathscr{C}(M)$. Whereas Lemma 19 states an equivalence for c to satisfies assumption (C) on M and $\mathrm{d}^2_{\mathscr{C}}$ to satisfies assumption (C) on a specific class of $\mathrm{d}^2_{\mathscr{C}}$-convex functions of $\mathscr{C}(M)$. Both these conditions imply the weak Ma-Trudinger-Wang condition $MTW(0)$. Therefore assumption (C) or regularity for $\mathrm{d}^2_{\mathscr{C}}$ on a subdomain on these specific sets are enough to ensure that $MTW_c(0)$ holds true on a subdomain on the base space. To prove Theorem 24 we also used the reverse results that assumption (C) or regularity for $\mathrm{d}^2_{\mathscr{C}}$ on a totally $\mathrm{d}^2_{\mathscr{C}}$-convex set D are implied by $MTW_{\mathrm{d}^2_{\mathscr{C}}}(0)$.*

Using the link between c-segments on the cone and on the base manifold, we could prove Theorem 24. We can also use such a strategy to derive a result on cross-curvature. Cross-curvature is essentially the curvature tensor of the Kim-McCann metric without the orthogonality condition, see [24]. It is also referred to as $MTW(0,0)$ [**?**, Section 1.5.b,c,d]. Thus, asking nonnegativity of the cross-curvature is a stronger condition than asking for $MTW(0)$ to hold true. However, this condition is known, as proven in [24], to pass to Riemannian submersions and products of manifolds, *i.e.* nonnegativity of cross-curvature is preserved, which may not be the case for the nonnegativity of the MTW tensor.

Theorem 25. *If the cross-curvature on the cone $\mathscr{C}(M)$ is nonpositive, it is also the case on M for the cost $-\log(\cos^2_+(d(x,y)))$.*

Proof. A synthetic formulation for the sign of the MTW_c tensor is given given by the convexity/concavity of the support function along a c-segment [**?**, Section 1.5.b,c,d] or [23, Theorem 2.10]. The convexity is equivalent to a nonnegative cross curvature whereas the concavity is equivalent to a nonpositive cross curvature. Using Lemma 22 and the fact that log is a concave increasing function we get that $t \mapsto H_{x,r}(t)$ concave implies $t \mapsto h_x(t)$ is also concave and prove the first part of the Lemma. □

Obviously, this result is not of direct interest for smoothness of unbalanced optimal transport since it requires nonnegativity of the cross-curvature tensor rather than nonpositivity.

Remark 9. *As* log *is concave we cannot prove here a result similar to Theorem 24, that would pushed the nonnegative cross curvature from the cone towards the base space. More precisely a consequence of Lemma* (19) *would be if the cross-curvature on the cone is nonnegative,* $\log(te^{f_0(x)} + (1-t)e^{f_1}(x))$ *is c-convex if f_0, f_1 are c-convex. However, we do not know if it implies nonnegativity of the cross-curvature on the base manifold, i.e. $tf_0 + (1-t)f_1$ is c-convex.*

5. Future directions

We have shown, not unsurprisingly, that regularity for unbalanced optimal transport can be reduced to the one of optimal transport through linearization of the dual problem. Regularity, being a structural result in itself, is interesting outside analysis. For instance, regularity of optimal transport maps is the key to be able to mitigate the curse of dimensionality of statistical optimal transport as done in [40] and to obtain minimax rate of convergence for the statistical estimation of optimal potentials [33]. Our results should allow similar gains in the statistical estimation of unbalanced optimal transport. We focus on Wasserstein-Fisher-Rao metric since it is the natural length space associated with the problem. This particular case leads us to examine the MTW condition of the induced cost. Interestingly, we showed that when the weak MTW condition on the cone is satisfied, the same holds true for the MTW condition for the induced cost on the base manifold. A similar result holds for cross-curvature, whose nonnepositivity on the cone implies nonpositivity of the corresponding cost on the manifold. This is an example of answer to a question

[4] We tried unsuccessfully to prove this result relying on symbolic computations.

formulated in [22]. Another open application of polar factorization can lead to new numerical scheme for the Camassa-Holm equation as done for incompressible Euler in [17].

References

[1] L. Ambrosio and N. Gigli. *A user's guide to optimal transport*. Lecture Notes in Mathematics. Springer Berlin Heidelberg, 2013.

[2] Martin Bauer, Emmanuel Hartman, and Eric Klassen. The square root normal field distance and unbalanced optimal transport, 2021.

[3] J-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[4] Benamou, Jean-David. Numerical resolution of an "unbalanced" mass transport problem. *ESAIM: M2AN*, 37(5):851–868, 2003.

[5] Kristian Bredies, Marcello Carioni, and Silvio Fanzon. A superposition principle for the inhomogeneous continuity equation with hellinger-kantorovich-regular coefficients, 2020.

[6] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.

[7] L. Caffarelli. The regularity of mappings with a convex potential. *J. Amer. Math. Soc.*, 5:99–104, 1992.

[8] L. Chizat, B. Schmitzer, G. Peyré, and F.-X. Vialard. An Interpolating Distance between Optimal Transport and Fisher-Rao. *Found. Comp. Math.*, 2016.

[9] Lenaic Chizat, Gabriel Peyre, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: dynamic and Kantorovich formulations. *J. Funct. Anal.*, 274(11):3090–3123, 2018.

[10] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *Mathematics of Computation*, 2018.

[11] G. De Philippis and A. Figalli. The Monge-Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc.*, 51:527–580, 2014.

[12] Guido De Philippis and Alessio Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[13] Guido De Philippis and Alessio Figalli. The Monge-Ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc. (N.S.)*, 51(4):527–580, 2014.

[14] Philippe Delanoë. Differential geometric heuristics for riemannian optimal mass transportation. In Boris Kruglikov, Valentin Lychagin, and Eldar Straume, editors, *Differential Equations - Geometry, Symmetries and Integrability*, volume 5 of *Abel Symposia*, pages 49–73. Springer Berlin Heidelberg, 2009.

[15] Jean Feydy, Benjamin CHARLIER, François-Xavier Vialard, and Gabriel Peyré. Optimal transport for diffeomorphic registration. In *MICCAI 2017*, Quebec, Canada, September 2017.

[16] D. S. Freed and D. Groisser. The basic geometry of the manifold of riemannian metrics and of its quotient by the diffeomorphism group. *Michigan Math. J.*, 36(3):323–344, 1989.

[17] Thomas Gallouët and Quentin Mérigot. A lagrangian scheme for the incompressible euler equation using optimal transport, 2016.

[18] Thomas Gallouët, Andrea Natale, and François-Xavier Vialard. Generalized compressible fluid flows and solutions of the camassa-holm variational model. *ARMA*, June 2019. working paper or preprint.

[19] Thomas Gallouët and François-Xavier Vialard. The camassa–holm equation as an incompressible euler equation: A geometric point of view. *Journal of Differential Equations*, 264(7):4199 – 4234, 2018.

[20] Tryphon T. Georgiou, Johan Karlsson, and Mir Shahrouz Takyar. Metrics for power spectra: An axiomatic approach. *IEEE Transactions on Signal Processing*, 57(3):859–867, 2009.

[21] B. Khesin and R. Wendt. *The geometry of infinite-dimensional groups*, volume 51. Springer Science & Business Media, 2008.

[22] Young-Heon Kim and Robert J. McCann. Continuity, curvature, and the general covariance of optimal transportation. *J. Eur. Math. Soc.*, 12, 2007.

[23] Young-Heon Kim and Robert J McCann. Towards the smoothness of optimal maps on riemannian submersions and riemannian products (of round spheres in particular). *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2012(664):1–27, 2012.

[24] Young-Heon Kim and Robert J. McCann. Towards the smoothness of optimal maps on riemannian submersions and riemannian products (of round spheres in particular). *Journal für die reine und angewandte Mathematik (Crelles Journal)*, 2012(664):1–27, 2012.

[25] S. Kondratyev, L. Monsaingeon, and D. Vorotnikov. A new optimal trasnport distance on the space of finite Radon measures. *Adv. Differential Equations*, 21(11):1117–1164, 2016.

[26] Stanislav Kondratyev and Dmitry Vorotnikov. Convex sobolev inequalities related to unbalanced optimal transport, 2019.

[27] Stanislav Kondratyev and Dmitry Vorotnikov. Spherical hellinger-kantorovich gradient flows, 2019.

[28] Vaios Laschos and Alexander Mielke. Geometric properties of cones with applications on the hellinger-kantorovich space, and a new distance on the space of probability measures, 2018.

[29] Paul W. Y. Lee and Jiayong Li. New examples satisfying Ma-Trudinger-Wang conditions. *SIAM J. Math. Anal.*, 44(1):61–73, 2012.
[30] M. Liero, A. Mielke, and G. Savaré. Optimal Entropy-Transport problems and a new Hellinger-Kantorovich distance between positive measures. *Inventiones Math.*, August 2018.
[31] Xi-Nan Ma, Neil S. Trudinger, and Xu-Jia Wang. Regularity of potential functions of the optimal transportation problem. *Arch. Ration. Mech. Anal.*, 177(2):151–183, 2005.
[32] R.J. McCann. Polar factorization of maps on riemannian manifolds. *Geometric & Functional Analysis GAFA*, 11(3):589–608, 2001.
[33] Boris Muzellec, Adrien Vacher, Francis Bach, François-Xavier Vialard, and Alessandro Rudi. Near-optimal estimation of smooth transport maps with kernel sums-of-squares, 2021.
[34] Nicoló De Ponti and Andrea Mondino. Entropy-transport distances between unbalanced metric measure spaces, 2020.
[35] R.T. Rockafellar. Integrals which are convex functionals. II. *Pacific Journal of Mathematics*, 39(2):439–469, 1971.
[36] Zhengyang Shen, Jean Feydy, Peirong Liu, Ariel Hernán Curiale, Ruben San Jose Estepar, Raul San Jose Estepar, and Marc Niethammer. Accurate point cloud registration with robust optimal transport, 2021.
[37] Bernd Sturmfels, Simon Telen, François-Xavier Vialard, and Max von Renesse. Toric geometry of entropic regularization, 2023.
[38] Thibault Séjourné, Jean Feydy, François-Xavier Vialard, Alain Trouvé, and Gabriel Peyré. Sinkhorn divergences for unbalanced optimal transport. 2019.
[39] Thibault Séjourné, François-Xavier Vialard, and Gabriel Peyré. The unbalanced gromov wasserstein distance: Conic formulation and relaxation, Sep 2020.
[40] Adrien Vacher, Boris Muzellec, Alessandro Rudi, Francis Bach, and Francois-Xavier Vialard. A dimension free computational upper-bound for smooth optimal transport estimation, Jan 2021.
[41] François-Xavier Vialard and Andrea Natale. Embedding camassa-holm equations in incompressible euler. *Journal of Geometric Mechanics*, page arXiv:1804.11080, Apr 2018.
[42] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[43] C. Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009.
[44] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.
[45] Zihao Wang, Datong Zhou, Ming Yang, Yong Zhang, Chenglong Rao, and Hao Wu. Robust document distance with wasserstein-fisher-rao metric. In Sinno Jialin Pan and Masashi Sugiyama, editors, *Proceedings of The 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pages 721–736. PMLR, 18–20 Nov 2020.

Appendix A. Proofs

Proof of Proposition 16 (Approximate differentiability). The proof is an adaptation of [30, Theorem 6.7] using arguments in [32, 42]. In particular we use the notation of [30]. Let (z_0, z_1) be a generalized optimal potential pair for $\mathrm{WF}^2(\rho_0, \rho_1)$ and γ an optimal coupling [30, Theorem 6.3]. We define the associated densities $\sigma_i = e^{-z_i}$, $i = 0, 1$. Since ρ_0 and ρ_1 are admissible [30, Theorem 6.3,b] implies $\mathrm{Supp}\left(p^1_*(\gamma) = \gamma_0\right) = \mathrm{Supp}(\rho_0)$ and $\mathrm{Supp}\left(p^2_*(\gamma) = \gamma_1\right) = \mathrm{Supp}(\rho_1)$. Therefore, there exist Borel sets $A_i \subset \mathrm{Supp}(\rho_i)$ with $\rho_i(M \setminus A_i) = 0$ such that

$$\sigma_0(x)\sigma_1(y) \geq \cos^2(d_{\pi/2}(x,y)) \qquad \text{in } A_0 \times A_1\,, \tag{A.1}$$

$$\sigma_0(x)\sigma_1(y) = \cos^2(d_{\pi/2}(x,y)) \qquad \gamma - \text{a. e. in } A_0 \times A_1\,. \tag{A.2}$$

To construct the set of approximate differentiability let

$$A_{1,n} = \{y \in M\,;\, \sigma_1(y) \geq 1/n\}$$

and consider, the function

$$s_{0,n} = \sup_{y \in A_{1,n}} \frac{\cos^2(d_{\pi/2}(x,y))}{\sigma_1(y)}\,.$$

By construction, $s_{0,n}$ is bounded, Lipschitz and thus differentiable vol a.e. Still by definition, we have $\sigma_0 \geq s_{0,n}$ and thus the sets $A_{0,n} = \{x \in M\,;\, \sigma_0(x) = s_{0,n}(x)\}$ are increasing. Since (A.2) is valid γ a.e. the set $\bigcap_{n=1}^{\infty}(X \setminus A_{0,n})$ is ρ_0 negligible. Let

$$A'_{0,n} = \left\{ x \in A_{0,n}\,;\, \lim_{r \to 0} \frac{\mathrm{vol}(B(x,r) \cap A_{0,n})}{\mathrm{vol}(B(x,r))} = 1 \text{ and } s_{0,n} \text{is differentiable at } x \right\}$$

be the set of points of $A_{0,n}$ with vol density 1. Remark that $\bigcap_{n=1}^{\infty}(X \setminus A'_{0,n})$ is also ρ_0 negligible. Let $(\bar{x},\bar{y}) \in A'_{0,n} \times A_{1,n}$ be such that

$$s_{0,n}(\bar{x})\sigma_1(\bar{y}) = \cos^2(d_{\pi/2}(\bar{x},\bar{y})) = \sigma_0(\bar{x})\sigma_1(\bar{y})\,.$$

Using (A.1), it holds, for all $x \in A_1$

$$\sigma_1(y) \geq \cos^2(d_{\pi/2}(x,\bar{y}))/s_{0,n}(x)\,.$$

In particular, $\cos^2(d_{\pi/2}(x,\bar{y}))/s_{0,n}(x)$ achieves its maximum at $\bar{x}$, implying $0 \in \nabla^+_{\bar{x}}(\cos^2(d_{\pi/2}(\cdot,\bar{y}))/s_{0,n}(\cdot))$. Since $s_{0,n}$ is differentiable at $\bar{x}$, it yields that $d^2(\cdot,y)$ is super-differentiable. By Lemma 15, it is also sub-differentiable and thus differentiable at $\bar{x}$. It holds

$$0 = \nabla \cos^2\left(\sqrt{2}\sqrt{\frac{1}{2}d^2_{\pi/2}(\bar{x},\bar{y})}\right)/s_{0,n}(\bar{x}) - \cos^2(d_{\pi/2}(\bar{x},\bar{y}))\nabla s_{0,n}(\bar{x})/s^2_{0,n}(\bar{x}) \tag{A.3}$$

$$= -2\sqrt{2}\tan(d_{\pi/2}(\bar{x},\bar{y}))\frac{\sqrt{2}}{2d_{\pi/2}(\bar{x},\bar{y})}\nabla\left(\frac{1}{2}d^2_{\pi/2}(\bar{x},\bar{y})\right) - \nabla \ln s_{0,n}(\bar{x})\,. \tag{A.4}$$

Let $-\nabla\left(\frac{1}{2}d^2_{\pi/2}(\bar{x},\bar{y})\right) = v_{\bar{x}\to\bar{y}} \in T_{\bar{x}}M$ be the unique vector such that $\bar{y} = \exp^M_{\bar{x}}(v_{\bar{x}\to\bar{y}})$, the last equality reads

$$\tilde{\nabla} z_0(\bar{x}) = -\tilde{\nabla}\ln\sigma_0(\bar{x}) = -\nabla \ln s_{0,n}(\bar{x}) = -2\tan(\|v_{\bar{x}\to\bar{y}}\|)\frac{v_{\bar{x}\to\bar{y}}}{\|v_{\bar{x}\to\bar{y}}\|}\,.$$

Therefore, $\bar{y}$ is unique ρ_1 a.e. and given by

$$\bar{y} = \exp^M_{\bar{x}}\left(v_{\bar{x}\to\bar{y}}\right) = \exp^M_{\bar{x}}\left(-\arctan\left(\frac{\|\tilde{\nabla} z_0(\bar{x})\|}{2}\right)\frac{\tilde{\nabla} z_0(\bar{x})}{\|\tilde{\nabla} z_0(\bar{x})\|}\right) = \varphi(\bar{x})\,.$$

It implies that γ is concentrated on the graph of φ in particular $\gamma = (Id,\varphi)_*\gamma_0$ and $\varphi_*\gamma_0 = \gamma_1$. The strict convexity of KL implies that the marginals γ_0 and γ_1 are unique [30, Theorem 6.7] thus

$$z_0 = -\log(\sigma_0) = -\log(\frac{\mathrm{d}\gamma_0}{\mathrm{d}\rho_0})$$

is unique ρ_0 a.e. and γ is also unique. Note that we used the admissible condition to say that σ_0 is ρ_0 a.e. positive. In order to prove (3.28), we start from (13) and a direct computation yields

(A.5)

$$\begin{aligned}
\mathrm{WFR}^2(\rho_0,\rho_1) &= \mathrm{KL}(\gamma_0,\rho_0) + \mathrm{KL}(\gamma_1,\rho_1) + \int_{M^2} c(x,y)\,\mathrm{d}\gamma(x,y)\\
&= \int_M \log\left(e^{-z_0}\right)e^{-z_0}\,\mathrm{d}\rho_0 + \int_M (1-e^{-z_0})\,\mathrm{d}\rho_0 + \int_M \log\left(e^{-z_1}\right)e^{-z_1}\,\mathrm{d}\rho_1 + \int_M (1-e^{-z_1})\,\mathrm{d}\rho_1\\
&+ \int_{M^2} c(x,\varphi(x))\,\mathrm{d}\gamma(x)\\
&= \int_M (1-e^{-z_0})\,\mathrm{d}\rho_0 + \int_M (1-e^{-z_1})\,\mathrm{d}\rho_1 + \int_M \left[c(x,\varphi(x)) - z_0(x) - z_1(\varphi(x))\right]\,\mathrm{d}\gamma_0(x)\\
&= \int_M (1-e^{-z_0})\,\mathrm{d}\rho_0 + \int_M (1-e^{-z_1})\,\mathrm{d}\rho_1.
\end{aligned}$$

□

INRIA, Project team Mokaplan, Université Paris-Dauphine, PSL Research University, UMR CNRS 7534-Ceremade.
Email address: thomas.gallouet@inria.fr

Dipartimento di Matematica, Università degli Studi di Roma "Tor Vergata", Rome Italy
Email address: ghezzi@mat.uniroma2.it

Université Gustave Eiffel, LIGM, CNRS, INRIA, Project team Mokaplan
Email address: fxvialard@normalesup.org

3.2 Unbalanced gradient flows and more general reaction diffusion PDE

Articles:

- **A JKO splitting scheme for Kantorovich-Fisher-Rao gradient flows.** *SIAM Journal on Mathematical Analysis, Vol. 49, Issue 2.* (2017) https://arxiv.org/abs/1602.04457. Gallouët T.O. et Monsaingeon L.
- **An unbalanced optimal transport splitting scheme for general advection-reaction-diffusion problems.** *Journal of Differential Equations ESAIM: Control, Optimisation and Calculus of Variations* (2018) https://hal.science/hal-01508911. Gallouët T.O., Laborde M. and Monsaingeon L.

Collaborators: The first article is written with L. Monsaingeon and the second one with both L. Monsaingeon et Maxime Laborde. We were somehow at the same career level.

Main contributions:

- Using the inf-convolution structure we proposed and proved that a splitting scheme made of one JKO step in the Wasserstein space followed by one in the euclidian converge towards Unbalanced Wasserstein Gradient flows.
- In the second paper we proved that the previous approach works well for more general reaction-diffusions equations, where the energy used in the Wasserstein JKO step and the Euclidian step are different. The Unbalanced metric is used as a common metric that can handle all estimates together.
- We construct and implement a numerical scheme for this splitting method.

Research directions: Thanks to the spitting scheme approach detailed in this section many of the technics developed for Wasserstein Gradient flows such as higher order scheme can be used for more general reaction diffusion equation. We need to understand how in interplay with the reaction step of the scheme.

A JKO SPLITTING SCHEME FOR KANTOROVICH-FISHER-RAO GRADIENT FLOWS *

THOMAS O. GALLOUËT† AND LÉONARD MONSAINGEON‡

Abstract. In this article we set up a splitting variant of the Jordan-Kinderlehrer-Otto scheme in order to handle gradient flows with respect to the Kantorovich-Fisher-Rao metric, recently introduced and defined on the space of positive Radon measure with varying masses. We perform successively a time step for the quadratic Wasserstein/Monge-Kantorovich distance, and then for the Hellinger/Fisher-Rao distance. Exploiting some inf-convolution structure of the metric we show convergence of the whole process for the standard class of energy functionals under suitable compactness assumptions, and investigate in details the case of internal energies. The interest is double: On the one hand we prove existence of weak solutions for a certain class of reaction-advection-diffusion equations, and on the other hand this process is constructive and well adapted to available numerical solvers.

Key words. Unbalanced Optimal transport, Wasserstein-Fisher-Rao, Hellinger-Kantorovich, Gradient flows, JKO scheme

AMS subject classifications. 35K15, 35K57, 35K65, 47J30

1. Introduction.

A new Optimal Transport distance on the space of positive Radon measures has been recently introduced independently by three different teams [13, 14, 25, 28, 29]. Contrarily to the classical Wasserstein-Monge-Kantorovich distances, which are restricted to the space of measures with fixed mass (typically probability measures), this new distance has the advantage of allowing for mass variations, can be computed between arbitrary measures, and does not require decay at infinity (such as finite moments). In [13, 14] the distance is called Wasserstein-Fisher-Rao and is introduced with imaging applications in mind. In [28, 29] the distance is referred to as the Hellinger-Kantorovich one, and was studied as a particular case of a larger class of Optimal Transport problems including primal/dual and static formulations. The second author introduced the same distance in [25], with applications to population dynamics and gradient flows in mind. In this paper we propose the name Kantorovich-Fisher-Rao for this metric (`KFR` in the sequel), taking into account all contributions.

On one side we aim here at understanding the local behavior of the `KFR` metric with respect to the by now classical quadratic Monge-Kantorovich/Wasserstein metric `MK` and the Hellinger/Fisher-Rao metric `FR`. On the other side we want to use this information to prove existence of weak solutions to gradient flows while avoiding to look too closely into the geometry of the `KFR` space. Moreover our constructive approach is naturally adapted to available numerical schemes and Monge-Ampère solvers.

A possible way to formalize abstract gradient flow structures is to prove convergence of the corresponding Minimizing Movement scheme, as introduced by De Giorgi [15] and later exploited by Jordan-Kinderlehrer-Otto for the `MK` metric [21]. Given a metric space (X, d) and a functional $F : X \to \mathbb{R}$, the JKO scheme with time-step $\tau > 0$ writes

$$x^{n+1} \in \operatorname*{Argmin}_{x \in X} \left\{ \frac{1}{2\tau} d^2(x, x^n) + F(x) \right\}. \tag{1}$$

*Submitted to the editors DATE.

Funding: L. Monsaingeon was partially supported by the Portuguese National Science Foundation through fellowship BPD/88207/2012 and by the UT Austin/Portugal CoLab program *Phase Transitions and Free Boundary Problems*. T. O. Gallouët was supported by the ANR project ISOTACE (ANR-12-MONU-013) hosted at CMLS, École polytechnique, CNRS, Université Paris-Saclay and by the fond de la Recherche Scientifique-FNRS under grant MIS F.4539.16..

†Département de mathématiques, Université de Liège, , Allée de la découverte 12, B-4000 Liège, Belgique. (thomas.gallouet@ulg.ac.be, http://thomas.gallouet.fr).

‡Institut Élie Cartan de Lorraine, Université de Lorraine, Site de Nancy, B.P. 70239, F-54506 Vandoeuvre-lès-Nancy Cedex, France. (leonard.monsaingeon@univ-lorraine.fr).

Letting $\tau \to 0$ one should expect to recover a weak solution of the gradient flow

$$\dot{x}(t) = -\operatorname{grad}_d F(x(t)). \tag{2}$$

Looking at (2), which is a differential equality between infinitesimal variations, we guess that only the local behavior of the metric d matters in (1).

The starting point of our analysis is therefore the local structure of the Kantorovich-Fisher-Rao metric, which endows the space of positive Radon measures $\rho \in \mathcal{M}^+$ with a formal Riemannian structure [25]. Based on some inf-convolution structure, our heuristic considerations will suggest that, infinitesimally, KFR should be the orthogonal sum of MK and FR:

$$\mathtt{KFR}^2 \approx \mathtt{MK}^2 + \mathtt{FR}^2.$$

More precisely, we will show that in the tangent plane there holds

$$\|\operatorname{grad}_{\mathtt{KFR}} \mathcal{F}(\rho)\|^2 = \|\operatorname{grad}_{\mathtt{MK}} \mathcal{F}(\rho)\|^2 + \|\operatorname{grad}_{\mathtt{FR}} \mathcal{F}(\rho)\|^2 \tag{3}$$

at least formally for reasonable functionals $\mathcal{F}$, and this is in fact the key point in this work. The notion of metric gradients and tangent norms appearing in (3) will be precised in section 2. This naturally leads to a splitting approach for KFR Minimizing Movements: we successively run a first time step for MK, leading to the diffusion term in the associated PDE, and then a second step for FR, leading to the reaction term in the PDE. This can also be viewed as replacing the direct approximation "by hypotenuses" in the JKO scheme (with the KFR distance) by a double approximation "by legs" (each of the legs corresponding to one of the FR, MK metrics). Formula (3) also indicates that the energy dissipation $D(t) := -\frac{dF}{dt} = |\dot{x}|^2 = |\operatorname{grad} F|^2$ will be correctly approximated in (2). One elementary Monge-Kantorovich JKO step is now well known, see for instance [38] and references therein. On the other hand the Fisher-Rao metric enjoys a Riemannian structure that can be recast, up to a change of variable, into a convex Hilbertian setting, and therefore the reaction step should be easy to handle numerically.

Here we show that the classical estimates (energy monotonicity, total square distance, mass control, BV...) propagate along each MK and FR substeps, and nicely fit together in the unified KFR framework. This allows us to prove existence of weak solutions for a whole class of reaction-advection-diffusion PDEs

$$\partial_t \rho = \operatorname{div}(\rho \nabla (U'(\rho) + \Psi + K * \rho)) - \rho (U'(\rho) + \Psi + K * \rho)$$

associated with KFR gradient flows

$$\partial_t \rho = -\operatorname{grad}_{\mathtt{KFR}} \mathcal{F}(\rho), \qquad \mathcal{F}(\rho) = \int_\Omega \big\{ U(\rho) + \Psi(x)\rho + \frac{1}{2} \rho K \star \rho \big\}.$$

The structural conditions on the *internal energy* U, *external potential* Ψ, *interaction kernel* K, and the meaning of the *metric gradient* $\operatorname{grad}_{\mathtt{KFR}}$ will be precised later on. Moreover we retrieve a natural Energy Dissipation Inequality at least in some particular cases, which is well known [3] to completely characterize metric gradient flows.

Our splitting method has several interests: First we avoid a possibly delicate geometrical analysis of the KFR space, in particular we do not need to differentiate the squared KFR distance. This is usually required to derive the Euler-Lagrange equations in the JKO scheme, but might not be straightforward here (see Section 3 for discussions). Secondly, the approach leads to a new constructive existence proof for weak solutions to the above class of PDEs, and can be implemented numerically (see [24] for an early application of this idea). For one elementary MK step many discretizations are now available, such as the semi-discrete scheme [32, 6], the augmented Lagrangian procedure [5], or the Entropic relaxation [36]. The Fisher-Rao minimizing step should not be difficult to implement, since the problem is convex with the good choice of variables.

Finally it is worth stressing that the KFR distance is, by construction, well adapted to handle general transport and reaction processes in a unified framework. One very natural extension of this work would be to consider two separate energy functionals $\mathcal{F}_1, \mathcal{F}_2$, to be used respectively in the diffusion and reaction parts. This natural approach is the purpose of our ongoing works [17, 26] and should allow to treat more general equations (not necessarily gradient flows). However, the rigorous analysis requires suitable compatibility conditions between the two driving functionals and becomes quite technical (see e.g. Remark 4.1). For the sake of exposition we chose to restrict here to the case of pure gradient flows $\mathcal{F}_1 = \mathcal{F} = \mathcal{F}_2$, when the technical estimates are more straightforward and allow to recover dissipation estimates (see Section 5.2).

The paper is structured as follows. In Section 2 we recall some basic facts on the three metrics involved: the quadratic Monge-Kantorovich MK, the Fisher-Rao FR, and the Kantorovich-Fisher-Rao KFR distances. We highlight the three differential Riemannian structures and gradient flow interpretations. Section 3 details the local relation between the three metrics, in particular the infinitesimal uncoupling of the inf-convolution. For the sake of exposition we deliberately remain formal in order to motivate the rigorous analysis in the next sections. In section 4 we define the splitting minimizing movement scheme for the KFR distance and prove, under natural compactness assumptions, the convergence towards a weak solution of the expected PDE. As an example in section 5 we work out all the technical details for the particular case of internal energies, and show that the previous abstract compactness hypothesis holds.

2. Preliminaries. From now on we always assume that $\Omega \subset \mathbb{R}^d$ is a convex subset, possibly unbounded. In this section we recall some facts about the Wasserstein-Monge-Kantorovich and Hellinger-Fisher-Rao distances MK, FR, and introduce the Kantorovich-Fisher-Rao distance KFR. We also present the differential points of view for each of them, allowing to retrieve the three corresponding pseudo Riemannian structures and compute gradients of functionals with respect to the MK, FR, KFR metrics.

2.1. The quadratic Monge-Kantorovich distance MK. We refer to [41] for an introduction and to [42] for a complete overview of the Wasserstein-Monge-Kantorovich distances.

Definition 2.1. *For any nonnegative Radon measures $\rho_0, \rho_1 \in \mathcal{M}_2^+$ with same mass $|\rho_0| = m = |\rho_1|$ and finite second moments, the quadratic Monge-Kantorovich distance is*

$$\mathtt{MK}^2(\rho_0, \rho_1) = \min_{\gamma \in \Gamma[\rho_0, \rho_1]} \int_{\Omega \times \Omega} |x - y|^2 \mathrm{d}\gamma(x, y), \tag{4}$$

where the admissible set of transference plans $\Gamma[\rho_0, \rho_1]$ *consists of nonnegative measures $\gamma \in \mathcal{M}^+(\Omega \times \Omega)$ with mass $|\gamma| = m$ and prescribed marginals $\Pi_x(\gamma) = \rho_0(x)$ and $\Pi_y(\gamma) = \rho_1(y)$.*

The minimizer is unique and is called an optimal plan. When ρ_0 does not charge small sets we have the characterization in terms of transport *maps*:

Theorem 2.1 (Brenier, Gangbo-McCann, [11, 19]). *With the same assumptions as in Definition 2.1, assume that ρ_0 does not give mass to $\mathcal{H}^{d-1}$ sets. Then*

$$\mathtt{MK}^2(\rho_0, \rho_1) = \min_{\rho_1 = \mathbf{t}\#\rho_0} \int_{\Omega} |x - \mathbf{t}(x)|^2 \mathrm{d}\rho_0(x), \tag{5}$$

and the optimal transport map $\mathbf{t}$ is unique $\mathrm{d}\rho_0$ almost everywhere.

We recall the definition of pushforwards by maps $\mathbf{t} : \Omega \to \Omega$

$$\rho_1 = \mathbf{t}\#\rho_0 \quad \Leftrightarrow \quad \int_{\Omega} \phi(y) \mathrm{d}\rho_1(y) = \int_{\Omega} \phi(\mathbf{t}(x)) \mathrm{d}\rho_0(x) \quad \text{for all } \phi \in \mathcal{C}_c(\Omega).$$

As first pointed out by Benamou and Brenier [4] we also have the following dynamic representation of the Wasserstein distance:

THEOREM 2.2 (Benamou-Brenier formula, [3, 4]). *There holds*

$$\mathtt{MK}^2(\rho_0,\rho_1) = \min_{(\rho,\mathbf{v})\in\mathcal{A}_{\mathtt{MK}}[\rho_0,\rho_1]} \int_0^1 \int_\Omega |\mathbf{v}_t|^2 \mathrm{d}\rho_t \mathrm{d}t, \tag{6}$$

where the admissible set $\mathcal{A}_{\mathtt{MK}}[\rho_0,\rho_1]$ *consists of curves* $[0,1] \ni t \mapsto (\rho_t,\mathbf{v}_t) \in \mathcal{M}^+(\Omega)\times L^2(\Omega,\mathrm{d}\rho_t)^d$ *such that* $t\mapsto\rho_t$ *is narrowly continuous with endpoints* ρ_0,ρ_1 *and solving the continuity equation*

$$\partial_t\rho_t + \operatorname{div}(\rho_t\mathbf{v}_t) = 0$$

in the sense of distributions $\mathcal{D}'((0,1)\times\Omega)$.

REMARK 2.1. *Note that, since we are minimizing the kinetic energy in (6), the admissible velocity fields* $\mathbf{v}$ *are implicitly taken in the varying weighted space* $\mathbf{v}\in L^2(0,1;L^2(\mathrm{d}\rho_t))$. *For such velocities in this energy space, the action of the product* $\rho_t\mathbf{v}_t$ *is well defined against any smooth test-function* $\varphi\in\mathcal{C}_c^\infty((0,1)\times\Omega)\subset L^2(0,1;L^2(\mathrm{d}\rho_t))$ *in the distributional formulation of the continuity equation, i-e*

$$-\langle\operatorname{div}(\rho\mathbf{v}),\varphi\rangle_{\mathcal{D}',\mathcal{D}} = \langle\rho\mathbf{v},\nabla\varphi\rangle_{\mathcal{D}',\mathcal{D}} = (\mathbf{v},\nabla\varphi)_{L^2(0,1;L^2(\mathrm{d}\rho_t))} = \int_0^1\int_\Omega \mathbf{v}_t\cdot\nabla\varphi\,\mathrm{d}\rho_t\,\mathrm{d}t.$$

In (6) a minimizing curve $t\mapsto\rho_t$ is of course a geodesics, with constant metric speed $\|\mathbf{v}_t\|^2_{L^2(\mathrm{d}\rho_t)} = cst = \mathtt{MK}^2(\rho_0,\rho_1)$. Note that we allow here for any arbitrary mass $|\rho_0| = m = |\rho_1| > 0$, and that the distance scales as $\mathtt{MK}^2(\alpha\rho_0,\alpha\rho_1) = \alpha\mathtt{MK}^2(\rho_0,\rho_1)$. This is apparent in all three formulations (4)(5)(6), which are linear in γ, ρ_0,ρ_1, and ρ_t respectively.

As is now well-known from the works of Otto [34], we can view the set of measures with fixed mass as a pseudo-Riemannian manifold, endowing the tangent plane

$$T_\rho\mathcal{M}^+_{\mathtt{MK}} = \{\partial_t\rho = -\operatorname{div}(\rho\mathbf{v}) \qquad \text{evaluated at } t=0\}$$

with the metrics

$$\|\partial_t\rho\|^2_{T_\rho\mathcal{M}^+_{\mathtt{MK}}} := \inf\left\{\|\mathbf{v}\|^2_{L^2(\mathrm{d}\rho)} : \quad \partial_t\rho = -\operatorname{div}(\rho\mathbf{v})\right\}.$$

It is easy to see that, among all possible velocities $\mathbf{v}$ representing the same tangent vector $\partial_t\rho = -\operatorname{div}(\rho\mathbf{v})$, there is a unique one with minimal $L^2(\mathrm{d}\rho)$ norm. A standard computation [41] shows that this particular velocity is necessarily potential, $\mathbf{v}=\nabla p$ for a pressure function p uniquely defined up to constants (see the proof of Proposition 2.2 below at least for smooth positive densities ρ). As a consequence we always choose to represent

$$\|\partial_t\rho\|^2_{T_\rho\mathcal{M}^+_{\mathtt{MK}}} = \|\nabla p\|^2_{L^2(\mathrm{d}\rho)} \quad \text{with the identification } \partial_t\rho = -\operatorname{div}(\rho\nabla p).$$

Here we remained formal and refer again to [41, 42] for details. Now metric gradients $\operatorname{grad}_{\mathtt{MK}}$ can be computed by the chain rule as follows: If $\partial_t\rho_t = -\operatorname{div}(\rho_t\nabla p_t)$ is a smooth curve passing through $\rho_t(0)=\rho$ with arbitrary initial velocity $\zeta = \partial_t\rho(0) = -\operatorname{div}(\rho\nabla p)$ then for functionals $\mathcal{F}(\rho) = \int_\Omega F(\rho(x),x)\mathrm{d}x$

$$\begin{aligned}
\langle\operatorname{grad}_{\mathtt{MK}}\mathcal{F}(\rho),\zeta\rangle_{T_\rho\mathcal{M}^+_{\mathtt{MK}}} &= \frac{d}{dt}\mathcal{F}(\rho_t)\bigg|_{t=0} = \frac{d}{dt}\left(\int_\Omega F(\rho_t(x),x)\mathrm{d}x\right)\bigg|_{t=0}\\
&= \int_\Omega F'(\rho)\times\{-\operatorname{div}(\rho\nabla p)\} = \int_\Omega \nabla F'(\rho)\cdot\nabla p\,\mathrm{d}\rho\\
&= (\nabla F'(\rho),\nabla p)_{L^2(\mathrm{d}\rho)},
\end{aligned}$$

where $F'(\rho) = \frac{\delta F}{\delta \rho}$ stands for the standard first variation with respect to ρ. For the classical case $\mathcal{F}(\rho) = \int_\Omega \{U(\rho) + \Psi\rho + \frac{1}{2}\rho K \star \rho\}$ considered here this means $F'(\rho) = U'(\rho) + \Psi(x) + K \star \rho$. This shows that one should identify gradients

$$\operatorname{grad}_{\mathtt{MK}} \mathcal{F}(\rho) = -\operatorname{div}(\rho \nabla F'(\rho))$$

through the $L^2(\mathrm{d}\rho)$ action in the tangent plane, and as a consequence the Monge-Kantorovich gradients flows read

$$\partial_t \rho = -\operatorname{grad}_{\mathtt{MK}} \mathcal{F}(\rho) \qquad \leftrightarrow \qquad \partial_t \rho = \operatorname{div}(\rho F'(\rho)). \tag{7}$$

2.2. The Fisher-Rao distance FR. The classical Hellinger-Kakutani distance [20, 22], or Fisher-Rao metric, was first introduced for probability measures and is well known in statistics and information theory for its connections with the Kullback's divergence and Fisher information [9]. It can be extended to arbitrary nonnegative measures as

Definition 2.2. *The Fisher-Rao distance between measures $\rho_0, \rho_1 \in \mathcal{M}^+$ is given by*

$$\mathtt{FR}^2(\rho_0, \rho_1) \stackrel{def}{=} \min_{(\rho, r) \in \mathcal{A}_{\mathtt{FR}}[\rho_0, \rho_1]} \int_0^1 \int_\Omega |r_t(x)|^2 \mathrm{d}\rho_t(x)\, \mathrm{d}t = 4 \int_\Omega \left| \sqrt{\frac{\mathrm{d}\rho_0}{\mathrm{d}\lambda}} - \sqrt{\frac{\mathrm{d}\rho_1}{\mathrm{d}\lambda}} \right|^2 \mathrm{d}\lambda. \tag{8}$$

The admissible set $\mathcal{A}_{\mathtt{FR}}[\rho_0, \rho_1]$ consists of curves $[0,1] \ni t \mapsto (\rho_t, r_t) \in \mathcal{M}^+(\Omega) \times L^2(\Omega, \mathrm{d}\rho_t)$ such that $t \mapsto \rho_t$ is narrowly continuous with endpoints ρ_0, ρ_1, and

$$\partial_t \rho_t = \rho_t r_t$$

in the sense of distributions $\mathcal{D}'((0,1) \times \Omega)$.

As in Remark 2.1 the reaction term r implicitly belongs to the energy space $L^2(0,1; L^2(\mathrm{d}\rho_t))$, so that ρr is a well-defined distribution $\mathcal{D}'((0,1) \times \Omega)$ through the $(r, .)_{L^2(0,1;L^2(\mathrm{d}\rho_t))}$ scalar product. In the last explicit formula λ is any reference measure such that ρ_0, ρ_1 are both absolutely continuous with respect to λ, with Radon-Nikodym derivatives $\frac{\mathrm{d}\rho_i}{\mathrm{d}\lambda}$. By 1-homogeneity this expression doe not depend on the choice of λ, and the normalizing factor 4 is chosen so that the metric for the pivot space in the first dynamic formulation is exactly $L^2(\mathrm{d}\rho_t)$ and not some other multiple $\beta L^2(\mathrm{d}\rho_t)$.

At least for absolutely continuous measures $\mathrm{d}\rho_0, \mathrm{d}\rho_1 \ll \mathrm{d}x$ one can check that the minimum in the first definition is attained along the geodesic

$$\rho_t = [(1-t)\sqrt{\rho_0} + t\sqrt{\rho_1}]^2 \quad \text{and} \quad r_t := 2\frac{\sqrt{\rho_1} - \sqrt{\rho_0}}{\sqrt{\rho_t}} \in L^2(\mathrm{d}\rho_t).$$

Moreover this optimal curve $\partial_t \rho_t = \rho_t r_t$ has constant metric speed $\|r_t\|^2_{L^2(\mathrm{d}\rho_t)} = 4\int_\Omega |\sqrt{\rho_1} - \sqrt{\rho_0}|^2 = \mathtt{FR}^2(\rho_0, \rho_1)$, which should be expected for geodesics.

More importantly, the first Lagrangian formulation in (8) suggests to view the metric space $(\mathcal{M}^+, \mathtt{FR})$ as a Riemannian manifold, endowing the tangent plane

$$T_\rho \mathcal{M}^+_{\mathtt{FR}} = \left\{ \partial_t \rho_t = \rho_t r_t \qquad \text{evaluated at } t = 0 \right\}$$

with the metrics

$$\|\partial_t \rho\|^2_{T_\rho \mathcal{M}^+_{\mathtt{FR}}} = \|r\|^2_{L^2(\mathrm{d}\rho)} \quad \text{with the identification } \partial_t \rho = \rho r.$$

Metric gradients $\operatorname{grad}_{\mathtt{FR}}$ can then be computed by the chain rule as follows: If $\partial_t \rho_t = \rho_t r_t$ is a smooth curve passing through $\rho_t(0) = \rho$ with arbitrary initial velocity $\zeta = \partial_t \rho = \rho r$ then

for functionals $\mathcal{F}(\rho) = \int_\Omega F(\rho(x), x)\mathrm{d}x$ we can compute

$$\langle \operatorname{grad} \mathcal{F}(\rho), \zeta \rangle_{T_\rho \mathcal{M}^+_{\mathtt{FR}}} = \frac{d}{dt}\mathcal{F}(\rho_t)\Big|_{t=0} = \frac{d}{dt}\left(\int_\Omega F(\rho_t(x), x)\mathrm{d}x\right)\Big|_{t=0}$$
$$= \int_\Omega F'(\rho)\rho r = \langle F'(\rho), r\rangle_{L^2(\mathrm{d}\rho)},$$

where $F'(\rho) = \frac{\delta F}{\delta \rho}$ as before. This shows that

$$\operatorname{grad}_{\mathtt{FR}} \mathcal{F}(\rho) = \rho F'(\rho) \tag{9}$$

with identification through the $L^2(\mathrm{d}\rho)$ action in the tangent plane, and as a consequence gradients flows with respect to the Hellinger-Fisher-Rao metrics read

$$\partial_t \rho = -\operatorname{grad}_{\mathtt{FR}} \mathcal{F}(\rho) \qquad \leftrightarrow \qquad \partial_t \rho = -\rho F'(\rho). \tag{10}$$

2.3. The Kantorovich-Fisher-Rao distance KFR. As introduced in [14], we have

Definition 2.3. *The Fisher-Rao-Hellinger-Kantorovich-Wasserstein distance between measures $\rho_0, \rho_1 \in \mathcal{M}^+(\Omega)$ is*

$$\mathtt{KFR}^2(\rho_0, \rho_1) = \inf_{(\rho, \mathbf{v}, r) \in \mathcal{A}_{\mathtt{KFR}}[\rho_0, \rho_1]} \int_0^1 \int_\Omega (|\mathbf{v}_t(x)|^2 + |r_t(x)|^2)\mathrm{d}\rho_t(x)\,\mathrm{d}t \tag{11}$$

The admissible set $\mathcal{A}_{\mathtt{KFR}}[\rho_0, \rho_1]$ is the set of curves $[0,1] \ni t \mapsto (\rho_t, \mathbf{v}_t, r_t) \in \mathcal{M}^+(\Omega) \times L^2(\Omega, \mathrm{d}\rho_t)^d \times L^2(\Omega, \mathrm{d}\rho_t)$ such that $t \mapsto \rho_t$ is narrowly continuous with endpoints ρ_0, ρ_1 and solves the continuity equation with source

$$\partial_t \rho_t + \operatorname{div}(\rho_t \mathbf{v}_t) = \rho_t r_t$$

in the sense of distributions $\mathcal{D}'((0,1)\times\Omega)$.

As in Remark 2.1 the velocity fields and reaction term implicitly belong to the energy space $L^2(0,1; L^2(\mathrm{d}\rho_t))$, so that both products $\rho\mathbf{v}, \rho r$ are well-defined as distributions $\mathcal{D}'((0,1)\times\Omega)$. Comparing (11) with (6) and (8), this dynamic formulation *à la Benamou-Brenier* [4] shows that the KFR distance can be viewed as an inf-convolution of the Monge-Kantorovich and Fisher-Rao distances MK, FR. By the results of [14, 13, 28] the infimum in the definition is always a minimum, and the corresponding minimizing curves $t \mapsto \rho_t$ are of course called geodesics. As shown in [25, 14, 28] geodesics need not be unique, see also the brief discussion in section 4. Interestingly, there are other possible formulations of the distance in terms of static unbalanced optimal transportation, primal-dual characterizations with relaxed marginals, lifting to probability measures on a cone over Ω, and duality with subsolutions of Hamilton-Jacobi equations. See also [28, 29] as well as [37] for a related version with mass penalization.

As an immediate consequence of the definition 11 we have a first interplay between the distances KFR, MK, FR:

Proposition 2.1. *Let $\rho_0, \rho_1 \in \mathcal{M}^+_2$ such that $|\rho_0| = |\rho_1|$. Then*

$$\mathtt{KFR}^2(\rho_0, \rho_1) \le \mathtt{MK}^2(\rho_0, \rho_1).$$

Similarly for all $\mu_0, \mu_1 \in \mathcal{M}^+$ (with possibly different masses) there holds

$$\mathtt{KFR}^2(\mu_0, \mu_1) \le \mathtt{FR}^2(\mu_0, \mu_1).$$

Proof. If $|\rho_0| = |\rho_1|$ then the optimal Monge-Kantorovich geodesics $\partial_t \rho_t + \operatorname{div}(\rho_t \mathbf{v}_t) = 0$ from ρ_0 to ρ_1 gives an admissible path in (11) with $r \equiv 0$ and cost exactly $\mathtt{MK}^2(\rho_0, \rho_1)$. Likewise for arbitrary measures μ_0, μ_1 one can follow the Fisher-Rao geodesics $\partial_r \rho_t = \rho_t r_t$, which gives an admissible path with $\mathbf{v} \equiv 0$ and cost $\mathtt{FR}^2(\mu_0, \mu_1)$. ∎

PROPOSITION 2.2. *The definition* (11) *of the* KFR *distance can be restricted to the subclass of admissible paths* $(\mathbf{v}_t, r_t)$ *such that* $\mathbf{v}_t = \nabla r_t$.

Proof. By [14, thm. 2.1] there exists a minimizing curve $(\rho_t, \mathbf{v}_t, r_t)$ in (11), which by definition is a KFR-geodesic between ρ_0, ρ_1 (we also refer to [25, thm. 6] and [29] for the existence of geodesics). Here we stay at the formal level and assume that $\rho, \mathbf{v}, r$ are smooth with $\rho > 0$ everywhere.

Consider first an arbitrary smooth vector-field $\mathbf{w}$ such that $\operatorname{div} \mathbf{w}_t = 0$ for all $t \in [0,1]$, and let $\mathbf{v}^\varepsilon := \mathbf{v} + \varepsilon \frac{\mathbf{w}}{\rho}$. Then $\operatorname{div}(\rho \mathbf{v}^\varepsilon) = \operatorname{div}(\rho \mathbf{v}) + 0$ and the triplet $(\rho_t, \mathbf{v}^\varepsilon_t, r_t)$ is an admissible competitor in (11). Writing the optimality condition we compute

$$\begin{aligned} 0 &= \frac{d}{d\varepsilon}\left(\frac{1}{2}\int_0^1 \int_\Omega (|\mathbf{v}^\varepsilon_t(x)|^2 + |r_t(x)|^2) \mathrm{d}\rho_t(x)\,\mathrm{d}t\right)\Big|_{\varepsilon=0} \\ &= \int_0^1 \int_\Omega \mathbf{v}_t(x) \cdot \frac{\mathbf{w}_t(x)}{\rho_t(x)} \mathrm{d}\rho_t(x)\,\mathrm{d}t = \int_0^1 \int_\Omega \mathbf{v}_t(x) \cdot \mathbf{w}_t(x)\,\mathrm{d}x\mathrm{d}t. \end{aligned}$$

This L^2 orthogonality with all divergence-free vector fields classically implies that $\mathbf{v}_t$ is potential for all times, i-e $\mathbf{v}_t = \nabla u_t$ for some u_t.

Fix now any smooth $\phi \in \mathcal{C}^\infty_c((0,1)\times\Omega)$, and define $\tilde{\mathbf{v}}^\varepsilon_t := \mathbf{v}_t + \varepsilon \nabla \phi_t = \nabla(u_t + \varepsilon \phi_t)$. Defining s_t by $\rho_t s_t = \operatorname{div}(\rho_t \nabla \phi_t)$ and $\tilde{r}^\varepsilon_t := r_t + \varepsilon s_t$ it is easy to check that $(\rho_t, \tilde{\mathbf{v}}^\varepsilon_t, \tilde{r}^\varepsilon_t)$ solves the continuity equation, and this triplet is again an admissible competitor in (11). Writing the optimality condition we get now

$$\begin{aligned} 0 = \frac{d}{d\varepsilon}\left(\frac{1}{2}\int_0^1 \int_\Omega (|\tilde{\mathbf{v}}^\varepsilon_t(x)|^2 + |\tilde{r}^\varepsilon_t(x)|^2) \mathrm{d}\rho_t(x)\,\mathrm{d}t\right)\Big|_{\varepsilon=0} \\ = \int_0^1 \int_\Omega \Big(\nabla u_t(x) \cdot \nabla \phi_t + r_t(x) s_t(x)\Big) \mathrm{d}\rho_t(x)\,\mathrm{d}t \\ = \int_0^1 \int_\Omega \nabla\Big(u_t - r_t\Big)(x) \cdot \nabla \phi_t \mathrm{d}\rho_t(x)\,\mathrm{d}t, \end{aligned}$$

where we used the identity $r_t s_t \rho_t = r_t \operatorname{div}(\rho_t \nabla \phi_t)$ to integrate by parts in the last equality. As ϕ was arbitrary this implies $\operatorname{div}(\rho_t \nabla u_t) = \operatorname{div}(\rho_t \nabla r_t)$ and $\|\mathbf{v}_t\|^2_{L^2(\mathrm{d}\rho_t)} = \|\nabla u_t\|^2_{L^2(\mathrm{d}\rho_t)} = \|\nabla r_t\|^2_{L^2(\mathrm{d}\rho_t)}$. In particular the triplet $(\rho_t, \nabla r_t, r_t)$ is admissible and has the same cost as the optimal $(\rho_t, \mathbf{v}_t, r_t)$, which concludes the proof. ∎

As a consequence we have the alternative definition of the KFR distance as introduced in [25], which couples the reaction and velocity:

THEOREM 2.3. *For all* $\rho_0, \rho_1 \in \mathcal{M}^+(\Omega)$ *there holds*

$$\mathtt{KFR}^2(\rho_0, \rho_1) = \inf_{(\rho,u)\in \tilde{A}_{\mathtt{KFR}}[\rho_0,\rho_1]} \int_0^1 \int_\Omega (|\nabla u_t(x)|^2 + |u_t(x)|^2) \mathrm{d}\rho_t(x)\,\mathrm{d}t, \tag{12}$$

where $\tilde{A}_{\mathtt{KFR}}[\rho_0, \rho_1]$ *is the set of curves* $[0,1] \ni t \mapsto (\rho_t, \nabla u_t, u_t) \in \mathcal{M}^+(\Omega) \times L^2(\Omega, \mathrm{d}\rho_t)^d \times L^2(\Omega, \mathrm{d}\rho_t)$ *such that* $t \mapsto \rho_t$ *is narrowly continuous with endpoints* ρ_0, ρ_1 *and solves*

$$\partial_t \rho_t + \operatorname{div}(\rho_t \nabla u_t) = \rho_t u_t$$

in the sense of distributions $\mathcal{D}'((0,1)\times\Omega)$.

The potentials u belong now implicitly to the energy space $L^2(0,1; H^1(\mathrm{d}\rho_t))$ with obviously $\|u_t\|^2_{H^1(\mathrm{d}\rho)} := \int_\Omega (|\nabla u_t|^2 + |u_t|^2) \mathrm{d}\rho_t$, and both products $\rho_t \nabla u_t, \rho_t u_t$ define distributions as before. Note that Theorem 2.3 shows that the KFR distance constructed in [14], based on the uncoupled $(\mathbf{v}, r)$ formulation, is indeed the same as that in [25], modeled on the $(\nabla u, u)$ potential framework.

In order to define now the Riemannian structure on $(\mathcal{M}^+, \texttt{KFR})$ inherited from the Lagrangian minimization, we endow the tangent plane

$$T_\rho \mathcal{M}^+_{\texttt{KFR}} = \Big\{ \partial_t \rho = -\operatorname{div}(\rho v) + \rho r \quad \text{evaluated at } t = 0 \Big\}$$

with the Riemannian metrics

$$\|\partial_t \rho\|^2_{T_\rho \mathcal{M}^+_{\texttt{KFR}}} := \inf \Big\{ \|\mathbf{v}\|^2_{L^2(\mathrm{d}\rho)} + \|r\|^2_{L^2(\mathrm{d}\rho)} : \quad \partial_t \rho = -\operatorname{div}(\rho \mathbf{v}) + \rho r \Big\}.$$

Then Theorem 2.3 also allows to construct the one-to-one correspondence between tangent vectors $\partial_t \rho$ and potentials u, such that

$$\|\partial_t \rho\|^2_{T_\rho \mathcal{M}^+_{\texttt{KFR}}} = \|u\|^2_{H^1(\mathrm{d}\rho)} \quad \text{with the identification } \partial_t \rho = -\operatorname{div}(\rho \nabla u) + \rho u.$$

With this one-to-one correspondence at hand, metric gradients $\operatorname{grad}_{\texttt{KFR}} \mathcal{F}$ can be computed by the chain rule as earlier: If $\partial_t \rho_t + \operatorname{div}(\rho_t \nabla u_t) = \rho_t u_t$ is a smooth curve passing through $\rho_t(0) = \rho$ with arbitrary initial velocity $\zeta = \partial_t \rho_t(0) = -\operatorname{div}(\rho \nabla u) + \rho u$ then for functionals $\mathcal{F}(\rho) = \int_\Omega F(\rho(x), x) \mathrm{d}x$ we have

$$\begin{aligned}
\langle \operatorname{grad}_{\texttt{KFR}} \mathcal{F}(\rho), \zeta \rangle_{T_\rho \mathcal{M}^+_{\texttt{KFR}}} &= \frac{d}{dt} \mathcal{F}(\rho_t)\Big|_{t=0} = \frac{d}{dt} \left(\int_\Omega F(\rho_t(x), x) \mathrm{d}x \right)\Big|_{t=0} \\
&= \int_\Omega F'(\rho) \times \{ -\operatorname{div}(\rho \nabla u) + \rho u \} \\
&= \int_\Omega \{ \nabla F'(\rho) \cdot \nabla u + F'(\rho) u \} \, \mathrm{d}\rho = \langle F'(\rho), u \rangle_{H^1(\mathrm{d}\rho)},
\end{aligned}$$

where $F'(\rho) = \frac{\delta F}{\delta \rho}$ as before. This shows that

$$\operatorname{grad}_{\texttt{KFR}} \mathcal{F}(\rho) = -\operatorname{div}\left(\rho \nabla F'(\rho)\right) + \rho F'(\rho)$$

through the canonical $H^1(\mathrm{d}\rho)$ action in the tangent plane. In particular KFR gradient flows read

$$\partial_t \rho = -\operatorname{grad}_{\texttt{KFR}} \mathcal{F}(\rho) \qquad \leftrightarrow \qquad \partial_t \rho = \operatorname{div}(\rho \nabla F'(\rho)) - \rho F'(\rho), \tag{13}$$

which should be compared with (7) and (10).

3. Infinitesimal uncoupling of the inf-convolution. Let us first summarize the previous informal discussion on each of the three metrics: the quadratic Monge-Kantorovich distance is modeled on the homogeneous $\dot{H}^1(\mathrm{d}\rho)$ space, the Fisher-Rao distance is based on $L^2(\mathrm{d}\rho)$, and the KFR metrics is constructed on the full $H^1(\mathrm{d}\rho)$ structure. Each of these Riemannian structures are defined via identification of tangent vectors as

$$\begin{array}{lll}
\texttt{MK}: & \|\partial_t \rho\|^2_{T_\rho \mathcal{M}^+_{\texttt{MK}}} = \|\nabla p\|^2_{L^2(\mathrm{d}\rho)} = \int_\Omega |\nabla p|^2 \mathrm{d}\rho, & \partial_t \rho + \operatorname{div}(\rho \nabla p) = 0, \\
\texttt{FR}: & \|\partial_t \rho\|^2_{T_\rho \mathcal{M}^+_{\texttt{FR}}} = \|r\|^2_{L^2(\mathrm{d}\rho)} = \int_\Omega |r|^2 \mathrm{d}\rho, & \partial_t \rho = \rho r, \\
\texttt{KFR}: & \|\partial_t \rho\|^2_{T_\rho \mathcal{M}^+_{\texttt{KFR}}} = \|u\|^2_{H^1(\mathrm{d}\rho)} = \int_\Omega (|\nabla u|^2 + u^2) \mathrm{d}\rho, & \partial_t \rho + \operatorname{div}(\rho \nabla u) = \rho u.
\end{array}$$

Given a tangent vector $\zeta^u_{\texttt{KFR}} = -\operatorname{div}(\rho \nabla u) + \rho u \in T_\rho \mathcal{M}^+_{\texttt{KFR}}$ we can naturally define a Monge-Kantorovich tangent vector $\zeta^u_{\texttt{MK}} := -\operatorname{div}(\rho \nabla u) \in T_\rho \mathcal{M}^+_{\texttt{MK}}$, and a Fisher-Rao tangent vector $\zeta^u_{\texttt{FR}} := \rho u \in T_{\texttt{FR}} \mathcal{M}^+_{\texttt{FR}}$. Observing that by construction

$$\|\zeta^u_{\texttt{KFR}}\|^2_{T_\rho \mathcal{M}^+_{\texttt{KFR}}} = \|\zeta^u_{\texttt{MK}}\|^2_{T_\rho \mathcal{M}^+_{\texttt{MK}}} + \|\zeta^u_{\texttt{FR}}\|^2_{T_\rho \mathcal{M}^+_{\texttt{FR}}}, \tag{14}$$

this suggests to view the tangent plane as the orthogonal sum

$$T_\rho \mathcal{M}^+_{\texttt{KFR}} = T_\rho \mathcal{M}^+_{\texttt{MK}} \oplus^\perp T_\rho \mathcal{M}^+_{\texttt{FR}}, \qquad \zeta^u_{\texttt{KFR}} = \zeta^u_{\texttt{MK}} + \zeta^u_{\texttt{FR}}. \tag{15}$$

More precisely, let us define an equivalence relation $\sim$ on $T_\rho\mathcal{M}^+_{\mathtt{MK}} \oplus T_\rho\mathcal{M}^+_{\mathtt{FR}}$ by $(\mathbf{v}, r) \sim (\tilde{\mathbf{v}}, \tilde{r})$ if $-\operatorname{div}(\rho\mathbf{v}) + \rho r = -\operatorname{div}(\rho\tilde{\mathbf{v}}) + \rho\tilde{r}$. Each $(\mathbf{v}, r)$ lies in an equivalence class $[(\nabla u, u)] = [u]$ on which we define the norm

$$\|[u]\|^2_\sim = \|\nabla u\|^2_{L^2(\mathrm{d}\rho)} + \|u\|^2_{L^2(\mathrm{d}\rho)} = \|\zeta^u_{\mathtt{MK}}\|^2_{T_\rho\mathcal{M}^+_{\mathtt{MK}}} + \|\zeta^u_{\mathtt{FR}}\|^2_{T_\rho\mathcal{M}^+_{\mathtt{FR}}}.$$

Then the orthogonality in (14) should be understood as

$$\Big(T_\rho\mathcal{M}^+_{\mathtt{KFR}}, \|\cdot\|^2_{T_\rho\mathcal{M}^+_{\mathtt{KFR}}}\Big) = \Big(\big(T_\rho\mathcal{M}^+_{\mathtt{MK}} \oplus T_\rho\mathcal{M}^+_{\mathtt{FR}}\big) / \sim, \|\cdot\|^2_\sim\Big).$$

Thus infinitesimally $\mathtt{KFR}^2 \approx \mathtt{MK}^2 + \mathtt{FR}^2$, and this will motivate later on replacing the approximation "by hypotenuses" by an approximation "by legs" in the JKO scheme - see section 4 and in particular (23)(24). The orthogonality between the transport/$\mathtt{MK}$ and reaction/$\mathtt{FR}$ processes also yields a natural strategy to send a measure ρ_0 to another ρ_1: one can send first ρ_0 to the renormalized $\tilde{\rho}_0 := \frac{|\rho_0|}{|\rho_1|}\rho_1$ by pure Monge-Kantorovich transport (which is possible since $|\tilde{\rho}_0| = |\rho_0|$), and then send $\tilde{\rho}_0$ to ρ_1 by pure Fisher-Rao reaction. This amounts to following separately and successively the two orthogonal directions in the decomposition (15).

An immediate consequence of this observation is

PROPOSITION 3.1. *For arbitrary measures $\rho_0, \rho_1 \in \mathcal{M}^+$ let $\tilde{\rho}_0 := \frac{|\rho_0|}{|\rho_1|}\rho_1$. Then*

$$\mathtt{KFR}^2(\rho_0, \rho_1) \le 2\big(\mathtt{MK}^2(\rho_0, \tilde{\rho}_0) + \mathtt{FR}^2(\tilde{\rho}_0, \rho_1)\big). \tag{16}$$

Proof. It suffices to follow first a pure Monge-Kantorovich geodesics ($r \equiv 0$) from ρ_0 to $\tilde{\rho}_0$ scaled in time $t \in [0, 1/2]$, and then a pure Fisher-Rao geodesic ($\mathbf{v} \equiv 0$) from $\tilde{\rho}_0$ to ρ_1 scaled in time $t \in [1/2, 1]$. Because of the rescaling in time each of these half-paths have an extra factor 2, amounting to a total cost of $2\mathtt{MK}^2(\rho_0, \tilde{\rho}_0) + 2\mathtt{FR}^2(\tilde{\rho}_0, \rho_1)$ for this admissible path. The result then follows from the definition (11) of $\mathtt{KFR}^2$ as an infimum over all paths. □

Note that estimate (16) holds for any arbitrary measure $\rho_0, \rho_1 \in \mathcal{M}^+$, but has a multiplicative factor 2 which in view of (14)(15) is certainly not optimal at short range $\mathtt{KFR}(\rho_0, \rho_1) \ll 1$. Consider now two very close measures $\mathtt{KFR}(\rho_0, \rho_1) \ll 1$. Then the above transformation from ρ_0 to ρ_1 can essentially be considered as occurring infinitesimally in the tangent plane $T_\rho\mathcal{M}^+_{\mathtt{KFR}} = T_\rho\mathcal{M}^+_{\mathtt{MK}} \oplus^\perp T_\rho\mathcal{M}^+_{\mathtt{FR}}$. Roughly speaking, this means that the two transport and reaction processes from ρ_0 to $\tilde{\rho}_0$ and from $\tilde{\rho}_0$ to ρ_1 in the previous proof can be considered as occurring *simultaneously and independently* at the infinitesimal level. Thus the factor 2 in (16) is unnecessary, and one should expect in fact

$$\mathtt{KFR}^2(\rho_0, \rho_1) \approx \mathtt{MK}^2(\rho_0, \tilde{\rho}_0) + \mathtt{FR}^2(\tilde{\rho}_0, \rho_1) \tag{17}$$

for nearby measures $\mathtt{KFR}(\rho_0, \rho_1) \ll 1$. This can be made rigorous at least for one-point particles

$$\rho_0 = k_0\delta_{x_0}, \qquad \rho_1 = k_1\delta_{x_1}$$

at close distance, i-e $|x_1 - x_0| \ll 1$ and $k_1 \approx k_0$. In this setting it was shown in [25, Section 3.3] and proved rigorously [14, thm. 4.1] and [29, thm. 3.1] that the geodesics ρ_t from ρ_0 to ρ_1 is a moving one-point mass of the form $\rho_t = k_t\delta_{x_t}$ for some suitable curve $t \mapsto (x_t, k_t) \in \Omega \times \mathbb{R}^+$.

REMARK 3.1. *The one-point ansatz $\rho_t = k_t\delta_{x_t}$ is in fact correct not only for short distances $|x_1 - x_0| \ll 1$, but also as long as $|x_1 - x_0| < \pi$. Past this threshold $|x_1 - x_0| \ge \pi$ it is more efficient to virtually displace mass from x_0 to x_1 by pure reaction, i-e by killing mass at x_0 while simultaneously creating some at x_1.*

In the continuity equation $\partial_t\rho_t + \operatorname{div}(\rho_t\mathbf{v}_t) = \rho_t r_t$ the advection moves particles around according to $\frac{d}{dt}x_t = \mathbf{v}_t$ and the reaction reads $\frac{d}{dt}k_t = k_t r_t$, each with infinitesimal cost $k_t|\mathbf{v}_t|^2$ and $k_t|r_t|^2$. The optimal $(\mathbf{v}_t, r_t)$ for the one-point ansatz $\rho_t = k_t\delta_{x_t}$ can be computed

explicitly by looking at the coupled formulation (12) with $\mathbf{v}_t = \nabla u_t, r_t = u_t$, and optimizing the cost with respect to admissible potentials u_t. Omitting the details (see again [13, 14, 25, 28, 29]), the optimal cost can be computed explicitly as

$$\texttt{KFR}^2(\rho_0,\rho_1) = 4\left(k_0+k_1-2\sqrt{k_0k_1}\cos\left(\frac{|x_1-x_0|}{2}\right)\right) \qquad \text{for } \begin{cases} \rho_i = k_i\delta_{x_i} \\ |x_1-x_0|<\pi. \end{cases} \tag{18}$$

REMARK 3.2. *It was shown in [13, 28, 29] that the* KFR *distance can be recovered by means of a suitable Riemannian submersion* $(\mathcal{P}_2(C_\Omega),\texttt{MK}) \to (\mathcal{M}^+(\Omega),\texttt{KFR})$. *Here* $\mathrm{C}_\Omega = \{[x,r]\in\Omega\times\mathbb{R}^+\}/\sim$ *is a cone overlying* Ω *obtained by identification of all the tips* $[x,0]$ *into a single point* $\diamond\in C_\Omega$, *and is suitably endowed with the cone distance* $d_C^2([x_0,r_0],[x_1,r_1]) = r_0^2+r_1^2-2r_0r_1\cos(|x_1-x_0|/2\wedge\pi)$. *In formula (18) one sees in fact, up to the normalizing factor* 4, *the natural Monge-Kantorovich distance* $\texttt{KFR}^2\left(\delta_{[x_0,k_0]},\delta_{[x_1,k_1]}\right) = \texttt{MK}^2\left(\delta_{[x_0,\sqrt{k_0}]},\delta_{[x_1,\sqrt{k_1}]}\right) = d_C^2([x_0,\sqrt{k_0}],[x_1,\sqrt{k_1}])$ *between unit Dirac masses in the overlying space* $\mathcal{P}_2(C_\Omega)$. *We refrain from discussing further the Riemannian submersion and the corresponding static formulations of* KFR, *and refer again to [13, 28, 29, 18] for rigorous statements.*

In this setting and with the previous notation $\tilde\rho_0 = \frac{|\rho_0|}{|\rho_1|}\rho_1 = k_0\delta_{x_1}$ we have here

$$\texttt{MK}^2(\rho_0,\tilde\rho_0) = \texttt{MK}^2(k_0\delta_{x_0},k_0\delta_{x_1}) = k_0|x_1-x_0|^2,$$

and by (8)

$$\texttt{FR}^2(\tilde\rho_0,\rho_1) = \texttt{FR}^2(k_0\delta_{x_1},k_1\delta_{x_1}) = 4\int_\Omega\left|\sqrt{\frac{\mathrm{d}\rho_1}{\mathrm{d}\delta_{x_1}}}-\sqrt{\frac{\mathrm{d}\tilde\rho_0}{\mathrm{d}\delta_{x_1}}}\right|^2\mathrm{d}\delta_{x_1} = 4\left|\sqrt{k_1}-\sqrt{k_0}\right|^2.$$

Taylor-expanding (18) at order two in $|x_1-x_0|, |\sqrt{k_1}-\sqrt{k_0}| \ll 1$ gives

$$\begin{aligned}\texttt{KFR}^2(\rho_0,\rho_1) &= k_0|x_1-x_0|^2+4|\sqrt{k_1}-\sqrt{k_0}|^2+\mathcal{O}\Big(|x_1-x_0|^2|\sqrt{k_1}-\sqrt{k_0}|\Big) \\ &= \texttt{MK}^2(\rho_0,\tilde\rho_0)+\texttt{FR}^2(\tilde\rho_0,\rho_1)+\text{lower order},\end{aligned} \tag{19}$$

which shows that our claim (17) holds true at least for one-point particles and at order one in the squared distances.

REMARK 3.3. *Due to* $4|\sqrt{k_1}-\sqrt{k_0}|^2 = \texttt{FR}^2(\tilde\rho_0,\rho_1) \ll 1$ *we have* $k_1 = k_0+\mathcal{O}(|\sqrt{k_1}-\sqrt{k_0}|)$. *The previous expression can therefore be rewritten as*

$$\texttt{KFR}^2(\rho_0,\rho_1) = \frac{k_0+k_1}{2}|x_1-x_0|^2+4|\sqrt{k_1}-\sqrt{k_0}|^2+\textit{lower order}$$

and the apparent loss of symmetry in k_0,k_1 *in (19) is thus purely artificial.*

REMARK 3.4. *An interesting question would be to determine how much information on the transport/reaction coupling is encoded in the remainder, and this is also related to the curvature of the* KFR *space.*

Justifying and/or quantifying the above discussion and (17) for general measures with $\texttt{KFR}(\rho_0,\rho_1) \ll 1$ is an interesting question left for future work. One can think that the superposition principle should apply: viewing any measure as a continuum of one-point Lagrangian particles and taking for granted that the infinitesimal uncoupling holds for single particles, it seems natural that the result should also hold for all measures.

4. Minimizing scheme. We turn now our attention to gradient-flows

$$\partial_t\rho = -\operatorname{grad}_{\texttt{KFR}}\mathcal{F}(\rho) \tag{20}$$

of functionals

$$\mathcal{F}(\rho) = \begin{cases} \int_\Omega \left\{ U(\rho) + \Psi(x)\rho + \frac{1}{2}\rho K \star \rho \right\} \mathrm{d}x & \text{if } \mathrm{d}\rho \ll \mathrm{d}x \\ \infty & \text{otherwise} \end{cases}$$

with respect to the KFR distance. Without further mention we implicitly restrict to absolutely continuous measures (with respect to Lebesgue), and still denote their Radon-Nikodym derivatives $\rho = \frac{\mathrm{d}\rho}{\mathrm{d}x}$ with a slight abuse of notations. According to (13) this corresponds to PDEs of the form

$$\partial_t \rho = \operatorname{div}(\rho \nabla (U'(\rho) + \Psi + K \star \rho)) - \rho (U'(\rho) + \Psi + K \star \rho), \tag{21}$$

appearing for example in the tumor growth model studied in [35].

The natural minimizing movement for (20) should be

$$\rho^{n+1} \in \operatorname*{Argmin}_{\rho \in \mathcal{M}^+} \left\{ \frac{1}{2\tau} \mathtt{KFR}^2(\rho, \rho^n) + \mathcal{F}(\rho) \right\} \tag{22}$$

for some small time step $\tau > 0$. In order to obtain an Euler-Lagrange equation, a classical and natural strategy would be to consider perturbations $\varepsilon \mapsto \rho_\varepsilon$ of the minimizer $\rho_\varepsilon(0) = \rho^{n+1}$ starting with velocity $\partial_\varepsilon \rho(0) = -\operatorname{div}(\rho^{n+1} \nabla \phi) + \rho^{n+1}\phi$ for any arbitrary smooth ϕ, corresponding to choosing all possible directions of perturbation in the tangent plane $T_{\rho^{n+1}} \mathcal{M}^+_{\mathtt{KFR}}$. Writing down the optimality criterion $\frac{d}{d\varepsilon}\left(\frac{1}{2\tau} \mathtt{KFR}^2(\rho_\varepsilon, \rho^n) + \mathcal{F}(\rho_\varepsilon) \right)\big|_{\varepsilon=0} = 0$ should then give the sought Euler-Lagrange equation. In order to exploit this, one should in particular know how to differentiate the squared distance $\rho \mapsto \mathtt{KFR}^2(\rho, \mu)$ with respect to such perturbations ρ_ε of the minimizer. At this stage the theory does not provide yet the necessary tools, even though what the formula should read is quite clear: For any reasonably smooth Riemannian manifold and smooth curve $x(t)$ with $x(0) = x$ we have

$$\frac{d}{dt}\left(\frac{1}{2} d^2(x(t), y) \right)\Big|_{t=0} = \langle x'(0), \zeta \rangle_{T_x \mathcal{M}},$$

where ζ is the terminal velocity $y'(1) \in T_x \mathcal{M}$ of the geodesics from y to x. Here the KFR-geodesics $(\mu_s)_{s \in [0,1]}$ from ρ^n to ρ^{n+1} should solve $\partial_s \mu_s + \operatorname{div}(\mu_s \nabla u_s) = \mu_s u_s$ and the terminal velocity $\zeta = \partial_s \mu(1) \in T_{\rho^{n+1}} \mathcal{M}^+_{\mathtt{KFR}}$ should be identified with some potential $u^{n+1} = u|_{s=1} \in H^1(\mathrm{d}\rho^{n+1})$ through $\zeta = -\operatorname{div}(\rho^{n+1} \nabla u^{n+1}) + \rho^{n+1} u^{n+1}$, see section 2.3. We should therefore expect

$$\frac{d}{d\varepsilon}\left(\frac{1}{2} \mathtt{KFR}^2(\rho_\varepsilon, \rho^n) \right)\Big|_{\varepsilon=0} = \langle \partial_\varepsilon \rho(0), \zeta \rangle_{T_{\rho^{n+1}} \mathcal{M}^+_{\mathtt{KFR}}} = \int_\Omega (\nabla \phi \cdot \nabla u^{n+1} + \phi u^{n+1}) \mathrm{d}\rho^{n+1}.$$

However, this can raise delicate technical issues at the cut-locus, where geodesics cease to be minimizing and prevent any differentiability of the squared distance. Indeed, it was shown in [29, section 5.2], [14, thm. 4.1], and [25, section 3.5] that such cut-loci do exist for $\Omega = \mathbb{R}^d$, and even that the set of non-unique geodesics generically spans an infinite-dimensional convex set. This is related to the threshold $|x_1 - x_0| = \pi$ for one-point measures, see Remark 3.1. In other words the squared distance may very well not be differentiable, even in the case of the simplest geometry $\Omega = \mathbb{R}^d$ of the underlying space. This is in sharp contrast with classical mass conservative optimal transportation, where the cut-locus in $\mathcal{P}(X)$ is intimately related to the geometry of the underlying Riemannian manifold X [42].

In the context of minimizing movements one should expect two successive steps to be extremely close, typically $\mathtt{KFR}(\rho^{n+1}, \rho^n) = \mathcal{O}(\sqrt{\tau})$ as $\tau \to 0$. It seems reasonable to hope that geodesics then become unique at short distance, and one might therefore think that the previous cut-locus issue should not arise here for small $\tau > 0$. However, even assuming that we could somehow compute a unique minimizing geodesics $(\rho_s)_{s \in [0,1]}$ from ρ^n to ρ^{n+1} and safely evaluate the terminal velocity $\partial_s \rho(1) = -\operatorname{div}(\rho^{n+1} \nabla u^{n+1}) + \rho^{n+1} u^{n+1}$

at $s = 1$ in order to differentiate the squared distance, it would remain to derive a (possibly approximated) relation between the Riemannian point of view and the more classical PDE framework, e.g. by proving an estimate like

$$\int_\Omega (\nabla u^{n+1} \cdot \nabla \phi + u^{n+1}\phi) \mathrm{d}\rho^{n+1} \approx \int_\Omega \frac{\rho^{n+1} - \rho^n}{\tau}\phi + \text{remainder}.$$

In this last display we see the interplay between the forward tangent vector $u^{n+1} \in H^1(\mathrm{d}\rho^{n+1}) \cong T_{\rho^{n+1}}\mathcal{M}^+_{\texttt{KFR}}$, encoding the Riemannian variation from ρ^n to ρ^{n+1}, and the standard difference quotient $\frac{\rho^{n+1}-\rho^n}{\tau} \approx \partial_t \rho$. One should then typically prove that the remainder is quadratic $\mathcal{O}\left(\texttt{KFR}^2(\rho^{n+1}, \rho^n)\right)$. Within the framework of classical optimal transport this is usually done exploiting the explicit representation of the MK metrics in terms of optimal transport maps (or transference plans, or Kantorovich potentials), which are in turn related to some static formulations of the problem. See later on section 4.1 and in particular the Taylor expansion (32) for details, and also remark 4.2. However, and even though static formulations of the KFR distance have been derived in [28], the current theory does not provide yet such an asymptotic expansion.

In order to circumvent these technical issues, let us recall from the discussion in section 3 that the inf-convolution formally uncouples at short distance. This strongly suggests replacing $\texttt{KFR}^2$ by the approximation $\texttt{MK}^2 + \texttt{FR}^2 \approx \texttt{KFR}^2$, and as a consequence we naturally substitute the direct one-step minimizing scheme (22) by a sequence of two elementary substeps

$$\rho^n \xrightarrow{\texttt{MK}^2} \rho^{n+\frac{1}{2}} \xrightarrow{\texttt{FR}^2} \rho^{n+1}.$$

Each of these substeps are pure Monge-Kantorovich/transport and Fisher-Rao/reaction variational steps, respectively and successively

$$\rho^{n+\frac{1}{2}} \in \operatorname*{Argmin}_{\rho \in \mathcal{M}_2^+,\, |\rho| = |\rho^n|} \left\{ \frac{1}{2\tau}\texttt{MK}^2(\rho, \rho^n) + \mathcal{F}(\rho) \right\} \tag{23}$$

$$\rho^{n+1} \in \operatorname*{Argmin}_{\rho \in \mathcal{M}^+} \left\{ \frac{1}{2\tau}\texttt{FR}^2(\rho, \rho^{n+\frac{1}{2}}) + \mathcal{F}(\rho) \right\}. \tag{24}$$

Note that the first Monge-Kantorovich step is mass preserving by construction, while the second will account for mass variations.

The underlying idea is that the scheme follows alternatively the two privileged directions in $T_\rho \mathcal{M}^+_{\texttt{KFR}} = T_\rho \mathcal{M}^+_{\texttt{MK}} \oplus T_\rho \mathcal{M}^+_{\texttt{FR}}$, corresponding to pure Monge-Kantorovich transport and pure Fisher-Rao reaction respectively. Another possible interpretation is that of an operator-splitting method: from (7)(9)(13) we get

$$\begin{aligned} -\operatorname{grad}_{\texttt{KFR}} \mathcal{F}(\rho) &= \operatorname{div}(\rho \nabla (U'(\rho) + \Psi + K \star \rho)) - \rho(U'(\rho) + \Psi + K \star \rho) \\ &= -\operatorname{grad}_{\texttt{MK}} \mathcal{F}(\rho) - \operatorname{grad}_{\texttt{FR}} \mathcal{F}(\rho). \end{aligned}$$

Viewing the same functional $\mathcal{F}(\rho)$ through distinct "differential lenses" (i-e using respectively the MK and FR differential structures) gives the two transport and reaction terms in the PDE (21). Thus it is very natural to split the PDE in two separate transport/reaction operators and treat separately each of them in their own and intrinsic differential framework. This idea of hybrid variational structures has been successfully applied e.g. in [23, 7, 8] for systems of equations where each component is viewed from separate differential perspectives, but not to the splitting of one single equation as it is the case here. A related splitting scheme was employed in [10] to construct weak solutions of fractional Fokker-Planck equations $\partial_t \rho = \Delta^{2s}\rho + \operatorname{div}(\rho \nabla \Psi)$, using a Monge-Kantorovich variational scheme in order to handle the transport term. However the discretization of the fractional Laplacian was treated in a non metric setting, the PDE cannot be viewed as the sum of gradient-flows of the same

functional for two different "orthogonal" metrics, and the approach therein is thus more a technical tool than an intrinsic variational feature.

Another natural consequence of this formal point of view is the following: From the orthogonality (14) in $T_\rho\mathcal{M}^+_{\mathtt{KFR}} = T_\rho\mathcal{M}^+_{\mathtt{MK}} \oplus T_\rho\mathcal{M}^+_{\mathtt{FR}}$ we can compute

$$\mathcal{D}(t) := -\frac{d}{dt}\mathcal{F}(\rho(t)) = -\|\operatorname{grad}_d \mathcal{F}\|^2_{T_\rho\mathcal{M}^+_{\mathtt{KFR}}} = -\|\operatorname{grad}_{\mathtt{MK}} \mathcal{F}\|^2_{T_\rho\mathcal{M}^+_{\mathtt{MK}}} - \|\operatorname{grad}_{\mathtt{FR}} \mathcal{F}\|^2_{T_\rho\mathcal{M}^+_{\mathtt{FR}}},$$

which really means that the total dissipation for the coupled KFR metrics is just the sum of the two elementary MK, FR dissipations. One can of course check this formula by computing $\frac{d}{dt}\mathcal{F}(\rho_t)$ along solutions of the PDE. This may be useful at the discrete level, since regularity is essentially related to dissipation. For example λ-convexity ensures that the energy is dissipated at a minimum rate, which in turn can be viewed as a quantifiable regularization in the spirit of Brézis-Pazy. This will be illustrated in Proposition 5.4, where we show that one indeed recovers an Energy Dissipation Inequality with respect to KFR from the two elementary MK, FR geodesic convexity and dissipation.
We first collect some general properties of our two-steps MK/FR splitting scheme, which share common features with the intrinsic one-step scheme (22) and only exploit the metric structure regardless of any PDE considerations.

LEMMA 4.1 (Total-square distance estimate). *Let $\rho^n, \rho^{n+\frac12}$ be recursive solutions of (23)(24). Then*

$$\frac{1}{\tau}\sum_{n\geq 0} \mathtt{KFR}^2(\rho^{n+1},\rho^n) \leq 4\left(\mathcal{F}(\rho^0) - \inf_{\mathcal{M}^+}\mathcal{F}\right). \tag{25}$$

Note that this estimate is useful only if $\mathcal{F}(\rho^0) < \infty$ and $\mathcal{F}$ is bounded from below. The former condition is a natural restriction to finite-energy initial data, and the latter is a reasonable assumption which holds true e.g. if $U(\rho) = \rho^m$ for some $m > 1$ and the external potential $\Psi(x) \geq 0$ outside of a finite measure set.

Proof. Testing $\rho = \rho^n$ in (23) and $\rho = \rho^{n+\frac12}$ in (24) we get

$$\begin{aligned}
\frac{1}{2\tau}\mathtt{MK}^2(\rho^{n+\frac12},\rho^n) + \mathcal{F}(\rho^{n+\frac12}) &\leq \mathcal{F}(\rho^n),\\
\frac{1}{2\tau}\mathtt{FR}^2(\rho^{n+1},\rho^{n+\frac12}) + \mathcal{F}(\rho^{n+1}) &\leq \mathcal{F}(\rho^{n+\frac12}).
\end{aligned}$$

Summing over $n \geq 0$ and noticing that the energy contributions are telescopic, we get the mixed total-square distance estimate

$$\frac{1}{\tau}\sum_{n\geq 0}\left\{\mathtt{FR}^2(\rho^{n+1},\rho^{n+\frac12}) + \mathtt{MK}^2(\rho^{n+\frac12},\rho^n)\right\} \leq 2\left(\mathcal{F}(\rho^0) - \inf_{\mathcal{M}^+}\mathcal{F}\right). \tag{26}$$

By triangular inequality and Proposition 2.1 it is easy to check that

$$\mathtt{KFR}^2(\rho^{n+1},\rho^n) \leq 2\left(\mathtt{FR}^2(\rho^{n+1},\rho^{n+\frac12}) + \mathtt{MK}^2(\rho^{n+\frac12},\rho^n)\right), \tag{27}$$

and our statement follows. ∎

REMARK 4.1. *It is worth stressing that, when trying to handle two different functionals $\partial_t\rho = \operatorname{div}(\rho\nabla F_1'(\rho)) - \rho F_2'(\rho)$ in the diffusion and reaction, the distance estimate for the two successive MK, FR steps would not result in a telescopic sum $\left[\mathcal{F}(\rho^{n+1}) - \mathcal{F}(\rho^{n+\frac12})\right] + \left[\mathcal{F}(\rho^{n+\frac12}) - \mathcal{F}(\rho^n)\right]$ as above, but rather in $\left[\mathcal{F}_1(\rho^{n+1}) - \mathcal{F}_1(\rho^{n+\frac12})\right] + \left[\mathcal{F}_2(\rho^{n+\frac12}) - \mathcal{F}_2(\rho^n)\right]$. This can in fact be controlled with suitable compatibility conditions on $\mathcal{F}_1, \mathcal{F}_2$ and estimating the crossed dissipations as in [26, 17], but we decided to focus here on $\mathcal{F}_1 = \mathcal{F} = \mathcal{F}_2$ in order to illustrate the general idea in a simpler variational setting.*

As already discussed the factor 2 in (27) is not optimal, and from the infinitesimal decoupling we should expect $\mathtt{KFR}^2(\rho^{n+1},\rho^n) \approx \mathtt{FR}^2(\rho^{n+1},\rho^{n+\frac12}) + \mathtt{MK}^2(\rho^{n+\frac12},\rho^n)$. Thus our estimate (25) should have a factor 2 instead of 4 in the right-hand side, which is exactly the classical total square distance estimate that one would get applying the direct one-step minimizing scheme (22) with respect to the full KFR metric.

Assuming that we can solve recursively (23)-(24) for some given initial datum

$$\rho_0 \in \mathcal{M}^+, \qquad \mathcal{F}(\rho^0) < \infty,$$

we construct two piecewise-constant interpolating curves

$$t \in ((n-1)\tau, n\tau],\ n \geq 0 : \qquad \left\{ \begin{array}{l} \tilde\rho^\tau(t) = \rho^{n+\frac12}, \\ \rho^\tau(t) = \rho^{n+1}. \end{array} \right.$$

By construction we have the energy monotonicity

$$\forall\, 0 \leq t_1 \leq t_2 : \qquad \mathcal{F}(\rho^\tau(t_2)) \leq \mathcal{F}(\tilde\rho^\tau(t_2)) \leq \mathcal{F}(\rho^\tau(t_1)) \leq \mathcal{F}(\tilde\rho^\tau(t_1)) \leq \mathcal{F}(\rho^0),$$

and an easy application of the Cauchy-Schwarz inequality with the total square-distance estimate (25) gives moreover the classical $\frac12$-Hölder estimate

$$\forall\, 0 \leq t_1 \leq t_2 : \qquad \left\{ \begin{array}{l} \mathtt{KFR}(\rho^\tau(t_2),\rho^\tau(t_1)) \leq C|t_2 - t_1 + \tau|^{\frac12} \\ \mathtt{KFR}(\tilde\rho^\tau(t_2),\tilde\rho^\tau(t_1)) \leq C|t_2 - t_1 + \tau|^{\frac12} \end{array} \right. . \tag{28}$$

Moreover for all $t > 0$ we have $\tilde\rho^\tau(t) = \rho^{n+\frac12}$ and $\rho^\tau(t) = \rho^{n+1}$ for some $n \geq 0$. From the total square estimate (26) we have therefore $\mathtt{FR}^2(\tilde\rho^\tau(t),\rho^\tau(t)) \leq C\tau$, and by Proposition 2.1 we conclude that the two curves $\rho^\tau, \tilde\rho^\tau$ stay close

$$\forall\, t \geq 0 : \qquad \mathtt{KFR}(\tilde\rho^\tau(t),\rho^\tau(t)) \leq \mathtt{FR}(\tilde\rho^\tau(t),\rho^\tau(t)) \leq C\sqrt{\tau} \tag{29}$$

uniformly in τ.

As a fairly general consequence of the total-square distance estimate (25), we retrieve an abstract convergence (pointwise in time) when $\tau \to 0$ for a weak topology:

Corollary 4.1. *Assume that $\mathcal{F}(\rho^0) < \infty$ and $\mathcal{F}$ is bounded from below on $\mathcal{M}^+$. Then there exists a* KFR*-continuous curve $\rho \in \mathcal{C}^{\frac12}([0,\infty);\mathcal{M}^+_{\mathtt{KFR}})$ and a discrete subsequence $\tau \to 0$ (not relabeled here) such that*

$$\forall\, t \geq 0 : \qquad \rho^\tau(t), \tilde\rho^\tau(t) \to \rho(t) \quad \textit{weakly-}* \ \textit{when } \tau \to 0.$$

Note that our statement is again unrelated to any PDE consideration, and merely exploits the metric structure. We recall that the weak-$*$ convergence of measures is defined in duality with $\mathcal{C}_0(\Omega)$ test-functions. Observe that the two interpolated curves converge to the *same* limit, and note that because $\rho \in \mathcal{C}([0,\infty);\mathcal{M}^+_{\mathtt{KFR}})$ the initial datum $\rho(0) = \rho^0$ is taken continuously in the KFR metric sense. In particular since KFR metrizes the narrow convergence of measures [25, thm. 3] the initial datum $\rho(0) = \rho^0$ will be taken at least in the narrow sense, which is stronger than weak-$*$ or distributional convergence.

Proof. From the proof of [25, lem. 2.2] it is easy to see that we have mass control

$$\forall\, \mu,\nu \in \mathcal{M}^+ : \qquad |\nu| \leq |\mu| + \mathtt{KFR}^2(\nu,\mu).$$

Applying this with $\nu = \rho^\tau(t), \tilde\rho^\tau(t)$ and $\mu = \rho^0$, and noting that the square-distance estimate (25) controls $\mathtt{KFR}^2(\rho^\tau(t),\rho^0), \mathtt{KFR}^2(\tilde\rho^\tau(t),\rho^0) \leq C(t+\tau)$, we see that the masses are controlled as $|\rho^\tau(t)| + |\tilde\rho^\tau(t)| \leq C(1+T)$ uniformly in τ in any finite time interval $t \in [0,T]$. By the Banach-Alaoglu in $\mathcal{M} = \mathcal{C}_0^*$ we see that $\rho^\tau(t), \tilde\rho^\tau(t)$ lie in the fixed weakly-$*$ relatively

compact set $\mathcal{K}_T = \{|\rho| \leq C(1+T)\}$ for all $t \in [0,T]$. By [25, thm. 5] we know that the $\mathtt{KFR}$ distance is lower semi-continuous with respect to the weak-$*$ convergence of measures, and the metric space $(\mathcal{M}^+, \mathtt{KFR})$ is complete [25, thm. 3]. Exploiting the time equicontinuity (28), the lower semi-continuity, and the completeness, we can apply a refined version of the Arzelà-Ascoli theorem [3, prop. 3.3.1] to conclude that, up to extraction of a discrete subsequence if needed, $\rho^\tau(t) \to \rho(t)$ and $\tilde\rho^\tau(t) \to \tilde\rho(t)$ pointwise in $t \in [0,T]$ for the weak-$*$ convergence and for some limit curves $\rho, \tilde\rho \in \mathcal{C}^{\frac12}([0,T]; \mathcal{M}^+_{\mathtt{KFR}})$. Moreover $\rho(t), \tilde\rho(t) \in \mathcal{K}_T$ for all $t \in [0,T]$, and by diagonal extraction we can assume that this holds for all $T > 0$. Finally as we already know that $\rho^\tau(t)$ and $\tilde\rho^\tau(t)$ converge weakly-$*$ to $\rho(t)$ and $\tilde\rho(t)$ respectively, we conclude by (29) and lower semi-continuity that $\mathtt{KFR}(\rho(t), \tilde\rho(t)) \leq \liminf_{\tau\to 0} \mathtt{KFR}(\rho^\tau(t), \tilde\rho^\tau(t)) = 0$ for any arbitrary $t \geq 0$. Thus $\rho = \tilde\rho$ as desired and the proof is complete. ∎

In order to connect now the previous abstract metric considerations with the PDE framework, we detail each of the substeps (23)(24) and exploit the particular $\mathtt{MK}, \mathtt{FR}$ Riemannian structures to retrieve the corresponding Euler-Lagrange equations.

In order to keep our notations light we write μ for the previous step and ρ^* for the minimizer. Thus $\mu = \rho^n$ and $\rho^* = \rho^{n+\frac12}$ in the first $\mathtt{MK}$ step $\rho^n \to \rho^{n+\frac12}$, while $\mu = \rho^{n+\frac12}$ and $\rho^* = \rho^{n+1}$ in the next $\mathtt{FR}$ step $\rho^{n+\frac12} \to \rho^{n+1}$.

4.1. The Monge-Kantorovich substep. For some fixed absolutely continuous measure $\mu \in \mathcal{M}^+_2$ (finite second moment) and mass $|\mu| = m$, let us consider here an elementary minimization step

$$\rho^* \in \operatorname*{Argmin}_{\rho \in \mathcal{M}^+_2,\, |\rho| = m} \left\{ \frac{1}{2\tau} \mathtt{MK}^2(\rho, \mu) + \mathcal{F}(\rho) \right\}. \tag{30}$$

Note that, if Ω is bounded, the restriction on finite second moments can be relaxed. Further assuming that $\mathcal{F}$ is lower semi-continuous with respect to the weak L^1 convergence (which is typically satisfied for the classical models), it is easy to obtain an absolutely continuous minimizer $\rho^* \in \mathcal{M}^+_2$ with mass $|\rho^*| = m = |\mu|$. Additional assumptions (e.g. strict convexity) sometimes guarantee uniqueness. Here we do not take interest in optimal conditions guaranteeing existence and/or uniqueness of minimizers, and this should be checked on a case-to-case basis depending on the structure of U, Ψ, K.

From the classical theory of optimal transportation we know that there exists a (backward) optimal map $\mathbf{t}$ from ρ^* to μ, such that

$$\mathtt{MK}^2(\rho^*, \mu) = \int_\Omega |x - \mathbf{t}(x)|^2 \, \mathrm{d}\rho^*(x).$$

A by-now standard computation [38, 41] shows that the Euler-Lagrange equation associated with (30) can be written in the form

$$\forall\, \zeta \in \mathcal{C}^\infty_c(\Omega; \mathbb{R}^d) : \qquad \int_\Omega \frac{\mathrm{id} - \mathbf{t}}{\tau} \cdot \zeta \, \mathrm{d}\rho^* + \int_\Omega \nabla(U'(\rho^*) + \Psi + K \star \rho^*) \cdot \zeta \, \mathrm{d}\rho^* = 0. \tag{31}$$

Using the definition of the pushforward $\mu = \mathbf{t}\#\rho^*$ we recall the classical Taylor expansion

$$\begin{aligned} \int_\Omega (\rho^* - \mu)\phi &= \int_\Omega (\rho^* - \mathbf{t}\#\rho^*)\phi = \int_\Omega \big(\phi(x) - \phi(\mathbf{t}(x))\big)\rho^*(x) \\ &= \int_\Omega \Big(x - \mathbf{t}(x)) \cdot \nabla\phi(x) + \mathcal{O}\left(\|D^2\phi\|_\infty |x - \mathbf{t}(x)|^2 \right) \Big) \, \mathrm{d}\rho^*(x) \\ &= \int_\Omega (\mathrm{id} - \mathbf{t}) \cdot \nabla\phi \, \mathrm{d}\rho^* + \mathcal{O}\left(\|D^2\phi\|_\infty \mathtt{MK}^2(\rho^*, \mu) \right) \end{aligned} \tag{32}$$

for all $\phi \in \mathcal{C}^\infty_c(\Omega)$. Taking $\zeta = \nabla\phi$ in (31) and substituting finally yields

$$\int_\Omega (\rho^* - \mu)\phi = -\tau \int_\Omega \nabla(U'(\rho^*) + \Psi + K \star \rho^*) \cdot \nabla\phi \, \mathrm{d}\rho^* + \mathcal{O}\left(\|D^2\phi\|_\infty \mathtt{MK}^2(\rho^*, \mu) \right) \tag{33}$$

for all smooth test functions ϕ. This is of course an approximation of the implicit implicit Euler scheme

$$\frac{\rho^* - \mu}{\tau} = \mathrm{div}(\rho^* \nabla(U'(\rho^* + \Psi + K \star \rho^*)),$$

the approximate error being controlled quadratically in the $\mathtt{MK}$ distance. Note that this corresponds to the pure transport part $\partial_t \rho = \mathrm{div}(\rho \nabla(U'(\rho) + \Psi + K \star \rho^*)) + (\dots)$ in the PDE (21).

4.2. The Fisher-Rao substep. Let us fix as before an arbitrary measure $\mu \in \mathcal{M}^+$ (no restriction on the second moment), and assume that there exists somehow an absolutely continuous minimizer

$$\rho^* \in \operatorname*{Argmin}_{\rho \in \mathcal{M}^+} \left\{ \frac{1}{2\tau} \mathtt{FR}^2(\rho, \mu) + \mathcal{F}(\rho) \right\}. \tag{34}$$

The existence and uniqueness of minimizers can again be obtained under suitable superlinearity, lower semi-continuity, and convexity assumptions on U, Ψ, K, and we do not worry about this issue.

Let us start by differentiating the squared distance for suitable perturbations ρ_ε of the minimizer ρ^*. According to section 2.2 an arbitrary $\psi \in \mathcal{C}_c^\infty(\Omega)$ is identified to a tangent vector in $T_{\rho^*} \mathcal{M}^+_{\mathtt{FR}}$ through

$$\left\{ \begin{array}{l} \partial_\varepsilon \rho_\varepsilon = \rho_\varepsilon \psi \\ \rho_\varepsilon(0) = \rho^* \end{array} \right. \qquad \Leftrightarrow \qquad \rho_\varepsilon = \rho^* e^{\varepsilon \psi}.$$

Denoting by $\mu_s = [(1-s)\sqrt{\mu} + s\sqrt{\rho^*}]^2$ the Fisher-Rao geodesics from μ to ρ^*, the terminal velocity $\partial_s \mu(1) = 2\sqrt{\rho^*}(\sqrt{\rho^*} - \sqrt{\mu})$ can be represented by the $L^2(\mathrm{d}\rho^*)$ action of $r = 2\frac{\sqrt{\rho^*} - \sqrt{\mu}}{\sqrt{\rho^*}}$. Using the first variation formula $\frac{d}{dt}\left(\frac{1}{2} d^2(x(t), y)\right)\big|_{t=0} = \langle x'(0), y'(1) \rangle_{x(0)}$ and our $L^2(\mathrm{d}\rho)$ identification of the tangent spaces in section 2.3 we can guess that

$$\begin{aligned} \frac{d}{d\varepsilon}\left(\frac{1}{2}\mathtt{FR}^2(\rho_\varepsilon, \mu)\right)\Big|_{\varepsilon=0} &= \langle \partial_\varepsilon \rho(0), \partial_s \mu(1) \rangle_{T_{\rho^*} \mathcal{M}^+_{\mathtt{FR}}} \\ &= (\psi, r)_{L^2(\mathrm{d}\rho^*)} = 2 \int_\Omega (\sqrt{\rho^*} - \sqrt{\mu}) \sqrt{\rho^*} \psi, \end{aligned}$$

which can be checked by differentiating w.r.t. ε in the explicit representation (8). Using the same Riemannian formalism we similarly anticipate that

$$\begin{aligned} \frac{d}{d\varepsilon} \mathcal{F}(\rho_\varepsilon)\Big|_{\varepsilon=0} &= \langle \mathrm{grad}_{\mathtt{FR}} \mathcal{F}, \partial_\varepsilon \rho(0) \rangle_{T_{\rho^*} \mathcal{M}^+_{\mathtt{FR}}} \\ &= \langle F'(\rho^*), \psi \rangle_{L^2(\mathrm{d}\rho^*)} = \int_\Omega \rho^* (U'(\rho^*) + \Psi + K \star \rho^*) \psi, \end{aligned}$$

and this can be checked again by differentiating $\frac{d}{d\varepsilon} \mathcal{F}(\rho_\varepsilon) = \int_\Omega \partial_\varepsilon(\dots)$ under the integral sign. Writing the the optimality condition $\frac{d}{d\varepsilon}\left(\frac{1}{2\tau}\mathtt{FR}^2(\rho_\varepsilon, \mu) + \mathcal{F}(\rho_\varepsilon)\right)\big|_{\varepsilon=0} = 0$ thus gives the Euler-Lagrange equation

$$\forall\, \psi \in \mathcal{C}_c^\infty(\Omega) : \qquad \int_\Omega (\sqrt{\rho^*} - \sqrt{\mu}) \sqrt{\rho^*} \psi = -\frac{\tau}{2} \int_\Omega \left\{ U'(\rho^*) + \Psi + K \star \rho^* \right\} \rho^* \psi. \tag{35}$$

In order to relate this with the more standard Euclidean difference quotient, we first assume that $U'(\rho^*) + \Psi + K \star \rho^* \in L^2(\mathrm{d}\rho^*)$, or in other words that $\mathrm{grad}_{\mathtt{FR}} \mathcal{F}(\rho^*)$ can indeed be considered as a tangent vector of $T_{\rho^*} \mathcal{M}^+_{\mathtt{FR}}$. This should be natural, but may require a case-to-case analysis depending on the structure of U, Ψ, K. Then an easy density argument shows that the previous equality holds for all $\psi \in L^2(\mathrm{d}\rho^*)$. Taking in particular $\psi =$

$\frac{\sqrt{\rho^*}+\sqrt{\mu}}{\sqrt{\rho^*}}\phi \in L^2(\mathrm{d}\rho^*)$ for arbitrary $\phi \in \mathcal{C}_c^\infty(\Omega)$, we obtain a slight variant of the previous Euler-Lagrange equation (35) in the form

$$\forall\, \phi \in \mathcal{C}_c^\infty(\Omega) : \qquad \int_\Omega (\rho^* - \mu)\phi = -\tau \int_\Omega \frac{\sqrt{\rho^*}(\sqrt{\rho^*}+\sqrt{\mu})}{2}\big\{U'(\rho^*) + \Psi + K \star \rho^*\big\}\phi. \tag{36}$$

Recalling that in the minimizing scheme we only deal with measures at short $\mathcal{O}(\sqrt{\tau})$ distance, one should essentially think of this as if $\rho^* \approx \mu$ in the right-hand side, and (36) is thus an approximation of the implicit Euler scheme

$$\frac{\rho^* - \mu}{\tau} = -\rho^*(U'(\rho^*) + \Psi + K \star \rho^*).$$

Note that this is the reaction part $\partial_t \rho = (\dots) - \rho(U'(\rho) + \Psi + K \star \rho^*)$ in the PDE (21).

Remark 4.2. *Contrarily to the corresponding approximate Euler-Lagrange equation (33) for one elementary Monge-Kantorovich substep, (36) does not involve any quadratic remainder $\mathcal{O}(\mathtt{FR}^2(\rho^*, \mu))$. The price to pay for this is that the right-hand side appears now as a slight "twist" of the more natural and purely Riemannian object $-\rho^*(U'(\rho^*) + \Psi + K \star \rho^*) = -\operatorname{grad}_{\mathtt{FR}} \mathcal{F}(\rho^*)$ in (35), the twist occurring through the approximation $\frac{\sqrt{\rho^*}(\sqrt{\rho^*}+\sqrt{\mu})}{2} \approx \rho^*$.*

Remark 4.3. *A technical issue might arise here for unbounded domains. Indeed since we construct recursively $\rho^n \xrightarrow{\mathtt{MK}^2} \rho^{n+\frac{1}{2}} \xrightarrow{\mathtt{FR}^2} \rho^{n+1}$ one should make sure that, in the second reaction substep, the minimizer ρ^{n+1} keeps finite second moment so that the scheme can be safely iterated afterward. This should be generally guaranteed if the external potential Ψ is quadratically confining, but may require once again a delicate analysis depending on the structure of U, Ψ, K (we will show in section 5 that this holds e.g. in the simple case $\Psi, K \equiv 0$).*

4.3. Convergence to a weak solution. We can now show that, under some strong compactness assumptions, the limit $\rho = \lim \rho^\tau = \lim \tilde\rho^\tau$ is generically a weak solution to the original PDE.

Theorem 4.1. *Let $\rho^\tau, \tilde\rho^\tau, \rho$ as in Corollary 4.1, and assume that*

$$\left\{ \begin{array}{rcl} \tilde\rho^\tau \nabla\left(U'(\tilde\rho^\tau) + \Psi + K \star \tilde\rho^\tau\right) & \rightharpoonup & \rho \nabla\left(U'(\rho) + \Psi + K \star \rho\right) \\ \sqrt{\rho^\tau}\frac{\sqrt{\rho^\tau}+\sqrt{\tilde\rho^\tau}}{2}\left(U'(\rho^\tau) + \Psi + K \star \rho^\tau\right) & \rightharpoonup & \rho\left(U'(\rho) + \Psi + K \star \rho\right) \end{array} \right. \tag{37}$$

at least weakly in $L^1_{\mathrm{loc}}((0,\infty)\times\Omega)$. Then ρ is a nonnegative weak solution of

$$\left\{ \begin{array}{ll} \partial_t \rho = \operatorname{div}(\rho \nabla(U'(\rho) + \Psi + K \star \rho)) - \rho(F'(\rho) + \Psi + K \star \rho) & \text{in } (0,\infty)\times\Omega \\ \rho_{|t=0} = \rho^0 & \text{in } \mathcal{M}^+(\Omega) \end{array} \right.$$

For the sake of generality we simply assumed here that the nonlinear terms pass to the limit as in (37). This is of course a strong hypothesis to be checked in each case of interest, and usually requires *strong convergence* $\rho^\tau, \tilde\rho^\tau \to \rho$ (e.g. pointwise a.e.). We shall discuss in section 5 some strategies to retrieve such compactness.

Proof. As already discussed after Corollary 4.1, the initial datum $\rho(0) = \rho^0$ is taken continuously at least in the metric sense $(\mathcal{M}^+, \mathtt{KFR})$. Moreover, any limit $\rho = \lim\limits_{\tau\to 0} \rho^\tau$ in any weak sense will automatically be nonnegative.

Fix now any $0 < t_1 < t_2$ and $\phi \in \mathcal{C}_c^\infty(\Omega)$. For fixed τ we have $\rho^\tau(t_i) = \rho^{N_i}$ for $N_i = \lceil t_i/\tau \rceil$, and $T_i = N_i\tau \to t_i$ as $\tau \to 0$. Moreover for fixed $n \geq 0$ we have by construction the two Euler-Lagrange equations (33)(36), one for each Monge-Kantorovich and Fisher-Rao substep as in section 4.1 and section 4.2 respectively. More explicitly, there holds

$$\int_\Omega (\rho^{n+\frac{1}{2}} - \rho^n)\phi = -\tau \int_\Omega \rho^{n+\frac{1}{2}} \nabla(U'(\rho^{n+\frac{1}{2}}) + \Psi + K \star \rho^{n+\frac{1}{2}}) \cdot \nabla\phi \\ + \mathcal{O}\left(\|D^2\phi\|_\infty \mathtt{MK}^2(\rho^{n+\frac{1}{2}}, \rho^n)\right)$$

and

$$\int_\Omega (\rho^{n+1}-\rho^{n+\frac12})\phi = -\tau\int_\Omega \frac{\sqrt{\rho^{n+1}}(\sqrt{\rho^{n+1}}+\sqrt{\rho^{n+\frac12}})}{2}\big\{U'(\rho^{n+1})+\Psi+K\star\rho^{n+1}\big\}\phi.$$

Summing from $n=N_1$ to $n=N_2-1$, using the square-distance estimate (26) to control the remainder term in the first Euler-Lagrange equation above, and recalling that the interpolated curves are piecewise constant, we immediately get

$$\begin{aligned}
\int_\Omega (\rho^\tau(t_2)-\rho^\tau(t_1))\,\phi &= \sum_{n=N_1}^{N_2-1}\int_\Omega \Big\{(\rho^{n+1}-\rho^{n+\frac12})+(\rho^{n+\frac12}-\rho^n)\Big\}\,\phi\\
&= -\sum_{n=N_1}^{N_2-1}\tau\int_\Omega \frac{\sqrt{\rho^{n+1}}(\sqrt{\rho^{n+1}}+\sqrt{\rho^{n+\frac12}})}{2}\big\{U'(\rho^{n+1})+\Psi+K\star\rho^{n+1}\big\}\phi\\
&\quad -\sum_{n=N_1}^{N_2-1}\tau\int_\Omega \rho^{n+\frac12}\nabla(U'(\rho^{n+\frac12})+\Psi+K\star\rho^{n+\frac12})\cdot\nabla\phi\\
&\quad +\mathcal{O}\left(\|D^2\phi\|_\infty\sum_{n=N_1}^{N_2-1}\mathtt{MK}^2(\rho^{n+\frac12},\rho^n)\right)\\
&= -\int_{T_1}^{T_2}\int_\Omega \frac{\sqrt{\rho^\tau}(\sqrt{\rho^\tau}+\sqrt{\tilde\rho^\tau})}{2}\big\{U'(\rho^\tau)+\Psi+K\star\rho^\tau\big\}\phi\\
&\quad -\int_{T_1}^{T_2}\int_\Omega \tilde\rho^\tau\nabla(U'(\tilde\rho^\tau)+\Psi+K\star\tilde\rho^\tau)\cdot\nabla\phi \quad +\mathcal{O}\left(\|D^2\phi\|_\infty\tau\right).
\end{aligned}$$

From Corollary 4.1 we know that $\rho^\tau(t)$ converge weakly-$*$ to $\rho(t)$ pointwise in time, so the left-hand side passes to the limit when $\tau\to0$. Due to our strong assumption (37) and because $T_i\to t_i$ the right-hand side also passes to the limit. As a consequence we get

$$\int_\Omega (\rho(t_2)-\rho(t_1))\,\phi = -\int_{t_1}^{t_2}\int_\Omega \rho\Big(\nabla(U'(\rho)+\Psi+K\star\rho)\cdot\nabla\phi+\big(U'(\rho)+\Psi+K\star\rho\big)\phi\Big)$$

for all $0<t_1<t_2$ and $\phi\in\mathcal{C}^\infty_c(\Omega)$, which is clearly an admissible weak formulation of $\partial_t\rho=\mathrm{div}(\rho\nabla(U'(\rho)+\Psi+K\star\rho))-\rho(U'(\rho)+\Psi+K\star\rho)$. ∎

If $\Omega\neq\mathbb{R}^d$ some further work may be needed to retrieve the homogeneous Neumann condition $\rho\nabla(U'(\rho)+\Psi+K\star\rho)\cdot\nu=0$ on $\partial\Omega$. This amounts to extending the class of $\mathcal{C}^\infty_c(\Omega)$ test functions to $\mathcal{C}^1_{loc}(\overline\Omega)$ and should generically hold with just enough regularity on the solution, but we will disregard this technical issue for the sake of simplicity.

5. Compactness issues: an illustrative example. In Theorem 4.1 we assumed for simplicity that the nonlinear terms pass to the limit, mainly in the distributional sense. In order to prove this, the usual strategy is to obtain first some energy/dissipation-type estimates to show that the nonlinear terms have a weak limit, and then prove pointwise convergence $\rho^\tau(t,x)\to\rho(t,x)$ a.e. $(t,x)\in\mathbb{R}^+\times\Omega$ to identify the weak limit (typically as weak-strong products of limits). Thus the problem should amount to retrieving enough compactness on the interpolating curves $\rho^\tau,\tilde\rho^\tau$. With the help of any Aubin-Lions-Simon type results this essentially requires compactness in time and space, which can be handled separately for different topologies in a flexible way. Compactness in space usually follows from the aforementioned energy/dissipation estimates, and the energy monotonicity should of course help: if e.g. the total energy $\mathcal{F}(\rho)=\int_\Omega U(\rho)+(\dots)$ controls any $L^q(\Omega)$ norm then $\mathcal{F}(\rho^\tau(t))\le\mathcal{F}(\rho^0)$ immediately controls $\|\rho^\tau\|_{L^\infty(0,\infty;L^q)}$ uniformly in τ. A rule of thumbs for parabolic equations is usually that space regularity can be transferred to time regularity. Thus the parabolic nature of the scheme should allow here to transfer space estimates,

if any, to time estimates. Note also that some sort of time compactness (approximate equicontinuity) is already guaranteed by (28), but in a very weak metric sense for which the standard Aubin-Lions-Simon theory does not apply directly.

A slight modification of the usual arguments should however be required here, because the scheme is decomposed in two separate substeps. The first Monge-Kantorovich substep (30) encodes the higher order part of the PDE, which is parabolic and should therefore be smoothing. This regularization can often be quantified using by-now classical methods in (Monge-Kantorovich) optimal transport theory, such as BV estimates [16], the *flow-interchange* technique from [30], regularizing λ-displacement convexity in the spirit of [3, 31], or any other strategy. On the other hand the second Fisher-Rao substep (34) encodes the reaction part of the PDE, hence we cannot expect any smoothing at this stage. One should therefore make sure that, in the step $\rho^{n+\frac{1}{2}} \xrightarrow{\texttt{FR}} \rho^{n+1}$, the regularity of $\rho^{n+\frac{1}{2}}$ inherited from the previous step $\rho^n \xrightarrow{\texttt{MK}} \rho^{n+\frac{1}{2}}$ propagates to ρ^{n+1} at least to some extent.

At this stage we would like to point out one other possible advantage of our splitting scheme: it is well known [3] that λ-geodesic convexity is a central tool in the theory of gradient flows in abstract metric spaces, and leads to quantified regularization properties at the discrete level. Second order differential calculus *à la Otto* [34] with respect to the KFR Riemannian structure was discussed in [25, 29] (also earlier suggested in [27]) and allows to determine at least formally if a given functional $\mathcal{F}$ is λ-geodesically convex for the distance KFR. However, in our scheme each step only sees either one of the differential MK/FR structures and therefore only separate geodesic convexity comes into play. Consider for example the case of internal energies $\mathcal{F}(\rho) = \int_\Omega U(\rho)$. Then the celebrated condition for McCann's displacement convexity [31] with respect to MK reads $\rho P'(\rho) - \left(1 - \frac{1}{d}\right) P(\rho) \geq 0$ in space dimension d, where the pressure $P(\rho) := \rho U'(\rho) - U(\rho)$. On the other hand using the Riemannian formalism in section 2.2 it is easy to see that, at least formally, this same functional is λ-geodesically convex with respect to FR if and only if $\rho U''(\rho) + \frac{U'(\rho)}{2} \geq \lambda$. This condition can be interpreted as $s \mapsto U(s^2)$ being $\lambda/4$-convex in $s = \sqrt{\rho}$, the latter change of variables naturally arising through (8) and $\texttt{FR}^2(\rho_0, \rho_1) = 4\|\sqrt{\rho_1} - \sqrt{\rho_0}\|_{L^2}^2$. Those two conditions are very easy to check separately and, in the light of the infinitesimal uncoupling, it seems likely that simultaneous convexity with respect to each of the MK, FR metrics is equivalent to convexity with respect to the coupled KFR structure. See [25, section 3] and [29, section 5.1] for related discussions.

The rest of this section is devoted to the illustration of possible compactness strategies in the simple case

$$\text{(H)} \qquad \begin{cases} \Psi, K \equiv 0, \\ U \in \mathcal{C}^1([0,\infty)) \cap \mathcal{C}^2(0,\infty) \quad \text{with } U(0) = 0, \\ U', U'' \geq 0, \\ \rho U''(\rho) \quad \text{is bounded for small } \rho \in (0, \rho_0], \end{cases}$$

which from now will be assumed without further mention. We would like to stress here that (H) holds for any Porous-Medium-type nonlinearity $U_m(\rho) = \frac{1}{m-1}\rho^m$ at least in the slow diffusion regime $m > 1$, but *not* for the Boltzmann entropy $H(\rho) = \rho \log \rho - \rho$. Even though the latter is well behaved (displacement convex) with respect to the Monge-Kantorovich structure [21, 41], it is *not* with respect to the Fisher-Rao one. Indeed it is easy to check that $H(\rho)$ is not convex in $\sqrt{\rho}$, so that the Boltzmann entropy is not λ-displacement convex with respect to FR for any $\lambda \in \mathbb{R}$. This would require $\rho H''(\rho) + \frac{H'(\rho)}{2} = 1 + \frac{\log \rho}{2} \geq \lambda$ for some constant λ, which obviously fails for small ρ (this can be related to $\rho = 0$ being an extremal point in $\mathcal{M}^+$, where all the Riemannian formalism from section 2.3 degenerates). Since the purpose of this section is to illustrate that strong compactness can be retrieved at least in some particular cases, we chose to set $\Psi \equiv 0$ to make the computations and estimates as light as possible. The case $\Psi \not\equiv 0$ follows with only minor modifications at least for reasonable potentials (see e.g. remark 5.1 and [26, 17]). Including interaction

terms $K \not\equiv 0$ may be more involved and require additional assumptions, and we shall not comment further on this.

5.1. Propagation of regularity at the discrete level. Whenever $U', U'' \geq 0$, the PDE $\partial_t \rho = \mathrm{div}(\rho \nabla U'(\rho)) - \rho U'(\rho) = \mathrm{div}(\rho U''(\rho)\nabla\rho) - \rho U'(\rho)$ is formally parabolic, satisfies the maximum principle $\|\rho(t)\|_\infty \leq \|\rho^0\|_\infty$, and initial regularity should propagate. We prove below that this holds at the discrete level:

PROPOSITION 5.1 (BV and L^∞ estimates). *Assume that the initial datum $\rho^0 \in BV \cap L^\infty(\Omega)$. Then for any $\tau < 2/U'(\|\rho^0\|_\infty)$ there holds*

$$\forall\, t \geq 0 : \qquad \|\rho^\tau(t)\|_{\mathrm{BV}(\Omega)} \leq \|\tilde{\rho}^\tau(t)\|_{\mathrm{BV}(\Omega)} \leq \|\rho^0\|_{\mathrm{BV}(\Omega)}$$

and

$$\forall\, t \geq 0 : \qquad \|\rho^\tau(t)\|_{L^\infty(\Omega)} \leq \|\tilde{\rho}^\tau(t)\|_{L^\infty(\Omega)} \leq \|\rho^0\|_{L^\infty(\Omega)}.$$

Proof. We argue at the discrete level by showing that the estimates propagate in each substep. We shall actually prove a more precise result, namely

$$\|\rho^{n+\frac{1}{2}}\|_{\mathrm{BV}} \leq \|\rho^n\|_{\mathrm{BV}}, \qquad \|\rho^{n+\frac{1}{2}}\|_{L^\infty} \leq \|\rho^n\|_{L^\infty} \tag{38}$$

and

$$\|\rho^{n+1}\|_{\mathrm{BV}} \leq \|\rho^{n+\frac{1}{2}}\|_{\mathrm{BV}}, \qquad \rho^{n+1}(x) \leq \rho^{n+\frac{1}{2}}(x) \text{ a.e.} \tag{39}$$

The propagation (38) in the first `MK` step only requires convexity $U'' \geq 0$ and no smallness condition on the time step τ. This should be expected since the `MK` step is an *implicit* discretization of $\partial_t \rho = \mathrm{div}(\rho \nabla U'(\rho)) = \mathrm{div}(\rho U''(\rho)\nabla\rho)$, which is formally parabolic as soon as $U'' \geq 0$. We recall first that by construction the step is mass preserving, $\|\rho^{n+\frac{1}{2}}\|_{L^1} = \|\rho^n\|_{L^1}$. With our assumption $U'' \geq 0$ we can directly apply [16, thm. 1.1] to obtain $\|\rho^{n+\frac{1}{2}}\|_{\mathrm{TV}} \leq \|\rho^n\|_{\mathrm{TV}}$, which immediately entails the BV estimate. An early proof of $\|\rho^{n+\frac{1}{2}}\|_{L^\infty} \leq \|\rho^n\|_{L^\infty}$ can be found in [33] for the particular case $U(\rho) = \rho^2$, and the case of general convex U is covered by [38, prop. 7.32] (see also [12, 39]). For the propagation (39) in the `FR` step we show below that the minimizer ρ^{n+1} can be written as

$$\rho^{n+1}(x) = R(\rho^{n+\frac{1}{2}}(x)) \qquad \text{a.e. } x \in \Omega$$

for some 1-Lipschitz function $R : \mathbb{R}^+ \to \mathbb{R}^+$ with $R(0) = 0$. This will ensure that $0 \leq \rho^{n+1}(x) \leq \rho^{n+\frac{1}{2}}(x)$ and entail the L^∞ and L^1 bounds as well as the total variation estimate (see [1] for the Lip $\circ$ BV composition of maps). Note that $\rho^{n+1}(x) \leq \rho^{n+\frac{1}{2}}(x)$ shows in particular that the second moments propagate to the next step, which should require further assumptions on U, Ψ in the general case. In the rest of the proof we write $\rho^* = \rho^{n+1}$ and $\mu = \rho^{n+\frac{1}{2}}$ for simplicity, in agreement with our previous notations in section 4.2.

By (35) with $\Psi, K \equiv 0$ we see that

$$(\sqrt{\rho^*} - \sqrt{\mu})\sqrt{\rho^*} = -\frac{\tau}{2}\rho^* U'(\rho^*) \tag{40}$$

at least in $L^1_{\mathrm{loc}}(\Omega)$, hence a.e. $x \in \Omega$. From $U' \geq 0$ we immediately get that either $\rho^* = 0$ or $\sqrt{\rho^*} \leq \sqrt{\mu}$, which gives in any case $\rho^*(x) \leq \mu(x)$ a.e.

We show now that if the CFL condition $\tau \leq 2/U'(\|\rho^0\|_\infty)$ holds then ρ^* and μ share the same support, i-e $\rho^*(x) > 0 \Leftrightarrow \mu(x) > 0$. From the previous inequality $\rho^* \leq \mu$ we only have to show that $\rho^*(x) > 0$ as soon as $\mu(x) > 0$. Assume by contradiction that there is some subset $E \subset \Omega$ with positive Lebesgue measure such that $\rho^*(x) = 0$ but $\mu(x) > 0$ in E. We claim that

$$\overline{\rho} := \rho^* \chi_{E^\complement} + \mu \chi_E$$

is then a strictly better competitor than the minimizer ρ^*. In order to check this we first compute the square distance

$$\begin{aligned}\frac{1}{4}\left(\mathtt{FR}^2(\overline{\rho},\mu)-\mathtt{FR}^2(\rho^*,\mu)\right) &= \int_\Omega \left|\sqrt{\overline{\rho}}-\sqrt{\mu}\right|^2 - \int_\Omega \left|\sqrt{\rho^*}-\sqrt{\mu}\right|^2 \\ &= \left(\int_{E^\complement} \left|\sqrt{\rho^*}-\sqrt{\mu}\right|^2 + \int_E |\sqrt{\mu}-\sqrt{\mu}|^2\right) \\ &\qquad - \left(\int_{E^\complement} \left|\sqrt{\rho^*}-\sqrt{\mu}\right|^2 + \int_E |0-\sqrt{\mu}|^2\right) = -\int_E \mu < 0.\end{aligned}$$

For the energy contribution we have by convexity

$$\begin{aligned}\mathcal{F}(\overline{\rho})-\mathcal{F}(\rho^*) = \int_\Omega U(\overline{\rho})-U(\rho^*) &\le \int_\Omega U'(\overline{\rho})(\overline{\rho}-\rho^*) \\ &= \int_E U'(\overline{\rho})(\mu-0) \le U'(\|\rho^0\|_\infty)\int_E \mu.\end{aligned}$$

Note that $0 \le \rho^*, \overline{\rho}, \mu \le \|\rho^0\|_\infty$ almost everywhere, so that all these integrals are well-defined. Gathering these two inequalities we obtain

$$\frac{1}{2\tau}\left(\mathtt{FR}^2(\overline{\rho},\mu)-\mathtt{FR}^2(\rho^*,\mu)\right) + \left(\mathcal{F}(\overline{\rho})-\mathcal{F}(\rho^*)\right) \le \left(-\frac{2}{\tau}+U'(\|\rho^0\|_\infty)\right)\int_E \mu < 0$$

because $\int_E \mu > 0$ and $\tau < 2/U'(\|\rho^0\|_\infty)$. This shows that $\overline{\rho}$ is a strictly better competitor and yields the desired contradiction, thus $\rho^* > 0 \Leftrightarrow \mu > 0$.

Now inside the common support of ρ^*, μ we can divide (40) by $\sqrt{\rho^*} > 0$, and $\rho = \rho^*(x)$ is a solution of the implicit equation

$$f(\rho,\mu) = 0 \quad \text{with} \quad f(\rho,\mu) := \sqrt{\rho}\left(1+\frac{\tau}{2}U'(\rho)\right) - \sqrt{\mu}$$

with $\mu = \mu(x)$ and a.e. $x \in \operatorname{supp}\rho^* = \operatorname{supp}\mu$. An easy application of the implicit functions theorem shows that, for any $\mu > 0$, this has a unique solution $\rho = R(\mu)$ for a $\mathcal{C}^1(0,\infty)$ function R satisfying $0 < R(\mu) \le \mu$ for $\mu > 0$. Moreover one can compute explicitly for all $\mu > 0$

$$\begin{aligned}0 < \frac{dR}{d\mu}(\mu) = -\left.\frac{\partial_\mu f}{\partial_\rho f}\right|_{\rho=R(\mu)} &= \frac{\frac{1}{2\sqrt{\mu}}}{\frac{1}{2\sqrt{\rho}}\left(1+\frac{\tau}{2}U'(\rho)\right)+\frac{\tau}{2}\sqrt{\rho}U''(\rho)} \\ &\le \frac{\frac{1}{2\sqrt{\mu}}}{\frac{1}{2\sqrt{\rho}}\left(1+\frac{\tau}{2}U'(\rho)\right)} = \frac{1}{\frac{\sqrt{\mu}}{\sqrt{\rho}}\left(1+\frac{\tau}{2}U'(\rho)\right)} = \frac{\rho}{\mu} \le 1,\end{aligned}$$

where we used successively $U'' \ge 0$, $f(\rho,\mu) = 0 \Leftrightarrow 1+\frac{\tau}{2}U'(\rho) = \frac{\sqrt{\mu}}{\sqrt{\rho}}$, and $\rho = R(\mu) \le \mu$. Extending by continuity $R(0) = 0$, we have shown that $\rho^*(x) = R(\mu(x))$ a.e. $x \in \Omega$ for some 1-Lipschitz function $R : \mathbb{R}^+ \to \mathbb{R}^+$ with $R(0) = 0$, and the proof is complete. ∎

Remark 5.1. *A closer analysis of the implicit functions theorem above reveals that the argument only requires $U' \ge 0$ and $\rho U''(\rho) + U'(\rho)/2 \ge 0$, which is less stringent than our assumption $U', U'' \ge 0$ as in (H). As already suggested this former condition corresponds to convexity of $s \mapsto U(s^2)$ in the $s = \sqrt{\rho}$ variable, or more intrinsically to geodesic convexity of $\mathcal{F}(\rho) = \int_\Omega U(\rho)$ with respect to the Fisher-Rao distance. We also point out that the same approach works with external potentials $\Psi \not\equiv 0$ under suitable structural assumptions: one shows first that strict positivity is preserved in the sense that* $\operatorname{supp}\rho^{n+1} = \operatorname{supp}\rho^{n+\frac{1}{2}}$, *which is to be expected since the ODE $\partial_t\rho = -\rho(U'(\rho)+\Psi(x))$ formally preserves positivity. Exploiting the Euler-Lagrange equations (35)(36), an implicit functions theorem $f(\rho,\mu,\Psi) = 0 \Leftrightarrow \rho = R(\mu,\Psi)$ then applies inside the common support to propagate the regularity. This still controls $\nabla\rho = \partial_\mu R\,\nabla\mu + \partial_\Psi R\nabla\Psi$ provided that Ψ is smooth enough, see [17, 26] for details.*

5.2. Compactness and Energy Dissipation Inequality. In this section we check that our strong assumption (37) in Theorem 4.1 holds in the particular case of internal energies only, i-e that the nonlinear terms in the PDE pass to the limit. We start by improving the weak convergence in Corollary 4.1:

PROPOSITION 5.2. *Assume* (H). *Then*

$$\rho^\tau, \tilde\rho^\tau \to \rho \qquad \text{in } L^1_{\rm loc}([0,\infty);L^1)$$

for some (discrete) subsequence $\tau \to 0$.

We give two proofs: the first one is elementary and fully exploits the uniform-in-time compactness estimates from Proposition 5.1, which were derived here for the particular case $\Psi \equiv K \equiv 0$ only. The second proof is less straightforward but enlightens the general idea of transferring space regularity to time regularity through the PDE itself, and should apply to non-trivial potentials and interactions with minor modifications.

First proof of Proposition 5.2. Let us recall from Proposition 4.1 that $\rho^\tau(t), \tilde\rho^\tau(t)$ both converge weakly-$*$ to the same limit $\rho(t)$ pointwise in time. We claim that this weak-$*$ convergence can be strengthened into strong $L^1(\Omega)$ convergence. Indeed for any fixed $t \geq 0$ we have $\|\rho^\tau(t)\|_{\rm BV}, \|\tilde\rho^\tau(t)\|_{\rm BV} \leq \|\rho^0\|_{\rm BV}$ so by compactness $\mathrm{BV}(\Omega) \subset\subset L^1(\Omega)$ we see that $\{\rho^\tau(t)\}_{\tau>0}, \{\tilde\rho^\tau(t)\}_{\tau>0}$ are L^1 relatively compact for fixed $t \geq 0$. Because strong L^1 convergence implies in particular weak-$*$ convergence of measures, and because we already know that these sequences are weakly-$*$ convergent, uniqueness of the limit shows in fact that the whole sequences are strongly converging in L^1 to the same limit

$$\forall\, t \geq 0: \qquad \lim_{L^1} \rho^\tau(t) = \lim_{\mathrm{w}-*} \rho^\tau(t) = \rho(t) = \lim_{\mathrm{w}-*} \tilde\rho^\tau(t) = \lim_{L^1} \tilde\rho^\tau(t).$$

From this strong pointwise-in time L^1 convergence and the uniform $L^\infty(0,\infty;L^1)$ bounds from Proposition 5.1, an easy application of Lebesgue's dominated convergence theorem in any finite time interval $[0,T]$ finally gives strong $L^1((0,T);L^1)$ convergence for all $T > 0$. ∎

Before giving the second proof we need a well known technical result:

LEMMA 5.1. *Let* μ_0, μ_1 *be any absolutely continuous measures with finite second moments, same mass* $|\mu_0| = |\mu_1|$, *and bounded in* $L^p(\Omega)$ *for some* $1 \leq p \leq \infty$ *by the same constant* C_p. *Then*

$$\forall\, \phi \in W^{1,2p'}(\Omega): \qquad \left|\int_\Omega (\mu_1 - \mu_0)\phi\right| \leq \sqrt{C_p}\mathtt{MK}(\mu_0,\mu_1)\|\nabla\phi\|_{L^{2p'}},$$

with the convention $1' = \infty$ *and* $\infty' = 1$.

Proof. Let $(\mu_t, \mathbf{v}_t)_{t\in[0,1]}$ be the unique Monge-Kantorovich geodesics from μ_0 to μ_1, satisfying $\partial_t\mu_t + \mathrm{div}(\mu_t\mathbf{v}_t) = 0$ with constant metric speed $\|\mathbf{v}_t\|_{L^2(\mathrm{d}\mu_t)} = cst = \mathtt{MK}(\mu_0,\mu_1)$. We first claim that $\|\mu_t\|_{L^p} \leq C_p$ as well along this geodesics. Indeed if $p = 1$ this is simply the mass conservation, and the proof for $p = \infty$ can be found in [33]. For $1 < p < \infty$ this is a simple consequence of the displacement convexity of $\mathcal{E}_p[\mu] = \int_\Omega \frac{\mu^p}{p-1}$, [41, thm. 5.15]. Using the weak formulation of the continuity equation, we compute by Hölder's inequality

$$\begin{aligned}\left|\int_\Omega (\mu_1-\mu_0)\phi\right| &= \left|\int_0^1\int_\Omega \mathbf{v}_t\cdot\nabla\phi\,\mathrm{d}\mu_t\mathrm{d}t\right| \leq \int_0^1 \left(\int_\Omega |\mathbf{v}_t|^2\mathrm{d}\mu_t\right)^{\frac12}\left(\int_\Omega |\nabla\phi|^2\mu_t\right)^{\frac12}\mathrm{d}t\\ &\leq \mathtt{MK}(\mu_0,\mu_1)\int_0^1 (\|\mu_t\|_{L^p}\||\nabla\phi|^2\|_{L^{p'}})^{\frac12}\mathrm{d}t \leq \sqrt{C_p}\mathtt{MK}(\mu_0,\mu_1)\|\nabla\phi\|_{L^{2p'}}\end{aligned}$$

and the proof is complete. ∎

Second proof of Proposition 5.2. Here we assume that Ω is bounded for simplicity, but the same argument would actually work for unbounded domains simply replacing all the functional spaces by their local counterparts ($\mathrm{BV}_{\mathrm{loc}}$, H^1_{loc}, $L^p_{\mathrm{loc}}\dots$).

We first control the difference quotient $\|\rho^{n+1}-\rho^n\|_Y$ in the dual space $Y:=H^1(\Omega)^*$. For the Monge-Kantorovich step we can apply the previous Lemma 5.1 with $p=\infty$, $2p'=2$, $\|\rho^{n+\frac12}\|_{L^\infty}\le\|\rho^n\|_{L^\infty}\le\|\rho^0\|_{L^\infty}$ and obtain by duality

$$\|\rho^{n+\frac12}-\rho^n\|_Y\le C\,\mathtt{MK}(\rho^{n+\frac12},\rho^n).$$

For the reaction step we recall the Euler-Lagrange equation (36), which reads for $\Psi,K\equiv0$

$$\forall\,\phi\in\mathcal{C}^\infty_c(\Omega):\qquad \int_\Omega(\rho^{n+1}-\rho^{n+\frac12})\phi=-\tau\int_\Omega\frac{\sqrt{\rho^{n+1}}(\sqrt{\rho^{n+1}}+\sqrt{\rho^{n+\frac12}})}{2}U'(\rho^{n+1})\phi.$$

Because in the right-hand side $\rho^{n+\frac12},\rho^{n+1}$ are bounded in $L^1\cap L^\infty(\Omega)$ uniformly in n this gives

$$\|\rho^{n+1}-\rho^{n+\frac12}\|_Y\le\|\rho^{n+\frac12}-\rho^{n+1}\|_{L^2}\le C\tau.$$

By triangular inequality we deduce from the previous two estimates that

$$\|\rho^{n+1}-\rho^n\|_Y\le C(\tau+\mathtt{MK}(\rho^{n+1},\rho^n)),$$

and using the square distance estimate (26) and Cauchy-Schwarz inequality we obtain the approximate equicontinuity

$$\forall\,0\le t_1\le t_2:\qquad \|\rho^\tau(t_2)-\rho^\tau(t_1)\|_Y\le C(|t_2-t_1+\tau|+|t_2-t_1+\tau|^{\frac12}).$$

Because the embedding $H^1\subset\subset L^2$ is compact we have $L^2\subset\subset Y$ as well. Thanks to the $L^1\cap L^\infty(\Omega)$ bounds from Proposition 5.1 we have τ-uniform bounds $\|\rho^\tau(t)\|_{L^2}\le C$, and we see that there is a Y-relatively compact set $\mathcal{K}_Y=\{\|\rho\|_{L^2}\le C\}$ such that $\rho^\tau(t)\in\mathcal{K}_Y$ for all $t\ge0$. Exploiting the above equicontinuity we can apply again the same variant of the Arzelá-Ascoli theorem [3, prop. 3.3.1] in any bounded time interval to deduce that there exists a subsequence (not relabeled) and $\rho\in\mathcal{C}([0,T];Y)$ such that $\rho^\tau(t)\to\rho(t)$ in Y for all $t\in[0,T]$. A further application of Lebesgue's dominated convergence theorem with $\|\rho^\tau(t)\|_Y\le C$ shows that $\rho^\tau\to\rho$ in $L^p([0,T];Y)$ for all $p\ge1$ and fixed $T>0$, and by Cantor's procedure

$$\rho^\tau\to\rho\qquad\text{in } L^p_{\mathrm{loc}}([0,\infty);Y).$$

Let now $X:=\mathrm{BV}\cap L^\infty(\Omega)\subset\subset L^2(\Omega)=:B$. We just proved that

$$X\subset\subset B\subset Y\quad\text{and}\quad\begin{cases}\rho^\tau\text{ is bounded in }L^\infty(0,\infty;X),\\ \rho^\tau\text{ is relatively compact in }L^p_{\mathrm{loc}}([0,\infty);Y)\end{cases}$$

for all $p\ge1$. By standard Aubin-Lions-Simon theory [40, lem. 9] we get that ρ^τ is relatively compact in $L^p_{\mathrm{loc}}([0,\infty);B)$ for all $p\ge1$. In particular we get pointwise a.e. convergence $\rho^\tau(t,x)\to\rho(t,x)$ (up to extraction of a further subsequence), and a last application of Lebesgue's dominated convergence allows to conclude. The argument is identical for $\tilde\rho^\tau$. ∎

In order to show that the nonlinear terms pass to the limit as in (37) we shall need the following variant of the Banach-Alaoglu theorem with varying measures:

Lemma 5.2 (compactness for vector-fields). *Let $\mathcal{O}\subset\mathbb{R}^m$ be an open set (not necessarily bounded), $\{\sigma_n\}_{n\ge0}\subset\mathcal{M}^+(\mathcal{O})$ a sequence of finite nonnegative Radon measures narrowly converging to $\sigma\in\mathcal{M}^+(\mathcal{O})$, and $\mathbf{v}_n$ a sequence of vector fields on $\mathcal{O}$. If*

$$\|\mathbf{v}_n\|_{L^2(\mathcal{O},\mathrm{d}\sigma_n;\mathbb{R}^m)}\le C$$

then there exists $\mathbf{v}\in L^2(\mathcal{O},\mathrm{d}\sigma;\mathbb{R}^m)$ such that, up to extraction of some subsequence,

$$\forall\,\zeta\in\mathcal{C}^\infty_c(\mathcal{O};\mathbb{R}^m):\qquad\lim_{n\to\infty}\int_{\mathcal{O}}\mathbf{v}_n\cdot\zeta\mathrm{d}\sigma_n=\int_{\mathcal{O}}\mathbf{v}\cdot\zeta\mathrm{d}\sigma$$

and

$$\|\mathbf{v}\|_{L^2(\mathcal{O},\mathrm{d}\sigma;\mathbb{R}^m)} \le \liminf_{n\to\infty} \|\mathbf{v}_n\|_{L^2(\mathcal{O},\mathrm{d}\sigma_n;\mathbb{R}^m)}.$$

The proof can be found in [3, thm. 5.4.4] for probability measures, see also [25, prop. 5.3] for an abstract version. As anticipated, we have now

PROPOSITION 5.3. *Assume* (H). *Then $\rho^\tau, \tilde\rho^\tau$ satisfy the compactness assumption* (37) *in Theorem 4.1.*

Proof. From the strong $L^1_{\mathrm{loc}}([0,\infty);L^1)$ convergence in Proposition 5.2 and the uniform $L^1\cap L^\infty(\Omega)$ bounds in Proposition 5.1, a straightforward application of Lebesgue's dominated convergence theorem yields strong convergence $\sqrt{\rho^\tau}\frac{\sqrt{\tilde\rho^\tau}+\sqrt{\rho^\tau}}{2}U'(\rho^\tau)\to\rho U'(\rho)$ at least in $L^1_{\mathrm{loc}}((0,\infty)\times\Omega)$. Therefore the reaction terms pass to the limit as in (37), and we only have to check that the diffusion part does too.

Let $\mathbf{t}^{n+\frac12}$ be the (backwards) optimal map from $\rho^{n+\frac12}$ to ρ^n, and recall that the Euler-Lagrange equation (31) holds with $\mu=\rho^n$ and minimizer $\rho^*=\rho^{n+\frac12}$. An easy density argument shows that (31) can in fact be written as $\frac{\mathrm{id}-\mathbf{t}^{n+\frac12}}{\tau}=-\nabla U'(\rho^{n+\frac12})$ in $L^2(\mathrm{d}\rho^{n+\frac12})$, which should be interpreted as an equality in the tangent plane $T_{\rho^{n+\frac12}}\mathcal{M}^+_{\mathtt{MK}}$. Taking thus the $L^2(\mathrm{d}\rho^{n+\frac12})$ norm we obtain

$$\tau\|\nabla U'(\rho^{n+\frac12})\|^2_{L^2(\mathrm{d}\rho^{n+\frac12})}=\frac{1}{\tau}\|\mathrm{id}-\mathbf{t}^{n+\frac12}\|^2_{L^2(\mathrm{d}\rho^{n+\frac12})}=\frac{1}{\tau}\mathtt{MK}^2(\rho^{n+\frac12},\rho^n).$$

Recalling that the interpolated curve $\tilde\rho^\tau(t)$ is piecewise constant and summing from $n=0$ to $n=\lceil T/\tau\rceil+1$ for fixed any $T>0$, we obtain from the total square-distance estimate (26)

$$\int_0^T\int_\Omega |\nabla U'(\tilde\rho^\tau(t))|^2\mathrm{d}\tilde\rho^\tau(t)\,\mathrm{d}t\le C \quad\Leftrightarrow\quad \int_{\mathcal{O}}|\nabla U'(\tilde\rho^\tau)|^2\mathrm{d}\sigma^\tau\le C \tag{41}$$

with $\mathcal{O}=(0,T)\times\Omega\subset\mathbb{R}^{1+d}$ and $\mathrm{d}\sigma^\tau(t,x)=\mathrm{d}\tilde\rho^\tau_t(x)\otimes\mathrm{d}t$. Recall that $\|\tilde\rho^\tau(t)\|_{L^1(\Omega)}\le\|\rho^0\|_\Omega$, so that σ^τ is really a *finite* measure on $\mathcal{O}$ for finite $T>0$. From the strong $L^1_{\mathrm{loc}}([0,\infty);L^1)$ convergence $\tilde\rho^\tau\to\rho$ (Proposition 5.2) it is easy to check that σ^τ converges narrowly to $\mathrm{d}\sigma(t,x)=\mathrm{d}\rho_t(x)\otimes\mathrm{d}t=\rho(t,x)\mathrm{d}x\mathrm{d}t$. Applying Lemma 5.2 we see that there is a vector-field $\mathbf{v}\in L^2(\mathcal{O},\mathrm{d}\sigma)=L^2(0,T;L^2(\mathrm{d}\rho_t))$ such that, up to extraction of a subsequence,

$$\int_0^T\int_\Omega\tilde\rho^\tau\nabla U'(\tilde\rho^\tau)\cdot\zeta\to\int_0^T\int_\Omega\rho(t,x)\mathbf{v}(t,x)\cdot\zeta(t,x)\,\mathrm{d}x\mathrm{d}t$$

for all $\zeta\in\mathcal{C}^\infty_c((0,T)\times\Omega;\mathbb{R}^n)$. In order to identify the weak limit $\mathbf{v}$, recall that the thermodynamic pressure $P(\rho):=\rho U'(\rho)-U(\rho)$. Since $P'(\rho)=\rho U''(\rho)$ our assumptions on U show that P is Lipschitz in any bounded interval $\rho\in[0,M]$. With the strong convergence $\rho^\tau\to\rho$ and the uniform $L^1\cap L^\infty(\Omega)$ bounds one immediately gets $P(\tilde\rho^\tau)\to P(\rho)$ in $L^1_{\mathrm{loc}}((0,\infty)\times\Omega)$, and as a consequence $\nabla P(\tilde\rho^\tau)\rightharpoonup\nabla P(\rho)$ in the sense of distributions $\mathcal{D}'((0,T)\times\Omega)$. Note that the measure $\mathrm{d}\sigma(t,x)=\mathrm{d}\rho_t(x)\otimes\mathrm{d}t$ is finite on any subdomain $(0,T)\times\Omega$, hence $\mathbf{v}\in L^2(\mathcal{O},\mathrm{d}\sigma)\subset L^1(\mathcal{O},\mathrm{d}\sigma)$ and $\rho\mathbf{v}\in L^1((0,T)\times\Omega)$. Writing $\nabla P(\rho)=P'(\rho)\nabla\rho=\rho U''(\rho)\nabla\rho=\rho\nabla U'(\rho)$ we conclude that $\rho\mathbf{v}=\nabla P(\rho)=\rho\nabla U'(\rho)$, thus $\mathbf{v}=\nabla U'(\rho)$ at least in $L^2(\mathrm{d}\rho)$. A further diagonal extraction shows that the limit $\mathbf{v}$ can be chosen independent of T, and the proof is complete. ∎

As an immediate consequence, we get

THEOREM 5.1. *Assume* (H). *Then, up to extraction of a discrete subsequence not relabeled here, the solution of the* `MK-FR` *splitting scheme ρ^τ converges to a weak solution ρ of the PDE* (21).

Proof. Simply use Proposition 5.3 to apply Theorem 4.1. ∎

Our next and final result illustrates perhaps even better the deep interplay between our two-steps variational discretization and the full KFR metric:

PROPOSITION 5.4. *In addition to (H), assume that $\mathcal{F}(\rho)$ is geodesically convex with respect to the* MK *structure, i-e $\rho P'(\rho) \geq \left(1 - \frac{1}{d}\right) P(\rho)$ with $P(\rho) = \rho U'(\rho) - U(\rho)$ [41]. Then we have*

$$\mathcal{F}(\rho(t_2)) + \int_{t_1}^{t_2} \int_\Omega (|\nabla U'(\rho)|^2 + |U'(\rho)|^2) \, \mathrm{d}\rho \, \mathrm{d}t \leq \mathcal{F}(\rho(t_1)) \tag{42}$$

and for all $0 \leq t_1 \leq t_2$.

From the discussion in section 2.3 we known that $\|U'(\rho)\|^2_{H^1(\mathrm{d}\rho)}$ can be interpreted either as the metric slope $|\partial \mathcal{F}(\rho)|^2 = \|\operatorname{grad}_{\mathtt{KFR}} \mathcal{F}(\rho)\|^2_{\mathtt{KFR}}$ or, through the continuity equation $\partial_t \rho = \operatorname{div}(\rho \nabla U'(\rho)) - \rho U'(\rho)$, as the metric speed $|\rho'(t)|^2$ with respect to our distance KFR. Hence (42) can be rephrased as the Energy Dissipation Inequality (EDI)

$$\mathcal{F}(\rho(t_2)) + \int_{t_1}^{t_2} \left\{ \frac{1}{2}|\rho'(t)|^2 + \frac{1}{2}|\partial \mathcal{F}(\rho(t))|^2 \right\} \mathrm{d}t \leq \mathcal{F}(\rho(t_1)),$$

which is one of the possible formulations of gradient flows in abstract metric spaces. We refer the reader to [2, 3] for the connection between EDIs in abstract metric spaces and gradient flow formulations. However, and to the best of our knowledge, no full and tractable characterizations of metric speeds $|\rho'(t)|$ and metric slopes $|\partial \mathcal{F}(\rho)|$ are available at this early stage of the general KFR theory (see however [25] for the characterization of Lipschitz curves). For the sake of rigor we thus prefer to state the dissipation inequality in the PDE-oriented form (42), rather than in the abstract metric setting.

Note that (H) already implies $\rho U''(\rho) + U'(\rho)/2 \geq 0$, which is equivalent to geodesic convexity with respect to FR. Thus we essentially assumed here that $\mathcal{F}$ is separately geodesically convex with respect to each of the MK, FR structures, respectively, and it is not surprising that we recover in the end a dissipation inequality for the full KFR metrics.

Proof. Let $\mathbf{t}^{n+\frac{1}{2}}$ be the optimal map from $\rho^{n+\frac{1}{2}}$ to ρ^n. By the above-tangent characterization of the displacement convexity with respect to MK [41, prop. 5.29] we have

$$\begin{aligned}
\mathcal{F}(\rho^n) &\geq \mathcal{F}(\rho^{n+\frac{1}{2}}) + \int_\Omega (\mathbf{t}^{n+\frac{1}{2}} - \mathrm{id}) \cdot \nabla U'(\rho^{n+\frac{1}{2}}) \mathrm{d}\rho^{n+\frac{1}{2}} \\
&= \mathcal{F}(\rho^{n+\frac{1}{2}}) + \tau \int_\Omega |\nabla U'(\rho^{n+\frac{1}{2}})|^2 \mathrm{d}\rho^{n+\frac{1}{2}},
\end{aligned}$$

where the last equality follows by reinterpreting the Euler-Lagrange (31) as $\mathbf{t}^{n+\frac{1}{2}} - \mathrm{id} = \tau \nabla U'(\rho^{n+\frac{1}{2}})$ in $L^2(\mathrm{d}\rho^{n+\frac{1}{2}})$.

For the reaction part let us recall that $\rho U''(\rho) + \frac{U'(\rho)}{2} \geq 0$ corresponds to the convexity of $s \mapsto U(s^2)$ in $s = \sqrt{\rho}$. Using this convexity we obtain

$$\begin{aligned}
\mathcal{F}(\rho^{n+\frac{1}{2}}) &\geq \mathcal{F}(\rho^{n+1}) + \int_\Omega 2\sqrt{\rho^{n+1}} U'(\rho^{n+1}) \left(\sqrt{\rho^{n+\frac{1}{2}}} - \sqrt{\rho^{n+1}} \right) \\
&= \mathcal{F}(\rho^{n+1}) + \tau \int_\Omega |U'(\rho^{n+1})|^2 \mathrm{d}\rho^{n+1},
\end{aligned}$$

where the last equality follows now by reinterpreting the Euler-Lagrange equation (35) as $2\frac{\sqrt{\rho^{n+1}} - \sqrt{\rho^{n+\frac{1}{2}}}}{\sqrt{\rho^{n+1}}} = -\tau U'(\rho^{n+1})$ in $L^2(\mathrm{d}\rho^{n+1})$. We get altogether

$$\mathcal{F}(\rho^{n+1}) + \tau \left(\int_\Omega |\nabla U'(\rho^{n+\frac{1}{2}})|^2 \mathrm{d}\rho^{n+\frac{1}{2}} + \int_\Omega |U'(\rho^{n+1})|^2 \mathrm{d}\rho^{n+1} \right) \leq \mathcal{F}(\rho^n).$$

For any $0 \le t_1 \le t_2$ let now $N_1, N_2 \in \mathbb{N}$ such that $\rho^\tau(t_i) = \rho^{N_i}$, and $T_i = N_i\tau$. Summing the previous inequality from $n = N_1$ to $n = N_2 - 1$ gives

$$\mathcal{F}(\rho^\tau(t_2)) + \int_{T_1}^{T_2}\int_\Omega |\nabla U'(\tilde\rho^\tau)|^2 \,\mathrm{d}\tilde\rho^\tau \mathrm{d}t + \int_{T_1}^{T_2}\int_\Omega |U'(\rho^\tau)|^2 \,\mathrm{d}\rho^\tau \mathrm{d}t \le \mathcal{F}(\rho^\tau(t_1)). \tag{43}$$

We proved in Proposition 5.3 that $\tilde\rho^\tau \nabla U'(\tilde\rho^\tau) \rightharpoonup \rho\nabla U'(\rho)$, and observe that $T_i \to t_i$ as $\tau \to 0$. From the energy estimate (41) and the lower semi-continuity in Lemma 5.2 we deduce that

$$\int_{t_1}^{t_2}\int_\Omega |\nabla U'(\rho)|^2 \,\mathrm{d}\rho \,\mathrm{d}t \le \liminf_{\tau\to 0} \int_{T_1}^{T_2}\int_\Omega |\nabla U'(\tilde\rho^\tau)|^2 \,\mathrm{d}\tilde\rho^\tau \mathrm{d}t,$$

and from the strong convergence in Proposition 5.2 with the uniform $L^1 \cap L^\infty(\Omega)$ bounds (Proposition 5.1) it is easy to see that

$$\int_{t_1}^{t_2}\int_\Omega |U'(\rho)|^2 \,\mathrm{d}\rho \,\mathrm{d}t = \lim_{\tau\to 0} \int_{T_1}^{T_2}\int_\Omega |U'(\rho^\tau)|^2 \,\mathrm{d}\rho^\tau \mathrm{d}t.$$

Similarly one can verify that

$$\forall\, t \ge 0: \qquad \mathcal{F}(\rho^\tau(t)) = \int_\Omega U(\rho^\tau(t)) \to \int_\Omega U(\rho(t)) = \mathcal{F}(\rho(t)).$$

Indeed with our assumptions U is Lipschitz in any bounded interval $\rho \in [0, M]$, $\|\rho^\tau(t)\|_{L^\infty} \le M = \|\rho^0\|_{L^\infty}$ uniformly in τ, and in the first proof of Proposition 5.2 we obtained strong $L^1(\Omega)$ convergence $\rho^\tau(t) \to \rho(t)$ pointwise in time. As a consequence we can pass to the lim inf in (43) to retrieve (42) and the proof is complete. ∎

Acknowledgments. LM was partially supported by the Portuguese National Science Foundation through fellowship BPD/88207/2012 and by the UT Austin/Portugal CoLab program *Phase Transitions and Free Boundary Problems*. T. O. Gallouët was supported by the ANR project ISOTACE (ANR-12-MONU-013) hosted at CMLS, École polytechnique, CNRS, Université Paris-Saclay and by the fond de la Recherche Scientifique-FNRS under grant MIS F.4539.16. We wish to thank the anonymous referees for their useful comments and suggestions.

REFERENCES

[1] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*, volume 254. Clarendon Press Oxford, 2000.

[2] Luigi Ambrosio and Nicola Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.

[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] 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.

[5] Jean-David Benamou, Guillaume Carlier, and Maxime Laborde. An augmented Lagrangian approach to Wasserstein gradient flows and applications. working paper or preprint, December 2015.

[6] Jean-David Benamou, Guillaume Carlier, Quentin Mérigot, and Edouard Oudet. Discretization of functionals involving the Monge-Ampère operator. *arXiv preprint arXiv:1408.4536, to appear in Numerische Mathematik*, 2014.

[7] A Blanchet, JA Carrillo, D Kinderlehrer, M Kowalczyk, P Laurençot, and S Lisini. A hybrid variational principle for the Keller-Segel system in $\mathbb{R}^2$. *ESAIM M2AN*, 2015.

[8] Adrien Blanchet and Philippe Laurençot. The parabolic-parabolic Keller-Segel system with critical diffusion as a gradient flow in $\mathbb{R}^d, d \ge 3$. *Communications in Partial Differential Equations*, 38(4):658–686, 2013.

[9] Vladimir I Bogachev. *Measure theory*, volume 1 & 2. Springer Science & Business Media, 2007.

[10] Malcolm Bowles and Martial Agueh. Weak solutions to a fractional Fokker–Planck equation via splitting and Wasserstein gradient flow. *Applied Mathematics Letters*, 42:30–35, 2015.

[11] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.
[12] Guillaume Carlier and Filippo Santambrogio. A variational model for urban planning with traffic congestion. *ESAIM: Control, Optimisation and Calculus of Variations*, 11(04):595–613, 2005.
[13] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.
[14] Lenaic Chizat, Bernhard Schmitzer, Gabriel Peyré, and François-Xavier Vialard. An interpolating distance between optimal transport and Fischer-Rao. *arXiv preprint arXiv:1506.06430*, 2015.
[15] Ennio De Giorgi. New problems on minimizing movements. In *Boundary Value Problems for PDEs and their Applications, eds. Massons*, pages 81–98, 1993.
[16] Guido De Philippis, Alpár Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. Bv estimates in optimal transportation and applications. *arXiv preprint arXiv:1503.06389*, 2015.
[17] Thomas Gallouët, Maxime Laborde, and Léonard Monsaingeon. A splitting scheme for very degenerate advection-reaction-diffusion equations. *In preparation*, 2016.
[18] Thomas Gallouët and François-Xavier Vialard. From unbalanced optimal transport to the camassa-holm equation. *arXiv preprint arXiv:1609.04006*, 2016.
[19] Wilfrid Gangbo and Robert J McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.
[20] Ernst Hellinger. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 136:210–271, 1909.
[21] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.
[22] Shizuo Kakutani. On equivalence of infinite product measures. *Annals of Mathematics*, pages 214–224, 1948.
[23] David Kinderlehrer and Michal Kowalczyk. The Janossy effect and hybrid variational principles. *Discrete Contin. Dyn. Syst. Ser. B*, 11(1):153–176, 2009.
[24] David Kinderlehrer and Noel J Walkington. Approximation of parabolic equations using the wasserstein metric. *ESAIM: Modélisation Mathématique et Analyse Numérique*, 33(4):837–852, 1999.
[25] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite radon measures. *arXiv preprint arXiv:1505.07746*, 2015.
[26] Maxime Laborde. *Systèmes de particules en interaction, approche par flot de gradient dans l'espace de Wasserstein*. PhD thesis, Université Paris-Dauphine, 2016.
[27] Matthias Liero and Alexander Mielke. Gradient structures and geodesic convexity for reaction–diffusion systems. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 371(2005):20120346, 2013.
[28] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal Entropy-Transport problems and a new Hellinger-Kantorovich distance between positive measures. *arXiv preprint arXiv:1508.07941*, 2015.
[29] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger-Kantorovich distance and geodesic curves. *arXiv preprint arXiv:1509.00068*, 2015.
[30] Daniel Matthes, Robert J McCann, and Giuseppe Savaré. A family of nonlinear fourth order equations of gradient flow type. *Communications in Partial Differential Equations*, 34(11):1352–1397, 2009.
[31] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.
[32] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.
[33] Felix Otto. Dynamics of labyrinthine pattern formation in magnetic fluids: A mean-field theory. *Archive for Rational Mechanics and Analysis*, 141(1):63–103, 1998.
[34] Felix Otto. The geometry of dissipative evolution equations: the Porous Medium Equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[35] Benoît Perthame, Fernando Quirós, and Juan Luis Vázquez. The Hele–Shaw asymptotics for mechanical models of tumor growth. *Archive for Rational Mechanics and Analysis*, 212(1):93–127, 2014.
[36] Gabriel Peyré. Entropic approximation of wasserstein gradient flows. *SIAM Journal on Imaging Sciences*, 8(4):2323–2351, 2015.
[37] Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.
[38] Filippo Santambroggio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Number 58. Birkhaüser, 2015.
[39] Filippo Santambrogio. Transport and concentration problems with interaction effects. *Journal of Global Optimization*, 38(1):129–141, 2007.
[40] Jacques Simon. Compact sets in the space $L^p(0,T;B)$. *Annali di Matematica pura ed applicata*, 146(1):65–96, 1986.
[41] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.
[42] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

An unbalanced Optimal Transport splitting scheme for general advection-reaction-diffusion problems

T.O. Gallouët, M. Laborde, L. Monsaingeon

September 26, 2023

Abstract

In this paper, we show that unbalanced optimal transport provides a convenient framework to handle reaction and diffusion processes in a unified metric framework. We use a constructive method, alternating minimizing movements for the Wasserstein distance and for the Fisher-Rao distance, and prove existence of weak solutions for general scalar reaction-diffusion-advection equations. We extend the approach to systems of multiple interacting species, and also consider an application to a very degenerate diffusion problem involving a Gamma-limit. Moreover, some numerical simulations are included.

1 Introduction

Since the seminal works of Jordan-Kinderlehrer-Otto [19], it is well known that certain diffusion equations can be interpreted as gradient flows in the space of probability measures, endowed with the quadratic Wasserstein distance $\mathtt{W}$. The well-known JKO scheme (a.k.a. minimizing movement), which is a natural implicit Euler scheme for such gradient flows, naturally leads to constructive proofs of existence for weak solutions to equations or systems with mass conservation such as, for instance, Fokker-Planck equations [19], Porous Media Equations [32], aggregation equation [9], double degenerate diffusion equations [31], general degenerate parabolic equation [1] etc. We refer to the classical textbooks of Ambrosio, Gigli and Savaré [4] and to the books of Villani [43, 44] for a detailed account of the theory and extended bibliography. Recently, this theory has been extended to study the evolution of interacting species with mass-conservation, see for examples [15, 45, 23, 20, 8].

Nevertheless in biology, for example for diffusive prey-predator models, the conservation of mass may not hold, and the classical optimal transport theory does not apply. An unbalanced optimal transport theory was recently introduced simultaneously in [11, 12, 21, 25, 26], and the resulting Wasserstein-Fisher-Rao ($\mathtt{WFR}$) metrics (also referred to as the Hellinger-Kantorovich distance $\mathtt{HK}$) allows to compute distances between measures with variable masses while retaining a convenient Riemannian structure. See section 2 for the definition and a short discussions on this $\mathtt{WFR}$ metric. We also refer to [37, 16] for earlier attempts to account for mass variations within the framework of optimal transport.

The $\mathtt{WFR}$ metrics can be seen as an *inf-convolution* between Wasserstein/transport and Fisher-Rao/reaction processes, and is therefore extremely convenient to control both in a unified metric setting. This allows to deal with non-conservative models of population dynamics, see e.g. [21, 22]. In [18], the first and third authors proposed a variant of the JKO scheme for $\mathtt{WFR}$-gradient flows corresponding to some particular class of reaction-diffusion PDEs: roughly speaking, the reaction and diffusion were handled separately in two separate $\mathtt{FR}, \mathtt{W}$ metrics, and then patched together using a particular uncoupling of the inf-convolution, namely $\mathtt{WFR}^2 \approx \mathtt{W}^2 + \mathtt{FR}^2$ in some sense (see [18, section 3] for a thorough discussion). However, the analysis was restricted to very particular structures for the PDE, corresponding to pure $\mathtt{WFR}$ gradient-flows.

In this work we aim at extending this splitting scheme in order to handle more general reaction-diffusion problems, not necessarily corresponding to gradient flows. Roughly speaking, the structure of our splitting scheme is the following: the transport/diffusion part of the PDE is treated by a

single Wasserstein JKO step

$$\rho^k \xrightarrow[\text{transport}]{\mathtt{W}} \rho^{k+1/2},$$

and the next Fisher-Rao JKO step

$$\rho^{k+1/2} \xrightarrow[\text{reaction}]{\mathtt{FR}} \rho^{k+1}$$

handles the reaction part of the evolution. As already mentioned, the WFR metric will allow to suitable control both steps in a unified metric framework. We will first state a general convergence result for scalar reaction-diffusion equations, and then illustrate on a few particular examples how the general idea can be adapted to treat e.g. prey-predator systems or very degenerate Hele-Shaw diffusion problems. In this work we do not focus on optimal results and do not seek full generality, but rather wish to illustrate the efficiency of the general approach.

Another advantage of the splitting scheme is that is well adapted to existing Monge/Kantorovich/Wasserstein numerical solvers, and the Fisher-Rao step turns out to be a simple pointwise convex problem which can be implemented in a very simple way. See also [10, 13] for a more direct numerical approach by entropic regularization. Throughout the paper we will illustrate the theoretical results with a few numerical tests. All the numerical simulations were implemented with the augmented Lagrangian ALG2-JKO scheme from [6] for the Wasserstein step, and we used a classical Newton algorithm for the Fisher-Rao step.

The paper is organized as follows. In section 2 we recall the basic definitions and useful properties of the Wasserstein-Fisher-Rao distance WFR. Section 3 contains the precise description of the splitting scheme and a detailed convergence analysis for a broad class of reaction-diffusion equations. In section 4 we present an extension to some prey-predator multicomponent systems with nonlocal interactions. In section 5 we extend the general result from section 3 to a very degenerate tumor growth model studied in [34], corresponding to a pure WFR gradient flow: we show that the splitting scheme captures fine properties of the model, particularly the Γ-convergence of discrete gradient flows as the degenerate diffusion parameter of Porous Medium type $m \to \infty$. The last section 6 contains an extension to a tumor-growth model coupled with an evolution equation for the nutrients.

2 Preliminaries

Let us first fix some notations. Throughout the whole paper, Ω denotes a possibly unbounded convex subset of $\mathbb{R}^d$, Q_T represents the product space $[0,T] \times \Omega$, for $T > 0$, and we write $\mathcal{M}^+ = \mathcal{M}^+(\Omega)$ for the set of nonnegative finite Radon measures on Ω. We say that a curve of measures $t \mapsto \rho_t \in \mathcal{C}_w([0,1]; \mathcal{M}^+)$ is narrowly continuous if it is continuous with respect to the narrow convergence of measures, namely for the duality with $\mathcal{C}_b(\Omega)$ test-functions.

Definition 2.1. *The Fisher-Rao distance between $\rho_0, \rho_1 \in \mathcal{M}^+$ is*

$$\mathtt{FR}(\rho_0, \rho_1) := \min_{(\rho_t, r_t) \in \mathcal{A}_{\mathtt{FR}}[\rho_0, \rho_1]} \int_0^1 \int_\Omega |r_t|^2 \, d\rho_t(x) dt,$$

where the admissible set $A_{\mathtt{FR}}[\rho_0, \rho_1]$ consists in curves $[0,1] \ni t \mapsto (\rho_t, r_t) \in \mathcal{M}^+ \times \mathcal{M}$ such that $t \mapsto \rho_t \in \mathcal{C}_w([0,1]; \mathcal{M}^+)$ is narrowly continuous with endpoints $\rho_t(0) = \rho_0, \rho_t(1) = \rho_1$, and

$$\partial_t \rho_t = \rho_t r_t$$

in the sense of distributions $\mathcal{D}'((0,1) \times \Omega)$.

The Monge-Kantorovich-Wasserstein admits several equivalent definitions and formulations, and we refer e.g. to [43, 44, 4, 41] for a complete description. For our purpose we shall only need the dynamical Benamou-Brenier formula:

Theorem 2.2 (Benamou-Brenier formula, [5, 4]). *There holds*

$$\mathtt{W}^2(\rho_0,\rho_1)=\min_{(\rho,\mathbf{v})\in\mathcal{A}_{\mathtt{W}}[\rho_0,\rho_1]}\int_0^1\int_\Omega|\mathbf{v}_t|^2\mathrm{d}\rho_t\mathrm{d}t, \tag{2.1}$$

where the admissible set $\mathcal{A}_{\mathtt{W}}[\rho_0,\rho_1]$ *consists in curves* $(0,1)\ni t\mapsto(\rho_t,\mathbf{v}_t)\in\mathcal{M}^+\times\mathcal{M}(\Omega;\mathbb{R}^d)$ *such that* $t\mapsto\rho_t$ *is narrowly continuous with endpoints* $\rho_t(0)=\rho_0$, $\rho_t(1)=\rho_1$ *and solving the continuity equation*

$$\partial_t\rho_t+\mathrm{div}(\rho_t\mathbf{v}_t)=0$$

in the sense of distributions $\mathcal{D}'((0,1)\times\Omega)$.

According to the original definition in [11] we have

Definition 2.3. *The Wasserstein-Fisher-Rao distance between* $\rho_0,\rho_1\in\mathcal{M}^+(\Omega)$ *is*

$$\mathtt{WFR}^2(\rho_0,\rho_1):=\inf_{(\rho,\mathbf{v},r)\in\mathcal{A}_{\mathtt{WFR}}[\rho_0,\rho_1]}\int_0^1\int_\Omega(|\mathbf{v}_t(x)|^2+|r_t|^2)\,d\rho_t(x)dt, \tag{2.2}$$

where the admissible set $\mathcal{A}_{\mathtt{WFR}}[\rho_0,\rho_1]$ *is the set of curves* $t\in[0,1]\mapsto(\rho_t,v_t,r_t)\in\mathcal{M}^+\times\mathcal{M}(\Omega;\mathbb{R}^d)\times\mathcal{M}$ *such that* $t\mapsto\rho_t\in\mathcal{C}_w([0,1],\mathcal{M}^+)$ *is narrowly continuous with endpoints* $\rho_{|t=0}=\rho_0$, $\rho_{|t=1}=\rho_1$ *and solves the continuity equation with source*

$$\partial_t\rho_t+\mathrm{div}(\rho_t v_t)=\rho_t r_t.$$

Comparing definition 2.3 with definition 2.1 and Theorem 2.2, this dynamical formulation *à la Benamou-Brenier* shows that the WFR distance can be viewed as an inf-convolution of the Wasserstein and Fisher-Rao distances W, FR. From [11, 12, 21, 25] the infimum in (2.2) is always a minimum, and the corresponding minimizing curves $t\mapsto\rho_t$ are of course constant-speed geodesics $\mathtt{WFR}(\rho_t,\rho_s)=|t-s|\mathtt{WFR}(\rho_0,\rho_1)$. Then $(\mathcal{M}^+,\mathtt{WFR})$ is a complete metric space, and WFR metrizes the narrow convergences of measures (see again [11, 12, 21, 25]). Interestingly, there are other possible formulations of the distance in terms of static unbalanced optimal transportation, primal-dual characterizations with relaxed marginals, lifting to probability measures on a cone over Ω, duality with subsolutions of Hamilton-Jacobi equations, and we refer to [11, 12, 21, 26, 25] for more details.

As a first useful interplay between the distances WFR, W, FR we have

Proposition 2.4 ([18]). *Let* $\rho_0,\rho_1\in\mathcal{M}_2^+$ *such that* $|\rho_0|=|\rho_1|$. *Then*

$$\mathtt{WFR}^2(\rho_0,\rho_1)\leqslant\mathtt{W}^2(\rho_0,\rho_1).$$

Similarly for all $\mu_0,\mu_1\in\mathcal{M}^+$ *(with possibly different masses) there holds*

$$\mathtt{WFR}^2(\mu_0,\mu_1)\leqslant\mathtt{FR}^2(\mu_0,\mu_1).$$

Finally, for all $\nu_0,\nu_1\in\mathcal{M}_2^+$ *such that* $|\nu_0|=|\nu_1|$ *and all* $\nu\in\mathcal{M}^+$, *there holds*

$$\mathtt{WFR}^2(\nu_0,\nu)\leqslant2(\mathtt{W}^2(\nu_0,\nu_1)+\mathtt{FR}^2(\nu_1,\nu)).$$

Moreover, we have the following link between the reaction and the velocity in (2.2), which was the original definition in [21]:

Proposition 2.5 ([18]). *The definition* (2.3) *of the* WFR *distance can be restricted to the subclass of admissible paths* $(\mathbf{v}_t,r_t)=(\nabla u_t,u_t)$ *for potentials* $u_t\in H^1(\mathrm{d}\rho_t)$ *and continuity equations*

$$\partial_t\rho_t+\mathrm{div}(\rho_t\nabla u_t)=\rho_t u_t.$$

This shows that $(\mathcal{M}^+,\mathtt{WFR})$ can be endowed with the formal Riemannian structure constructed as follow: any two tangent vectors $\xi^1=\partial_t\rho^1,\xi^2=\partial_t\rho^2$ can be uniquely identified with potentials u^i by solving the elliptic equations

$$\xi^i=-\mathrm{div}(\rho\nabla u^i)+\rho u^i.$$

Then the Riemaniann tensor is naturally constructed on the $H^1(\mathrm{d}\rho)$ scalar product, i-e

$$g_\rho(\xi^1,\xi^2) := \langle u^1, u^2\rangle_{H^1(\mathrm{d}\rho)} = \int_\Omega (\nabla u^1 \cdot \nabla u^2 + u^1 u^2)\mathrm{d}\rho.$$

This is purely formal, and we refer again to [18] for discussions. Given a functional

$$\mathcal{F}(\rho) := \int_\Omega F(\rho) + \int_\Omega \rho V + \frac{1}{2}\int_\Omega (K*\rho)\rho,$$

this Riemannian structure also allows to compute WFR gradients as

$$\mathrm{grad}_{\mathtt{WFR}}\,\mathcal{F}(\rho) = -\,\mathrm{div}\left(\rho\nabla\frac{\delta\mathcal{F}}{\delta\rho}\right) + \rho\frac{\delta\mathcal{F}}{\delta\rho} = \mathrm{grad}_{\mathtt{W}}\,\mathcal{F}(\rho) + \mathrm{grad}_{\mathtt{FR}}\,\mathcal{F}(\rho),$$

where $\frac{\delta\mathcal{F}}{\delta\rho} = F'(\rho) + V + K*\rho$ denotes the Euclidean first variation of $\mathcal{F}$ with respect to ρ. In other words, the Riemannian tangent vector $\mathrm{grad}_{\mathtt{WFR}}\,\mathcal{F}(\rho)$ is represented in the previous $H^1(\mathrm{d}\rho)$ duality by the scalar potential $u = \frac{\delta\mathcal{F}}{\delta\rho}$.

3 An existence result for general parabolic equations

In this section, we propose to solve scalar parabolic equations of the form

$$\left\{\begin{array}{l} \partial_t\rho = \mathrm{div}(\rho\nabla(F_1'(\rho)+V_1)) - \rho(F_2'(\rho)+V_2) \\ \rho|_{t=0} = \rho_0 \\ \rho\nabla(F_1'(\rho)+V_1)|_{\partial\Omega}\cdot\nu = 0 \end{array}\right. \tag{3.1}$$

in a bounded domain $\Omega\subset\mathbb{R}^d$ with Neumann boundary condition and suitable initial conditions. Our goal is to extend to the case $F_1\neq F_2, V_1\neq V_2$ the method initially introduced in [18] for variational WFR-gradient flows, i-e (3.1) with $F_1 = F_2$ and $V_1 = V_2$.

We assume for simplicity that $F_1 : \mathbb{R}\to\mathbb{R}$ is given by

$$F_1(z) = \left\{\begin{array}{ll} z\log z - z & \text{(linear diffusion)} \\ \text{or} & \\ \frac{1}{m_1-1}z^{m_1} & \text{(Porous Media diffusion)} \end{array}\right. , \tag{3.2}$$

and $F_2 : \mathbb{R}\to\mathbb{R}$ is given by

$$F_2(z) = \frac{1}{m_2-1}z^{m_2}, \qquad \text{for some } m_2 > 1. \tag{3.3}$$

Note that we cannot take $F_2(z) = z\log z - z$ because the Boltzmann entropy is not well behaved (neither regular nor convex) with respect to the Fisher-Rao metric in the reaction step, see [18, 26, 25] for discussions. In addition, we assume that

$$V_1\in W^{1,\infty}(\Omega) \qquad \text{and} \qquad V_2\in L^\infty(\Omega).$$

We denote $\mathcal{E}_1,\mathcal{E}_2 : \mathcal{M}^+\to\mathbb{R}$ the energy functionals

$$\mathcal{E}_i(\rho) := \mathcal{F}_i(\rho) + \mathcal{V}_i(\rho),$$

where

$$\mathcal{F}_i(\rho) := \left\{\begin{array}{ll} \int_\Omega F_i(\rho) & \text{if } \rho\ll\mathcal{L}_{|\Omega} \\ +\infty & \text{otherwise,} \end{array}\right. \qquad \text{and} \qquad \mathcal{V}_i(\rho) := \int_\Omega V_i\rho.$$

Although more general statements with suitable structural assumptions could certainly be proved, we do not seek full generality here and choose to restrict from the beginning to the above simple (but nontrivial) setting for the sake of exposition.

Definition 3.1. *A weak solution of* (3.1) *is a curve* $[0,+\infty) \ni t \mapsto \rho(t,\cdot) \in L^1_+ \cap L^\infty(\Omega)$ *such that for all* $T < \infty$ *the pressure* $P_1(\rho) := \rho F_1'(\rho) - F_1(\rho)$ *satisfies* $\nabla P_1(\rho) \in L^2([0,T]\times\Omega)$*, and*

$$\int_0^{+\infty} \left(\int_\Omega (\rho \partial_t \phi - \nabla V_1 \cdot \nabla \phi \rho - \nabla P_1(\rho) \cdot \nabla \phi - \rho(F_2'(\rho) + V_2)\phi)\, dx \right) dt = - \int_\Omega \phi(0,x)\rho_0(x)\, dx$$

for every $\phi \in \mathcal{C}_c^\infty([0,+\infty)\times\mathbb{R}^d)$.

Note that the pressure P_1 is defined so that the diffusion term $\mathrm{div}(\rho\nabla F_1'(\rho)) = \Delta P_1(\rho)$, at least for smooth solutions.

The starting point of our analysis is that (3.1) can be written, at least formally as,

$$\partial_t \rho = \mathrm{div}(\rho\nabla(F_1'(\rho)+V_1)) - \rho(F_2'(\rho)+V_2) \quad \leftrightarrow \quad \partial_t \rho = -\,\mathrm{grad}_{\mathtt{W}}\,\mathcal{E}_1(\rho) - \mathrm{grad}_{\mathtt{FR}}\,\mathcal{E}_2(\rho).$$

Our splitting scheme is a variant of that originally introduced in [18], and can be viewed as an operator splitting method: each part of the PDE above is discretized (in time) in its own $\mathtt{W},\mathtt{FR}$ metric, and corresponds respectively to a $\mathtt{W}$/transport/diffusion step and to a $\mathtt{FR}$/reaction step. More precisely, let $h > 0$ be a small time step. Starting from the initial datum $\rho_h^0 := \rho_0$, we construct two recursive sequences $(\rho_h^k)_k$ and $(\rho_h^{k+1/2})_k$ such that

$$\left\{ \begin{array}{l} \rho_h^{k+1/2} \in \displaystyle\operatorname*{argmin}_{\rho\in\mathcal{M}^+, |\rho|=|\rho_h^k|} \left\{ \tfrac{1}{2h}\mathtt{W}^2(\rho,\rho_h^k) + \mathcal{E}_1(\rho) \right\}, \\ \\ \rho_h^{k+1} \in \displaystyle\operatorname*{argmin}_{\rho\in\mathcal{M}^+} \left\{ \tfrac{1}{2h}\mathtt{FR}_2^2(\rho,\rho_h^{k+1/2}) + \mathcal{E}_2(\rho) \right\}. \end{array} \right. \tag{3.4}$$

With our structural assumptions on F_i, V_i and arguing as in [18], the direct method shows that this scheme is well-posed, i-e that each minimizing problem in (3.4) admits a unique minimizer. We construct next two piecewise-constant interpolating curves

$$\left\{ \begin{array}{l} \rho_h(t) = \rho_h^{k+1}, \\ \tilde\rho_h(t) = \rho_h^{k+1/2}, \end{array} \right. \quad \text{for all } t \in (kh,(k+1)h]. \tag{3.5}$$

Our main results in this section is the constructive existence of weak solutions to (3.1):

Theorem 3.2. *Assume that* $\rho_0 \in L^1_+ \cap L^\infty(\Omega)$*. Then, up to a discrete subsequence (still denoted* $h \to 0$ *and not relabeled here),* ρ_h *and* $\tilde\rho_h$ *converge strongly in* $L^1((0,T)\times\Omega)$ *to a weak solution* ρ *of* (3.1)*.*

Note that any uniqueness for (3.1) would imply convergence of the whole (continuous) sequence $\rho_h, \tilde\rho_h \to \rho$ as $h \to 0$, but for the sake of simplicity we shall not address this issue here.

The main technical obstacle in the proof of Theorem 3.2 is to retrieve compactness in time. For the classical minimizing scheme of any energy $\mathcal{E}$ on any metric space (X,d), suitable time compactness is usually retrieved in the form of the *total-square distance estimate* $\frac{1}{2h}\sum\limits_{k\geq 0} d^2(x^k,x^{k+1}) \leqslant \mathcal{E}(x_0) - \inf\mathcal{E}$. This usually works because only one functional is involved, and $\mathcal{E}(x_0) - \inf\mathcal{E}$ is obtained as a telescopic sum of one-step energy dissipations $\mathcal{E}(x^{k+1}) - \mathcal{E}(x^k)$. Here each of our elementary step in (3.1) involves one of the $\mathtt{W},\mathtt{FR}$ metrics, and we will use the $\mathtt{WFR}$ distance to control both simultaneously: this strongly leverages the inf-convolution structure, the $\mathtt{WFR}$ distance being precisely built on a compromise between $\mathtt{W}$/transport and $\mathtt{FR}$/reaction. On the other hand we also have two different functionals $\mathcal{E}_1, \mathcal{E}_2$, and we will have to carefully estimate the dissipation of $\mathcal{E}_1$ during the $\mathtt{FR}$ reaction step (driven by $\mathcal{E}_2$) as well as the dissipation of $\mathcal{E}_2$ during the $\mathtt{W}$ transport/diffusion step (driven by $\mathcal{E}_1$).

We start by collecting one-step estimates, exploiting the optimality conditions for each elementary minimization procedure, and postpone the proof of Theorem 3.2 to the end of the section.

3.1 Optimality conditions and pointwise L^∞ estimates

The optimality conditions for the first Wasserstein step $\rho^k \to \rho^{k+1/2}$ in (3.4) are by now classical, and can be written for example

$$\frac{-\nabla\varphi_h^{k+1/2}}{h}\rho_h^{k+1/2} = \nabla P_1(\rho_h^{k+1/2}) + \rho_h^{k+1/2}\nabla V_1 \qquad \text{a.e.} \tag{3.6}$$

Here $\varphi_h^{k+1/2}$ is an optimal (backward) Kantorovich potential from $\rho_h^{k+1/2}$ to ρ_h^k.

Lemma 3.3. *For all $k \geqslant 0$,*

$$\|\rho_h^{k+1/2}\|_{L^1} = \|\rho_h^k\|_{L^1} \tag{3.7}$$

and for all constant C such that $V_1 \leqslant C$,

$$\rho_h^k(x) \leqslant (F_1')^{-1}(C - V_1(x)) \text{ a.e.} \qquad \Rightarrow \qquad \rho_h^{k+1/2}(x) \leqslant (F_1')^{-1}(C - V_1(x)) \text{ a.e.} \tag{3.8}$$

Proof. The Wasserstein step is mass conservative by construction, so the first part is obvious. The second part is a direct consequence of a generalization [36, lemma 2] of Otto's maximum principle [32]. □

Remark 3.4. *Note that if $\rho_h^k \leqslant M$, we may take $C = F_1'(M) + \|V_1\|_{L^\infty}$ in (3.8). Formally, this corresponds to taking $\overline{\rho}(x) := (F_1')^{-1}(C - V_1(x))$ as a stationary Barenblatt supersolution for $\partial_t\rho = \operatorname{div}(\rho\nabla(F_1'(\rho) + V_1))$ at the continuous level. In addition, if $V_1 \equiv 0$ we recover Otto's maximum principle [32] in the form $\|\rho^{k+1/2}\|_{L^\infty} \leqslant \|\rho^k\|_{L^\infty}$.*

For the second Fisher-Rao reaction step, the optimality condition has been obtained in [18, section 4.2] in the form

$$\left(\sqrt{\rho_h^{k+1}} - \sqrt{\rho_h^{k+1/2}}\right)\sqrt{\rho_h^{k+1}} = -\frac{h}{2}\rho_h^{k+1}\left(F_2'(\rho_h^{k+1}) + V_2\right) \qquad \text{a.e.} \tag{3.9}$$

As a consequence we have

Lemma 3.5. *There is $C \equiv C(V_2) > 0$ such that for $h \leqslant h_0(V_2)$ small enough we have*

$$\rho_h^{k+1}(x) \leqslant (1 + Ch)\rho_h^{k+1/2}(x) \qquad \text{a.e.}, \tag{3.10}$$

and for all $M > 0$ there is $c \equiv c(M, V_2)$ such that if $\|\rho_h^{k+1/2}\|_\infty \leqslant M$ then

$$(1 - ch)\rho_h^{k+1/2}(x) \leqslant \rho_h^{k+1}(x) \qquad \text{a.e.} \tag{3.11}$$

Note in particular that this immediately implies

$$\operatorname{supp}\,\rho_h^{k+1} = \operatorname{supp}\,\rho_h^{k+1/2}, \tag{3.12}$$

which was to be expected since the reaction part $\partial_t\rho = -\rho(F_2'(\rho) + V_2)$ of the PDE (3.1) preserves strict positivity.

Proof. We start with the upper bound: inside $\operatorname{supp}\rho_h^{k+1}$, (3.9) and $F_2' \geqslant 0$ give

$$\begin{aligned}\sqrt{\rho_h^{k+1}(x)} - \sqrt{\rho_h^{k+1/2}(x)} &= -h\sqrt{\rho_h^{k+1}(x)}(F_2'(\rho_h^{k+1}(x)) + V_2(x)) \\ &\leqslant -hV_2(x)\sqrt{\rho_h^{k+1}(x)} \leqslant h\|V_2\|_\infty\sqrt{\rho_h^{k+1}(x)}\end{aligned}$$

whence

$$\sqrt{\rho_h^{k+1}(x)} \leqslant \frac{1}{1 - h\|V_2\|_\infty}\sqrt{\rho_h^{k+1/2}(x)}.$$

Taking squares and using

$$\frac{1}{(1-h\|V_2\|_\infty)^2} = 1 + 2\|V_2\|_{L^\infty} h + \mathcal{O}(h^2) \leqslant 1 + 3\|V_2\|_{L^\infty} h$$

for small h gives the desired inequality.

For the lower bound (3.11), we first observe that since $F_2'' \geqslant 0$ and from (3.10) we have $F_2'(\rho_h^{k+1}) \leqslant F_2'((1+Ch)\rho_h^{k+1/2}) \leqslant F_2'(2M)$ if h is small enough. Then (3.9) gives inside $\operatorname{supp}\rho^{k+1}$

$$\begin{aligned}\sqrt{\rho_h^{k+1}(x)} - \sqrt{\rho_h^{k+1/2}(x))} &= -h\sqrt{\rho_h^{k+1}(x)}(F_2'(\rho_h^{k+1}(x)) + V_2(x)) \\ &\geqslant -h(F_2'(2M) + \|V_2\|_\infty)\sqrt{\rho_h^{k+1}(x)},\end{aligned}$$

hence

$$\rho_h^{k+1}(x) \geqslant \frac{1}{(1+h(F_2'(2M)+\|V_2\|_\infty))^2}\rho_h^{k+1/2}(x) \geqslant (1-ch)\rho_h^{k+1/2}(x)$$

for small h. □

Combining Lemma 3.3 and Lemma 3.5, we obtain at the continuous level

Proposition 3.6. *For all $T > 0$ there exist constants M_T, M_T' such that for all $t \in [0,T]$,*

$$\|\rho_h(t)\|_{L^1\cap L^\infty}, \|\tilde\rho_h(t)\|_{L^1\cap L^\infty} \leqslant M_T$$

and

$$\|\rho_h(t) - \tilde\rho_h(t)\|_{L^1} \leqslant hM_T'$$

uniformly in $h \geqslant 0$.

Note from the second estimate that strong $L^1((0,T)\times\Omega)$ convergence of ρ_h will immediately imply convergence of $\tilde\rho_h$ to the same limit.

Proof. By induction combining (3.8) and (3.10), we obtain, for all $t \in [0,T]$,

$$\|\rho_h(t)\|_{L^\infty}, \|\tilde\rho_h(t)\|_{L^\infty} \leqslant C_T,$$

where C_T is a constant depending on $\|V_1\|_{L^\infty}$, see [36, lemma 2]. The L^1 bound is even easier: since the Wasserstein step is mass preserving, we can integrate (3.10) in space to get

$$\|\rho_h^{k+1}\|_{L^1} \leqslant (1+Ch)\|\rho_h^{k+1/2}\|_{L^1} = (1+Ch)\|\rho_h^{k+1}\|_{L^1}.$$

For $t \leqslant T \Leftrightarrow k \leqslant \lfloor T/h \rfloor$ the L^1 bounds immediately follow by induction, with $(1+Ch)^{\lfloor T/h\rfloor} \lesssim e^{CT}$. and we conclude again by induction.

In order to compare now ρ_h and $\tilde\rho_h$, we take advantage of the above upper bound to write $\rho_h^{k+1/2} \leqslant M_T$ as long as $kh \leqslant T$. Taking $c = c(M_T)$ in (3.11) and combining with (3.10), we have

$$-ch\rho_h^{k+1/2} \leqslant \rho_h^{k+1/2} - \rho_h^{k+1} \leqslant Ch\rho_h^{k+1/2} \qquad \text{a.e.}$$

Integrating in Ω we conclude that

$$\|\rho_h(t) - \tilde\rho_h(t)\|_1 = \|\rho_h^{k+1} - \rho_h^{k+1/2}\|_1 \leqslant h\max\{c,C\}\|\rho_h^{k+1/2}\|_1 \leqslant h\max\{c,C\}M_T = hM_T'$$

and the proof is complete. □

3.2 Energy dissipation

Our goal is here to estimate the crossed dissipation along each elementary $\mathtt{W}, \mathtt{FR}$ step.

Testing $\rho = \rho_h^k$ in the first Wasserstein step in (3.4), we get as usual

$$\frac{1}{2h}\mathtt{W}^2(\rho_h^{k+1/2}, \rho_h^k) \leqslant \mathcal{F}_1(\rho_h^k) - \mathcal{F}_1(\rho_h^{k+1/2}) + \int_\Omega V_1(\rho_h^k - \rho_h^{k+1/2}). \tag{3.13}$$

Since V_1 is Globally Lipschitz we can first use standard methods from [15, 23] to control $\int_\Omega V_1(\rho_h^k - \rho_h^{k+1/2})$ in terms of $\mathtt{W}^2(\rho_h^{k+1/2}, \rho_h^k)$, and suitably reabsorb in the left-hand side to obtain

$$\frac{1}{4h}\mathtt{W}^2(\rho_h^{k+1/2}, \rho_h^k) \leqslant \mathcal{F}_1(\rho_h^k) - \mathcal{F}_1(\rho_h^{k+1/2}) + C_T h. \tag{3.14}$$

The dissipation of $\mathcal{F}_1$ along the Fisher-Rao step is controlled as

Proposition 3.7. *For all $T > 0$ there exists a constant $C_T > 0$ such that, for all $k \geqslant 0$ and $k \leq \lfloor T/h \rfloor$,*

$$\mathcal{F}_1(\rho_h^{k+1}) \leqslant \mathcal{F}_1(\rho_h^{k+1/2}) + C_T h. \tag{3.15}$$

Proof. We first treat the case of $F_1(z) = \frac{1}{m_1 - 1} z^{m_1}$ with $m_1 > 1$. Since F_1 is increasing, we use (3.10) to obtain

$$\begin{aligned} \mathcal{F}_1(\rho_h^{k+1}) - \mathcal{F}_1(\rho_h^{k+1/2}) &\leqslant \frac{((1+Ch)^{m_1} - 1)}{m_1 - 1} \int_\Omega (\rho_h^{k+1/2})^{m_1} \\ &\leqslant Ch \|\rho^{k+1/2}\|_{L^\infty}^{m_1 - 1} \|\rho^{k+1/2}\|_{L^1}, \end{aligned}$$

and we conclude from Proposition 3.6.

In the second case $F_1(z) = z\log(z) - z$, we have

$$\mathcal{F}_1(\rho_h^{k+1}) = \int_{\{\rho_h^{k+1} \leqslant e^{-1}\}} \rho_h^{k+1} \log(\rho_h^{n+1}) + \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} \rho_h^{k+1} \log(\rho_h^{k+1}) - \int_\Omega \rho_h^{k+1}.$$

Note from Proposition 3.6 that the z contribution in $F_1(z) = z\log z - z$ is immediately controlled by $|\int \rho_h^{k+1} - \int \rho_h^{k+1/2}| \leqslant \|\rho_h^{k+1} - \rho_h^{k+1/2}\|_{L^1} \leqslant hM_T'$, so we only have to estimate the $z\log z$ contribution. Since $z \mapsto z\log z$ is increasing on $\{z \geqslant e^{-1}\}$ and using (3.10), the second term in the right hand side becomes

$$\begin{aligned} \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} \rho_h^{k+1} \log(\rho_h^{k+1}) &\leqslant \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} (1+Ch)\rho_h^{k+1/2} \log((1+Ch)\rho_h^{k+1/2}) \\ &\leqslant \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} \rho_h^{k+1/2} \log(\rho_h^{k+1/2}) + Ch \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} \rho_h^{k+1/2} \log(\rho_h^{k+1/2}) \\ &\qquad + (1+Ch) \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} \rho_h^{k+1/2} \log(1+Ch) \\ &\leqslant \int_{\{\rho_h^{k+1} \geqslant e^{-1}\}} \rho_h^{k+1/2} \log(\rho_h^{k+1/2}) + C_T h, \end{aligned}$$

where we used $\|\rho_h^{k+1/2}\|_{L^1} \leqslant M_T$ from Proposition 3.6 as well as $\log(1+Ch) \leqslant Ch$ in the last inequality. Using the same method with the bound from below (3.11) on $\{\rho_h^{k+1} \leqslant e^{-1}\}$ (where $z \mapsto z\log z$ is now decreasing), we obtain similarly

$$\int_{\{\rho_h^{k+1} \leqslant e^{-1}\}} \rho_h^{k+1} \log(\rho_h^{k+1}) \leqslant \int_{\{\rho_h^{k+1} \leqslant e^{-1}\}} \rho_h^{k+1/2} \log(\rho_h^{k+1/2}) + C_T h.$$

Combining both inequalities gives

$$\int_\Omega \rho_h^{k+1} \log(\rho_h^{k+1}) \leqslant \int_\Omega \rho_h^{k+1/2} \log(\rho_h^{k+1/2}) + C_T h$$

and the proof is complete.

□

Summing (3.14) and (3.15) over k we obtain

$$\frac{1}{2h}\sum_{k=0}^{N-1}\mathtt{W}^2(\rho_h^{k+1/2},\rho_h^k)\leqslant\mathcal{F}_1(\rho_0)-\mathcal{F}_1(\rho_h^N)+C_T, \tag{3.16}$$

where $N=\lfloor\frac{T}{h}\rfloor$.

In the above estimate we just controlled the dissipation of $\mathcal{F}_1$ along the $\mathtt{FR}$/reaction steps, and the goal is now to similarly estimate the dissipation of $\mathcal{F}_2$ along the Wasserstein step. Testing $\rho=\rho_h^{k+1/2}$ in the second Fisher-Rao step in (3.4), we obtain

$$\frac{1}{2h}\mathtt{FR}_2(\rho_h^{k+1},\rho_h^{k+1/2})\leqslant\mathcal{F}_2(\rho_h^{k+1/2})-\mathcal{F}_2(\rho_h^{k+1})+\int_\Omega V_2(\rho_h^{k+1/2}-\rho_h^{k+1}). \tag{3.17}$$

Since we assumed $V_2\in L^\infty(\Omega)$ and because $\rho_h(t)=\rho_h^{k+1}$ remains close to $\tilde\rho_h(t)=\rho_h^{k+1/2}$ in L^1 uniformly in t,h by Proposition 3.6, we immediately control the potential part as

$$\int_\Omega V_2(\rho_h^{k+1/2}-\rho_h^{k+1})\leqslant\|V_2\|_\infty C_T h. \tag{3.18}$$

For the internal energy we argue exactly as in the proof Proposition 3.7 (for the Porous Media part, since we chose here $F_2(z)=\frac{1}{m_2-1}z^{m_2}$), and obtain

$$\mathcal{F}_2(\rho_h^{k+1/2})-\mathcal{F}_2(\rho_h^{k+1})\leqslant C_T h. \tag{3.19}$$

Combining (3.17), (3.18) and (3.19), we immediately deduce that

$$\frac{1}{2h}\sum_{k=0}^{N-1}\mathtt{FR}^2(\rho_h^{k+1/2},\rho_h^{k+1})\leqslant C_T, \tag{3.20}$$

where $N=\lfloor\frac{T}{h}\rfloor$ as before.

Finally, we recover an approximate compactness in time in the form

Proposition 3.8. *There exists a constant $C_T>0$ such that for all h small enough and $k\leqslant N=\lfloor T/h\rfloor$,*

$$\frac{1}{h}\sum_{k=0}^{N-1}\mathtt{WFR}^2(\rho_h^k,\rho_h^{k+1})\leqslant 4\mathcal{F}_1(\rho_0)+C_T. \tag{3.21}$$

Proof. Adding (3.16) and (3.20) gives

$$\frac{1}{h}\sum_{k=0}^{N-1}\mathtt{W}^2(\rho_h^k,\rho_h^{k+1/2})+\mathtt{FR}^2(\rho_h^{k+1/2},\rho_h^{k+1})\leqslant 2\left(\mathcal{F}_1(\rho_0)-\mathcal{F}_1(\rho_h^N)+C_T\right)+2C_T\leqslant 2\mathcal{F}_1(\rho_0)+C_T,$$

since in any case $F_1(z)=\frac{1}{m_1-1}z^{m_1}\geqslant 0$ and $F_1(z)=z\log z-z\geqslant -1$ is bounded from below on the bounded domain Ω, hence $\mathcal{F}_1(\rho_h^N)\geqslant -C_\Omega$ uniformly. It then follows from Proposition 2.4 that $\mathtt{W}^2(\rho_h^k,\rho_h^{k+1/2})+\mathtt{FR}^2(\rho_h^{k+1/2},\rho_h^{k+1})\geqslant\frac{1}{2}\mathtt{WFR}^2\rho_h^k,\rho_h^{k+1}$ in the left-hand side, and the result immediately follows.

□

3.3 Estimates and convergences

From the total-square distance estimate (3.21) we recover as usual the approximate $\frac{1}{2}$-Hölder estimate

$$\mathtt{WFR}(\rho_h(t),\rho_h(s))+\mathtt{WFR}(\tilde\rho_h(t),\tilde\rho_h(s))\leqslant C_T|t-s+h|^{1/2} \tag{3.22}$$

for all fixed $T > 0$ and $t, s \in [0, T]$. From (3.20) and Proposition 2.4 we have moreover

$$\mathtt{WFR}(\rho_h(t), \tilde{\rho}_h(t)) \leqslant \mathtt{FR}(\rho_h(t), \tilde{\rho}_h(t)) \leqslant C\sqrt{h}. \tag{3.23}$$

Using a refined version of Ascoli-Arzelà theorem, [4, prop. 3.3.1] and arguing exactly as in [18, prop. 4.1], we see that for all $T > 0$ and up to extraction of a discrete subsequence, ρ_h and $\tilde{\rho}_h$ converge uniformly to the same $\mathtt{WFR}$-continuous curve $\rho \in \mathcal{C}^{1/2}([0, T], \mathcal{M}^+_{\mathtt{WFR}})$ as

$$\sup_{t\in[0,T]} \left(\mathtt{WFR}(\rho_h(t), \rho(t)) + \mathtt{WFR}(\tilde{\rho}_h(t), \rho(t))\right) \to 0.$$

In order to pass to the limit in the nonlinear terms, we first strengthen this $\mathtt{WFR}$-convergence into a more tractable L^1 convergence. The first step is to retrieve compactness in space:

Proposition 3.9. *For all $T > 0$, ρ_h and $\tilde{\rho}_h$ satisfies*

$$\|P_1(\tilde{\rho}_h)\|_{L^2([0,T];H^1(\Omega))} \leqslant C_T. \tag{3.24}$$

Proof. From (3.6) and the $L^1 \cap L^\infty$ bounds from Proposition 3.6 we see that

$$\begin{aligned}
\int_\Omega |\nabla P_1(\rho_h^{k+1/2})|^2 &\leqslant \frac{1}{2h^2}\int_\Omega |\nabla\varphi_h^{k+1/2}|^2(\rho_h^{k+1/2})^2 + \frac{1}{2}\int_\Omega |\nabla V_1|^2(\rho_h^{k+1/2})^2 \\
&\leqslant \frac{C_T}{2h^2}\int_\Omega |\nabla\varphi_h^{k+1/2}|^2\rho_h^{k+1/2} + \frac{1}{2}\|\nabla V_1\|^2_\infty \int_\Omega (\rho_h^{k+1/2})^2 \\
&\leqslant C_T\left(\frac{\mathtt{W}^2(\rho_h^{k+1/2}, \rho_h^k)}{h^2} + 1\right)
\end{aligned}$$

since $\varphi_h^{k+1/2}$ is the optimal (backward) Kantorovich potential from $\rho_h^{k+1/2}$ to ρ_h^k. Multiplying by $h > 0$, summing over k, and exploiting (3.16) gives

$$\|P_1(\tilde{\rho}_h)\|^2_{L^2([0,T];H^1(\Omega))} \leqslant \sum_{k=0}^{N-1} h\|P_1(\rho_h^{k+1/2})\|^2_{H^1} \leqslant C_T(\mathcal{F}_1(\rho_0) - \mathcal{F}_1(\rho_h^N) + 1) \leqslant C_T,$$

where we used as before $\mathcal{F}_1(\rho_h^N) \geqslant -C_\Omega$ in the last inequality. □

We are now in position of proving our main result:

Proof of Theorem 3.2. Exploiting (3.21) and (3.24), we can apply the extension of the Aubin-Lions lemma established by Rossi and Savaré in [39] to obtain that $\tilde{\rho}_h$ converges to ρ strongly in $L^1(Q_T)$ (see [23]). By diagonal extraction if needed, we can assume that the convergence holds in $L^1(Q_T)$ for all fixed $T > 0$. Then by Proposition 3.6 we have

$$\|\rho_h - \rho\|_{L^1(Q_T)} \leqslant \|\rho_h - \tilde{\rho}_h\|_{L^1(Q_T} + \|\tilde{\rho}_h - \rho\|_{L^1(Q_T)} \leqslant C_T h + \|\tilde{\rho}_h - \rho\|_{L^1(Q_T)} \to 0$$

hence $\rho_h \to \rho$ as well.

Moreover, since $P_1(\tilde{\rho}_h)$ is bounded in $L^2((0, T), H^1(\Omega))$ we can assume that $\nabla P_1(\tilde{\rho}_h) \rightharpoonup \nabla P_1(\rho)$ in $L^2((0, T), H^1(\Omega))$ for all $T > 0$. Exploiting the Euler-Lagrange equations (3.6)(3.9) and arguing exactly as in [18, Theorem 4], it is easy to pass to the limit to conclude that

$$\int_\Omega \rho(t_2)\varphi - \rho(t_1)\varphi = -\int_{t_1}^{t_2}\int_\Omega \Big\{\nabla P(\rho)\cdot\nabla\varphi + \rho\nabla V_1\cdot\nabla\varphi - \rho(F_2'(\rho) + V_2)\varphi\Big\}$$

for all $0 < t_1 < t_2$ and $\varphi \in \mathcal{C}^1_b(\Omega)$. Since $\rho \in \mathcal{C}([0, T]; \mathcal{M}^+_{\mathtt{WFR}})$ takes the initial datum $\rho(0) = \rho_0$ and $\mathtt{WFR}$ metrizes the narrow convergence of measures, this is well-known to be equivalent to our weak formulation in Definition 3.1, and the proof is complete. □

Remark 3.10. *In the above proofs one can check that Theorem 3.2 extends in fact to all $\mathcal{C}^1$ nonlinearities F_2 such that $F_2' \geqslant C$ for some $C \in \mathbb{R}$. Likewise, we stated and proved our main result in bounded domains for convenience: all the above arguments immediately extend to $\Omega = \mathbb{R}^d$ at least for $F_1(z) = \frac{1}{m_1 - 1} z^{m_1} \geqslant 0$. The only place where we actually used the boundedness of Ω was in the proof of Proposition 3.8, when we bounded from below $\mathcal{F}_1(\rho_h^N) \geqslant -C_\Omega$ in order to retrieve the total-square distance estimate. When $\Omega = \mathbb{R}^d$ and $F_1(z) = z \log z - z$ a lower bound $\mathcal{F}_1(\rho_h^N) \geqslant -C_T$ still holds, but the proof requires a tedious control of the second moments $\mathfrak{m}_2(\rho) = \int_{\mathbb{R}^d} |x|^2 \rho$ hence we did not address this technical issue for the sake of brevity.*

4 Application to systems

In this section we shall try to illustrate that the previous scheme is very tractable and allows to solve systems of the form

$$\begin{cases} \partial_t \rho_1 = \operatorname{div}(\rho_1 \nabla(F_1'(\rho_1) + V_1[\rho_1, \rho_2])) - \rho_1(G_1'(\rho_1) + U_1[\rho_1, \rho_2]), \\ \partial_t \rho_2 = \operatorname{div}(\rho_2 \nabla(F_2'(\rho_2) + V_2[\rho_1, \rho_2])) - \rho_2(G_2'(\rho_2) + U_2[\rho_1, \rho_2]), \\ \rho_{1|t=0} = \rho_{1,0},\ \rho_{2|t=0} = \rho_{2,0}. \end{cases} \tag{4.1}$$

For simplicity we assume again that Ω is a smooth, bounded subset of $\mathbb{R}^d$. Then the system (4.1) is endowed with Neumann boundary conditions,

$$\rho_1 \nabla(F_1'(\rho_1) + V_1[\rho_1, \rho_2]) \cdot \nu = 0 \text{ and } \rho_2 \nabla(F_2'(\rho_2) + V_2[\rho_1, \rho_2]) \cdot \nu = 0 \qquad \text{on } \mathbb{R}^+ \times \partial\Omega,$$

where ν is the outward unit normal to $\partial\Omega$. In system of the form (4.1), we allow interactions between densities in the potential terms $V_i[\rho_1, \rho_2]$ and $U_i[\rho_1, \rho_2]$. In the mass-conservative case (without reaction terms), this system has already been studied in [15, 23, 8], using a semi-implicit JKO scheme introduced by Di Francesco and Fagioli, [15]. This section combines the splitting scheme introduced in the previous section and semi-implicit schemes both for the Wasserstein JKO step and for the Fisher-Rao JKO step.

For the ease of exposition we keep the same assumptions for F_i and G_i as in the previous section, i.e the diffusion terms F_i satisfy (3.2) and the reaction terms G_i satisfy (3.3). Moreover, since the potentials depend now on the densities ρ_1 and ρ_2, we need stronger hypotheses: we assume that $V_i : L^1(\Omega; \mathbb{R}^+)^2 \to \mathcal{C}^1(\Omega)$ are continuous and verify, uniformly in $\rho_1, \rho_2 \in L^1(\Omega; \mathbb{R}^+)$,

$$\begin{gathered} \|V_i[\rho_1, \rho_2]\|_{W^{1,\infty}(\Omega)} \leqslant K(1 + \|\rho_1\|_{L^1(\Omega)} + \|\rho_2\|_{L^1(\Omega)}), \\ \|\nabla(V_i[\rho_1, \rho_2]) - \nabla(V_i[\mu_1, \mu_2])\|_{L^\infty(\Omega)} \leqslant K(\|\rho_1 - \mu_1\|_{L^1(\Omega)} + \|\rho_2 - \mu_2\|_{L^1(\Omega)}). \end{gathered} \tag{4.2}$$

The interacting potentials we have in mind are of the form $V_i[\rho_1, \rho_2] = K_{i,1} * \rho_1 + K_{i,2} * \rho_2$, where $K_{i,1}, K_{i,2} \in W^{1,\infty}(\Omega)$ and then V_i satisfies (4.2). For the reaction, we assume that the potentials U_i are continuous from $L^1(\Omega)_+^2$ to L^1 with moreover

$$U_i[\rho_1, \rho_2] \geqslant -K, \qquad \forall\, \rho_1, \rho_2 \in L^1(\Omega; \mathbb{R}^+) \tag{4.3}$$

for some $K \in \mathbb{R}$, and

$$\|U_i[\rho_1, \rho_2]\|_{L^\infty(\Omega)} \leqslant K_M, \qquad \forall \|\rho_1\|_{L^1(\Omega)}, \|\rho_2\|_{L^1(\Omega)} \leqslant M \tag{4.4}$$

for some nondecreasimg function $K_M \geqslant 0$ of M. The examples we have in mind are of the form

$$U_1[\rho_1, \rho_2] = C_1 \frac{\rho_2}{1 + \rho_1}, \quad U_2[\rho_1, \rho_2] = -C_2 \frac{\rho_1}{1 + \rho_1}$$

for some constants $C_i \geq 0$, or nonlocal reactions

$$U_i[\rho_1, \rho_2](x) = \int_\Omega K_{i,1}(x, y) \rho_1(y)\, dy + \int_\Omega K_{i,2}(x, y) \rho_2(y)\, dy$$

for some nonnegative kernels $K_{i,j} \in L^1 \cap L^\infty$. Such reaction models appear for example in biological adaptive dynamics [33].

Definition 4.1. *We say that* $(\rho_1, \rho_2) : \mathbb{R}^+ \to L^1_+ \cap L^\infty_+(\Omega)$ *is a weak solution of* (4.1) *if, for* $i \in \{1, 2\}$ *and all* $T < +\infty$*, the pressure* $P_i(\rho_i) := \rho_i F_i'(\rho_i) - F_i(\rho_i)$ *satisfies* $\nabla P_i(\rho_i) \in L^2([0, T] \times \Omega)$*, and*

$$\int_0^{+\infty} \left(\int_\Omega (\rho \partial_t \phi_i - \rho_i \nabla V_i[\rho_1, \rho_2] \cdot \nabla \phi_i - \nabla P_i(\rho_i) \cdot \nabla \phi_i - \rho_i (G_i'(\rho_i) + U_i[\rho_1, \rho_2]) \phi_i) \, dx \right) dt = - \int_\Omega \phi_i(0, x) \rho_{i,0}(x) \, dx, \quad (4.5)$$

for all $\phi_i \in \mathcal{C}^\infty_c([0, +\infty) \times \mathbb{R}^d)$.

Then, the following result holds,

Theorem 4.2. *Assume that* $\rho_{1,0}, \rho_{2,0} \in L^1 \cap L^\infty_+(\Omega)$ *and that* V_i, U_i *satisfy* (4.2)(4.3)(4.4)*. Then* (4.1) *admits at least one weak solution.*

Note that this result can be easily adapted to systems with an arbitrary number of species $N \geqslant 2$, coupled by nonlocal terms $V_i[\rho_1, \dots, \rho_N]$ and $U_i[\rho_1, \dots, \rho_N]$.

Remark 4.3. *A refined analysis shows that our approach would allow to handle systems of the form*

$$\begin{cases} \partial_t \rho_1 - \operatorname{div}(\rho_1 \nabla (F_1'(\rho_1) + V_1)) = -\rho_1 h_1(\rho_1, \rho_2), \\ \partial_t \rho_2 - \operatorname{div}(\rho_2 \nabla (F_2'(\rho_2) + V_2)) = +\rho_2 h_2(\rho_1), \end{cases}$$

where h_1 *is a nonnegative continuous function and* h_2 *is a continuous functions.*

Indeed since $h_1 \geq 0$ *the reaction term is the first equation is nonpositive, hence* $\|\rho_1(t)\|_{L^\infty(\Omega)} \leqslant C_T$*. Then it follows that* $-h_2(\rho_1)$ *satisfies assumptions* (4.3) *and* (4.4)*. A classical example is* $h_2(\rho_1) = \rho_1^\alpha$ *and* $h_1(\rho_1, \rho_2) = \rho_1^{\alpha-1} \rho_2$*, where* $\alpha \geqslant 1$*, see for example [38] for more discussions.*

As already mentioned, the proof of theorem 4.2 is based on a semi-implicit splitting scheme. More precisely, we construct four sequences $\rho_{1,h}^{k+1/2}, \rho_{1,h}^{k+1}, \rho_{2,h}^{k+1/2}, \rho_{2,h}^{k+1}$ defined recursively as

$$\begin{cases} \rho_{i,h}^{k+1/2} \in \underset{\rho \in \mathcal{M}^+, |\rho| = |\rho_{i,h}^k|}{\operatorname{argmin}} \left\{ \frac{1}{2h} \mathtt{W}^2(\rho, \rho_{i,h}^k) + \mathcal{F}_i(\rho) + \mathcal{V}_i(\rho | \rho_{1,h}^k, \rho_{2,h}^k) \right\} \\ \\ \rho_{i,h}^{k+1} \in \underset{\rho \in \mathcal{M}^+}{\operatorname{argmin}} \left\{ \frac{1}{2h} \mathtt{FR}^2(\rho, \rho_{i,h}^{k+1/2}) + \mathcal{G}_i(\rho) + \mathcal{U}_i(\rho | \rho_{1,h}^k, \rho_{2,h}^k) \right\} \end{cases}, \quad (4.6)$$

where the fully implicit terms

$$\mathcal{F}_i(\rho) := \begin{cases} \int_\Omega F_i(\rho) & \text{if } \rho \ll \mathcal{L}_{|\Omega} \\ +\infty & \text{otherwise} \end{cases} \quad \text{and} \quad \mathcal{G}_i(\rho) := \begin{cases} \int_\Omega G_i(\rho) & \text{if } \rho \ll \mathcal{L}_{|\Omega} \\ +\infty & \text{otherwise} \end{cases},$$

and the semi-implicit terms

$$\mathcal{V}_i(\rho | \mu_1, \mu_2) := \int_\Omega V_i[\mu_1, \mu_2] \rho \quad \text{and} \quad \mathcal{U}_i(\rho | \mu_1, \mu_2) := \int_\Omega U_i[\mu_1, \mu_2] \rho.$$

In the previous section, the proof of theorem 3.2 for scalar equations strongly leveraged the uniform $L^\infty(\Omega)$-bounds on the discrete solutions. Here an additional difficulty arises due to the nonlocal terms $\nabla V_i[\rho_1, \rho_2]$ and $U_i[\rho_1, \rho_2]$, which are a priori not uniformly bounded in $L^\infty(\Omega)$. Using assumption (4.3) we will first obtain a uniform $L^1(\Omega)$-bound on ρ_1, ρ_2, and then extend proposition 3.6 to the system (4.1). This in turn will give a uniform $W^{1,\infty}$ control on $V_i[\rho_1, \rho_2]$ and L^∞ control on $U_i[\rho_1, \rho_2]$ through our assumptions (4.2)-(4.3)-(4.4), which will finally allow to argue as in the previous section and give L^∞ control on ρ_1, ρ_2.

Numerical simulations for a diffusive prey-predator system are presented at the end of this section.

4.1 Properties of discrete solutions

Arguing as in the case of one equation, the optimality conditions for the Wasserstein step and for the Fisher-Rao step first give

Lemma 4.4. *For all $k \geqslant 0$ and $i \in \{1,2\}$, we have*

$$\|\rho_{i,h}^{k+1/2}\|_{L^1} = \|\rho_{i,h}^{k}\|_{L^1}. \tag{4.7}$$

Moreover, there exists $C_i \equiv C(U_i) > 0$ (uniform in k) such that

$$\rho_{i,h}^{k+1}(x) \leqslant (1 + C_i h)\rho_{i,h}^{k+1/2}(x) \qquad a.e. \tag{4.8}$$

Proof. The first part is simply the mass conservation in the Wasserstein step, and the second part follows the lines of the proof of (3.10) in Lemma 3.5 using assumption (4.3). □

As a direct consequence we have uniform control on the L^1-norms:

Lemma 4.5. *For all $T > 0$ there exist constants $C_T, C'_T > 0$ such that, for all $t \in [0,T]$,*

$$\|\rho_{i,h}(t)\|_{L^1}, \|\tilde{\rho}_{i,h}(t)\|_{L^1} \leqslant C_T$$

and

$$\|V_i[\rho_{1,h}(t), \rho_{2,h}(t)]\|_{W^{1,\infty}}, \|V_i[\tilde{\rho}_{1,h}(t), \tilde{\rho}_{2,h}(t)]\|_{W^{1,\infty}} \leqslant C'_T. \tag{4.9}$$

Proof. Integrating (4.8) and iterating with (4.7), we obtain for all $t \leqslant T$ and $k \leqslant \lfloor T/h \rfloor$

$$\|\rho_{i,h}^{k+1}\|_{L^1} \leqslant (1 + C_i h)\|\rho_{i,h}^{k}\|_{L^1} \leqslant (1 + C_i h)^k \|\rho_{i,0}\|_{L^1} \leqslant e^{C_i T}\|\rho_{i,0}\|_{L^1}.$$

Then (4.9) follows from our assumption (4.2) on the interactions. □

Combining (4.8) and (4.9), we deduce

Proposition 4.6. *For all $T > 0$, there exists M_T such that for all $t \in [0,T]$,*

$$\|\rho_{i,h}(t)\|_{L^\infty}, \|\tilde{\rho}_{i,h}(t)\|_{L^\infty} \leqslant M_T.$$

Then, there exists $c_i \equiv c(M_T, U_i) \geq 0$, such that, for all $k \leqslant \lfloor T/h \rfloor$ and $h \leqslant h_0(U_1, U_2)$,

$$(1 - c_i h)\rho_{i,h}^{k+1/2} \leqslant \rho_{i,h}^{k+1}.$$

In particular, there exist $M'_T > 0$ such that for all $t \in [0,T]$,

$$\|\rho_{i,h}(t) - \tilde{\rho}_{i,h}(t)\|_{L^1} \leqslant h M'_T.$$

Proof. The first L^∞ estimate can be found in [36, Lemma 2], and the rest of our statement can be proved exactly as in Lemma 3.5 and Proposition 3.6. □

4.2 Estimates and convergences

Since we proved that $V_1[\rho_{1,h}, \rho_{2,h}]$ and $V_2[\rho_{1,h}, \rho_{2,h}]$ are bounded in $L^\infty([0,T], W^{1,\infty}(\Omega))$, we can argue exactly as in the previous section for the Wasserstein step and obtain

$$\frac{1}{4h}\mathtt{W}^2(\rho_{i,h}^{k+1/2}, \rho_{i,h}^{k}) \leqslant \mathcal{F}_i(\rho_{i,h}^{k}) - \mathcal{F}_i(\rho_{i,h}^{k+1/2}) + C_T h, \tag{4.10}$$

see (3.13)-(3.14) for details. Since $\tilde{\rho}_{1,h}$ and $\tilde{\rho}_{2,h}$ are uniformly bounded in $L^1(\Omega)$ (Lemma 4.5), our assumption (4.4) ensures that $U_1[\rho_{1,h}^{k+1/2}, \rho_{2,h}^{k+1/2}]$ and $U_2[\rho_{1,h}^{k+1/2}, \rho_{2,h}^{k+1/2}]$ are uniformly bounded in $L^\infty(\Omega)$. Proposition 4.6 then allows to argue exactly as in (3.17)-(3.18)-(3.19) for the Fisher-Rao step, and we get

$$\frac{1}{2h}\mathtt{FR}^2(\rho_{h}^{k+1}, \rho_{h}^{k+1/2}) \leqslant \mathcal{G}_i(\rho_{i,h}^{k+1/2}) - \mathcal{G}_i(\rho_{i,h}^{k+1}) + C_T h. \tag{4.11}$$

The dissipation of $\mathcal{F}_i$ along the Fisher-Rao step is obtained in the same way as Proposition 3.7 and we omit the details:

Proposition 4.7. *For all $T > 0$ and $i \in \{1, 2\}$, there exist constants $C_T, C'_T > 0$ such that, for all $k \geqslant 0$ with $hk \leqslant T$,*

$$\mathcal{F}_i(\rho_{i,h}^{k+1}) \leqslant \mathcal{F}_i(\rho_{i,h}^{k+1/2}) + C_T h,$$
$$\mathcal{G}_i(\rho_{i,h}^{k+1/2}) \leqslant \mathcal{G}_i(\rho_{i,h}^{k+1}) + C'_T h.$$

From (4.10) and (4.11) this immediately gives a telescopic sum

$$\frac{1}{2h}\left(\mathtt{W}^2(\rho_{i,h}^k, \rho_{i,h}^{k+1/2}) + \mathtt{FR}^2(\rho_h^{k+1/2}, \rho_h^k)\right) \leqslant 2[\mathcal{F}_i(\rho_{i,h}^k) - \mathcal{F}_i(\rho_{i,h}^{k+1})] + C_T h$$

which in turn yields an approximate $\frac{1}{2}$-Hölder estimate (with respect to the WFR distance) as in Proposition 3.8. The rest of the proof of Theorem 4.2 is then identical to section 3 and we omit the details.

4.3 Numerical application: prey-predator systems

Our constructive scheme can be implemented numerically, by simply discretizing (4.6) in space. We use the augmented Lagrangian method ALG-JKO from [6] to solve the Wasserstein step, and the Fisher-Rao step is just a convex pointwise minimization problem. Indeed, it is known [18, 27] that $\mathtt{FR}^2(\rho, \mu) = 4\|\sqrt{\rho} - \sqrt{\mu}\|_{L^2}^2$, hence the Fisher-Rao step in (4.6) is a mere convex pointwise minimization problem of the form: for all $x \in \Omega$ (and omitting all indexes $\rho_{i,h}$),

$$\rho^{k+1}(x) = \operatorname*{argmin}_{\rho \geq 0}\left\{4\left|\sqrt{\rho} - \sqrt{\rho^{k+1/2}(x)}\right|^2 + 2hF(\rho)\right\}.$$

This is easily solved using any simple Newton procedure.

Figure (1) shows the numerical solution of the following diffusive prey-predator system

$$\begin{cases} \partial_t \rho_1 - \Delta\rho_1 - \operatorname{div}(\rho_1 \nabla V_1[\rho_1, \rho_2]) = A\rho_1(1 - \rho_1) - B\frac{\rho_1\rho_2}{1+\rho_1}, \\ \partial_t \rho_2 - \Delta\rho_2 - \operatorname{div}(\rho_2 \nabla V_2[\rho_1, \rho_2]) = \frac{B\rho_1\rho_2}{1+\rho_1} - C\rho_2, \end{cases}.$$

Here the ρ_1 species are preys and ρ_2 are predators, see for example [30], the parameters $A = 10, C = 5, B = 70$, and the interactions are chosen as

$$V_1[\rho_1, \rho_2] = |x|^2 * \rho_1 - |x|^2 * \rho_2, \quad V_2[\rho_1, \rho_2] = |x|^2 * \rho_1 + |x|^2 * \rho_2.$$

In (4.1) this corresponds to

$$G_1(\rho_1) = A\frac{\rho_1^2}{2}, \quad G_2(\rho_2) = 0, \quad U_1[\rho_1, \rho_2] = \frac{B\rho_2}{1+\rho_1} - A, \quad U_2[\rho_1, \rho_2] = -\frac{B\rho_1}{1+\rho_1} + C.$$

Of course, U_1 and U_2 satisfy assumptions (4.3) and (4.4), and then Theorem 4.2 gives a solution of the prey-predator system. As before, we shall disregard the uniqueness issue for the sake of simplicity. Figure (2) depicts the mass evolution of the prey and predator species: we observe the usual oscillations in time with phase opposition, a characteristic behaviour for Lotka-Volterra types of systems.

5 Application to a tumor growth model with very degenerate enery

In this section we take interest in the equation

$$\begin{cases} \partial_t \rho = \operatorname{div}(\rho \nabla p) + \rho(1 - p), \\ p \geqslant 0 \quad \text{and} \quad p(1 - \rho) = 0 \\ 0 \leqslant \rho \leqslant 1, \\ \rho_{|t=0} = \rho_0. \end{cases} \tag{5.1}$$

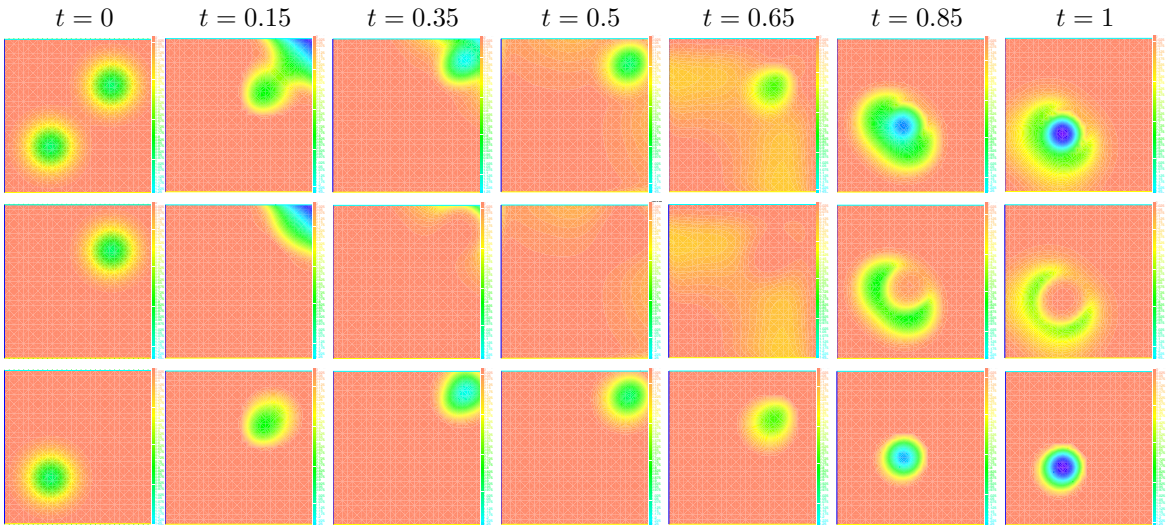


Figure 1: *Evolution of two species with prey-predator interactions. First row: display of* $\rho_1 + \rho_2$. *Second row: display of the prey* ρ_1. *Third row: display of the predator* ρ_1.

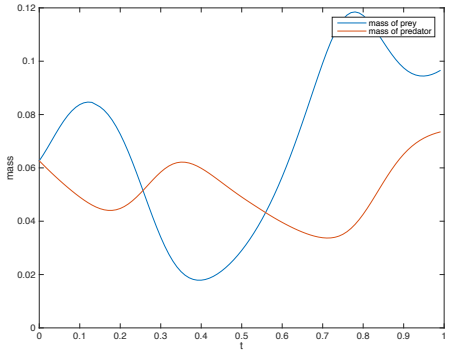


Figure 2: *Mass evolution for two-species prey-predator interactions.*

This equation is motivated by tumor growth models [34, 35] and exhibits a Hele-Shaw patch dynamics: if $\rho_0 = \chi_{\Omega_0}$ then the solution remains an indicator $\rho(t) = \chi_{\Omega(t)}$ and the boundary moves with normal velocity $V = -\nabla p|_{\partial\Omega(t)}$, see [2] for a rigorous analysis in the framework of viscosity solutions.

At least formally, we remark that (5.1) is the Wasserstein-Fisher-Rao gradient flow of the singular functional

$$\mathcal{F}(\rho) := \mathcal{F}_\infty(\rho) - \int_\Omega \rho,$$

where

$$\mathcal{F}_\infty(\rho) := \begin{cases} 0 & \text{if } \rho \leqslant 1 \text{ a.e,} \\ +\infty & \text{otherwise.} \end{cases}$$

Indeed, the compatibility conditions $p \geqslant 0$ and $p(1-\rho) = 0$ in (5.1) really mean that the pressure p belongs to the subdifferential $\partial\mathcal{F}_\infty(\rho)$, and (5.1) thus reads as the gradient flow

$$\partial_t \rho = \operatorname{div}(\rho \nabla u) - \rho u, \qquad u = p - 1 \in -\partial\mathcal{F}(\rho).$$

However, this functional is too singular for the previous splitting scheme to correctly capture the

very degenerate diffusion. Indeed, the naive and direct approach from section 3 would lead to

$$\left\{\begin{array}{l} \rho_h^{k+1/2} \in \underset{\rho\leqslant 1,\,|\rho|=|\rho_h^k|}{\operatorname{argmin}} \left\{\frac{1}{2h}\mathtt{W}^2(\rho,\rho_h^k) - \int_\Omega \rho\right\}, \\ \\ \rho_h^{k+1} \in \underset{\rho\leqslant 1}{\operatorname{argmin}} \left\{\frac{1}{2h}\mathtt{FR}^2(\rho,\rho_h^{k+1/2}) - \int_\Omega \rho\right\}. \end{array}\right.$$

Since the Wasserstein step is mass-conservative by definition, the $\int \rho$ term has no effect in the first step and the latter reads as "project ρ_h^k on $\{\rho \leqslant 1\}$ w.r.t to the $\mathtt{W}$ distance". Since the output of the reaction step $\rho_h^{k+1} \leqslant 1$, the Wasserstein step will never actually project anything, and the diffusion is completly shut down. As an example, it is easy to see that if the initial datum is an indicator $\rho_0 = \chi_{\Omega_0}$ then the above naive scheme leads to a stationary solution $\rho_h^{k+1} = \rho_h^{k+1/2} = \rho_0$ for all $k \geqslant 0$, while the real solution should evolve according to the aforementioned Hele-Shaw dynamics $\rho(t) = \chi_{\Omega(t)}$ [2, 34]. One could otherwise try to write a semi-implicit scheme as follows: 1) keep the projection on $\{\rho \leqslant 1\}$ in the first Wasserstein step. As in [29] a pressure term $p_h^{k+1/2}$ appears as a Lagrange multiplier in the Wasserstein projection. 2) in the $\mathtt{FR}$/reaction step, relax the constraint $\rho \leqslant 1$ and minimize instead $\rho^{k+1} \in \operatorname{argmin}\left\{\frac{1}{2h}\mathtt{FR}^2(\rho) + \int \rho p^{k+1/2} - \int \rho\right\}$, and keep iterating. This seems to correctly capture the diffusion at least numerically speaking, but raises technical issues in the rigorous proof of convergence and most importantly destroys the variational structure at the discrete level (due to the fact that the reaction step becomes semi-explicit).

We shall use instead an approximation procedure, which preserves the variational structure at the discrete level: it is well-known that the Porous-Medium functional

$$\mathcal{F}_m(\rho) := \left\{\begin{array}{ll} \int_\Omega \frac{\rho^m}{m-1} & \text{if } \rho^m \in L^1(\Omega) \\ +\infty & \text{otherwise} \end{array}\right.$$

Γ-converges to $\mathcal{F}_\infty$ as $m \to \infty$, see [7]. In the spirit of [40], one should therefore expect that the gradient flow ρ_m of $\mathcal{F}_m(\rho) - \int \rho$ converges to the gradient flow ρ_∞ of the limiting functional $\mathcal{F}(\rho) = \mathcal{F}_\infty(\rho) - \int \rho$. Implementing the splitting scheme for the regular energy functional $\mathcal{F}_m(\rho) - \int \rho$ gives a sequence $\rho_{h,m}$, and we shall prove below that $\rho_{h,m}$ converges to a solution of the limiting gradient flow as $m \to \infty$ and $h \to 0$. However, it is known [17] that the limit depends in general on the interplay between the time-step h and the regularization parameter ($m \to \infty$ here), and for technical reasons we shall enforce the condition

$$mh \to 0 \qquad \text{as } m \to \infty \text{ and } h \to 0.$$

Note that [34] already contained a similar approximation $m \to \infty$ but without exploiting the variational structure of the m- gradient flow, and our approach is thus different. The above gradient-flow structure was already noticed and fully exploited in the ongoing work [10], where existence and uniqueness of weak solutions is proved and numerical simulations are performed needless of any splitting an using directly the $\mathtt{WFR}$ structure. Here we rather emphasize the fact that the splitting does capture delicate Γ-convergence phenomena.

In order to make this rigorous, we fix a time step $h > 0$ and construct two sequences $(\rho_{h,m}^{k+1/2})_k$ and $(\rho_{h,m}^k)_k$, with $\rho_{h,m}^0 = \rho_0$, defined recursively as

$$\left\{\begin{array}{l} \rho_h^{k+1/2} \in \underset{\rho\in\mathcal{M}^+,\,|\rho|=|\rho_h^k|}{\operatorname{argmin}} \left\{\frac{1}{2h}\mathtt{W}^2(\rho,\rho_{h,m}^k) + \mathcal{F}_m(\rho) - \int_\Omega \rho\right\}, \\ \\ \rho_h^{k+1} \in \underset{\rho\in\mathcal{M}^+}{\operatorname{argmin}} \left\{\frac{1}{2h}\mathtt{FR}^2(\rho,\rho_h^{k+1/2}) + \mathcal{F}_m(\rho) - \int_\Omega \rho\right\}. \end{array}\right. \tag{5.2}$$

As is common in the classical theory of Porous Media Equations [42], we define the pressure as the first variation

$$p_m := F_m'(\rho) = \frac{m}{m-1}\rho^{m-1}.$$

We accordingly write

$$p_{h,m}^{k+1/2} := \frac{m}{m-1}(\rho_{h,m}^{k+1/2})^{m-1} \qquad \text{and} \qquad p_{h,m}^{k+1} := \frac{m}{m-1}(\rho_{h,m}^{k+1})^{m-1}$$

for the discrete pressures. As in section 3 we denote by $\rho_{h,m}(t), p_{h,m}(t)$ and $\tilde{\rho}_{h,m}(t), \tilde{p}_{h,m}(t)$ the piecewise constant interpolations of $\rho_{h,m}^{k+1}, p_{h,m}^{k+1}$ and $\rho_{h,m}^{k+1/2}, p_{h,m}^{k+1/2}$, respectively.

Our main result is

Theorem 5.1. *Assume that $\rho_0 \in BV(\Omega)$, $\rho_0 \leqslant 1$, and $mh \to 0$ as $h \to 0$ and $m \to \infty$. Then for all $T > 0$, $\rho_{h,m}, \tilde{\rho}_{h,m}$ both converge to some ρ strongly in $L^1((0,T)\times\Omega)$, the pressures $p_{h,m}, \tilde{p}_{h,m}$ both converge to some p weakly in $L^2((0,T), H^1(\Omega))$, and (ρ, p) is the unique weak solution of (5.1).*

Since we have a WFR gradient-flow structure, uniqueness should formally follows from the -1 geodesic convexity of the driving functional $\mathcal{E}_\infty(\rho) - \int_\Omega \rho$ with respect to the WFR distance [24, 26] and the resulting contractivity estimate $\mathtt{WFR}(\rho^1(t), \rho^2(t)) \leq e^t \mathtt{WFR}(\rho_0^1, \rho_0^2)$. This is proved rigorously in [10], and therefore we retrieve convergence of the whole sequence $\rho_{h,m} \to \rho$ in Theorem 5.1 (and not only for subsequences). Given this uniqueness, it is clearly enough to prove convergence along any discrete (sub)sequence, and this is exactly what we show below.

The strategy of proof for Theorem 5.1 is exactly as in section 3, except that we need now the estimates to be uniform in both in $h \to 0$ and $m \to \infty$.

5.1 Estimates and convergences

In this section, we improve the previous estimates from section 3. We start with an explicit L^∞-bound:

Lemma 5.2. *Assume that $\rho_0 \leqslant 1$, then for all $t \in \mathbb{R}^+$,*

$$\|\rho_{h,m}(t,\cdot)\|_\infty, \|\tilde{\rho}_{h,m}(t,\cdot)\|_\infty \leqslant 1.$$

Proof. We argue by induction at the discrete level, starting from $\rho_0 = \rho_{h,m}^0 \leqslant 1$ by assumption. If $\|\rho_{h,m}^k\|_\infty \leqslant 1$, Otto's maximum principle [31] implies that $\|\rho_{h,m}^{k+1/2}\|_\infty \leqslant \|\rho_{h,m}^k\|_\infty \leqslant 1$ in the Wasserstein step.

Assume now by contradiction that $E := \{\rho_{h,m}^{k+1} > 1\}$ has positive Lebesgue measure. The optimality condition (3.9) for the Fisher-Rao minimization step gives, dividing by $\sqrt{\rho_{h,m}^{k+1}} > 0$ in E,

$$\sqrt{\rho_{h,m}^{k+1}} - \sqrt{\rho_{h,m}^{k+1/2}} = \frac{h}{2}\sqrt{\rho_{h,m}^{k+1}}\left(1 - \frac{m}{m-1}(\rho_{h,m}^{k+1})^{m-1}\right)$$

Then $1 - \frac{m}{m-1}(\rho_{h,m}^{k+1})^{m-1} \leqslant 1 - \frac{m}{m-1} < 0$ in the right-hand side, hence the desired contradiction $\rho_{h,m}^{k+1} < \rho_{h,m}^{k+1/2} \leqslant 1$. □

Noticing that the functional $\frac{1}{m-1}\int \rho^m - \int \rho$ corresponds to taking explicitly $F_2(z) = z^m/m - 1$ and $V_2(x) \equiv -1$ in section 3, it is easy to reproduce the computations from the proof of Lemma 3.5 and carefully track the dependence of the constants w.r.t $m > 1$ to obtain

Lemma 5.3. *There exists $c > 0$ such that, for all $m > m_0$ large enough and all $h \leq h_0$ small enough,*

$$(1 - ch)\rho_{h,m}^{k+1/2}(x) \leqslant \rho_{h,m}^{k+1}(x) \leqslant (1+h)\rho_{h,m}^{k+1/2}(x) \qquad a.e. \tag{5.3}$$

Note that this holds regardless of any compatibility such as $hm \to 0$. The key point is here that the lower bound c previously depended on an upper bound M on $\rho^{k+1/2}$ in Lemma 3.5, but since we just obtained in Lemma 5.2 the universal upper bound $\rho^{k+1/2} \leqslant 1$ we end up with a lower bound which is also uniform in h, m. The proof is identical to that of Lemma 3.5 and we omit the details for simplicity.

Recalling that the Wasserstein step is mass-preserving, we obtain by immediate induction and for all $0 \leq t \leq T$

$$\|\rho_{h,m}(t)\|_{L^1}, \|\tilde{\rho}_{h,m}(t)\|_{L^1} \leqslant e^T \|\rho_0\|_{L^1}$$

as well as

$$\|\rho_{h,m}(t) - \tilde\rho_{h,m}(t)\|_{L^1} \leqslant C_T h. \tag{5.4}$$

Testing successively $\rho = \rho_{h,m}^k$ and $\rho = \rho_{h,m}^{k+1/2}$ in (5.2), we get

$$\frac{1}{2h}\left(\mathtt{W}^2(\rho_{h,m}^k, \rho_{h,m}^{k+1/2}) + \mathtt{FR}^2(\rho_{h,m}^{k+1/2}, \rho_{h,m}^{k+1})\right) \leqslant \mathcal{F}_m(\rho_{h,m}^k) - \mathcal{F}_m(\rho_{h,m}^{k+1}) + \int_\Omega (\rho_{h,m}^{k+1/2} - \rho_{h,m}^{k+1}).$$

Using Proposition 2.4 to control $\mathtt{WFR}^2 \lesssim 2(\mathtt{W}^2 + \mathtt{FR}^2)$ and the lower bound in (5.3) yields

$$\begin{aligned}
\frac{1}{4h}\mathtt{WFR}^2(\rho_{h,m}^{k+1}, \rho_{h,m}^k) &\leqslant \frac{1}{2h}\left(\mathtt{W}^2(\rho_{h,m}^k, \rho_{h,m}^{k+1/2}) + \mathtt{FR}^2(\rho_{h,m}^{k+1/2}, \rho_{h,m}^{k+1})\right) \\
&\leqslant \mathcal{F}_m(\rho_{h,m}^k) - \mathcal{F}_m(\rho_{h,m}^{k+1}) + \int_\Omega (\rho_{h,m}^{k+1/2} - \rho_{h,m}^{k+1}) \\
&\leqslant \mathcal{F}_m(\rho_{h,m}^k) - \mathcal{F}_m(\rho_{h,m}^{k+1}) + ch\int_\Omega \rho_{h,m}^{k+1/2} \\
&\leqslant \mathcal{F}_m(\rho_{h,m}^k) - \mathcal{F}_m(\rho_{h,m}^{k+1}) + che^T
\end{aligned}$$

for all $k \leqslant N := \lfloor T/h \rfloor$.

Summing over k we get

$$\begin{aligned}
\frac{1}{4h}\sum_{k=0}^{N-1} \mathtt{WFR}^2(\rho_{h,m}^k, \rho_{h,m}^{k+1}) &\leqslant \mathcal{F}_m(\rho_0) - \mathcal{F}_m(\rho_{h,m}^N) + C_T \\
&\leqslant \frac{1}{m-1}\int_\Omega \rho_0^m + C_T \leqslant \frac{1}{m-1}\int_\Omega \rho_0 + C_T \leqslant C_T,
\end{aligned}$$

where we used successively $F_m \geq 0$ to get rid of $\mathcal{F}_m(\rho_{h,m}^N)$, and $\rho_0^m \leq \rho_0$ for $\rho_0 \leq 1$ and $m > 1$.

Consequently, for all fixed $T > 0$ and any $t, s \in [0, T]$ we obtain the classical $\frac{1}{2}$-Hölder estimate

$$\begin{cases} \mathtt{WFR}(\rho_{h,m}(t), \rho_{h,m}(s)) \leqslant C_T|t - s + h|^{1/2}, \\ \mathtt{WFR}(\tilde\rho_{h,m}(t), \tilde\rho_{h,m}(s)) \leqslant C_T|t - s + h|^{1/2}. \end{cases} \tag{5.5}$$

Exploiting the explicit algebraic structure of $F_m(z) = \frac{1}{m-1}z^m$, compactness in space will be given here by

Lemma 5.4. *If $\rho_0 \in BV(\Omega)$ then*

$$\sup_{t\in[0,T]} \left\{\|\rho_{h,m}(t,\cdot)\|_{BV(\Omega)}, \|\tilde\rho_{h,m}(t,\cdot)\|_{BV(\Omega)}\right\} \leqslant e^T\|\rho_0\|_{BV(\Omega)}.$$

Proof. The argument closely follows the lines of [18, prop. 5.1]. We first note from [14, thm. 1.1] that the BV-norm is nonincreasing during the Wasserstein step,

$$\|\rho_{h,m}^{k+1/2}\|_{BV(\Omega)} \leqslant \|\rho_{h,m}^k\|_{BV(\Omega)}.$$

Using as before the implicit function theorem, we show below that $\rho_{h,m}^{k+1} = R(\rho_{h,m}^{k+1/2})$ for some suitable $(1+h)$-Lispchitz function R. By standard $Lip \circ BV$ composition [3] this will prove that

$$\|\rho_{h,m}^{k+1}\|_{BV(\Omega)} \leqslant (1+h)\|\rho_{h,m}^{k+1/2}\|_{BV(\Omega)}$$

and will conclude the proof by immediate induction.

Indeed, we already know from (5.3) that $\rho_{h,m}^{k+1/2}$ and $\rho_{h,m}^{k+1}$ share the same support. In this support and from (3.9) it is easy to see that $\rho = \rho_{h,m}^{k+1}(x)$ is the unique positive solution of $f(\rho, \rho_{h,m}^{k+1/2}(x)) = 0$ with

$$f(\rho, \mu) = \sqrt{\rho}\left(1 - \frac{h}{2}\left(1 - \frac{m}{m-1}\rho^{m-1}\right)\right) - \sqrt{\mu}.$$

For $\mu > 0$, the implicit function theorem gives the existence of a $\mathcal{C}^1$ map R such that $f(\rho,\mu) = 0 \Leftrightarrow \rho = R(\mu)$, with $R(0) = 0$. An algebraic computation shows moreover that $0 < \frac{dR}{d\mu} = -\frac{\partial_\mu f}{\partial_\rho f}_{|\rho=R(\mu)} \leqslant (1+h)$ uniformly in $m > 1$, hence R is $(1+h)$-Lipschitz as claimed and the proof is complete.

□

Proposition 5.5. *Up to extraction of a discrete sequence $h \to 0, m \to \infty$, there holds*

$$\rho_{h,m},\, \tilde\rho_{h,m} \to \rho \qquad \text{strongly in } L^1(Q_T)$$

$$p_{h,m} \rightharpoonup p \quad \text{and} \quad \tilde p_{h,m} \rightharpoonup \tilde p \qquad \text{weakly in all } L^q(Q_T)$$

for all $T > 0$. If in addition $mh \to 0$ then $p = \tilde p$.

Proof. The first part of the statement follows exactly as in section 3, exploiting the $\frac{1}{2}$-Hölder estimates (5.5) and the space compactness from Proposition 5.4 in order to apply the Rossi-Savaré theorem [39]. The fact that $\rho_{h,m}, \tilde\rho_{h,m}$ have the same limit comes from (5.4).

For the pressures, we simply note from $\rho_{h,m} \leqslant 1$ and $m \gg 1$ that $p_{h,m} = \frac{m}{m-1}\rho_{h,m}^{m-1} \leqslant 2\rho_{h,m}$ is bounded in $L^1 \cap L^\infty(Q_T)$ uniformly in h, m in any finite time interval $[0,T]$. Thus up to extraction of a further sequence we have $p_{h,m} \rightharpoonup p$ in all $L^q(Q_T)$, and likewise for $\tilde p_{h,m} \rightharpoonup \tilde p$.

Finally, we only have to check that $p = \tilde p$ if $hm \to 0$. Because $\rho_{h,m}, \tilde\rho_{h,m} \leqslant 1$ and $z \mapsto z^{m-1}$ is $(m-1)$-Lipschitz on $[0,1]$ we have for all fixed $t \geqslant 0$ that

$$\begin{aligned}\int_\Omega |p_{m,h}(t,\cdot) - \tilde p_{m,h}(t,\cdot)| &= \int_\Omega \frac{m}{m-1}|\rho_{h,m}^{m-1}(t,\cdot) - \tilde\rho_{h,m}^{m-1}(t,\cdot)| \\ &\leqslant m\int_\Omega |\rho_{h,m}(t) - \tilde\rho_h(t)| \leqslant C_T hm \longrightarrow 0,\end{aligned}$$

where we used (5.4) in the last inequality. Hence $p = \tilde p$ and the proof is complete.

□

In order to pass to the limit in the diffusion term $\mathrm{div}(\rho\nabla p)$ we first improve the convergence of $\tilde p_{h,m}$:

Lemma 5.6. *There exists a constant C_T, independent of h and m, such that*

$$\|\tilde p_{h,m}\|_{L^2((0,T),H^1(\Omega))} \leqslant C_T$$

for all $T > 0$. Consequently, up to a subsequence, $\tilde p_{h,m}$ converges weakly in $L^2((0,T),H^1(\Omega))$ to p.

Proof. The proof is based on the flow interchange technique developed by Matthes, McCann and Savaré in [28]. Let η be the (smooth) solution of

$$\begin{cases} \partial_t \eta = \Delta \eta^{m-1} + \varepsilon\Delta\eta, \\ \eta|_{t=0} = \rho_{h,m}^{k+1/2}. \end{cases}$$

It is well known [4] that η is the Wasserstein gradient flow of

$$\mathcal{G}(\rho) := \int_\Omega \frac{\rho^{m-1}}{m-2} + \varepsilon\int_\Omega \rho\log(\rho).$$

Since $\mathcal{G}$ is geodesically 0-convex, η satisfies the Evolution Variational Inequality (EVI)

$$\frac{1}{2}\frac{d^+}{dt}\Big|_{t=s} \mathtt{W}^2(\eta(s),\rho) \leqslant \mathcal{G}(\rho) - \mathcal{G}(\eta(s)),$$

for all $s > 0$ and for all $\rho \in \mathcal{P}^{\mathrm{ac}}(\Omega)$, where $\frac{d^+}{dt}f(t) := \limsup_{s\to 0^+} \frac{f(t+s)-f(t)}{s}$. By optimality of $\rho_{h,m}^{k+1/2}$ in (5.2), we obtain that

$$\frac{1}{2}\frac{d^+}{dt}\Big|_{t=s} \mathtt{W}^2(\eta(s),\rho_{h,m}^k) \geqslant -h\,\frac{d^+}{dt}\Big|_{t=s} \mathcal{F}_m(\eta(s)).$$

Since η is smooth due to the regularizing $\varepsilon\Delta$ term, we can legitimately integrate by parts for all $s>0$

$$\begin{aligned}\frac{d}{ds}\mathcal{F}_m(\eta(s)) &= \int_\Omega \frac{m}{m-1}\eta(s)^{m-1}(\Delta\eta(s)^{m-1}+\varepsilon\Delta\eta(s))\\ &= -\int_\Omega \frac{m}{m-1}|\nabla\eta(s)^{m-1}|^2-\varepsilon\int_\Omega m\eta(s)^{m-2}|\nabla\eta(s)|^2\\ &\leqslant -\int_\Omega \frac{m}{m-1}|\nabla\eta(s)^{m-1}|^2=-\frac{m-1}{m}\int_\Omega\left|\nabla\left(\frac{m}{m-1}\eta(s)^{m-1}\right)\right|^2\end{aligned}$$

Remarking that $\frac{m}{m-1}\eta(s)^{m-1}\to\frac{m}{m-2}\rho_{h,m}^{k+1/2}=p_{h,m}^{k+1/2}$ as $s\to 0$, an easy lower semi-continuity argument gives that

$$\int_\Omega\frac{m-1}{m}|\nabla p_{h,m}^{k+1/2}|^2=\int_\Omega\frac{m}{m-1}|\nabla(\rho_{h,m}^{k+1/2})^{m-1}|^2\leqslant\liminf_{s\searrow 0}\left.\frac{d^+}{dt}\right|_{t=s}\mathcal{F}_m(\eta(s)).$$

Then we have

$$\begin{aligned}h\int_\Omega\frac{m-1}{m}|\nabla p_{h,m}^{k+1/2}|^2\leqslant{}&\mathcal{F}_{m-1}(\rho_{h,m}^k)-\mathcal{F}_{m-1}(\rho_{h,m}^{k+1/2})\\ &+\varepsilon\left(\int_\Omega\rho_{h,m}^k\log(\rho_{h,m}^k)-\int_\Omega\rho_{h,m}^{k+1/2}\log(\rho_{h,m}^{k+1/2})\right).\end{aligned}$$

First arguing as in Proposition 3.7 to control

$$\mathcal{F}_{m-1}(\rho_{h,m}^{k+1})\leqslant\mathcal{F}_{m-1}(\rho_{h,m}^{k+1/2})+C_Th,$$

and then passing to the limit $\varepsilon\searrow 0$, we obtain

$$h\int_\Omega\frac{m-1}{m}|\nabla p_{h,m}^{k+1/2}|^2\leqslant\mathcal{F}_{m-1}(\rho_{h,m}^k)-\mathcal{F}_{m-1}(\rho_{h,m}^{k+1})+C_Th.$$

Summing over k gives

$$\int_0^T\int_\Omega|\nabla\tilde{p}_{h,m}(t,x)|^2\,dxdt\leqslant\frac{m}{m-1}(\mathcal{F}_{m-1}(\rho_0)-\mathcal{F}_{m-1}(\rho_{h,m}^N)+C_T)\leqslant 2\mathcal{F}_{m-1}(\rho_0)+C_T$$

for all $T<+\infty$. Due to $\rho_0\leqslant 1$ and $m\gg 1$ we can bound $\mathcal{F}_{m-1}(\rho_0)=\frac{1}{m-2}\int\rho_0^{m-1}\leqslant\frac{1}{m-2}\int\rho_0\leqslant\|\rho_0\|_{L^1(\Omega)}$ and the result finally follows. □

5.2 Properties of the pressure p and conclusion

We start by showing that the limits ρ,p satisfy the compatibility conditions in (5.1).

Lemma 5.7. *There holds*

$$0\leqslant\rho,p\leqslant 1\quad\textit{and}\quad p(1-\rho)=0\ \textit{ a.e. in } Q_T.$$

Proof. By Lemma 5.2 it is obvious that $0\leqslant\rho\leqslant 1$ and $0\leqslant p\leqslant 1$ are inherited from $0\leqslant\rho_{h,m}\leqslant 1$ and $0\leqslant p_{h,m}=\frac{m}{m-1}\rho_{h,m}^{m-1}\leqslant\frac{m}{m-1}$.

In order to prove that $p(1-\rho)=0$, we first observe that

$$p_{h,m}(1-\rho_{h,m})\to 0\qquad\text{a.e. in } Q_T.$$

Indeed, since $\rho_{h,m}\to\rho$ strongly in $L^1(Q_T)$ we have $\rho_{h,m}(t,x)\to\rho(t,x)$ a.e. If the limit $\rho(t,x)<1$ then $\rho_{h,m}(t,x)\leqslant(1-\varepsilon)$ for small h and large m. Hence $p_{h,m}(t,x)=\frac{m}{m-1}\rho_{h,m}^{m-1}\leqslant\frac{m}{m-1}(1-\varepsilon)^{m-1}\to 0$ while $1-\rho_{h,m}$ remains bounded, and therefore the product $p_{h,m}(1-\rho_{h,m})\to 0$. Now if the limit $\rho(t,x)=1$ then the pressure $p_{h,m}=\frac{m}{m-1}\rho_{h,m}^{m-1}\leqslant\frac{m}{m-1}$ remains bounded, while $1-\rho_{h,m}(t,x)\to 0$ hence the product goes to zero in this case too.

Thanks to the uniform L^∞ bounds $\rho_{h,m} \leqslant 1$ and $p_{h,m} \leqslant \frac{m}{m-1} \leqslant 2$ we can apply Lebesgue's convergence theorem to deduce from this pointwise a.e. convergence that, for all fixed nonnegative $\varphi \in C_c^\infty(Q_T)$, there holds

$$\lim \int_{Q_T} p_{h,m}(1-\rho_{h,m})\varphi = 0.$$

On the other hand since $\rho_{h,m} \to \rho$ strongly in $L^1(Q_T)$ hence a.e, and because $0 \leqslant \rho_{h,m} \leqslant 1$, we see that $(1-\rho_{h,m})\varphi \to (1-\rho)\varphi$ in all $L^q(Q_T)$. From Proposition 5.5 we also had that $p_{h,m} \rightharpoonup p$ in all $L^q(Q_T)$, hence by strong-weak convergence we have that

$$\int_{Q_T} p(1-\rho)\varphi = \lim \int_{Q_T} p_{h,m}(1-\rho_{h,m})\varphi = 0$$

for all $\varphi \geqslant 0$. Because $p(1-\rho) \geqslant 0$ we conclude that $p(1-\rho) = 0$ a.e. in Q_T and the proof is achieved.

□

We end this section with

Proof of Theorem 5.1. We only sketch the argument and refer to [18] for the details. Fix any $0 < t_1 < t_2$ and $\varphi \in \mathcal{C}_c^2(\mathbb{R}^d)$. Exploiting the Euler-Lagrange equations (3.6)(3.9) and summing from $k = k_1 = \lfloor t_1/h \rfloor$ to $k = k_2 - 1 = \lfloor t_2/h \rfloor - 1$, we first obtain

$$\int_{\mathbb{R}^d} \rho_{h,m}(t_2)\varphi - \rho_{h,m}(t_1)\varphi + \int_{k_1 h}^{k_2 h} \int_{\mathbb{R}^d} \tilde{\rho}_{h,m} \nabla \tilde{p}_{h,m} \cdot \nabla \varphi = -\int_{k_1 h}^{k_2 h} \int_{\mathbb{R}^d} \rho_{h,m}(1-p_{h,m})\varphi + R(h,m),$$

where the remainder $R(h,m) \to 0$ for fixed φ. The strong convergence $\rho_{h,m}, \tilde{\rho}_{h,m} \to \rho$ and the weak convergences $\nabla \tilde{p}_{h,m} \rightharpoonup \nabla \tilde{p} = \nabla p$ and $p_{h,m} \rightharpoonup p$ are then enough pass to the limit to get the corresponding weak formulation for all $0 < t_1 < t_2$. Moreover since the limit $\rho \in \mathcal{C}([0,T]; \mathcal{M}_{\texttt{WFR}}^+)$ the initial datum $\rho(0) = \rho_0$ is taken at least in the sense of measures. This gives an admissible weak formulation of (5.1), and the proof is complete. □

5.3 Numerical simulation

The constructive scheme (5.2) naturally leads to a fully discrete algorithm, simply discretizing the minimization problem in space for each $\texttt{W}, \texttt{FR}$ step. We use again the ALG2-JKO scheme [6] for the Wasserstein steps. As already mentioned the Fisher-Rao step is a mere convex pointwise minimization problem, here explicitly given by: for all $x \in \Omega$,

$$\rho_{h,m}^{k+1}(x) = \operatorname*{argmin}_{\rho \geq 0} \left\{ 4 \left| \sqrt{\rho} - \sqrt{\rho_{h,m}^{k+1/2}(x)} \right|^2 + 2h \left(\frac{\rho^m}{m-1} - 1 \right) \right\}$$

and poses no difficulty in the practical implementation using a standard Newton method.

Figure 3 depicts the evolution of the numerical solution $\rho_{h,m}$ for $m = 100$ and with a time step $h = 0.005$. We remark that the tumor first saturates the constraint ($\rho \nearrow 1$) in its initial support, and then starts diffusing outwards. This is consistent with the qualitative behaviour described in [34].

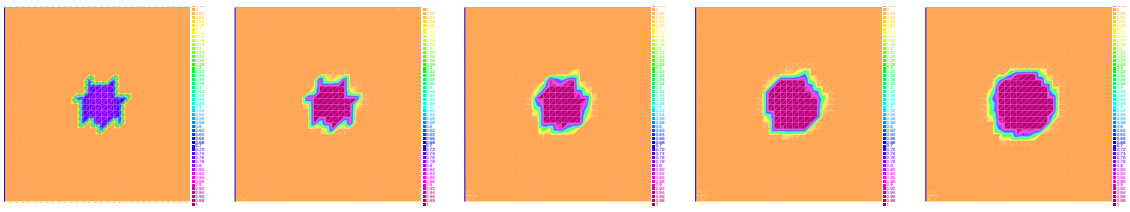


Figure 3: *Snapshot of the approximate solution* $\rho_{h,m}(t,.)$ *to* (5.1)*, with* $m = 100$*,* $h = 0.005$*.*

6 A tumor growth model with nutrient

In this section we use the same approach for the following tumor growth model with nutrients, appearing e.g. in [34]

$$\begin{cases} \partial_t \rho - \operatorname{div}(\rho \nabla p) = \rho\left((1-p)(c+c_1) - c_2\right), \\ \partial_t c - \Delta c = -\rho c, \\ 0 \leqslant \rho \leqslant 1, \\ p \geqslant 0 \text{ and } p(1-\rho) = 0, \\ \rho_{|t=0} = \rho_0,\ c_{|t=0} = c_0. \end{cases} \tag{6.1}$$

Here c_1 and c_2 are two positive constants, and the nutrient c is now diffusing in Ω in addition to begin simply consumed by the tumor ρ, according to the second equation. For technical convenience we work here on a convex bounded domain $\Omega \subset \mathbb{R}^d$, endowed with natural Neumann boundary conditions for both ρ and c.

Contrarily to section 5 this is not a `WFR` gradient flow anymore, and we therefore introduce a semi-implicit splitting scheme. Starting from the initial datum $\rho^0_{h,m} := \rho_0, c^0_{h,m} := c_0$ we construct four sequences $\rho^{k+1/2}_{h,m}, \rho^k_{h,m}, c^{k+1/2}_{h,m}, c^k_{h,m}$, defined recursively as

$$\begin{cases} \rho^{k+1/2}_{h,m} \in \underset{\rho\in\mathcal{M}^+, |\rho|=|\rho^k_{h,m}|}{\operatorname{argmin}} \left\{ \frac{1}{2h}\mathtt{W}^2(\rho, \rho^k_{h,m}) + \mathcal{F}_m(\rho) \right\}, \\ \\ c^{k+1/2}_{h,m} \in \underset{c\in\mathcal{M}^+, |c|=|c^k_{h,m}|}{\operatorname{argmin}} \left\{ \frac{1}{2h}\mathtt{W}^2(c, c^k_{h,m}) + \mathcal{E}(\rho) \right\}, \end{cases} \tag{6.2}$$

and

$$\begin{cases} \rho^{k+1}_{h,m} \in \underset{\rho\in\mathcal{M}^+}{\operatorname{argmin}} \left\{ \frac{1}{2h}\mathtt{FR}^2(\rho, \rho^{k+1/2}_{h,m}) + \mathcal{E}_{1,m}(\rho|c^{k+1/2}_{h,m}) \right\}, \\ \\ c^{k+1}_{h,m} \in \underset{c\in\mathcal{M}^+}{\operatorname{argmin}} \left\{ \frac{1}{2h}\mathtt{FR}^2(c, c^{k+1/2}_{h,m}) + \mathcal{E}_2(c|\rho^{k+1/2}_{h,m}) \right\}, \end{cases} \tag{6.3}$$

where

$$\mathcal{E}(\rho) := \int_\Omega \rho\log(\rho),$$

$$\mathcal{E}_{1,m}(\rho|c) := \int_\Omega (c+c_1)\,\frac{\rho^m}{m-1} + \int_\Omega (c_2 - c - c_1)\rho,$$

and

$$\mathcal{E}_2(c|\rho) := \int_\Omega \rho c.$$

As earlier it is easy to see that these sequences are well-defined (i-e there exists a unique minimizer for each step), and the pressures are defined as before as

$$p^{k+1/2}_{h,m} := \frac{m}{m-1}(\rho^{k+1/2}_{h,m})^{m-1} \quad \text{and} \quad p^{k+1}_{h,m} := \frac{m}{m-1}(\rho^{k+1}_{h,m})^{m-1}.$$

We denote again by $a_{h,m}(t), \tilde{a}_{h,m}(t)$ the piecewise constant interpolation of any discrete quantity $a^{k+1}_{h,m}, a^{k+1/2}_{h,m}$ respectively. Our main result reads:

Theorem 6.1. *Assume $\rho_0 \in BV(\Omega)$ with $\rho_0 \leqslant 1$ and $c_0 \in L^\infty(\Omega) \cap BV(\Omega)$. Then $\rho_{h,m}$ and $\tilde{\rho}_{h,m}$ strongly converge to ρ in $L^1((0,T)\times\Omega)$ and $c_{h,m}$ and $\tilde{c}_{h,m}$ strongly converge to c in $L^1((0,T)\times\Omega)$ when $h \searrow 0$ and $m \nearrow +\infty$. Moreover, if $mh \to 0$, then $p_{h,m}, \tilde{p}_{h,m}$ converge weakly in $L^2((0,T), H^1(\Omega))$ to a unique p, and (ρ, p, c) is a solution of (6.1).*

Note that uniqueness of solutions would result in convergence of the whole sequence. Uniqueness was proved in [34, thm. 4.2] for slightly more regular weak solutions, but we did not push in this direction for the sake of simplicity. The method of proof is almost identical to section 5 so we only sketch the argument and emphasize the main differences.

We start by recalling the optimality conditions for the scheme (6.2)-(6.3). The Euler-Lagrange equations for the tumor densities in the Wasserstein and Fisher-Rao steps are

$$\begin{cases} \rho_{h,m}^{k+1/2}\nabla p_{h,m}^{k+1/2} = \frac{\nabla\varphi}{h}\rho_{h,m}^{k+1/2}, \\ \sqrt{\rho_{h,m}^{k+1}} - \sqrt{\rho_{h,m}^{k+1/2}} = \frac{h}{2}\sqrt{\rho_{h,m}^{k+1}}\left((1-p_{h,m}^{k+1})(c_{h,m}^{k+1/2}+c_1)-c_2\right), \end{cases} \tag{6.4}$$

where φ is a (backward) Kantorovich potential for $\mathtt{W}(\rho_{h,m}^{k+1/2},\rho_{h,m}^{k})$. For the nutrient, the Euler-Lagrange equations are

$$\begin{cases} \nabla c_{h,m}^{k+1/2} = \frac{\nabla\psi}{h}c_{h,m}^{k+1/2}, \\ \sqrt{c_{h,m}^{k+1}} - \sqrt{c_{h,m}^{k+1/2}} = -\frac{h}{2}\sqrt{c_{h,m}^{k+1}}\rho_{h,m}^{k+1/2}, \end{cases} \tag{6.5}$$

with ψ a Kantorovich potential for $\mathtt{W}(c_{h,m}^{k+1/2},c_{h,m}^{k})$.

Using the optimality conditions for the Fischer-Rao steps, we obtain directly the following L^∞ bounds:

Lemma 6.2. *For all $k \geqslant 0$*

$$\|c_{h,m}^{k+1}\|_{L^\infty(\Omega)} \leqslant \|c_{h,m}^{k+1/2}\|_{L^\infty(\Omega)} \leqslant \|c_{h,m}^{k}\|_{L^\infty(\Omega)},$$

and at the continuous level

$$\|c_{h,m}(t,\cdot)\|_{L^\infty(\Omega)}, \|\tilde{c}_{h,m}(t,\cdot)\|_{L^\infty(\Omega)} \leqslant \|c_0\|_{L^\infty(\Omega)} \qquad \forall\, t \geq 0.$$

Moreover,

$$\|\rho_{h,m}(t,\cdot)\|_\infty, \|\tilde{\rho}_{h,m}(t,\cdot)\|_\infty \leqslant 1$$

and there exists $c_T \equiv c_T(\|c_0\|_{L^\infty}), C_T \equiv C_T(\|c_0\|_{L^\infty}) > 0$ such that

$$\begin{gathered} (1-c_Th)\rho_{h,m}^{k+1/2}(x) \leqslant \rho_{h,m}^{k+1}(x) \leqslant (1+C_Th)\rho_{h,m}^{k+1/2}(x) \qquad \text{a.e. in } \Omega. \\ (1-h)c_{h,m}^{k+1/2}(x) \leqslant c_{h,m}^{k+1}(x) \leqslant c_{h,m}^{k+1/2}(x) \qquad \text{a.e. in } \Omega. \end{gathered} \tag{6.6}$$

Proof. The proof of the estimates on $c_{h,m}$ and $\tilde{c}_{h,m}$ is obvious because one step of Wasserstein gradient flow with the Boltzmann entropy decreases the L^∞-norm in (6.2) (see [32, 1]), and, because the product $\sqrt{c_{h,m}^{k+1}}\rho_{h,m}^{k+1/2}$ is nonnegative in (6.5), the L^∞-norm is also nonincreasing during the Fischer-Rao step. The proof for $\rho_{h,m}$ and $\tilde{\rho}_{h,m}$ is the same as in lemma 5.2. Using the fact that $\|\tilde{\rho}_{h,m}(t,\cdot)\|_\infty \leqslant 1$, we see that the term $\Phi(p_{h,m}^{k+1},c_{h,m}^{k+1/2}) := (1-p_{h,m}^{k+1})(c_{h,m}^{k+1/2}+c_1)-c_2$ in (6.4) is bounded in L^∞ uniformly in k. This allows to argue exactly as in Lemma 3.5 to retrieve the estimate (6.6) and concludes the proof. □

With these bounds it is easy to prove as in proposition 3.15 that

$$\begin{gathered} \mathcal{F}_m(\rho_{h,m}^{k+1}) \leqslant \mathcal{F}_m(\rho_{h,m}^{k+1/2}) + C_Th, \\ \mathcal{E}_{1,m}(\rho_{h,m}^{k+1/2}|c_{h,m}^{k+1/2}) - \mathcal{E}_{1,m}(\rho_{h,m}^{k+1}|c_{h,m}^{k+1/2}) \leqslant C_Th, \\ \mathcal{E}(c_{h,m}^{k+1}) \leqslant \mathcal{E}(c_{h,m}^{k+1/2}) + C_Th, \\ \mathcal{E}_2(c_{h,m}^{k+1/2}|\rho_{h,m}^{k+1/2}) - \mathcal{E}_2(c_{h,m}^{k+1}|\rho_{h,m}^{k+1/2}) \leqslant C_Th, \end{gathered} .$$

for some C_T independent of m. Then we obtain the usual $\frac{1}{2}$-Hölder estimates in time with respect to the WFR distance, which in turn implies that $\rho_{h,m},\tilde{\rho}_{h,m}$ converge to some $\rho \in L^\infty([0,T],L^1(\Omega))$ and $c_{h,m},\tilde{c}_{h,m}$ converge to some $c \in L^\infty([0,T],L^1(\Omega))$ pointwise in time with respect to WFR, see (3.20), Proposition 3.8, and (3.22) for details.

As before we need to improve the convergence in order to pass to the limit in the nonlinear terms. For $\rho_{h,m}$ and $\tilde{\rho}_{h,m}$, this follows from

Lemma 6.3. *For all $T > 0$, if $\rho_0, c_0 \in BV(\Omega)$,*

$$\sup_{t\in[0,T]} \left\{ \|\rho_{h,m}(t,\cdot)\|_{BV(\Omega)} + \|c_{h,m}(t,\cdot)\|_{BV(\Omega)} \right\} \leqslant e^{C_T T}(\|\rho_0\|_{BV(\Omega)} + \|c_0\|_{BV(\Omega)})$$
$$\sup_{t\in[0,T]} \left\{ \|\tilde\rho_{h,m}(t,\cdot)\|_{BV(\Omega)} + \|\tilde c_{h,m}(t,\cdot)\|_{BV(\Omega)} \right\} \leqslant e^{C_T T}(\|\rho_0\|_{BV(\Omega)} + \|c_0\|_{BV(\Omega)}).$$

Proof. The argument is a generalization of Lemma 5.4, see [18, remark 5.1]. First, the BV-norm is nonincreasing during the Wasserstein step, [14, thm. 1.1],

$$\|\rho_{h,m}^{k+1/2}\|_{BV(\Omega)} \leqslant \|\rho_{h,m}^{k}\|_{BV(\Omega)} \text{ and } \|c_{h,m}^{k+1/2}\|_{BV(\Omega)} \leqslant \|c_{h,m}^{k}\|_{BV(\Omega)}.$$

Arguing as in Lemma 5.4, we observe that, inside $\operatorname{supp}\rho_{h,m}^{k+1/2} = \operatorname{supp}\rho_{h,m}^{k+1}$, the minimizer $\rho = \rho_{h,m}^{k+1}(x)$ is the unique positive solution of $f(\rho, \rho_{h,m}^{k+1/2}(x), c_{h,m}^{k+1/2}(x)) = 0$, with

$$f(\rho,\mu,c) = \sqrt{\rho}\left(1 - \frac{h}{2}\left(\left(1 - \frac{m}{m-1}\rho^{m-1}\right)(c+c_1) - c_2\right)\right) - \sqrt{\mu}.$$

For $\mu > 0$ the implicit function theorem gives as before a $\mathcal{C}^1$ map R such that $f(\rho,\mu,c) = 0 \Leftrightarrow \rho = R(\mu,c)$. An easy algebraic computation and (6.6) then gives $0 < \partial_\mu R(\mu,c) \leqslant (1 + C_T h)$ and $|\partial_c R(\mu,c)| \leqslant C_T h$ for some constant $C_T > 0$ independent of h, m, k. This implies that

$$\begin{aligned}\|\rho_{h,m}^{k+1}\|_{BV(\Omega)} &\leqslant (1+C_T h)\|\rho_{h,m}^{k+1/2}\|_{BV(\Omega)} + C_T h\|c_{h,m}^{k+1/2}\|_{BV(\Omega)} \\ &\leqslant (1+C_T h)\|\rho_{h,m}^{k}\|_{BV(\Omega)} + C_T h\|c_{h,m}^{k}\|_{BV(\Omega)}.\end{aligned}$$

The same argument shows that

$$\|c_{h,m}^{k+1}\|_{BV(\Omega)} \leqslant (1+C_T h)\|c_{h,m}^{k}\|_{BV(\Omega)} + C_T h\|\rho_{h,m}^{k}\|_{BV(\Omega)},$$

and a simple induction allows to conclude. □

Proposition 6.4. *Up to extraction of a discrete sequence $h \to 0, m \to +\infty$,*

$$\rho_{h,m},\ \tilde\rho_{h,m} \to \rho \qquad \textit{strongly in } L^1(Q_T)$$

$$p_{h,m} \rightharpoonup p \textit{ and } \tilde p_{h,m} \rightharpoonup \tilde p \qquad \textit{weakly in all } L^q(Q_T)$$

for all $T > 0$. If in addition $mh \to 0$ then $p = \tilde p \in L^2((0,T), H^1(\Omega))$ and (ρ, p) satisfies

$$0 \leqslant \rho, p \leqslant 1 \quad \textit{and} \quad p(1-\rho) = 0 \qquad \textit{a.e. in } Q_T.$$

Proof. The proof is the same as Proposition 5.5, Lemma 5.6, and Lemma 5.7. □

In order to conclude the proof of Theorem 6.1 we only need to check that ρ, p, c satisfy the weak formulation of (6.1): the strong convergence of $\rho_{h,m}, c_{h,m}$ and the weak convergence of $p_{h,m}$ are enough to take the limit in the nonlinear terms as in section 5.2, and we omit the details.

Acknowledgements

We warmly thank G. Carlier for fruitful discussions and suggesting us the problem in section 3

References

[1] Martial Agueh. Existence of solutions to degenerate parabolic equations via the Monge-Kantorovich theory. *Adv. Differential Equations*, 10(3):309–360, 2005.

[2] Damon Alexander, Inwon Kim, and Yao Yao. Quasi-static evolution and congested crowd transport. *Nonlinearity*, 27(4):823, 2014.

[3] Luigi Ambrosio, Nicola Fusco, and Diego Pallara. *Functions of bounded variation and free discontinuity problems*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, New York, 2000.

[4] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2005.

[5] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

[6] Benamou, Jean-David, Carlier, Guillaume, and Laborde, Maxime. An augmented lagrangian approach to wasserstein gradient flows and applications. *ESAIM: ProcS*, 54:1–17, 2016.

[7] Andrea Braides. Γ*-convergence for beginners*, volume 22 of *Oxford Lecture Series in Mathematics and its Applications*. Oxford University Press, Oxford, 2002.

[8] G. Carlier and M. Laborde. A splitting method for nonlinear diffusions with nonlocal, nonpotential drifts. *Nonlinear Analysis: Theory, Methods & Applications*, 150:1 – 18, 2017.

[9] J. A. Carrillo, M. DiFrancesco, A. Figalli, T. Laurent, and D. Slepčev. Global-in-time weak measure solutions and finite-time aggregation for nonlocal interaction equations. *Duke Math. J.*, 156(2):229–271, 2011.

[10] Lénaic Chizat and Simone Di Marino. A tumor growth hele-shaw problem as a gradient flow. *Work in progress*, 2017.

[11] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and Fischer-Rao. *arXiv preprint arXiv:1506.06430*, 2015.

[12] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: geometry and Kantorovich formulation. *arXiv preprint arXiv:1508.05216*, 2015.

[13] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced transport problems. *arXiv preprint arXiv:1607.05816*, 2016.

[14] Guido De Philippis, Alpár Richárd Mészáros, Filippo Santambrogio, and Bozhidar Velichkov. BV estimates in optimal transportation and applications. *Arch. Ration. Mech. Anal.*, 219(2):829–860, 2016.

[15] Marco Di Francesco and Simone Fagioli. Measure solutions for non-local interaction PDEs with two species. *Nonlinearity*, 26(10):2777–2808, 2013.

[16] Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with dirichlet boundary conditions. *Journal de mathématiques pures et appliquées*, 94(2):107–130, 2010.

[17] Florentine Fleißner. Gamma-convergence and relaxations for gradient flows in metric spaces: a minimizing movement approach. *arXiv preprint arXiv:1603.02822*, 2016.

[18] Thomas Gallouët and Leonard Monsaingeon. A JKO splitting scheme for kantorovich-fisher-rao gradient flows. working paper or preprint, February 2016.

[19] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.*, 29(1):1–17, 1998.

[20] David Kinderlehrer, Léonard Monsaingeon, and Xiang Xu. A wasserstein gradient flow approach to poisson-nernst-planck equations. *arXiv preprint arXiv:1501.04437*, 2015.

[21] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A new optimal transport distance on the space of finite radon measures. *arXiv preprint arXiv:1505.07746*, 2015.

[22] Stanislav Kondratyev, Léonard Monsaingeon, and Dmitry Vorotnikov. A fitness-driven cross-diffusion system from population dynamics as a gradient flow. *Journal of Differential Equations*, 261(5):2784 – 2808, 2016.

[23] M. Laborde. On some non linear evolution systems which are perturbations of Wasserstein gradient flows. *to appear in Radon Ser. Comput. Appl. Math.*, 2015.

[24] Matthias Liero and Alexander Mielke. Gradient structures and geodesic convexity for reaction–diffusion systems. *Phil. Trans. R. Soc. A*, 371(2005):20120346, 2013.

[25] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal Entropy-Transport problems and a new Hellinger-Kantorovich distance between positive measures. *arXiv preprint arXiv:1508.07941*, 2015.

[26] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal transport in competition with reaction: the Hellinger-Kantorovich distance and geodesic curves. *arXiv preprint arXiv:1509.00068*, 2015.

[27] Stefano Lisini, Daniel Matthes, and Giuseppe Savaré. Cahn-Hilliard and thin film equations with nonlinear mobility as gradient flows in weighted-Wasserstein metrics. *J. Differential Equations*, 253(2):814–850, 2012.

[28] Daniel Matthes, Robert J. McCann, and Giuseppe Savaré. A family of nonlinear fourth order equations of gradient flow type. *Comm. Partial Differential Equations*, 34(10-12):1352–1397, 2009.

[29] Bertrand Maury, Aude Roudneff-Chupin, Filippo Santambrogio, and Juliette Venel. Handling congestion in crowd motion modeling. *Netw. Heterog. Media*, 6(3):485–519, 2011.

[30] J. D. Murray. *Mathematical biology. II*, volume 18 of *Interdisciplinary Applied Mathematics*. Springer-Verlag, New York, third edition, 2003. Spatial models and biomedical applications.

[31] Felix Otto. Double degenerate diffusion equations as steepest descent, 1996.

[32] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Differential Equations*, 26(1-2):101–174, 2001.

[33] Benoît Perthame. *Transport equations in biology*. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2007.

[34] Benoît Perthame, Fernando Quirós, and Juan Luis Vázquez. The Hele-Shaw asymptotics for mechanical models of tumor growth. *Arch. Ration. Mech. Anal.*, 212(1):93–127, 2014.

[35] Benoît Perthame, Min Tang, and Nicolas Vauchelet. Traveling wave solution of the Hele-Shaw model of tumor growth with nutrient. *Math. Models Methods Appl. Sci.*, 24(13):2601–2626, 2014.

[36] Luca Petrelli and Adrian Tudorascu. Variational principle for general diffusion problems. *Appl. Math. Optim.*, 50(3):229–257, 2004.

[37] Benedetto Piccoli and Francesco Rossi. Generalized Wasserstein distance and its application to transport equations with source. *Archive for Rational Mechanics and Analysis*, 211(1):335–358, 2014.

[38] Michel Pierre. Global existence in reaction-diffusion systems with control of mass: a survey. *Milan J. Math.*, 78(2):417–455, 2010.

[39] Riccarda Rossi and Giuseppe Savaré. Tightness, integral equicontinuity and compactness for evolution problems in Banach spaces. *Ann. Sc. Norm. Super. Pisa Cl. Sci. (5)*, 2, 2003.

[40] Etienne Sandier and Sylvia Serfaty. Gamma-convergence of gradient flows with applications to ginzburg-landau. *Communications on Pure and Applied mathematics*, 57(12):1627–1672, 2004.

[41] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Progress in Nonlinear Differential Equations and Their Applications 87. Birkasauser Verlag, Basel, 2015.

[42] Juan Luis Vázquez. *The porous medium equation: mathematical theory*. Oxford University Press, 2007.

[43] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[44] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[45] Jonathan Zinsl. Geodesically convex energies and confinement of solutions for a multicomponent system of nonlocal interaction equations. Technical report, 2014.

3.3 Camassa-Holm

Articles:

- **The Camassa-Holm equation as an incompressible Euler equation: a geometric point of view.** *Journal of Differential Equations, Volume 264, Issue 7, Pages 4199-4234.* (2018) https://arxiv.org/abs/1609.04006. Gallouët T.O. and Vialard F.X.
- **Generalized compressible flows and solutions of the H(div) geodesic problem.** *Archive for Rational Mechanics and Analysis, Springer Verlag* (2020) https://hal.science/hal-01815531v3. Gallouët T.O., Natale A. et Vialard F.X.

Collaborators: These works are based on a collaboration with F.X. Vialard. The second paper was done also with Andrea Natale. At this moment he was a postdoc in our Inria team and under our shared supervision.

Main contributions:

- In the first paper we explain that the Camassa-Holm equation is for the Unbalanced Optimal Transport what the Incompressible Euler equation is for Optimal Transport.
- In particular we proved that Camassa-Holm equation is a particular solutions of the Incompressible Euler equation for a singular reference measure.
- In the second paper we used the previous observation in order to built generalized solutions for the Camassa-Holm geodesic problem.
- We proved the existence of such solutions and uniqueness of a pressure term. This quantity seems important and new.
- We proved that the relaxation is tight: for short time classical solutions of Camassa-Holm are solutions of the Camassa-Holm geodesic problem.

Research directions: One direction to pursue this work is to take advantage of this formulation in order to built numerical scheme that approximate solutions of the Camassa-Holm equation in the sprite of what we have done in Section 2.3. Another interesting question is to understand if the unique pressure term that we highlighted can be used in more classical PDE approach for this equation.

THE CAMASSA-HOLM EQUATION AS AN INCOMPRESSIBLE EULER EQUATION: A GEOMETRIC POINT OF VIEW

THOMAS GALLOUËT AND FRANÇOIS-XAVIER VIALARD

Abstract. The group of diffeomorphisms of a compact manifold endowed with the L^2 metric acting on the space of probability densities gives a unifying framework for the incompressible Euler equation and the theory of optimal mass transport. Recently, several authors have extended optimal transport to the space of positive Radon measures where the Wasserstein-Fisher-Rao distance is a natural extension of the classical L^2-Wasserstein distance. In this paper, we show a similar relation between this unbalanced optimal transport problem and the H^{div} right-invariant metric on the group of diffeomorphisms, which corresponds to the Camassa-Holm (CH) equation in one dimension. Geometrically, we present an isometric embedding of the group of diffeomorphisms endowed with this right-invariant metric in the automorphisms group of the fiber bundle of half densities endowed with an L^2 type of cone metric. This leads to a new formulation of the (generalized) CH equation as a geodesic equation on an isotropy subgroup of this automorphisms group; On S_1, solutions to the standard CH thus give radially 1-homogeneous solutions of the incompressible Euler equation on $\mathbb{R}^2$ which preserves a radial density that has a singularity at 0. An other application consists in proving that smooth solutions of the Euler-Arnold equation for the H^{div} right-invariant metric are length minimizing geodesics for sufficiently short times.

1. Introduction

In his seminal article [2], Arnold showed that the incompressible Euler equation can be viewed as a geodesic flow on the group of volume preserving diffeomorphisms of a Riemannian manifold M. His formulation had an important impact in the mathematical literature and it has led to many different works. Among others, let us emphasize two different points of view which have proven to be successful.

The first one has been investigated by Ebin and Marsden in [20] where the authors have taken an intrinsic point of view on the group of diffeomorphisms as an infinite dimensional weak Riemannian manifold. Formulating the geodesic equation as an ordinary differential equation in a Hilbert manifold of Sobolev diffeomorphisms, they proved, among others, local well-posedness of the geodesic equation for smooth enough initial conditions. Since then, many fluid dynamic equations, including the Camassa-Holm equation, have been written as a geodesic flow on a group of diffeomorphisms endowed with a right-invariant metric or connection [37, 32, 48, 23, 31] and analytical properties have been derived in the spirit of [20]. Note in particular that all these works assume a strong ambient topology such as H^s for s high enough and the topology given by the Riemannian metric is generically weaker, typically L^2 in the case of incompressible Euler.

Another point of view, motivated by the variational interpretation of geodesics as minimizers of the action functional, was initiated by Brenier. He developed an extrinsic approach by considering the group of volume preserving diffeomorphisms as a Riemannian submanifold embedded in the space of maps $L^2(M, M)$ which is particularly simple when M is the Euclidean space or torus. In particular, his polar factorization theorem [5] was motivated by a numerical scheme approximating geodesics on the group of volume preserving diffeomorphisms. Optimal transport then appeared as a key tool to project a map onto this group by minimizing the L^2 distance and it can be interpreted as a non-linear extension of the pressure in the incompressible Euler equation. Since then, optimal transport has witnessed an impressive development and found many important applications inside and outside mathematics, see for instance the gigantic monograph of Villani [56]. Brenier also used

optimal transport in order to define the notion of generalized geodesics for the incompressible Euler equation in [6].

In this article, we develop Brenier's point of view for a generalization in any dimension of the Camassa-Holm equation. Indeed, we present an isometric embedding of the group of diffeomorphisms endowed with the right-invariant H^{div} metric into a space of maps endowed with an L^2 metric. Moreover, the recently introduced Wasserstein-Fisher-Rao distance [14, 13], a generalization of optimal transport to measures that do not have the same total mass, plays the role of the L^2 Wasserstein distance for the incompressible Euler equation.

1.1. **Contributions.** The underlying key point for our work is the generalization of the (formal) Riemannian submersion already presented in [13], which unifies the unbalanced optimal problem and the H^{div} right-invariant metric. We rewrite the geodesic flow of the right-invariant H^{div} metric on the diffeomorphism group as a geodesic equation on a constrained submanifold of a semidirect product of group or equivalently on the automorphism group of the half-densities fibre bundle endowed with the cone metric (see Section 2.3 for its definition). This point of view has three applications: (1) We interpret solutions to the Camassa-Holm equation and one of its generalization in higher dimension as particular solutions of the incompressible Euler equation on the plane for a radial density which has a singularity at 0. This correspondence can be introduced via a sort of Madelung transform. (2) We generalize a result of Khesin et al. in [32] by computing the curvature of the group as a Riemannian submanifold. (3) Generalizing a result of Brenier to the case of Riemannian manifolds, which states that solutions of the incompressible Euler equation are length minimizing geodesic for sufficiently short times, we prove similar results for the Camassa-Holm equation.

Since the interpretation of the Camassa-Holm equation as an incompressible Euler equation is one of the main results of the paper, we present it below.

Theorem 1 (Camassa-Holm as incompressible Euler)**.** *Solutions to the Camassa-Holm equation on* S_1

$$\partial_t u - \frac{1}{4}\partial_{txx} u + 3\partial_x u\, u - \frac{1}{2}\partial_{xx} u\, \partial_x u - \frac{1}{4}\partial_{xxx} u\, u = 0 \tag{1.1}$$

are mapped to solutions of the incompressible Euler equation on $\mathbb{R}^2 \setminus \{0\}$ *for the density* $\rho = \frac{1}{r^4}\,\mathrm{Leb}$*, that is*

$$\begin{cases} \dot{v} + \nabla_v v = -\nabla P\,, \\ \nabla \cdot (\rho v) = 0\,, \end{cases} \tag{1.2}$$

by the map $u \mapsto \big(u(\theta), \frac{r}{2}\partial_x u(\theta)\big)$.

In other words, rewriting the Camassa-Holm equation in polar coordinates transforms it into an incompressible Euler equation. Obviously, the proof of the theorem can be reduced to a simple calculation. In this paper, we show the geometrical structures that underpin this formulation.

1.2. **Link to previous works.** Recently, several authors including the second author extended optimal transport to the case of unbalanced measures, i.e. measures that do not have the same total mass. Although several works extended optimal transport to this setting, surprisingly enough, the equivalent of the L^2-Wasserstein distance in this unbalanced setting has been introduced in 2015 simultaneously by [14, 13] motivated by imaging applications, [39, 40] motivated by gradient flows as well as [36] and by [54] for optimal transport of contact structures. In this paper, we show that, in the case of the Wasserstein-Fisher-Rao metric, the equivalent to the incompressible Euler equation is a generalization of the Camassa-Holm equation, namely the Euler-Arnold equation for the right-invariant metric H^{div} on the group of diffeomorphisms. In one dimension, geodesics for the right-invariant H^{div} metric are the solutions to the Camassa-Holm equation introduced in [12]. Since its introduction, the Camassa-Holm equation has attracted a lot of attention since it is a bi-Hamiltonian system as well as an integrable system, it exhibits peakon solutions and it is a model for waves in shallow water [18, 16, 38, 17, 9, 19, 30]. In particular, this equation is known for its well understood blow-up in finite time and is a model for wave breaking [44].

Although the title of [10], which refers to optimal transport and the Camassa-Holm equation, is seemingly close to our article, the authors introduce a metric based on optimal transport which gives Lipschitz estimates for the solutions of the Camassa-Holm equation and it is a priori completely different to our construction. Indeed, in our article, the optimal transport metric measures the discrepancy of not being in the stabilizer of the group action defined in Section 2.4 where the solutions of the Camassa-Holm equation lie.

Maybe more related to our results, homogeneous solutions of Euler equations have been studied for example in [21, 42], however the measure preserved in those works is not a singular measure, as in our work.

1.3. **Plan of the paper.** In Section 2, we recall the link between optimal transport and the incompressible Euler equation, then we introduce the Wasserstein-Fisher-Rao metric which generalizes the L^2 Wasserstein metric on the space of *probability* densities to the space of *integrable* densities, thus relaxing the mass constraint. We present the generalization of Otto's Riemannian submersion to this unbalanced case. This generalization uses a semidirect product of group which can be interestingly interpreted as the automorphism group of the principal fibre bundle of half-densities, as explained in Section 2.4. This semidirect product of group has a natural left action on the space of densities and it gives the Riemannian submersion between an L^2 type of metric on the group and the Wasserstein-Fisher-Rao metric on the space of densities.

In Section 3, we briefly review the result on the local well-posedness of the Camassa-Holm equation and its H^{div} generalization and the associated metric properties.

Section 4 presents the corresponding submanifold point of view corresponding to the Camassa-Holm equation (its generalization). The submanifold is the isotropy subgroup of the left action of the semidirect product of group and the ambient metric is the L^2 type of metric. As a direct consequence, it gives a generalization of a result on the sectional curvature written in [32, Theorem A.2].

The two main applications of our approach are detailed in Section 5. The one dimensional case is developed in section 5.1 where we show that solutions of the Camassa-Holm equation (its generalization) can be seen as particular solutions of an incompressible Euler equation for a particular density on the cone which has a singularity at 0. We improve a result of Ebin and Marsden in dimension 1 by extending Brenier's approach to show that every smooth geodesics are length minimizing on a sufficiently short time interval under mild conditions. Then, these result are generalized in 5.2.

1.4. **Notations.** Hereafter is a non exhaustive list of notations used throughout the paper.

- (M, g) is a smooth orientable Riemannian manifold which is assumed compact and without boundary. Its volume form is denoted by vol, TM and T^*M denote respectively the tangent and the cotangent bundle.
- The distance on (M, g) is sometimes denoted by d_M when a confusion might occur.
- For $x \in M$, the squared norm of a vector $v \in T_xM$ will be denoted by $\|v\|^2$ or $g(x)(v, v)$.
- For $x \in M$, we denote by $\exp_x^M : T_xM \to M$, the exponential map, the superscript being a reminder of the underlying manifold.
- $\mathcal{C}(M)$ is the Riemannian cone over (M, g) and is introduced in Definition 2.
- The operator div is the divergence w.r.t. the volume form on (M, g).
- The Lie bracket between two vector fields X, Y on M is denoted by $[X, Y]$.
- If $f \in C^1(M, \mathbb{R})$, then ∇f is the gradient of f w.r.t. the metric g. Sometimes, we use the notation ∇_x to make clear which variable we consider.
- The group of invertible linear maps on $\mathbb{R}^d$ is denoted by $\mathrm{GL}_d(\mathbb{R})$.
- For a quantity $f(t, x)$ that depends on time and space variable, we denote by $\dot{f}$ its time derivative.
- On $\mathbb{R}$ and $\mathbb{C}$, $|\cdot|$ denotes respectively the absolute value and the module.
- $M = S_n(r)$ the Euclidean sphere of radius r in $\mathbb{R}^{n+1}$.
- The Lebesgue measure is denoted by Leb.
- Sometimes, we use the notation $a \overset{\text{def.}}{=} b$ to define a as b.

2. A Geometric Point of View on Unbalanced Optimal Transport

Before presenting unbalanced optimal transport in more details, we give a brief overview of the link between optimal transport and the incompressible Euler equation.

2.1. Optimal transport and the incompressible Euler equation.

We first start from the usual static formulation of optimal transport and then present the dynamical formulation proposed by Benamou and Brenier. The link between the two formulations can be introduced via Otto's Riemannian submersion, which also provides a clear connection between incompressible Euler equation and the dynamical formulation of optimal transport. Our presentation closely follows the discussion in [34, Appendix A.5] and interesting complements can be found in [50, 32, 33]. In the rest of the section, unless otherwise mentioned, M denotes a smooth Riemannian manifold without boundary, for instance the flat torus.

Static formulation of optimal mass transport: The optimal mass transport problem as introduced by Monge in 1781 consists in finding, between two given probability measures ν_1 and ν_2, a map φ such that $\varphi_*\nu_1 = \nu_2$, i.e. the image measure of ν_1 by φ is equal to ν_2 and which minimizes a cost given by

$$\int_M c(x, \varphi(x))\, \mathrm{d}\nu_1(x)\,, \tag{2.1}$$

where c is a positive function that represents the cost of moving a particule of unit mass from location x to location y. This problem is ill-posed in the sense that solutions may not exist and the Kantorovich formulation of the problem is the correct relaxation of the Monge formulation, which can be presented as follows: On the space of probability measures on the product space $M \times M$, denoted by $\mathcal{P}(M \times M)$, find a minimizer to

$$\mathcal{I}(m) = \int_{M^2} c(x, y)\, \mathrm{d}m(x, y) \text{ such that } p^1_*(m) = \nu_1 \text{ and } p^2_*(m) = \nu_2\,, \tag{2.2}$$

where $p^1_*(m), p^2_*(m)$ denote respectively the image measure of $m \in \mathcal{P}(M \times M)$ under the projections on the first and second factors on $M \times M$. Most often in the litterature, the cost c is chosen as a power of a distance. From now on, we will only discuss the case $c(x, y) = d(x, y)^2$ where d is the distance associated with a Riemannian metric on M. In this case, the Kantorovich minimization problem defines the so-called L^2-Wasserstein distance on the space of probability measures. The Monge formulation can be expressed as a minimization problem as follows

$$W_2(\mu, \nu)^2 \overset{\text{def.}}{=} \inf_{\varphi \in \mathrm{Diff}(M)} \left\{ \int_M d(\varphi(x), x)^2 \, \mathrm{d}\nu_1(x) \,:\, \varphi_*\nu_1 = \nu_2 \right\}, \tag{2.3}$$

where $\mathrm{Diff}(M)$ denotes the group of smooth diffeomorphisms of M.

Dynamic formulation: In [3], Benamou and Brenier introduced a dynamical version of optimal transport which was inspired and motivated by the study of the incompressible Euler equation. Let $\rho_0, \rho_1 \in C^\infty(M, \mathbb{R}_+)$ be integrable densities, note that all the quantities will be implicitly time dependent. The dynamic formulation of the Wasserstein distance consists in minimizing

$$\mathcal{E}(v) = \int_0^1 \int_M \|v(t, x)\|^2 \rho(t, x)\, \mathrm{dvol}(x)\, \mathrm{d}t\ , \tag{2.4}$$

subject to the constraints $\dot\rho + \mathrm{div}(v\rho) = 0$ and initial condition $\rho(0) = \rho_0$ and final condition $\rho(1) = \rho_1$. The notation $\|\cdot\|$ stands for the Euclidean norm.

Equivalently, following [3], a convex reformulation using the momentum $\mathsf{m} = \rho v$ reads

$$\mathcal{E}(\mathsf{m}) = \int_0^1 \int_M \frac{\|\mathsf{m}(t, x)\|^2}{\rho(t, x)}\, \mathrm{dvol}(x)\, \mathrm{d}t\,, \tag{2.5}$$

subject to the constraints $\dot\rho + \mathrm{div}(\mathsf{m}) = 0$ and initial condition $\rho(0) = \rho_0$ and final condition $\rho(1) = \rho_1$. Let us underline that the functional $\mathcal{E}$ is convex in ρ, m and the continuity equation is linear in (ρ, m), therefore convex optimization methods can be applied for numerical purposes. Due to the continuity

equation, the problem is feasible if and only if the initial and final densities have the same total mass using Moser's lemma [51].

Otto's Riemannian submersion: The link between the static and dynamic formulations is made clear using Otto's Riemannian submersion [53] which emphasizes the idea of a group action on the space of probability densities. Let $\mathrm{Dens}_p(M)$ be the set of probability measures that have smooth positive densities with respect to the volume measure vol. We consider such a probability density denoted by ρ_0. Otto showed that the map

$$\begin{aligned}&\pi : \mathrm{Diff}(M) \to \mathrm{Dens}_p(M)\\ &\pi(\varphi) = \varphi_*(\rho_0)\end{aligned}$$

is a formal Riemannian submersion of the metric $L^2(\rho_0)$ on $\mathrm{Diff}(M)$ to the L^2-Wasserstein metric on $\mathrm{Dens}_p(M)$. For all the basic properties of Riemannian submersions, we refer the reader to [26]. The fiber of this Riemannian submersion at point $\rho_0 \equiv 1$ is the subgroup of diffeomorphisms preserving the volume measure vol, we denote it by $\mathrm{SDiff}(M)$ and we denote its tangent space at Id by $\mathrm{SVect}(M)$, the space of divergence free vector fields. The vertical space at a diffeomorphism $\varphi \in \mathrm{Diff}(M)$ for $\rho \overset{\text{def.}}{=} \varphi_*\rho_0$ is

$$\mathrm{Vert}_\varphi = \{v \circ \varphi\,;\, v \in \mathrm{Vect}(M) \text{ s.t. } \mathrm{div}(\rho v) = 0\}\,. \tag{2.6}$$

In particular, consider $\varphi \in \mathrm{SDiff}(M)$, the vertical space is $\mathrm{Vert}_\varphi = \{v \circ \varphi\,;\, v \in \mathrm{SVect}(M)\}$ and the horizontal space is

$$\mathrm{Hor}_\varphi = \{\nabla p \circ \varphi\,;\, p \in C^\infty(M, \mathbb{R})\}\,. \tag{2.7}$$

Incompressible Euler equation: On the fiber $\mathrm{SDiff}(M)$, the $L^2(\mathrm{vol})$ metric is right-invariant. In Arnold's seminal work [2], it is shown that the Euler-Lagrange equation associated with this metric is the incompressible Euler equation. Arnold derived this equation as a particular case of geodesic equations on a Lie group endowed with a right-invariant metric. In its Eulerian formulation, the incompressible Euler equation is, when $M = \mathbb{T}^d$ the flat torus for the Lebesgue measure,

$$\begin{cases}\partial_t v(t,x) + v(t,x)\cdot\nabla v(t,x) = -\nabla p(t,x), \quad t>0,\ x \in M\,,\\ \mathrm{div}(v) = 0\,,\\ v(0,x) = v_0(x)\,,\end{cases} \tag{2.8}$$

where $v_0 \in \mathrm{SVect}(M)$ is the initial condition and p is the pressure function. On a general Riemannian manifold (M, g) compact and without boundary, the formulation is similar, omitting the time and space variables, for the volume measure,

$$\begin{cases}\partial_t v + \nabla_v v = -\nabla p, \quad t>0,\ x \in M\,,\\ \mathrm{div}(v) = 0\,,\\ v(0,x) = v_0(x)\,,\end{cases} \tag{2.9}$$

where, in this case, the symbol ∇ denotes the Levi-Civita connection associated with the Riemannian metric on M and div denotes the divergence w.r.t. the volume measure. Another fruitful point of view consists in considering the group $\mathrm{SDiff}(M)$ as isometrically embedded in the group $\mathrm{Diff}(M)$ endowed with the $L^2(\mathrm{vol})$ (non right-invariant) metric. Therefore the geodesic equations are simply geodesic equations on the Riemannian submanifold $\mathrm{SDiff}(M)$ and the geodesic equations can be written as

$$\ddot{\phi} = -\nabla p \circ \phi\,, \tag{2.10}$$

where $\phi \in \mathrm{SDiff}(M)$ and p is still a pressure function. Using this Riemannian submanifold approach, Brenier was able to prove that solutions for which the Hessian of p is bounded in L^∞ are length minimizing for short times and several of his analytical results were derived from this formulation [4, 6].

Inviscid Burgers equation: The geodesic equation on the group of diffeomorphisms for the L^2 metric written in Eulerian coordinates is the compressible Burgers equation. Its formulation on $M = \mathbb{T}^d$ is

$$\partial_t u(t,x) + u(t,x) \cdot \nabla u(t,x) = 0\,, \tag{2.11}$$

or on a general Riemannian manifold

$$\partial_t u + \nabla_u u = 0\,. \tag{2.12}$$

This formulation is obviously related to the incompressible Euler equation where the pressure p can be interpreted as a Lagrange multiplier associated with the incompressibility constraint, which is not present in Burgers equation. Since the map π is a Riemannian submersion, geodesics on the space of densities can be lifted horizontally to geodesics on the group. These horizontal geodesics are potential solutions of the Burgers equation, if $u_0 = \nabla q_0$, i.e. u is a potential at the initial time, then u_t stays potential for all time (until it is not well defined any longer). The corresponding equation for the potential q is the Hamilton-Jacobi equation

$$\partial_t q(t,x) + \frac{1}{2}\|\nabla q(t,x)\|^2 = 0\,, \tag{2.13}$$

which, in this formulation, makes sense on a Riemannian manifold.

2.2. **The Wasserstein-Fisher-Rao metric, its dynamical formulation.** The continuity equation enforces the mass conservation property in the Benamou-Brenier formulation (2.4) (or (2.5) recalling that by definition $\mathsf{m} = \rho v$). This constraint can be relaxed by introducing a source term μ in the continuity equation,

$$\dot\rho = -\operatorname{div}(\rho v) + \mu \ = -\operatorname{div}(\mathsf{m}) + \mu\,. \tag{2.14}$$

For a given variation of the density $\dot\rho$, there exist a priori many couples (v,μ) that reproduce this variation. Following [55], it can be determined via the minimization of the norm of (v,μ), for a given choice of norm. The penalization of μ was chosen in [43] as the L^2 norm but a natural choice is rather the Fisher-Rao metric

$$\mathrm{FR}^2(\mu) = \int_M \frac{\mu(t,x)^2}{\rho(t,x)}\ \mathrm{dvol}(x)\,,$$

because it is homogeneous. In other words, this is the L^2 norm of the growth rate w.r.t. the density ρ since it can be written as $\int_M \alpha(t,x)^2\rho(t,x)\ \mathrm{dvol}(x)$ where α is the growth rate $\alpha(t,x) \overset{\text{def.}}{=} \frac{\mu(t,x)}{\rho(t,x)}$. Note in particular that this action is 1-homogeneous with respect to the couple (μ,ρ). This point is important for convex analysis properties and especially, in order to define the action functional on singular measures via the same formula. Obviously, there are many other choices of norms that satisfies this homogeneity property but this particular one can be related to the Camassa-Holm equation.

Thus, the Wasserstein-Fisher-Rao metric tensor denoted by WF_ρ is simply given by the infimal convolution, a standard tool in convex analysis, between the Wasserstein and the Fisher-Rao metric tensors. Indeed, the metric tensor at a density ρ is defined via the minimization

$$\mathrm{WF}_\rho(\dot\rho,\dot\rho) = \inf_{v,\alpha} \int_M \alpha(x)^2 + \|v(x)\|^2\ \mathrm{d}\rho(x)\ \text{ s.t. } \dot\rho = -\operatorname{div}(\rho v) + 2\alpha\rho\,. \tag{2.15}$$

The distance associated with this metric tensor has been named Wasserstein-Fisher-Rao [14], Hellinger-Kantorovich [39], Kantorovich-Fisher-Rao [28].

Definition 1 (WF metric)**.** Let (M,g) be a smooth Riemannian manifold compact and without boundary, $a,b \in \mathbb{R}_+^*$ be two positive real numbers and $\rho_0,\rho_1 \in \mathcal{M}_+(M)$ be two nonnegative Radon measures. The Wasserstein-Fisher-Rao metric is defined by

$$\mathrm{WF}^2(\rho_0,\rho_1) = \inf_{\rho,\mathsf{m},\mu} \mathcal{J}(\rho,\mathsf{m},\mu)\,, \tag{2.16}$$

where

$$\mathcal{J}(\rho,\mathsf{m},\mu)=a^2\int_0^1\int_M \frac{g^{-1}(x)(\tilde{\mathsf{m}}(t,x),\tilde{\mathsf{m}}(t,x))}{\tilde\rho(t,x)}\,\mathrm{d}\nu(t,x)\;+b^2\int_0^1\int_M\frac{\tilde\mu(t,x)^2}{\tilde\rho(t,x)}\,\mathrm{d}\nu(t,x) \tag{2.17}$$

over the set (ρ,m,μ) satisfying $\rho\in\mathcal{M}([0,1]\times M)$, $\mathsf{m}\in(\Gamma^0_M([0,1]\times M,TM))^*$ which denotes the dual of time dependent continuous vector fields on M (time dependent sections of the tangent bundle), $\mu\in\mathcal{M}([0,1]\times M)$ subject to the constraint

$$\int_0^1\int_M \partial_t f\,\mathrm{d}\rho+\int_0^1\int_M \mathsf{m}(\nabla_x f)-f\mu\,\mathrm{d}\nu=\int_M f(1,\cdot)\,\mathrm{d}\rho_1-\int_M f(0,\cdot)\,\mathrm{d}\rho_0 \tag{2.18}$$

satisfied for every test function $f\in C^1([0,1]\times M,\mathbb{R})$. Moreover, ν is chosen such that ρ,m,μ are absolutely continuous with respect to ν and $\tilde\rho,\tilde{\mathsf{m}},\tilde\mu$ denote their Radon-Nikodym derivative with respect to ν.

Remark 1. *Note that, in the previous definition, the divergence operator* $\mathrm{div}(\cdot)$ *is defined by duality on the space of* C^1 *functions. In addition, since the functions in the integrand of formula* (2.16) *are one homogeneous with respect to the triple of arguments* $(\tilde\rho,\tilde{\mathsf{m}},\tilde\mu)$*, the functional does not depend on the choice of* ν *which dominates the measures. Last, the Radon-Nikodym theorem applied to the measure* m *gives* $\mathsf{m}=\tilde{\mathsf{m}}\nu$ *where* $\tilde{\mathsf{m}}$ *is a measurable section of* T^*M*.*

This dynamical formulation enjoys most of the analytical properties of the initial Benamou-Brenier formulation (2.4) and especially convexity. Moreover, WF defines a distance on the space of nonnegative Radon measures which is continuous w.r.t. to the weak-* topology. An important consequence is the existence of optimal paths in the space of time-dependent measures [14] by application of the Fenchel-Rockafellar duality theorem. Note in particular that the Hamiltonian formulation of the geodesic flow can be formally derived as

$$\begin{cases}\partial_t\rho(t,x)+\mathrm{div}(\rho(t,x)\nabla_x q(t,x))-2q(t,x)\rho(t,x)=0\\ \partial_t q(t,x)+\|\nabla q(t,x)\|^2+q(t,x)^2=0\,,\end{cases}$$

where the second equation corresponds to the Hamilton-Jacobi equation (2.13). In fact, not only analytical properties of standard optimal transport are conserved but also some interesting geometrical properties such as the Riemannian submersion highlighted by Otto, as explained in the introduction. More precisely, the group of diffeomorphisms can be replaced by a semi-direct product of group between $\mathrm{Diff}(M)$ and the space $C^\infty(M,\mathbb{R}_+^*)$ which is a group under pointwise multiplication. In addition, this group acts on the space of densities $\mathrm{Dens}(M)$ and this action gives a Riemannian submersion between the group endowed with an L^2 type of metric, namely $L^2(M,\mathcal{C}(M))$ and the space of densities endowed with the Wasserstein-Fisher-Rao metric. The notation $\mathcal{C}(M)$ is the cone over M defined in the next section 2.3, it is the manifold $M\times\mathbb{R}_+^*$ endowed with the Riemannian metric given in Definition 2. Moreover, this semidirect product of groups is naturally identified as the automorphism group of the fibre bundle of half densities in section 2.4.

2.3. **A cone metric.** To motivate the introduction of the cone metric, let us first discuss informally what happens for a particle of mass $m(t)$ at a spatial position $x(t)$ in a Riemannian manifold (M,g) under the generalized continuity constraint (2.14); If the control variables $v(t,x)$ and $\alpha(t,x)$ are Lipschitz, then the solution of the continuity equation with initial data $m(0)\delta_{x(0)}$ has the form $m(t)\delta_{x(t)}$ where $m(t)\in\mathbb{R}_+^*$ is the mass of the Dirac measure and $x(t)\in M$ its location; The system reads

$$\begin{cases}\dot x(t)=v(t,x(t))\\ \dot m(t)=\alpha(t,x(t))m(t)\,,\end{cases} \tag{2.19}$$

which is directly obtained by duality since the flow map associated with (v,α) is well defined. This result would not hold if the vector field were not smooth enough, see [1]. Let us compute the action functional in the case where $\rho(t)=m(t)\delta_{x(t)}$. By the above result, (v,α) is completely determined at $(t,x(t))$ and it is sufficient to compute the action which reads $\int_0^1 a^2|v(x(t))|^2m(t)+b^2\frac{\dot m(t)^2}{m(t)}\,\mathrm{d}t$.

Thus, considering the particle as a point in $M \times \mathbb{R}_+^*$, the Riemannian metric seen by the particle is $a^2 mg + b^2 \frac{\mathrm{d}m^2}{m}$. Therefore, it will be of importance to study this Riemannian metric $M \times \mathbb{R}_+^*$. Actually, this space is isometric to the standard Riemannian cone defined below.

Definition 2 (Cone)**.** Let (M, g) be a Riemannian manifold. The cone over M denoted by $\mathcal{C}(M)$ is the quotient space $(M \times \mathbb{R}_+) \,/\, (M \times \{0\})$. The cone point $M \times \{0\}$ is denoted by $\mathcal{S}$. The cone will be endowed with the metric $g_{\mathcal{C}(M)} \overset{\text{def.}}{=} r^2 g + \mathrm{d}r^2$ defined on $M \times \mathbb{R}_+^*$ and r is the variable in $\mathbb{R}_+^*$.

The explicit formula for the distance on the Riemannian cone can be found in [11] and the isometry is given by the square root change of variable on the mass, as stated in the following proposition.

Proposition 1. *Let a, b be two positive real numbers and (M, g) be a Riemannian manifold. The distance on $(M \times \mathbb{R}_+^*, a^2 mg + \frac{b^2}{m}\,\mathrm{d}m^2)$ is given by*

$$d((x_1, m_1), (x_2, m_2))^2 = 4b^2 \left(m_2 + m_1 - 2\sqrt{m_1 m_2} \cos\left(\frac{a}{2b} d_M(x_1, x_2) \wedge \pi \right) \right) , \tag{2.20}$$

where the notation $\wedge$ stands for the minimum, that it $x \wedge y = \min(x, y)$ for $x, y \in \mathbb{R}$. The space $(M \times \mathbb{R}_+^*, mg + \frac{1}{4m}\,\mathrm{d}m^2)$ is isometric to $(\mathcal{C}(M), g_{\mathcal{C}(M)})$ by the change of variable $r = \sqrt{m}$. If c is a unit speed geodesic for the metric $\frac{a^2}{4b^2} g$, an isometry $S : \mathbb{C} \setminus \mathbb{R}_- \to M \times \mathbb{R}_+^*$ is defined by $S(re^{i\theta}) = (c(\theta), \frac{r^2}{4b^2})$.

In physical terms, it implies that mass can "appear" and "disappear" at finite cost. In other words, the Riemannian cone is not complete but adding the cone point, which represents $M \times \{0\}$, to $M \times \mathbb{R}_+^*$ turns it into a complete metric space when M is complete. Importantly, the distance associated with the cone metric (2.20) is 1-homogeneous in (m_1, m_2). In the rest of the paper, unless explicitly mentioned, we consider the case $a = 1$ and $b = 1/2$. We now collect known facts about Riemannian cones.

Proposition 2. *On the cone $\mathcal{C}(M)$, we denote by e the vector field defined by $\frac{\partial}{\partial r}$. The Levi-Civita connection on (M, g) will be denoted by ∇^g. For a given vector field X on M, define its lift as a vector field on $M \times \mathbb{R}_+^*$ by $\hat{X}(x, r) = (X(x), 0)$. The Levi-Civita connection on $\mathcal{C}(M)$ denoted by ∇ is given by*

$$\nabla_{\hat{X}} \hat{Y} = \widehat{\nabla^g_X Y} - rg(X, Y)e\,, \quad \nabla_e e = 0 \text{ and } \nabla_e \hat{X} = \nabla_{\hat{X}} e = \frac{1}{r}\hat{X}\,.$$

The curvature tensor R on the cone satisfies the following properties,

$$R(\hat{X}, e) = 0 \text{ and } R(\hat{X}, \hat{Y})\hat{Z} = (R_g(X, Y)Z - g(Y, Z)X + g(X, Z)Y, 0) \tag{2.21}$$

where R_g denotes the curvature tensor of (M, g). Let X, Y be two orthornormal vector fields on M,

$$K(\hat{X}, \hat{Y}) = K_g(X, Y) - 1 \tag{2.22}$$

where K and K_g denote respectively the sectional curvatures of $\mathcal{C}(M)$ and M.

Proof. Direct computations, see [25]. □

Let us give simple comments on Riemannian cones: Usual cones, embedded in $\mathbb{R}^3$ are cones over S_1 of length less than 2π. Although Riemannian cones over a segment in $\mathbb{R}$ are locally flat, the curvature still concentrates at the cone point. The cone over the sphere is isometric to the Euclidean space (minus the origin) and the cone over the Euclidean space has nonpositive curvature. In particular, the cone over S_1 is isometric to $\mathbb{R}^2 \setminus \{0\}$. We refer to [11] for more informations on cones from the point of view of metric geometry.

We need the explicit formulas for the geodesic equations on the cone.

Corollary 3. *The geodesic equations on the cone $\mathcal{C}(M)$ are given by*

$$\frac{D}{Dt}^g \dot{x} + 2\frac{\dot{r}}{r}\dot{x} = 0 \tag{2.23a}$$

$$\ddot{r} - rg(\dot{x}, \dot{x}) = 0\,, \tag{2.23b}$$

where $\frac{D}{Dt}^g$ *is the covariant derivative associated with* (M, g).
Alternatively, the geodesic equations on $(M \times \mathbb{R}_+^*, a^2 mg + \frac{b^2}{m}\mathrm{d}m^2)$ *can be written w.r.t. the initial "mass" coordinate as follows*

$$\frac{D}{Dt}^g \dot{x} + \frac{\dot{m}}{m}\dot{x} = 0 \tag{2.24a}$$

$$\ddot{m} - \frac{\dot{m}^2}{2m} - \frac{a^2}{2b^2} g(\dot{x}, \dot{x}) m = 0\,. \tag{2.24b}$$

Note that we used the isometry given in Proposition 1 to derive the equations and in particular, we implicitly used the equality $4b^2 m = r^2$. Since we have written the geodesic equations on the usual cone in polar coordinates, we used the square root of the "mass" coordinate, therefore we need to introduce below the space of square roots of densities to discuss the infinite dimensional setting.

2.4. **The automorphism group of the bundle of half-densities.** The cone can be seen as a trivial principal fibre bundle since $\mathcal{C}(M)$ is the direct product of M with the group $\mathbb{R}_+^*$. Let us denote $p_M : \mathcal{C}(M) \mapsto M$ the projection on the first factor. The group $\mathbb{R}_+^*$ induces a group action on $\mathcal{C}(M)$ defined by $\lambda \cdot (x, \lambda') \stackrel{\text{def.}}{=} (x, \lambda\lambda')$, for all $x \in M$ and $\lambda, \lambda' \in \mathbb{R}_+^*$. We now identify the trivial fibre bundle of half densities with the cone.

Definition 3. Let M be a smooth manifold without boundary and (U_α, u_α) be a smooth atlas. The bundle of s-densities is the line bundle given by the following cocycle

$$\Psi_{\alpha\beta} : U_\alpha \cap U_\beta \mapsto \mathrm{GL}_1(\mathbb{R}) = \mathbb{R}^*$$
$$\Psi_{\alpha\beta}(x) = |\det(\,\mathrm{d}(u_\beta \circ u_\alpha^{-1})(u_\alpha(x))|^s = \frac{1}{|\det(d(u_\alpha \circ u_\beta^{-1}))(u_\beta(x))|^s}\,.$$

We denote by $\mathrm{Dens}_s(M)$ the set of sections of this bundle and we use $\mathrm{Dens}(M)$ instead of $\mathrm{Dens}_1(M)$, the space of densities. This definition shows that this fibre bundle is also a principal fibre bundle over $\mathbb{R}_+^*$ and it will be the point of view adopted in the rest of the paper.

On any smooth manifold M, the fibre bundle of s-densities is a trivial principal bundle over $\mathbb{R}_+^*$ since there exists a smooth positive density on M. Note that this trivialization depends on the choice of this reference positive density. If one chooses such a positive density, then the 1/2-density bundle can be identified to the cone $\mathcal{C}(M)$. *Let us fix the reference volume form to be the volume measure* vol. *By this choice, we identify* $\mathrm{Dens}_{1/2}(M)$ *with the set of sections of the cone* $\mathcal{C}(M)$ *in the rest of the paper.* Thus every element of $\mathrm{Dens}_{1/2}(M)$ is a section of the cone $\mathcal{C}(M)$. We are now interested in transformations that preserve the group structure. Namely, one can define

$$\mathrm{Aut}(\mathcal{C}(M)) = \left\{\Phi \in \mathrm{Diff}(\mathcal{C}(M))\,;\, \Phi(x, r) = r \cdot \Phi(x, 1) \text{ for all } r \in \mathbb{R}_+^*\right\}\,, \tag{2.25}$$

which is the instantiation, in this particular case, of the definition of the automorphisms group of a principal fibre bundle. In other words, this is the subgroup of diffeomorphisms of the cone that preserve the group action on the fibers. In particular, $\mathrm{Aut}(\mathcal{C}(M))$ is a subgroup of $\mathrm{Diff}(\mathcal{C}(M))$. Of particular interest is the subgroup of $\mathrm{Aut}(\mathcal{C}(M))$ which is defined as

$$\mathrm{Gau}(\mathcal{C}(M)) = \{\Phi \in \mathrm{Aut}(\mathcal{C}(M))\,;\, p_M \circ \Phi = \mathrm{id}_M\}\,. \tag{2.26}$$

The set $\mathrm{Gau}(\mathcal{C}(M))$ is called the gauge group and it is a normal subgroup of $\mathrm{Aut}(\mathcal{C}(M))$. We now consider the injection $Inj : \mathrm{Diff}(M) \hookrightarrow \mathrm{Aut}(\mathcal{C}(M))$ defined by $Inj(\varphi) = (\varphi, \mathrm{id}_{\mathbb{R}_+^*})$. This is the standard situation of a semidirect product of groups between $i(\mathrm{Diff}(M))$ and $\mathrm{Gau}(\mathcal{C}(M))$ since the following sequence is exact

$$\mathrm{Gau}(\mathcal{C}(M)) \hookrightarrow \mathrm{Aut}(\mathcal{C}(M)) \to \mathrm{Diff}(M)\,, \tag{2.27}$$

where Inj defined above provides an associated section of the short exact sequence and the projection from $\mathrm{Aut}(\mathcal{C}(M))$ onto $\mathrm{Diff}(M)$ is given by $\Phi \mapsto p_M \circ \Phi(x, 1)$. Note that we could also have chosen the natural section associated to the natural bundle of half-densities. As is well-known for a trivial principal bundle, $\mathrm{Aut}(\mathcal{C}(M))$ is therefore equal to the semidirect product of group:

$$\mathrm{Aut}(\mathcal{C}(M)) = \mathrm{Diff}(M) \ltimes_\Psi \mathrm{Gau}(\mathcal{C}(M))\,, \tag{2.28}$$

where $\Psi : \mathrm{Diff}(M) \mapsto \mathrm{Aut}(\mathrm{Gau}(\mathcal{C}(M)))$ is defined by $\Psi(\varphi)(\lambda) = \varphi^{-1}\lambda\varphi$ being the associated inner automorphism of the group $\mathrm{Gau}(\mathcal{C}(M))$, where the composition is understood as composition of diffeomorphisms of $\mathcal{C}(M)$. Being a trivial principal fibre bundle, the gauge group can be identified with the space of positive functions on M. Let us denote $\Lambda_{1/2}(M) \stackrel{\text{def.}}{=} C^\infty(M, \mathbb{R}_+^*)$ which is a group under pointwise multiplication. The subscript $1/2$ is a reminder of the fact that $\Lambda_{1/2}(M)$ is the gauge group of $\mathcal{C}(M)$, the bundle of $1/2$-densities. Note that we do not use the standard left action but, instead, a right action for the inner automorphisms as presented in [35, Section 5.3], which fits better to our notations, although these two choices are equivalent. The identification of $\Lambda_{1/2}$ with the gauge group $\mathrm{Gau}(\mathcal{C}(M))$ is simply $\lambda \mapsto (\mathrm{id}_M, \lambda)$ where $(\mathrm{id}_M, \lambda) : (x, m) \mapsto (x, \lambda(x)m)$. The group composition law is given by

$$(\varphi_1, \lambda_1) \cdot (\varphi_2, \lambda_2) = (\varphi_1 \circ \varphi_2, (\lambda_1 \circ \varphi_2)\lambda_2) \tag{2.29}$$

and the inverse is

$$(\varphi, \lambda)^{-1} = (\varphi^{-1}, \lambda^{-1} \circ \varphi^{-1})\,. \tag{2.30}$$

By construction, the group $\mathrm{Aut}(\mathcal{C}(M))$ has a left action on the space $\mathrm{Dens}_{1/2}(M)$ as well as on $\mathrm{Dens}(M)$. The action on $\mathrm{Dens}(M)$ is explicitly defined by the map π defined by

$$\begin{aligned} &\pi : \big(\mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)\big) \times \mathrm{Dens}(M) \mapsto \mathrm{Dens}(M) \\ &\pi\left((\varphi, \lambda), \rho\right) \stackrel{\text{def.}}{=} \varphi_*(\lambda^2 \rho)\,. \end{aligned} \tag{2.31}$$

For particular choices of metrics, this left action is a Riemannian submersion as detailed below. Note that we will use both automorphism group and semidirect product notations equally, depending on the context.

2.5. **A Riemannian submersion between the automorphism group and the space of densities.** The semidirect product of group $\mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)$ will be endowed with the metric $L^2(M, \mathcal{C}(M))$ with respect to the reference measure on M. Let us recall it hereafter.

Definition 4 (L^2 metric)**.** Let M be a manifold endowed with a measure μ and (N, g) be a Riemannian manifold. Consider a measurable map $\varphi : M \to N$ and two measurable maps, $X, Y : M \mapsto TN$ such that $p_N \circ X = p_N \circ Y = \varphi$ where $p_N : TN \to N$ is the natural projection. Then, the L^2 Riemannian metric w.r.t. to the volume form μ and the metric g at point φ is defined by

$$\langle X, Y\rangle_\varphi = \int_M g(\varphi(x))(X(\varphi(x)), Y(\varphi(x)))\ \mathrm{d}\mu(x)\,. \tag{2.32}$$

This is probably the simplest type of (weak) Riemannian metrics on spaces of mappings and it has been studied in details in [20] in the case $L^2(M, M)$ and also in [24] where, in particular, the curvature is computed for $L^2(M, N)$ for N an other Riemannian manifold. Note in particular that this metric is *not* the right-invariant metric L^2 on the semidirect product of groups as in [31] or on the automorphism group which would lead to an EPDiff equation on a principal fibre bundle as developed in [29].

Proposition 4. *The geodesic equations on* $\mathrm{Aut}(\mathcal{C}(M))$ *endowed with the metric* $L^2(M, \mathcal{C}(M))$ *with respect to the reference measure on* ν *are given by the geodesic equations on the cone* (2.23)*, that is* $\frac{D}{Dt}(\dot\varphi, \dot\lambda) = 0$*, or more explicitely*

$$\frac{D}{Dt}^{g}\dot\varphi + 2\frac{\dot\lambda}{\lambda}\dot\varphi = 0 \tag{2.33a}$$

$$\ddot\lambda - \lambda g(\dot\varphi, \dot\varphi) = 0\,. \tag{2.33b}$$

Proof. This is a consequence of [24] or a direct adaptation of [20, Theorem 9.1] to the case $L^2(M, \mathcal{C}(M))$ and Corollary 3. □

We now state a crucial fact that arises from an elementary observation.

Proposition 5. *The automorphism group* $\mathrm{Aut}(\mathcal{C}(M))$ *is totally geodesic in* $\mathrm{Diff}(\mathcal{C}(M))$ *for the* $L^2(\mathcal{C}(M),\mathcal{C}(M))$ *metric.*

Proof. Note that the first equation (2.33a) is 0-homogeneous with respect to λ and the second equation (2.33b) is one homogeneous with respect to λ. This is a consequence of the fact that multiplication by positive reals acts as an affine isometry on $\mathcal{C}(M)$. Therefore, the path $\Phi(t) : (x,r) \mapsto (\varphi(t)(x), \lambda(t)r)$ also satisfies the geodesic equation in $\mathrm{Diff}(\mathcal{C}(M))$ for the $L^2(\mathcal{C}(M),\mathcal{C}(M))$ metric. □

Note that this property does not depend on the measure on $\mathcal{C}(M)$ used in the definition of the space $L^2(\mathcal{C}(M),\mathcal{C}(M))$.

Let us first recall some useful notions. From the point of view of fluid dynamics, the next definition corresponds to the change of variable between Lagrangian and Eulerian formulations.

Definition 5 (Right-trivialization)**.** Let H be a group and a smooth manifold at the same time, possibly of infinite dimensions, the right-trivialization of TH is the bundle isomorphism $\tau : TH \mapsto H \times T_{\mathrm{Id}}H$ defined by $\tau(h, X_h) \overset{\text{def.}}{=} (h, dR_{h^{-1}}X_h)$, where X_h is a tangent vector at point h and $\mathcal{R}_{h^{-1}} : H \to H$ is the right multiplication by h^{-1}, namely, $R_{h^{-1}}(f) = fh^{-1}$ for all $f \in H$.

In fluid dynamics, the right-trivialized tangent vector $dR_{h^{-1}}X_h$ corresponds to the spatial or Eulerian velocity and X_h is the Lagrangian velocity. Importantly, this right-trivialization map is continuous but not differentiable with respect to the variable h. Indeed, right-multiplication R_h is smooth, yet left multiplication is continuous and usually not differentiable, due to a loss of smoothness.

Example 6. *For the semi-direct product of groups defined above, we have*

$$\tau((\varphi,\lambda),(X_\varphi,X_\lambda)) = ((\varphi,\lambda),(X_\varphi\circ\varphi^{-1},(X_\lambda\lambda^{-1})\circ\varphi^{-1}))\,. \tag{2.34}$$

We will denote by (v,α) *an element of the tangent space of* $T_{(\mathrm{Id},1)}\,\mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)$.

As an immediate consequence of Proposition 4, we write the geodesic equation in Eulerian coordinates.

Corollary 7 (Geodesic equations in Eulerian coordinates)**.** *After right-trivialization, that is under the change of variable* $v \overset{\text{def.}}{=} \dot\varphi\circ\varphi^{-1}$ *and* $\alpha \overset{\text{def.}}{=} \frac{\dot\lambda}{\lambda}\circ\varphi^{-1}$*, the geodesic equations read*

$$\begin{cases} \dot v + \nabla_v v + 2\alpha v = 0 \\ \dot\alpha + \langle\nabla\alpha, v\rangle + \alpha^2 - g(v,v) = 0\,. \end{cases} \tag{2.35}$$

Recall now the infinitesimal action associated with a group action.

Definition 6 (Infinitesimal action)**.** For a smooth left action of H a Lie group on a manifold M and $q \in M$, the infinitesimal action is the map $T_{\mathrm{Id}}H \times M \mapsto TM$ defined by

$$\xi \cdot q \overset{\text{def.}}{=} \left.\frac{\mathrm{d}}{\mathrm{d}t}\right|_{t=0} (\exp(\xi t)\cdot q) \in T_qM \tag{2.36}$$

where $\cdot$ denotes the left action of H on M and $\exp(\xi t)$ is the Lie exponential, that is the solution to $\dot h = dR_h(\xi)$ and $h(0) = \mathrm{Id}$.

Example 8. *For* $\mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)$ *acting on* $\mathrm{Dens}(M)$*, the previous definition gives* $(v,\alpha)\cdot\rho = -\operatorname{div}(v\rho) + 2\alpha\rho$*. Indeed, one has*

$$(\varphi(t),\lambda(t))\cdot\rho = \mathrm{Jac}(\varphi(t)^{-1})(\lambda^2(t)\rho)\circ\varphi^{-1}(t)\,.$$

First recall that $\partial_t\varphi(t) = v\circ\varphi(t)$ *and* $\partial_t\lambda = \lambda(t)\alpha\circ\varphi(t)$*. Once evaluated at time* $t = 0$ *where* $\varphi(0) = \mathrm{Id}$ *and* $\lambda(0) = 1$*, the differentiation with respect to* φ *gives* $-\operatorname{div}(v\rho)$ *and the second term* $2\alpha\rho$ *is given by the differentiation with respect to* λ*.*

We now recall the result of [46, Claim of Section 29.21] in a finite dimensional setting. This result presents a standard construction to obtain Riemannian submersions from a transitive group action.

Proposition 9. *Consider a smooth left action of Lie group H on a manifold M which is transitive and such that for every $\rho \in M$, the infinitesimal action $\xi \mapsto \xi \cdot \rho$ is a surjective map. Let $\rho_0 \in M$ and a Riemannian metric G on H that can be written as:*

$$G(h)(X_h, X_h) = g(h \cdot \rho_0)(dR_{h^{-1}} X_h, dR_{h^{-1}} X_h) \tag{2.37}$$

for $g(h \cdot \rho_0)$ an inner product on $T_{\mathrm{Id}}H$. Let $X_\rho \in T_\rho M$ be a tangent vector at point $h \cdot \rho_0 = \rho \in M$, we define the Riemannian metric $\overline{g}$ on M by

$$\overline{g}(\rho)(X_\rho, X_\rho) \overset{\text{def.}}{=} \min_{\xi \in T_{\mathrm{Id}}H} g(\rho)(\xi, \xi) \text{ under the constraint } X_\rho = \xi \cdot \rho\,. \tag{2.38}$$

where $\xi = X_h \cdot h^{-1}$.

Then, the map $\pi_0 : H \to M$ defined by $\pi_0(h) = h \cdot \rho_0$ is a Riemannian submersion of the metric G on H to the metric $\overline{g}$ on M. Moreover a minimizer ξ in formula (2.38) is called an horizontal lift of X_ρ at Id.*)*

The formal application of this construction in our infinite dimensional situation leads to the result, stated in [13]:

Proposition 10. *Let $\rho_0 \in \mathrm{Dens}(M)$ and define the map*

$$\pi_0 : \mathrm{Aut}(\mathcal{C}(M)) \to \mathrm{Dens}(M)$$
$$\pi_0(\varphi, \lambda) = \varphi_*(\lambda^2 \rho_0)\,.$$

Then, the map π_0 is a Riemannian submersion of the metric $L^2(M, \mathcal{C}(M))$ on the group $\mathrm{Aut}(\mathcal{C}(M))$ to the Wasserstein-Fisher-Rao on the space of densities $\mathrm{Dens}(M)$.

The horizontal space and vertical space at $(\varphi, \lambda) \in \mathrm{Aut}(\mathcal{C}(M)) = \mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)$ such that $\varphi_*(\lambda^2 \rho_0) = \rho$ are then defined by,

$$\mathrm{Vert}_{(\varphi,\lambda)} = \{(v, \alpha) \circ (\varphi, \lambda)\,;\, (v, \alpha) \in \mathrm{Vect}(M) \times C^\infty(M, \mathbb{R}) \text{ s.t. } \mathrm{div}(\rho v) = 2\alpha\rho\}\,, \tag{2.39}$$

and

$$\mathrm{Hor}_{(\varphi,\lambda)} = \left\{ \left(\frac{1}{2} \nabla p, p \right) \circ (\varphi, \lambda)\,;\, p \in C^\infty(M, \mathbb{R}) \right\}\,. \tag{2.40}$$

Note that the minimization in (2.38) is taken on an affine space of direction the vertical space whereas the minimizer is an element of the horizontal space.

Note also that the fibers of the submersion are right-cosets of the subgroup H_0 in H. The proof of the previous proposition is in fact given by the change of variables associated with right-trivialization. Let ρ_0 be a reference density, the application of Proposition 9 gives

(2.41)

$$G(\varphi, \lambda)((X_\varphi, X_\lambda),(X_\varphi, X_\lambda)) = \int_M g(v, v)\rho \,\mathrm{d}x + \int_M \alpha^2 \rho \,\mathrm{d}x$$

$$= \int_M g(X_\varphi \circ \varphi^{-1}, X_\varphi \circ \varphi^{-1})\varphi_*(\lambda^2\rho_0)\mathrm{d}x + \int_M (X_\lambda \lambda^{-1})^2 \circ \varphi^{-1} \varphi_*(\lambda^2 \rho_0)\mathrm{d}x \tag{2.42}$$

$$= \int_M g(X_\varphi, X_\varphi)\lambda^2 \rho_0 \,\mathrm{d}x + \int_M X_\lambda^2 \,\rho_0 \,\mathrm{d}x\,. \tag{2.43}$$

Therefore, the metric G is the $L^2(M, \mathcal{C}(M))$ metric with respect to the density ρ_0. Moreover, in this particular situation, the horizontal lift is a minimizer of (2.38).

Proposition 11 (Horizontal lift)**.** *Let $\rho \in \mathrm{Dens}^s(\Omega)$ be a smooth density and $X_\rho \in H^s(\Omega, \mathbb{R})$ be a tangent vector at the density ρ. The horizontal lift at $(\mathrm{Id}, 1)$ of X_ρ is given by $(\frac{1}{2}\nabla\Phi, \Phi)$ where Φ is the solution to the elliptic partial differential equation:*

$$-\frac{1}{2}\,\mathrm{div}(\rho \nabla \Phi) + 2\Phi\rho = X_\rho\,. \tag{2.44}$$

By elliptic regularity, the unique solution Φ belongs to $H^{s+1}(M)$.

To prove Proposition 11, remark that equation (2.44) is the first order condition of the minimization problem (2.38) where the term X_ρ reads in this case $X_\rho = \xi \cdot \rho = (v, \alpha) \cdot \rho = -\operatorname{div}(\rho v) + 2\alpha\rho$.

A direct application of this Riemannian submersion viewpoint is the formal computation of the sectional curvature of the Wasserstein-Fisher-Rao in this smooth setting by applying O'Neill's formula see [26]. To recall it hereafter, we need the Lie bracket of right-invariant vector fields on $\mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)$.

Proposition 12. *Let (v_1, α_1) and (v_2, α_2) be two tangent vectors at identity in $\mathrm{Diff}(M) \ltimes_\Psi \Lambda_{1/2}(M)$. Then,*

$$[(v_1, \alpha_1), (v_2, \alpha_2)] = ([v_1, v_2], \nabla\alpha_1 \cdot v_2 - \nabla\alpha_2 \cdot v_1) \ , \tag{2.45}$$

where $[v_1, v_2]$ denotes the Lie bracket of vector fields.

Note that the application of this formula to horizontal vector fields gives $[(\frac{1}{2}\nabla\Phi_1, \Phi_1), (\frac{1}{2}\nabla\Phi_2, \Phi_2)] = (\frac{1}{4}[\nabla\Phi_1, \nabla\Phi_2], 0)$.

Proposition 13. *Let ρ be a smooth positive density on M and X_1, X_2 be two orthonormal tangent vectors at ρ and $\xi_{\Phi_1}, \xi_{\Phi_2}$ be their corresponding right-invariant horizontal lifts on the group. If O'Neill's formula can be applied, the sectional curvature of $\mathrm{Dens}(M)$ at point ρ is given by,*

$$K(\rho)(X_1, X_2) = \int_\Omega k(x,1)(\xi_1(x), \xi_2(x)) w(\xi_1(x), \xi_2(x)) \rho(x) \, \mathrm{d}\nu(x) + \frac{3}{4} \left\| [\xi_1, \xi_2]^V \right\|^2 \tag{2.46}$$

where

$$w(\xi_1(x), \xi_2(x)) = g_{\mathcal{C}(M)}(x,1)(\xi_1(x), \xi_1(x)) g_{\mathcal{C}(M)}(x)(\xi_2(x), \xi_2(x)) - \left(g_{\mathcal{C}(M)}(x,1)(\xi_1(x), \xi_2(x))\right)^2$$

and $[\xi_{\Phi_1}, \xi_{\Phi_2}]^V$ denotes the vertical projection of $[\xi_{\Phi_1}, \xi_{\Phi_2}]$ at identity, $\|\cdot\|$ denotes the norm at identity and $k(x,1)$ is the sectional curvature of the cone at point $(x,1)$ in the directions $(\xi_1(x), \xi_2(x))$.

This computation is only formal and we will not attempt here to give a rigorous meaning to this formula similarly to what has been done in [41] for the L^2 Wasserstein metric. Yet, it has interesting consequences: the curvature of the space of densities endowed with the WF metric is always greater or equal than the curvature of the cone $\mathcal{C}(M)$. In particular, it is non-negative if the curvature of (M, g) is bigger than 1, as a consequence of Proposition 2.

3. The H^{div} right-invariant metric on the diffeomorphism group

In this section, we summarize known results on the H^{div} right-invariant metric on the diffeomorphism group. We now define the H^{div} right-invariant metric.

Definition 7. Let (M, g) be a Riemannian manifold and $\mathrm{Diff}^s(M)$ be the group of diffeomorphisms which belong to $H^s(M)$ for $s > d/2 + 1$. The right-invariant H^{div} metric, implicitely dependent on two positive real parameters a, b, is defined by

$$G_\varphi(X_\varphi, X_\varphi) = \int_M a^2 |X_\varphi \circ \varphi^{-1}|^2 + b^2 \operatorname{div}(X_\varphi \circ \varphi^{-1})^2 \, \mathrm{dvol} \ . \tag{3.1}$$

The Euler-Arnold equation in one dimension (that is on the circle S^1 for instance) is the well-known Camassa-Holm equation (actually when $a = b = 1$):

$$a^2 \partial_t u - b^2 \partial_{txx} u + 3a^2 \partial_x u \, u - 2b^2 \partial_{xx} u \, \partial_x u - b^2 \partial_{xxx} u \, u = 0 \, . \tag{3.2}$$

On a general Riemannian manifold (M, g), the equation can be written as, with $n = a^2 u^\flat + b^2 \, \mathrm{d}\delta u^\flat$,

$$\partial_t n + a^2 \left(\operatorname{div}(u) u^\flat + \mathrm{d}\langle u, u \rangle + \iota_u \, \mathrm{d}u^\flat \right) + b^2 \left(\operatorname{div}(u) \, \mathrm{d}\delta u^\flat + \mathrm{d}\iota_u \, \mathrm{d}\delta u^\flat \right) = 0 \, , \tag{3.3}$$

where the notation $\flat$ corresponds to lowering the indices. More precisely, if $u \in \chi(M)$ then $u^\flat$ is the 1-form defined by $v \mapsto g(u, v)$. The notation δ is the formal adjoint to the exterior derivative d and ι is the insertion of vector fields which applies to forms.

On the well-posedness of the initial value problem. Although the theorem below is not stated in this particular form in [20], this result can be seen as a byproduct of their results as

explained in [49, Theorem 4.1]. For similar smoothness results in the case of smooth diffeomorphisms, we refer the reader to [15, Theorem 3].

Theorem 14. *On* $\mathrm{Diff}^s(S_1)$ *for* $s \geq 2$ *integer, the* H^1 *right-invariant metric is a smooth and weak Riemannian metric. Moreover, if* $s \geq 3$*, the exponential map is a smooth local diffeomorphism on* $T\,\mathrm{Diff}^s(S_1)$.

Global well-posedness does not hold in one dimension since there exist smooth initial conditions for the Camassa-Holm equation such that the solutions blow up in finite time.

In higher dimensions, the initial value problem has been studied by Michor and Mumford [52, Theorem 3]. This is not a direct result of [20] since the differential operator associated to the metric is not elliptic. They prove that the initial value problem on the space of vector fields is locally well posed for initial data in a Sobolev space of high enough order. Although the proof could probably be adapted to the case of a Riemannian manifold, in that case, the result of local well posedness is not proven yet.

On the metric properties of the H^{div} right-invariant metric. Michor and Mumford already had the following non-degeneracy result in [47].

Theorem 15 (Michor and Mumford)**.** *The distance on* $\mathrm{Diff}(M)$ *induced by the* H^{div} *right-invariant metric is non-degenerate. Namely, between two distinct diffeomorphisms the infimum of the lengths of the paths joining them is strictly positive.*

Due to the presence of blow up in the Camassa-Holm equation, metric completeness does not hold since it would imply geodesic completeness, that is global well posedness. However, it is still meaningful to ask whether geodesics are length minimizing for short times. Since the Gauss lemma is valid in a strong H^s topology, this ensures that geodesics are length minimizing among all curves that stay in a H^s neighborhood, see also [15]. However, this is *not* enough to prove that the associated geodesic distance is non degenerate since an almost minimizing geodesic can escape this neighborhood for arbitrarily small energy. This is what happens for the right-invariant metric $H^{1/2}$ on the circle S_1 where the metric is degenerate although there exists a smooth exponential map similarly to our case in 1D, see [22].

4. A Riemannian submanifold point of view on the H^{div} right-invariant metric

The starting point of this section is the following simple proposition whose proof is omitted.

Proposition 16. *Consider a Riemannian submersion constructed as in Proposition 9. Let* H_0 *be the isotropy subgroup of* ρ_0*, then, considering* H_0 *as a Riemannian submanifold of* H *and denoting* G_{H_0} *its induced metric,* G_{H_0} *is a right-invariant metric on* H_0.

The Riemannian submersion $\pi_0 : \mathrm{Aut}(\mathcal{C}(M)) \mapsto \mathrm{Dens}(M)$ defined in Proposition 10 enables to study the equivalent problem to the incompressible Euler equation. The fiber of the Riemannian submersion at vol is $\pi_0^{-1}(\{\mathrm{vol}\})$ and it will be denoted by $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$, it therefore corresponds to the group H_0 in the previous proposition. More explicitely, we have

$$\pi_0^{-1}(\{\mathrm{vol}\}) = \{(\varphi, \lambda) \in \mathrm{Aut}(\mathcal{C}(M)) \,:\, \varphi_*(\lambda^2\,\mathrm{vol}) = \mathrm{vol}\}\,. \tag{4.1}$$

The constraint $\varphi_*(\lambda^2\,\mathrm{vol}) = \mathrm{vol}$ can be made explicit as follows

$$\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M)) = \{(\varphi, \sqrt{\mathrm{Jac}(\varphi)}) \in \mathrm{Aut}(\mathcal{C}(M)) \,:\, \varphi \in \mathrm{Diff}(M)\}\,. \tag{4.2}$$

Note that this isotropy subgroup can be identified with the group of diffeomorphims of M since the map $\varphi \mapsto (\varphi, \sqrt{\mathrm{Jac}(\varphi)})$ is also a section of the short exact sequence (2.27). This shows that there is a natural identification between $\mathrm{Diff}(M)$ and $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$. Now, the vertical space at point $(\varphi, \sqrt{\mathrm{Jac}(\varphi)}) \in \mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ is

$$\mathrm{Ker}\left(d\pi_0(\varphi, \sqrt{\mathrm{Jac}(\varphi)})\right) = \{(v, \alpha) \circ (\varphi, \sqrt{\mathrm{Jac}(\varphi)}) \,:\, \mathrm{div}\, v = 2\alpha\,\}\,, \tag{4.3}$$

and equivalently

$$\operatorname{Ker}\left(d\pi_0(\varphi,\sqrt{\operatorname{Jac}(\varphi)})\right)=\left\{\left(v,\frac{1}{2}\operatorname{div}v\right)\circ(\varphi,\sqrt{\operatorname{Jac}(\varphi)})\ :\ v\in\operatorname{Vect}(M)\right\}. \tag{4.4}$$

It is now possible to apply equation (2.41) to obtain the explicit formula for the right-invariant metric on $\operatorname{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$. The metric $L^2(M,\mathcal{C}(M))$ on $\operatorname{Aut}(\mathcal{C}(M))$ restricted to $\operatorname{Diff}(M)\simeq\operatorname{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ reads

$$G_\varphi(X_\varphi,X_\varphi)=\int_M|v|^2\operatorname{dvol}+\frac{1}{4}\int_M|\operatorname{div}v|^2\operatorname{dvol}, \tag{4.5}$$

where $v=X_\varphi\circ\varphi^{-1}$. Therefore, on $\operatorname{Diff}(M)\simeq\operatorname{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$, the induced metric is a right-invariant H^{div} metric. In other words, we have

Theorem 17. *By its identification with $\operatorname{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$, the diffeomorphism group endowed with the H^{div} right-invariant metric, see Definition 7, is isometrically embedded in $L^2(M,\mathcal{C}(M))$.*

As a straightforward application, we retrieve theorem 15.

Corollary 18. *The distance on $\operatorname{Diff}(M)$ with the right-invariant metric H^{div} is non degenerate.*

Proof. Let $\varphi_0,\varphi_1\in\operatorname{Diff}(M)$ be two diffeomorphisms and c be a path joining them. The length of the path c for the right-invariant metric H^{div} is equal to the length of the lifted path $\tilde{c}$ in $\operatorname{Aut}(\mathcal{C}(M))$. Since $L^2(M,\mathcal{C}(M))$ is a Hilbert manifold, the length of the path $\tilde{c}$ is bounded below by the length of the geodesic joining the natural lifts of φ_0 and φ_1 in $L^2(M,\mathcal{C}(M))$. Therefore, it leads to

$$d_{H^{\mathrm{div}}}(\varphi_0,\varphi_1)\geq d_{L^2(M,\mathcal{C}(M))}\left((\varphi_0,\sqrt{\operatorname{Jac}(\varphi_0)}),(\varphi_1,\sqrt{\operatorname{Jac}(\varphi_1)})\right). \tag{4.6}$$

If $d_{H^{\mathrm{div}}}(\varphi_0,\varphi_1)=0$ then $d_{L^2(M,\mathcal{C}(M))}\left((\varphi_0,\sqrt{\operatorname{Jac}(\varphi_0)}),(\varphi_1,\sqrt{\operatorname{Jac}(\varphi_1)})\right)=0$ which implies $\varphi_0=\varphi_1$. □

Remark 2 (The Fisher-Rao metric)**.** *In* [33]*, it is shown that the $\dot{H}^1$ right-invariant metric descends to the Fisher-Rao metric on the space of densities. Let us explain why this situation differs from ours: It is well known that a left action of a group endowed with a right-invariant metric induces on the orbit a Riemannian metric for which the action is a Riemannian submersion. However, Khesin et al. do not consider a left action, but a right action on the space of densities: More precisely, if a reference density ρ is chosen, the map they considered is*

$$\begin{aligned}\operatorname{Diff}(M)&\to\operatorname{Dens}(M)\\ \varphi&\mapsto\varphi^*\rho\,.\end{aligned}$$

Obviously, this situation is equivalent to a left action of a group of diffeomorphisms endowed with a left-invariant metric. In such a situation, the descending metric property has to be checked [33, Proposition 2.3].

Their result can be read from our point of view: The $\dot{H}^1$ metric is $\frac{1}{4}\int_M|\operatorname{div}v|^2\,\mathrm{d}\mu$ and it corresponds to the case where $a=0$. It thus leads to a degenerate metric on the group. Viewed in the ambient space $L^2(M,\mathcal{C}(M))$, the projection on the bundle component is a (pseudo-) isometry from $L^2(M,\mathcal{C}(M))$ (endowed with this pseudo-metric) to the space of densities since $a=0$. Moreover, on the space of densities which lie in the image of the projection, that is, the set of probability densities, the projected metric is the Fisher-Rao metric.

We now use the identification between $\operatorname{Diff}(M)$ endowed with the right-invariant H^{div} metric and $\operatorname{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ as a submanifold of $\operatorname{Aut}(\mathcal{C}(M))$ and write the geodesic equations in this setting. As is standard for the incompressible Euler equation, the constraint is written in Eulerian coordinates and the corresponding geodesic are written hereafter.

Theorem 19. *The geodesic equations on the fiber* $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ *as a Riemannian submanifold of* $\mathrm{Aut}(\mathcal{C}(M))$ *endowed with the metric* $L^2(M,\mathcal{C}(M))$ *can be written in Lagrangian coordinates*

$$\begin{cases} \frac{D}{Dt}\dot\varphi + 2\frac{\dot\lambda}{\lambda}\dot\varphi = -\frac{1}{2}\nabla^g p\circ\varphi \\ \ddot\lambda - \lambda g(\dot\varphi,\dot\varphi) = -\lambda p\circ\varphi\,, \end{cases} \tag{4.7}$$

with a function $P : M \to \mathbb{R}$.
In Eulerian coordinates, the geodesic equations read

$$\begin{cases} \dot v + \nabla^g_v v + 2v\alpha = -\frac{1}{2}\nabla^g p \\ \dot\alpha + \langle\nabla\alpha, v\rangle + \alpha^2 - g(v,v) = -p\,, \end{cases} \tag{4.8}$$

where $\alpha = \frac{\dot\lambda}{\lambda}\circ\varphi^{-1}$ *and* $v = \partial_t\varphi\circ\varphi^{-1}$.

This submanifold point of view leads to a generalization of [32, Theorem A.2] on the sectional curvature of $\mathrm{Diff}(M)$ which has been computed and studied in [32]. The authors show that the curvature of $\mathrm{Diff}(S_1)$ can be written using the Gauss-Codazzi formula and they show the explicit embedding in a semi-direct product of groups similar to our situation.

As mentioned above, we consider $\mathrm{Diff}(M)$ as a submanifold of $L^2(M,\mathcal{C}(M))$. The second fundamental form can be computed as in the case of the incompressible Euler equation.

Proposition 20. *Let* U, V *be two smooth right-invariant vector fields on* $\mathrm{Aut}(\mathcal{C}(M))$ *that can be written as* $U(\varphi,\lambda) = (u,\alpha)\circ(\varphi,\lambda)$ *and* $V(\varphi,\lambda) = (v,\beta)\circ(\varphi,\lambda)$. *The second fundamental form for the isometric embedding* $\mathrm{Diff}(M)\hookrightarrow L^2(M,\mathcal{C}(M))$ *is*

$$\mathrm{II}(U,V) = \left(-\frac{1}{2}\nabla p\circ\varphi, -\lambda p\circ\varphi\right)\,, \tag{4.9}$$

where $p = (2\,\mathrm{Id} - \frac{1}{2}\Delta)^{-1}A(\nabla_{(u,\alpha)}(v,\beta))$ *is the unique solution of the elliptic PDE* (2.44)

$$(2\,\mathrm{Id} - \frac{1}{2}\Delta)(p) = A(\nabla_{(u,\alpha)}(v,\beta))\,, \tag{4.10}$$

where $A(w,\gamma) \stackrel{\text{def.}}{=} \mathrm{div}(w) - 2\gamma$. *Using the explicit expression of* $\nabla_{(u,\alpha)}(v,\beta)$ *the elliptic PDE reads*

$$(2\,\mathrm{Id} - \frac{1}{2}\Delta)(p) = \mathrm{div}(\nabla_u v + \beta u + \alpha v) - 2\langle\nabla\beta, u\rangle + 2g(u,v) - 2\alpha\beta\,. \tag{4.11}$$

Proof. By right-invariance of the metric, it suffices to treat the case $(\varphi,\lambda) = \mathrm{Id}$. The orthogonal projection is the horizontal lift defined in Proposition 11. Therefore, we compute the infinitesimal action of $\nabla_{(u,\alpha)}(v,\beta)$ on the volume form which is given by the linear operator A and we consider its horizontal lift $(-\frac{1}{2}\nabla p, -p)$ given by Proposition 11. By right-invariance, the orthogonal projection at (φ,λ) is given by $\left(-\frac{1}{2}\nabla p\circ\varphi, -\lambda p\circ\varphi\right)$.

From Proposition 2, one has

$$\nabla_{(u,\alpha)}(v,\beta) = (\nabla_u v + \beta u + \alpha v, \langle\nabla\beta, u\rangle - g(u,v) + \alpha\beta)\,, \tag{4.12}$$

and Formula (4.11) follows. □

We can then state the Gauss-Codazzi formula applied to our context.

Proposition 21. *Let* U, V *be two smooth right-invariant vector fields on* $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ *written as* $U(\varphi,\lambda) = (u,\alpha)\circ(\varphi,\lambda)$ *and* $V(\varphi,\lambda) = (v,\beta)\circ(\varphi,\lambda)$. *The sectional curvature of* $\mathrm{Diff}(M)$ *endowed with the right-invariant* H^{div} *metric is*

$$\langle R_{\mathrm{Diff}(M)}(U,V)V,U\rangle = \langle R_{L^2(M,\mathcal{C}(M))}(U,V)V,U\rangle + \langle\mathrm{II}(U,U),\mathrm{II}(V,V)\rangle - \langle\mathrm{II}(U,V),\mathrm{II}(U,V)\rangle\,, \tag{4.13}$$

where II *is the second fundamental form* (4.9) *and*

$$\langle R_{L^2(M,\mathcal{C}(M))}(U,V)V,U\rangle = \int_M \langle R_{\mathcal{C}(M)}(u,v)v,u\rangle\circ(\varphi,\lambda)\,\mathrm{d}\mu\,, \tag{4.14}$$

where $(\varphi,\lambda)\in\mathrm{Aut}(\mathcal{C}(M))$.

Proof. The only remaining point is the computation of the sectional curvature of $L^2(M, \mathcal{C}(M))$ which is done in Freed and Groisser's article [24]. □

Note that the sectional curvature of $L^2(M, \mathcal{C}(M))$ vanishes if $M = S_n$ since $\mathcal{C}(M) = \mathbb{R}^{n+1}$, which is the case for the one-dimensional Camassa-Holm equation. However, for $M = T_n$, $n \geq 2$, the flat torus, the sectional curvature of $\mathcal{C}(M)$ is non-positive and bounded below by -1 and thus the sectional curvature of $L^2(T_n, \mathcal{C}(T_n))$ is non-positive.

5. Applications

The point of view developed above provides an example of an isometric embedding of the group of diffeomorphisms endowed with the right-invariant H^{div} metric in an L^2 space such as $L^2(M, N)$, here with $N = \mathcal{C}(M)$. In this section, we develop two applications of this point of view. The first one consists in rewriting the Camassa-Holm equation as particular solutions of the incompressible Euler equation on the cone; the results hold in higher dimensions for the geodesics of the H^{div} metric. The second application is about minimizing properties of solutions of the Camassa-Holm equation and its generalization with H^{div}. We prove that, under mild conditions, smooth solutions are length minimizing for short times.

5.1. **The Camassa-Holm equation.** Let us consider the following Camassa-Holm equation,

$$\begin{cases} \partial_t u - \frac{1}{4}\partial_{txx} u + 3\partial_x u\, u - \frac{1}{2}\partial_{xx} u\, \partial_x u - \frac{1}{4}\partial_{xxx} u\, u = 0 \\ \partial_t \varphi(t,x) = u(t, \varphi(t,x))\,. \end{cases} \tag{5.1}$$

With respect to the standard Camassa-Holm equation, this equation has different coefficients that are chosen here to simplify the discussion. Unless otherwise mentioned, all the results still apply to the standard formulation of the equation. For such a choice of coefficients, the cone construction $\mathcal{C}(S_1)$ is isometric to $\mathbb{R}^2 \setminus \{0\}$ with the Euclidean metric. Following Theorem 17, we have the isometric injection

$$\mathcal{M} : \mathrm{Diff}(S_1) \to \mathrm{Aut}(\mathcal{C}(S_1)) \subset L^2(S_1, \mathbb{R}^2) \tag{5.2}$$

$$\varphi \mapsto (\varphi, \sqrt{\varphi'}) = \sqrt{\varphi'} e^{i\varphi}\,. \tag{5.3}$$

Then, solutions of the Camassa-Holm equation are geodesic for the flat metric $L^2(S_1, \mathbb{R}^2)$ on the constrained submanifold of maps (φ, λ) defined by the constraint $\varphi' = \lambda^2$. Note that the map $\mathcal{M}$ is very similar to a Madelung transform which maps solutions of the Schrödinger equation to solutions of a compressible Euler type of hydrodynamical equation. The geodesic equation on $\mathrm{Aut}(\mathcal{C}(S_1))$ reads

$$\begin{cases} \ddot{\varphi} + 2\frac{\dot{\lambda}}{\lambda}\dot{\varphi} = -\frac{1}{2}\partial_x p \circ \varphi \\ \ddot{\lambda} - \lambda\dot{\varphi}^2 = -\lambda p \circ \varphi\,, \end{cases} \tag{5.4}$$

where $p : S_1 \to \mathbb{R}$. Formula (5.4) looks similar to the incompressible Euler equation in Lagrangian coordinates. However, this geodesic equation is apparently written on the space of maps $S_1 \mapsto \mathcal{C}(S_1)$. Since $\mathrm{Aut}(\mathcal{C}(S_1)) \subset \mathrm{Diff}(\mathcal{C}(S_1))$, it can be expected to be a geodesic equation on the group of diffeomorphism of the cone. Indeed, we have

Theorem 22. *Solutions to the Camassa-Holm equation on S_1*

$$\partial_t u - \frac{1}{4}\partial_{txx} u + 3\partial_x u\, u - \frac{1}{2}\partial_{xx} u\, \partial_x u - \frac{1}{4}\partial_{xxx} u\, u = 0 \tag{5.5}$$

are mapped to solutions of the incompressible Euler equation on $\mathbb{R}^2 \setminus \{0\}$ for the density $\rho = \frac{1}{r^4}\,\mathrm{Leb}$, that is

$$\begin{cases} \dot{v} + \nabla_v v = -\nabla P\,, \\ \nabla \cdot (\rho v) = 0\,, \end{cases} \tag{5.6}$$

by the map $: \begin{bmatrix} u : S_1 \to \mathbb{R} \\ \theta \mapsto u(\theta) \end{bmatrix} \mapsto \begin{bmatrix} v : S_1 \times \mathbb{R}^+_* = \mathcal{C}(S_1) \to \mathbb{R}^2 \\ (\theta, r) \mapsto \big(u(\theta), \frac{r}{2}\partial_x u(\theta)\big) \end{bmatrix}$

Proof. We show that $\mathcal{M}(\varphi)$ provides solutions to the incompressible Euler equation written in Lagrangian coordinates. The second equation in (5.4) being linear in λ and the first equation being 0 homogeneous in λ, the geodesic equations can be rewritten as

$$\begin{cases} \ddot{\varphi} + 2\frac{\dot{\lambda}}{\lambda}\dot{\varphi} = -\frac{1}{2}\partial_x p \circ \varphi \\ \ddot{\lambda} r - \lambda r \dot{\varphi}^2 = -\lambda r p \circ \varphi \,. \end{cases} \tag{5.7}$$

Thus, the map $\Phi(t) : (x, r) \mapsto (\varphi(t, x), \lambda(t, x)r)$ satisfies

$$\ddot{\Phi}(t)(x, r) = -\nabla \Psi_p(t) \circ \Phi(t) \,, \tag{5.8}$$

where $\Psi_p(x, r) = \frac{1}{2}r^2 p(x)$. This formulation is close to the incompressible Euler equation, however, we need to check if the density $\rho(r, \theta) = 1/r^3 \, \mathrm{d}r \, \mathrm{d}\theta$ is preserved by pull-back by Φ, or equivalently due to the group structure, by pushforward. We first compute the Jacobian matrix, recalling that $\lambda = \sqrt{\partial_x \varphi}$,

$$D\Phi(x, r) = \begin{pmatrix} \partial_x \varphi & 0 \\ \frac{\partial_{xx}\varphi}{2\sqrt{\partial_x \varphi}} & \sqrt{\partial_x \varphi} \end{pmatrix} ,$$

whose determinant is $(\partial_x \varphi)^{3/2}$. We now compute the pushforward

$$\begin{aligned} \mathrm{Jac}(\Phi)\rho \circ \Phi(x, r) &= 1/(r\sqrt{\partial_x \varphi})^3 \, \mathrm{Jac}(\Phi) \\ &= 1/(r\sqrt{\partial_x \varphi})^3 (\partial_x \varphi)^{3/2} = \frac{1}{r^3} = \rho(x, r) \,. \end{aligned}$$

This proves the result in Lagrangian coordinates. To get the formulation in the theorem, one differentiates the map Φ at identity which gives $(u, \frac{r}{2}\partial_x u)$ for the vector field in polar coordinates. □

Remark 3 (About the blow-up)**.** *At this point, a natural question is about the difference between global well-posedness of incompressible Euler in 2D, whereas the Camassa-Holm equation has a well understood blow-up. Of course, there is no contradiction since the density for which the CH equation is similar to Euler has a singularity at zero, which allows for unbounded vorticity although we did not check this possibility. In a similar direction, we can cite* [21]*, since the authors mention that the singularity comes "from the vorticity amplification due to the presence of a density gradient". Note also that the typical situation of blow-up of the CH equation in the case of colliding peakons can be understood in this situation as the quantity $\sqrt{\partial_x \varphi}$ goes to zero in finite time.*

The second application consists in showing that smooth solutions of the Camassa-Holm equation are length minimizing for short times.

Theorem 23 (Smooth solutions to the Camassa-Holm equation (5.1) are length minimizing for short times.)**.** *Let $(\varphi(t), \lambda(t))$ be a smooth solution to the geodesic equations* (5.1) *(in the formulation* (5.4)*) on the time interval $[t_0, t_1]$. If $(t_1 - t_0)^2 |\langle w, \nabla^2 \Psi_p(x, r) w \rangle| < \pi^2 \|w\|^2$ holds for all $t \in [t_0, t_1]$ and $(x, r) \in \mathcal{C}(S_1)$ and $w \in T_{(x,r)}\mathcal{C}(S_1)$, then for every smooth curve $(\varphi_0(t), \lambda_0(t)) \in \mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(S_1))$ satisfying $(\varphi_0(t_i), \lambda_0(t_i)) = (\varphi(t_i), \lambda(t_i))$ for $i = 0, 1$ one has*

$$\int_{t_0}^{t_1} \|(\dot{\varphi}, \dot{\lambda})\|^2 \, \mathrm{d}t \leq \int_{t_0}^{t_1} \|(\dot{\varphi}_0, \dot{\lambda}_0)\|^2 \, \mathrm{d}t \,, \tag{5.9}$$

with equality if and only if the two paths coincide on $[t_0, t_1]$.

Remark 4. *This result only applies to this choice of coefficients and for other choices of coefficients the result still holds in an L^∞ neighborhood of the geodesic. This is done in the more general case of H^{div} in the next section. Since the proof is a direct adaptation of Brenier's* [7] *and it is simple in this particular case, we include it hereafter. It also helps to understand the proof in the general case of a Riemannian manifold.*

Proof. To alleviate notations, we denote $g_t = (\varphi(t), \lambda(t))$ and $h_t = (\varphi_0(t), \lambda_0(t))$. Since p can be chosen with zero mean, $\Psi_p(x,r) = \frac{1}{2}r^2 p(x)$ and $g_t = (\varphi(t), \sqrt{\mathrm{Jac}(\varphi(t))})$, by direct integration, for every $t \in [t_0, t_1]$

$$\int_{S_1} \Psi_p(g_t(x))\,\mathrm{d}x = 0\,. \tag{5.10}$$

The same equality holds for h_t. Let $s \in [0,1] \mapsto c(t,s,x)$ be a two parameters ($t \in [t_0,t_1]$ and $x \in S_1$) smooth family of geodesics on $\mathbb{R}^2$ such that $c(t,0,x) = g_t(x)$ and $c(t,1,x) = h_t(x)$ for every $t \in [t_0,t_1]$ and $x \in S_1$. Let us define $J(t,s,x) = \partial_t c(t,s,x)$, we have

$$J(t,0,x) = \partial_t g_t(x) \text{ and } J(t,1,x) = \partial_t h_t(x)\,. \tag{5.11}$$

Now, the result we want to prove can be reformulated as,

$$\int_{t_0}^{t_1}\int_{S_1} \|J(t,0,x)\|^2\,\mathrm{d}t\,\mathrm{d}x \le \int_{t_0}^{t_1}\int_{S_1} \|J(t,1,x)\|^2\,\mathrm{d}t\,\mathrm{d}x \tag{5.12}$$

with equality if and only if for almost every x, it holds $g_t(x) = h_t(x)$ for all $t \in [t_0,t_1]$. Using a second-order Taylor expansion of $\Psi_p(c(t,s,x))$ with respect to s at $s=0$ and denoting by $C \overset{\text{def.}}{=} \sup_{t\in[t_0,t_1]}\sup_{x\in S_1}\|\nabla^2\Psi_p(x)\|$, we have,

$$\Psi_p(h_t(x)) - \Psi_p(g_t(x)) - \langle \nabla\Psi_p(c(t,0,x)), \partial_s c(t,0,x)\rangle \le \frac{C}{2}\int_0^1 \|\partial_s c(t,s,x)\|^2\,\mathrm{d}s\,.$$

We will integrate in time t and apply the one dimensional Poincaré inequality in the t variable

$$\int_{t_0}^{t_1} \|\partial_s c(t,s,x)\|^2\,\mathrm{d}t \le \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1} |\partial_t\|\partial_s c(t,s,x)\||^2\,\mathrm{d}t\,, \tag{5.13}$$

for every s, x. Since $c(t,0,x)$ is a solution of the Camassa-Holm equation, one has $\partial_{tt}c = -\nabla\Psi_p(t)$. Thus, we have, integrating in time

$$\int_{t_0}^{t_1} \Psi_p(h_t(x)) - \Psi_p(g_t(x)) + \langle \partial_{tt}c(t,0,x), \partial_s c(t,0,x)\rangle\,\mathrm{d}t \le \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_0^1 |\partial_t\|\partial_s c(t,s,x)\||^2\,\mathrm{d}s\,\mathrm{d}t\,.$$

We also have $|\partial_t\|\partial_s c(t,s,x)\||^2 \le \|\partial_{ts}c(t,s,x)\|^2$. Then, integrating over S_1, the two first terms on the l.h.s. vanish and integrating by part in time, we get

$$\int_{t_1}^{t_2}\int_{S_1} -\langle \partial_t c(t,0,x), \partial_{st}c(t,0,x)\rangle\,\mathrm{d}t \le \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_{S_1}\int_0^1 \|\partial_{ts}c(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,, \tag{5.14}$$

where we used the fact that $\partial_s c(t,s,x)$ is constant in s since the geodesics on the plane are straight lines. Writing $f(s) = \frac{1}{2}\int_{t_0}^{t_1}\int_{S_1}\|J(t,s,x)\|^2\,\mathrm{d}t$, we want to prove $f(1) \ge f(0)$ and we have

$$-f'(0) \le \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_{S_1}\int_0^1 \|\partial_s J(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,.$$

Therefore, the result is proven if we can show that for some $\varepsilon > 0$

$$f(1) - f(0) - f'(0) \ge \varepsilon\int_{t_0}^{t_1}\int_{S_1}\int_0^1 \|\partial_s J(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,. \tag{5.15}$$

We have $f(1) - f(0) - f'(0) = \int_0^1 (1-s)f''(s)\,\mathrm{d}s$ and here $f''(s) = \int_{t_0}^{t_1}\int_{S_1}\|\partial_s J(t,s,x)\|^2\,\mathrm{d}t\,\mathrm{d}x$ since $\partial_{ss}J = 0$ because $\mathbb{R}^2$ has vanishing curvature, and also $\partial_s J = \mathrm{cste}(t,x)$, a constant w.r.t. s. Hence, we get

$$f(1) - f(0) - f'(0) = \frac{1}{2}\int_{t_0}^{t_1}\int_{S_1}\int_0^1 \|\partial_s J(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,. \tag{5.16}$$

Therefore,

$$f(1) - f(0) \ge \left(\frac{1}{2} - \frac{C(t_1-t_0)^2}{2\pi^2}\right)\int_{t_0}^{t_1}\int_{S_1}\int_0^1 \|\partial_s J(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,,$$

which is nonnegative if $t_1 - t_0 \leq \frac{\pi}{\sqrt{C}}$. □

Remark 5. *The condition on the Hessian is satisfied for smooth paths, see Remark 6. Moreover, similarly to Brenier's proof, the constant is sharp since the rotation at unit speed is a particular solution of the Camassa-Holm equation for which the Hessian is equal to* 1 *and it stops being a minimizer at the angle* π.

5.2. **The H^{div} case in higher dimensions.** In the general case, we are left with the geometry of the cone, and therefore, the map $\mathcal{M}$ maps solutions of the geodesic equation on the diffeomorphisms group for the right-invariant H^{div} metric to solutions of the incompressible Euler equation on the $\mathcal{C}(M)$ for a density which has a singularity at the cone point. In the general case, the geodesic equation is written as

$$\begin{cases} \frac{D}{Dt}\dot{\varphi} + 2\frac{\dot{\lambda}}{\lambda}\dot{\varphi} = -\frac{1}{2}\nabla^g p \circ \varphi \\ \ddot{\lambda} r - \lambda r g(\dot{\varphi},\dot{\varphi}) = -\lambda r p \circ \varphi \,. \end{cases} \tag{5.17}$$

Viewing the automorphisms (φ, λ) of the cone as diffeomorphisms of the cone, the geodesic equation is close to incompressible Euler equations, with the difference that the automorphisms do not preserve the Riemannian volume measure on $\mathcal{C}(M)$ but another density which has a singularity at the cone point.

Theorem 24. *On the group of diffeomorphisms of the cone, the geodesic equation can be written*

$$\frac{D}{Dt}(\dot{\varphi}, \dot{\lambda r}) = -\nabla \Psi_p \circ (\varphi, \lambda r)\,, \tag{5.18}$$

where $\Psi_p(x,r) \overset{\text{def.}}{=} \frac{1}{2} r^2 p(x)$. *Moreover, the diffeomorphisms of* $\mathcal{C}(M)$ (φ, λ) *preserve the measure* $\tilde{\nu} \overset{\text{def.}}{=} r^{-3}\,\mathrm{d}r\,\mathrm{dvol}$.
In other words, a solution (φ, λ) *of* (5.18) *is a solution of the incompressible Euler equation for the density* $r^{-3-d}\,\mathrm{dvol}_{\mathcal{C}(M)}$ *where* $\mathrm{dvol}_{\mathcal{C}(M)}$ *is the volume form on the cone* $\mathcal{C}(M)$ *and* d *is the dimension of* M.

Proof. The geodesic equations (5.17) can be rewritten in the form (5.18) since a direct computation gives $\nabla \Psi_p = (\frac{1}{2}\nabla^g p, rp)$.

The only remaining point is that (φ, λ) preserves the measure $r^{-3}\,\mathrm{d}\nu\,\mathrm{d}r$ on $\mathcal{C}(M)$, if the relation $\lambda = \sqrt{\mathrm{Jac}(\varphi)}$ holds. Indeed, the volume form $r^\theta\,\mathrm{d}\nu\,\mathrm{d}r$ is preserved by (φ, λ) if and only if the following equality is satisfied $(\lambda r)^\theta \lambda\,\mathrm{Jac}(\varphi) = r^\theta$, equivalently $\lambda^{\theta+3} = 1$. It is the case if and only if $\theta = -3$. □

In particular, this theorem underlines that $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M)) = \mathrm{Aut}(\mathcal{C}(M)) \cap \mathrm{SDiff}_{\tilde{\nu}}(\mathcal{C}(M))$. In remark 5, we mentioned that $\mathrm{Aut}(\mathcal{C}(M))$ is a totally geodesic subspace of $\mathrm{Diff}(\mathcal{C}(M))$, which explains the fact that the geodesic equation on $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ is actually a geodesic equation on $\mathrm{SDiff}_{\tilde{\nu}}(\mathcal{C}(M))$. We illustrate this situation in Figure 1.

The same result holds on more general Riemannian manifolds. We propose a straightforward generalization of Brenier's proof [7] in the case of Euler equation to a Riemannian setting. Note that, to our knowledge, no previous result was available on minimizing H^{div} geodesics. In the worst case of our theorem, we require only an L^∞ bound on the Jacobian and on the diffeomorphism.

Theorem 25. *Let* $(\varphi(t), \lambda(t))$ *be a smooth solution to the geodesic equations* (5.18) *on the time interval* $[t_0, t_1]$. *If* $(t_1 - t_0)^2 |\langle w, \nabla^2 \Psi_p(x,r) w\rangle| < \pi^2 \|w\|^2$ *holds for all* $t \in [t_0, t_1]$ *and* $(x,r) \in \mathcal{C}(M)$ *and* $w \in T_{(x,r)}\mathcal{C}(M)$, *then for every smooth curve* $(\varphi_0(t), \lambda_0(t)) \in \mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ *satisfying* $(\varphi_0(t_i), \lambda_0(t_i)) = (\varphi(t_i), \lambda(t_i))$ *for* $i = 0, 1$ *and the condition* $(*)$, *one has*

$$\int_{t_0}^{t_1} \|(\dot{\varphi}, \dot{\lambda})\|^2 \,\mathrm{d}t \leq \int_{t_0}^{t_1} \|(\dot{\varphi}_0, \dot{\lambda}_0)\|^2 \,\mathrm{d}t\,, \tag{5.19}$$

with equality if and only if the two paths coincide on $[t_0, t_1]$.

Define $\delta_0 \overset{\text{def.}}{=} \min\{r(x,t) \,:\, \text{injectivity radius at } (\varphi(t,x), \lambda(t,x))\}$, *then the condition* $(*)$ *is:*

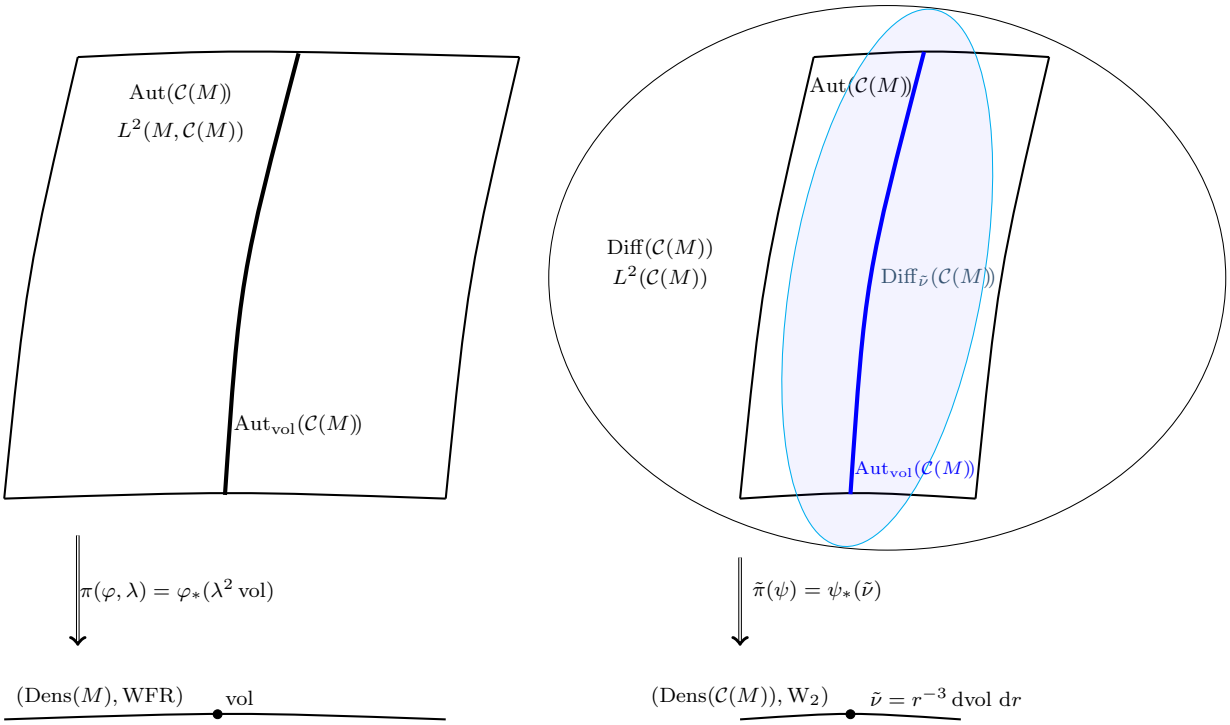


FIGURE 1. On the left, the picture represents the Riemannian submersion between $\mathrm{Aut}(\mathcal{C}(M))$ and the space of positive densities on M and the fiber above the volume form is $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$. On the right, the picture represents the automorphism group $\mathrm{Aut}(\mathcal{C}(M))$ isometrically embedded in $\mathrm{Diff}(\mathcal{C}(M))$ and the intersection of $\mathrm{Diff}_{\tilde{\nu}}(\mathcal{C}(M))$ and $\mathrm{Aut}(\mathcal{C}(M))$ is equal to $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$.

(1) *If the sectional curvature of* $\mathcal{C}(M)$ *can assume both signs or if* $\mathrm{diam}(M) \geq \pi$*, there exists* δ *satisfying* $0 < \delta < \delta_0$ *such that the curve* $(\varphi_0(t), \lambda_0(t))$ *has to belong to a* δ*-neighborhood of* $(\varphi(t), \lambda(t))$*, namely*

$$d_{\mathcal{C}(M)}\left((\varphi_0(t,x), \lambda_0(t,x)), (\varphi(t,x), \lambda(t,x))\right) \leq \delta$$

for all $(x,t) \in M \times [t_0, t_1]$ *where* $d_{\mathcal{C}(M)}$ *is the distance on the cone.*

(2) *If* $\mathcal{C}(M)$ *has non positive sectional curvature, then, for every* $\delta < \delta_0$*, there exists a short enough time interval on which the geodesic will be length minimizing.*

(3) *If* $M = S_d(1)$*, the result is valid for every path* $(\dot{\varphi}_0, \dot{\lambda}_0)$*.*

Remark 6. *Importantly, the condition on the Hessian is not empty, i.e. it is fulfilled in our case of interest: Indeed, when* p *is a* C^2 *function on* M*, the Hessian of* $\Psi_p(x,r) = \frac{1}{2}r^2 p(x)$ *is, in the orthonormal basis* $\partial_r, \frac{1}{r}e_1, \ldots, \frac{1}{r}e_d$ *where* $e_1, \ldots, e_d$ *is an orthornormal basis of* T_xM

$$\nabla^2 \Psi_p(x,r) = \begin{pmatrix} \frac{1}{2}\nabla^2 p(x) & \nabla p(x) \\ \nabla p^T(x) & p(x) \end{pmatrix}, \tag{5.20}$$

where ∇p *is the gradient of* p *in the orthornormal basis* $e_1, \ldots, e_d$*. Since* p *is smooth and* M *is compact, the Hessian of* p *is bounded uniformly on* $\mathcal{C}(M)$*.*

The proof is postponed in Appendix. The generalization of Brenier's proof that we propose is not completely satisfactory in positive curvature or, in the case of negative curvature, because of the injectivity radius bound. In the former case, the constructed interpolating paths have to pass through the cone point and therefore these paths $c(t,s,x)$ are not smooth any longer w.r.t. s and thus Jacobi fields are not smooth a priori. These two limitations could probably be overcome

using a different strategy than a geodesic homotopy between the two diffeomorphisms. We actually conjecture that the result holds true without the boundedness assumption.

6. Future directions

In this article, we have presented the geometric link between the Camassa-Holm equation and the new L^2 Wasserstein optimal transport metric between positive Radon measures. We presented an isometric embedding of the group of diffeomorphism group endowed with the right-invariant H^{div} metric in the space $L^2(M, \mathcal{C}(M))$. This isometric embedding enables to rewrite the Camassa-Holm equation, via a Madelung transform, as an incompressible Euler equation on the cone. In other words, the Camassa-Holm equation is a geodesic flow on $\mathrm{Aut}_{\mathrm{vol}}(\mathcal{C}(M))$ for the L^2 metric. As an application, this has also led to a result on the minimizing property of geodesics. The point of view developed in this paper can be taken to address the variational problem of shortest path for the H^{div} metric in the sense of Brenier [6, 8], which appears to be a non-trivial problem. Following Brenier, we will investigate elsewhere the uniqueness of the pressure as in [4]. This isometric embedding and the polar factorization theorem opens the way to design new numerical simulations of variational solutions of the Camassa-Holm equation, in the direction of [27, 45].

Following the point of view developed in this article, we plan to rewrite other fluid dynamic equations as geodesic equations on a submanifold of a space of maps endowed with an L^2 norm. The result may have, as shown for the Camassa-Holm equation, interesting analytical consequences.

Appendix A. Proof of Theorem 25

Proof. To alleviate notations, we denote $g_t = (\varphi(t), \lambda(t))$ and $h_t = (\varphi_0(t), \lambda_0(t))$. Since p can be choose with zero mean, $\Psi_p(x, r) = \frac{1}{2} r^2 p(x)$ and $g_t = (\varphi(t), \sqrt{\mathrm{Jac}(\varphi(t))})$, by direct integration, for every $t \in [t_0, t_1]$,

$$\int_M \Psi_p(g_t(x))\,\mathrm{d}x = 0\,. \tag{A.1}$$

The same equality holds for h_t.

Let $s \in [0,1] \mapsto c(t, s, x)$ be a two parameters ($t \in [t_0, t_1]$ and $x \in M$) family of geodesics on $\mathcal{C}(M)$ such that $c(t, 0, x) = g_t(x)$ and $c(t, 1, x) = h_t(x)$ for every $t \in [t_0, t_1]$ and $x \in M$. This family of geodesics is uniquely defined if one considers balls which do not intersect the cut locus. Uniformity of the radius of the balls can be obtained since $[t_0, t_1] \times M$ is compact, which defines δ_0. Consequently, the family of curves $c(t, s, x)$ is a smooth family of geodesics, at least as smooth as $g_t(x)$ and $h_t(x)$ are with respect to the parameters t, x. Since $\partial_t c(t, s, x)$ is a variation of geodesics, it is a Jacobi field as a function of s. Thus, we will use the notation $J(t, s, x) = \partial_t c(t, s, x)$. Consequently, we have

$$J(t, 0, x) = \partial_t g_t(x) \text{ and } J(t, 1, x) = \partial_t h_t(x)\,. \tag{A.2}$$

Now, the result we want to prove can be reformulated as,

$$\int_{t_0}^{t_1} \int_M \|J(t, 0, x)\|^2\,\mathrm{d}t\,\mathrm{d}x \leq \int_{t_0}^{t_1} \int_M \|J(t, 1, x)\|^2\,\mathrm{d}t\,\mathrm{d}x \tag{A.3}$$

with equality if and only if for almost every x, it holds $g_t(x) = h_t(x)$ for all $t \in [t_1, t_2]$. We now use a second-order Taylor expansion of $\Psi_p(c(t, s, x))$ with respect to s at $s = 0$. Denoting by $C \stackrel{\text{def.}}{=} \sup_{t \in [t_0, t_1]} \sup_{x \in M} \|\nabla^2 \Psi_{p_t}(x)\|$, we have,

$$\Psi_p(h_t(x)) - \Psi_p(g_t(x)) - \langle \nabla \Psi_p(c(t, 0, x), \partial_s c(t, 0, x) \rangle \leq \frac{C}{2} \int_0^1 \|\partial_s c(t, s, x)\|^2\,\mathrm{d}s\,.$$

Now, one has that $\partial_s c(t, s, x)$ vanishes at $t = t_0$ and $t = t_1$. We can therefore apply Poincaré inequality to $\|\partial_s c(t, s, x)\|$ to obtain

$$\int_{t_0}^{t_1} \|\partial_s c(t, s, x)\|^2\,\mathrm{d}t \leq \frac{C(t_1 - t_0)^2}{2\pi^2} \int_{t_0}^{t_1} |\partial_t \|\partial_s c(t, s, x)\||^2\,\mathrm{d}t\,. \tag{A.4}$$

Since $\partial_t\|\partial_s\| = \frac{1}{\|\partial_s c\|}\langle\nabla_t\partial_s c, \partial_s c\rangle$, we have the inequality $|\partial_t\|\partial_s\|| \leq \|\nabla_t\partial_s c\|$ and we get, exchanging derivatives,

$$\int_{t_0}^{t_1} \|\partial_s c(t,s,x)\|^2 \,\mathrm{d}t \leq \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\|\dot{J}(t,s,x)\|^2\,\mathrm{d}t\,,(t,0,x) \tag{A.5}$$

where $\dot{J}$ is the covariant derivative of J with respect to s. We thus have

$$\int_{t_0}^{t_1}\Psi_p(c(t,1,x)) - \Psi_p(c(t,0,x)) - \langle\nabla\Psi_p(c(t,0,x)), \partial_s c(t,0,x)\rangle \leq \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_0^1\|\dot{J}(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}t\,.$$

However, $g_t(x) = c(t,0,x)$ is a solution of $\nabla_t\partial_t c(t,0,x) = -\nabla\Psi_p(t,0,x)$, therefore, an integration by part w.r.t. t leads to

$$\int_{t_0}^{t_1}\Psi_p(c(t,1,x)) - \Psi_p(c(t,0,x)) - \langle\partial_t c(t,0,x), \nabla_t\partial_s c(t,0,x)\rangle\,\mathrm{d}t \leq \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_0^1\|\dot{J}(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}t\,.$$

Last, integrating over M and exchanging once again covariant derivatives gives

$$\int_{t_0}^{t_1}\int_M -\langle J(t,0,x), \dot{J}(t,0,x)\rangle\,\mathrm{d}x\,\mathrm{d}t \leq \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_M\int_0^1\|\dot{J}(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,.$$

Writing $f(s) = \frac{1}{2}\int_{t_0}^{t_1}\int_M\|J(t,s,x)\|^2\,\mathrm{d}t$, we want to prove $f(1) \geq f(0)$ and we have

$$-f'(0) \leq \frac{C(t_1-t_0)^2}{2\pi^2}\int_{t_0}^{t_1}\int_M\int_0^1\|\dot{J}(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,.$$

Therefore, the result is proven if we can show

$$f(1) - f(0) - f'(0) \geq \varepsilon\int_{t_0}^{t_1}\int_M\int_0^1\|\dot{J}(t,s,x)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,. \tag{A.6}$$

The left hand side can be reformulated using $f(1) - f(0) - f'(0) = \int_0^1(1-s)f''(s)\,\mathrm{d}s$ as

$$\int_{t_0}^{t_1}\int_M\int_0^1(1-s)(\|\dot{J}\|^2 - \langle R(\partial_s c, J)J, \partial_s c\rangle)\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \geq \varepsilon\int_{t_0}^{t_1}\int_M\int_0^1\|\dot{J}\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,, \tag{A.7}$$

with $\varepsilon = \frac{C(t_1-t_0)^2}{2\pi^2}$.

We now need to distinguish between two cases, the first one being when $\int_{t_0}^{t_1}\int_M\int_0^1\|\dot{J}\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \geq 1$. In this case, we use the inequality

$$\|J(t,1,x)\|^2 \leq 2\|J(t,0,x)\|^2 + 2\int_0^1\|\dot{J}(t,s,x)\|^2\,\mathrm{d}s\,, \tag{A.8}$$

in order to get

$$-\int_{t_0}^{t_1}\int_M\int_0^1(1-s)\langle R(\partial_s c, J)J, \partial_s c\rangle\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \leq \delta^2\int_{t_0}^{t_1}\int_M\int_0^1 K_{\sup}(2\|J(0)\|^2 + 2\|\dot{J}(s)\|^2)\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,, \tag{A.9}$$

where $\delta = \sup_{(x,t)\in M\times[t_0,t_1]}\|\partial_s c(t,0,x)\|$ and $K_{\sup}$ is a bound on $\max(K(y),0)$ with $K(y)$ is the maximum of the sectional curvatures at $y \in \mathcal{C}(M)$ for y in a bounded neighborhood of $\bigcup_{t\in[t_0,t_1]} g_t(M)$ which is compact. Then, there exists δ sufficiently small such that for every $(x,t) \in M\times[t_0,t_1]$,

$$\int_{t_0}^{t_1}\int_M\int_0^1(1-s)\langle R(\partial_s c, J)J, \partial_s c\rangle\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \leq 1 \leq \int_{t_0}^{t_1}\int_M\int_0^1\|\dot{J}\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,. \tag{A.10}$$

Now we study the second case, that is when $\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \leq 1$. Applying once again inequality (A.5), we obtain, using the Cauchy-Schwarz inequality,

$$\begin{aligned}\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\langle R(\partial_s c, J)J, \partial_s c\rangle\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t &\leq \varepsilon K_{\sup}\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\|J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\\ &\leq \varepsilon K_{\sup}\left(\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^4\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\right)^{1/2}\left(\int_{t_0}^{t_1}\int_M\int_0^1 \|J\|^4\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\right)^{1/2}.\end{aligned} \tag{A.11}$$

We now remark that for each t, x, the space of Jacobi fields is finite dimensional and consequently, norms are equivalent so that there exists a positive constant m that depends on t, x such that

$$\left(\int_0^1 \|\dot J\|^4\,\mathrm{d}s\right)^{1/2} \leq m\int_0^1 \|\dot J\|^2\,\mathrm{d}s \tag{A.12}$$

and

$$\left(\int_0^1 \|J\|^4\,\mathrm{d}s\right)^{1/2} \leq m\int_0^1 \|J\|^2\,\mathrm{d}s\,. \tag{A.13}$$

By compactness of $M\times[t_0,t_1]$, the constant m can be chosen independently of t, x and therefore, there exists a constant m' such that

$$\begin{aligned}\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\langle R(\partial_s c, J)J, \partial_s c\rangle\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \leq\\ \varepsilon K_{\sup}m'\left(\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\right)\left(\int_{t_0}^{t_1}\int_M\int_0^1 \|J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\right).\end{aligned} \tag{A.14}$$

Then, inequality (A.8) leads to

$$\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\langle R(\partial_s c, J)J, \partial_s c\rangle\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \;\leq\; \varepsilon K_{\sup}Cm'\left(\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\right), \tag{A.15}$$

with $M = \left(\int_{t_0}^{t_1}\int_M 2\|J(0)\|^2 + 2\int_0^1 \|\dot J(s)\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\right)$.

Let us recall that our goal is to prove the existence of $\varepsilon > 0$ such that

$$\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \geq \varepsilon\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2 + (1-s)\langle R(\partial_s c, J)J, \partial_s c\rangle\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,, \tag{A.16}$$

which, in the first case, reads

$$\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \geq 2\varepsilon\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,, \tag{A.17}$$

and in the second case

$$\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \geq \varepsilon(1+K_{\sup}Cm')\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,. \tag{A.18}$$

The existence of ε follows from the fact that the space of Jacobi fields is finite dimensional and the fact $M\times[t_0,t_1]$ is compact. It thus proves the result in the general case.

When the cone $\mathcal{C}(M)$ has non-positive sectional curvature, $K_{\sup} = 0$ therefore, we only have to prove the existence of ε such that

$$\int_{t_0}^{t_1}\int_M\int_0^1 (1-s)\|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t \geq \varepsilon\int_{t_0}^{t_1}\int_M\int_0^1 \|\dot J\|^2\,\mathrm{d}s\,\mathrm{d}x\,\mathrm{d}t\,, \tag{A.19}$$

which does not require an a priori bound on the neighborhood.

When $M = S_d(1)$, $\mathcal{C}(M)$ is flat and $\delta_0 = \infty$ and Jacobi fields are constant and the constant ε does not depend on the neighborhood and is equal to $1/2$ as in Brenier's proof. □

Acknowledgements

We would like to thank Yann Brenier and Klas Modin for stimulating discussions and a reviewer for his valuable comments which improved significantly this article.

References

[1] Luigi Ambrosio. *The Flow Associated to Weakly Differentiable Vector Fields: Recent Results and Open Problems*, pages 181–193. Springer US, Boston, MA, 2011.
[2] Vladimir Arnold. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Ann. Inst. Fourier (Grenoble)*, 16(fasc. 1):319–361, 1966.
[3] J-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.
[4] Y. Brenier. The dual least action problem for an ideal, incompressible fluid. *Archive for Rational Mechanics and Analysis*, 122(4):323–351, 1993.
[5] Yann Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.*, 44(4):375–417, 1991.
[6] Yann Brenier. Minimal geodesics on groups of volume-preserving maps and generalized solutions of the Euler equations. *Comm. Pure Appl. Math.*, 52(4):411–452, 1999.
[7] Yann Brenier. Topics on hydrodynamics and volume preserving maps. *Handbook of Mathematical Fluid Dynamics*, 2:55 – 86, 2003.
[8] Yann Brenier. Remarks on the minimizing geodesic problem in inviscid incompressible fluid mechanics. *Calc. Var. Partial Differential Equations*, 47(1-2):55–64, 2013.
[9] Alberto Bressan and Adrian Constantin. Global conservative solutions of the Camassa-Holm equation. *Arch. Ration. Mech. Anal.*, 183(2):215–239, 2007.
[10] Alberto Bressan and Massimo Fonte. An optimal transportation metric for solutions of the Camassa-Holm equation. *Methods Appl. Anal.*, 12(2):191–219, 2005.
[11] D. Burago, Y. Burago, and S. Ivanov. A course in metric geometry. *American Mathematical Soc.*, 2001.
[12] Roberto Camassa and Darryl D. Holm. An integrable shallow water equation with peaked solitons. *Phys. Rev. Lett.*, 71(11):1661–1664, 1993.
[13] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. Unbalanced Optimal Transport: Geometry and Kantorovich Formulation. *ArXiv e-prints*, August 2015.
[14] L. Chizat, B. Schmitzer, G. Peyré, and F.-X. Vialard. An Interpolating Distance between Optimal Transport and Fisher-Rao. *Found. Comp. Math.*, 2016.
[15] A. Constantin and B. Kolev. Geodesic flow on the diffeomorphism group of the circle. *Comment. Math. Helv.*, 78(4):787–804, 2003.
[16] Adrian Constantin. On the scattering problem for the Camassa-Holm equation. *R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci.*, 457(2008):953–970, 2001.
[17] Adrian Constantin and Joachim Escher. Wave breaking for nonlinear nonlocal shallow water equations. *Acta Math.*, 181(2):229–243, 1998.
[18] Adrian Constantin and David Lannes. The hydrodynamical relevance of the camassa–holm and degasperis–procesi equations. *Archive for Rational Mechanics and Analysis*, 192(1):165–186, 2008.
[19] Raphaël Danchin. A few remarks on the Camassa-Holm equation. *Differential Integral Equations*, 14(8):953–988, 2001.
[20] David G. Ebin and Jerrold Marsden. Groups of diffeomorphisms and the motion of an incompressible fluid. *Ann. of Math. (2)*, 92:102–163, 1970.
[21] T. M. Elgindi and I.-J. Jeong. Finite-time Singularity Formation for Strong Solutions to the Boussinesq System. *ArXiv e-prints*, August 2017.
[22] J. Escher and B. Kolev. Right-invariant Sobolev metrics of fractional order on the diffeomorphism group of the circle. *Journal of Geometric Mechanics*, 6(3):335 – 372, September 2014.
[23] Joachim Escher and Boris Kolev. The degasperis–procesi equation as a non-metric euler equation. *Mathematische Zeitschrift*, 269(3):1137–1153, 2011.
[24] D. S. Freed and D. Groisser. The basic geometry of the manifold of riemannian metrics and of its quotient by the diffeomorphism group. *Michigan Math. J.*, 36(3):323–344, 1989.
[25] S. Gallot. Équations différentielles caractéristiques de la sphère. *Annales scientifiques de l'Ecole Normale Superieure*, 12(2):235–267, 1979.
[26] S. Gallot, D. Hulin, and J. Lafontaine. *Riemannian Geometry*. Universitext. Springer, 2004.
[27] T. O. Gallouët and Q. Mérigot. A Lagrangian scheme for the incompressible Euler equation using optimal transport. ArXiv e-prints, May 2016.
[28] T. O. Gallouët and L. Monsaingeon. A JKO splitting scheme for Kantorovich-Fisher-Rao gradient flows, 2016.
[29] F. Gay-Balmaz, C. Tronci, and C. Vizman. Geometric dynamics on the automorphism group of principal bundles: geodesic flows, dual pairs and chromomorphism groups. *Journal of Geometric Mechanics*, 5:39–84, 2013.

[30] Katrin Grunert, Helge Holden, and Xavier Raynaud. Lipschitz metric for the periodic camassa–holm equation. *Journal of Differential Equations*, 250(3):1460 – 1492, 2011.
[31] D. D. Holm, J. E. Marsden, and T. S. Ratiu. The Euler-Poincaré equations and semidirect products with applications to continuum theories. *Adv. Math.*, 137:1–81, 1998.
[32] B. Khesin, J. Lenells, G. Misiolek, and S. C. Preston. Curvatures of Sobolev metrics on diffeomorphism groups. *Pure and Applied Mathematics Quarterly*, 9(2):291 – 332, 2013.
[33] B. Khesin, J. Lenells, G. Misiołek, and S. C. Preston. Geometry of Diffeomorphism Groups, Complete integrability and Geometric statistics. *Geom. Funct. Anal.*, 23(1):334–366, 2013.
[34] B. Khesin and R. Wendt. *The geometry of infinite-dimensional groups*, volume 51. Springer Science & Business Media, 2008.
[35] I. Kolář, P. W. Michor, and J. Slovák. *Natural operations in differential geometry*. Springer-Verlag, Berlin, 1993.
[36] S. Kondratyev, L. Monsaingeon, and D. Vorotnikov. A new optimal trasnport distance on the space of finite Radon measures. *Adv. Differential Equations*, 21(11):1117–1164, 2016.
[37] Shinar Kouranbaeva. The Camassa-Holm equation as a geodesic flow on the diffeomorphism group. *J. Math. Phys.*, 40(2):857–868, 1999.
[38] Jonatan Lenells. Conservation laws of the Camassa-Holm equation. *J. Phys. A, Math. Gen.*, 38(4):869–880, 2005.
[39] M. Liero, A. Mielke, and G. Savaré. Optimal Entropy-Transport problems and a new Hellinger-Kantorovich distance between positive measures. *ArXiv e-prints*, August 2015.
[40] M. Liero, A. Mielke, and G. Savaré. Optimal transport in competition with reaction: the Hellinger-Kantorovich distance and geodesic curves. *SIAM J. Image Analysis*, 48(4):2869–2911, 2016.
[41] J. Lott. Some geometric calculations on Wasserstein space. *Communications in Mathematical Physics*, 277(2):423–437, 2008.
[42] Xue Luo and Roman Shvydkoy. 2d homogeneous solutions to the euler equation. *Communications in Partial Differential Equations*, 40(9):1666–1687, 2015.
[43] J. Maas, M. Rumpf, C. Schönlieb, and S. Simon. A generalized model for optimal transport of images including dissipation and density modulation. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6), Apr 2015. arXiv:1504.01988.
[44] Henry P. McKean. Breakdown of the Camassa-Holm equation. *Comm. Pure Appl. Math.*, 57(3):416–418, 2004.
[45] Q. Mérigot and J.-M. Mirebeau. Minimal geodesics along volume preserving maps, through semi-discrete optimal transport. *ArXiv e-prints*, May 2015.
[46] P. W. Michor. *Topics in Differential Geometry*, volume 93 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2008.
[47] Peter W. Michor and David Mumford. Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. *Doc. Math.*, 10:217–245, 2005.
[48] G. Misiolek. Classical solutions of the periodic Camassa-Holm equation. *Geometric & Functional Analysis GAFA*, 12(5):1080–1104, 2002.
[49] Gerard Misiolek and Stephen C. Preston. Fredholm properties of riemannian exponential maps on diffeomorphism groups. *Inventiones mathematicae*, 179(1):191–227, 2010.
[50] K. Modin. Generalised Hunter-Saxton equations, optimal information transport, and factorisation of diffeomorphisms. *Journal of Geometric Analysis*, 25(2):1306–1334, April 2015.
[51] J. Moser. On the volume elements on a manifold. *Trans. Amer. Math. Soc.*, 120:286–294, 1965.
[52] D. Mumford and P. W. Michor. On Euler's equation and 'EPDiff'. *Journal of Geometric Mechanics*, 5:319 – 344, 2013.
[53] Felix Otto. The geometry of dissipative evolution equations: The porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[54] F Rezakhanlou. Optimal transport problems for contact structures, 2015.
[55] A. Trouvé and L. Younes. Metamorphoses through lie group action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005.
[56] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

CMLS, UMR 7640, École Polytechnique, FR-91128 Palaiseau Cedex.
E-mail address: thomas.gallouet@polytechnique.edu

Université Paris-Dauphine, PSL Research University, Ceremade, INRIA, Project team Mokaplan
E-mail address: fxvialard@normalesup.org

GENERALIZED COMPRESSIBLE FLOWS AND SOLUTIONS OF THE $H(\mathrm{div})$ GEODESIC PROBLEM

THOMAS GALLOUËT, ANDREA NATALE AND FRANÇOIS-XAVIER VIALARD

ABSTRACT. We study the geodesic problem on the group of diffeomorphism of a domain $M \subset \mathbb{R}^d$, equipped with the $H(\mathrm{div})$ metric. The geodesic equations coincide with the Camassa-Holm equation when $d = 1$, and represent one of its possible multi-dimensional generalizations when $d > 1$. We propose a relaxation à la Brenier of this problem, in which solutions are represented as probability measures on the space of continuous paths on the cone over M. We use this relaxation to prove that smooth $H(\mathrm{div})$ geodesics are globally length minimizing for short times. We also prove that there exists a unique pressure field associated to solutions of our relaxation. Finally, we propose a numerical scheme to construct generalized solutions on the cone and present some numerical results illustrating the relation between the generalized Camassa-Holm and incompressible Euler solutions.

1. INTRODUCTION

The $H(\mathrm{div})$ minimizing geodesic problem on the group of diffeomorphisms of a compact domain in $\mathbb{R}^d$ can be stated as follows:

Problem 1.1 ($H(\mathrm{div})$ geodesic problem)**.** *Let M be a compact domain in $\mathbb{R}^d$ and let $\mathrm{Diff}(M)$ be the group of smooth diffeomorphisms of M. Denote by ρ_0 the Lebesgue measure on M. Given $h \in \mathrm{Diff}(M)$, find a smooth curve $t \in [0,T] \mapsto \varphi_t \in \mathrm{Diff}(M)$ satisfying*

$$\varphi_0 = \mathrm{Id}\,, \quad \varphi_T = h\,, \tag{1.1}$$

and minimizing the action $\int_0^T l(u)\,\mathrm{d}t$, with Lagrangian given by

$$l(u) = \int_M |u|^2\,\mathrm{d}\rho_0 + \frac{1}{4}\int_M |\mathrm{div}\,u|^2\,\mathrm{d}\rho_0\,, \tag{1.2}$$

where $u : [0,T] \times M \to \mathbb{R}^d$ is the Eulerian velocity field defined by the equation $\partial_t\varphi_t(\cdot) = u(t,\varphi_t(\cdot))$.

Michor and Mumford proved in [31] that the $H(\mathrm{div})$ Lagrangian (1.2) defines a non-vanishing distance on the diffeomorphism group, in contrast to the L^2 case for which the metric is degenerate (i.e., there exist non trivial maps h for which the infimum of the action vanishes). Note also that in dimension $d \geq 2$ the distance induced by the H^s metric vanishes if and only if $s < 1$, as recently proved by Jerrard and Maor [23]. Local well-posedness and existence of $H(\mathrm{div})$ geodesics is guaranteed if h is close to the identity Id in a sufficiently strong topology, due to Ebin and Marsden [15].

In one dimension, the Lagrangian in (1.2) is equivalent to the square of the H^1 norm, and if we replace M by the real line, the Euler-Lagrange equations coincide with the Camassa-Holm (CH) equation. For the choice of coefficients in (1.2) the CH equation reads as follows:

$$\partial_t u - \frac{1}{4}\partial_{txx}u + 3u\partial_x u - \frac{1}{2}\partial_{xx}u\partial_x u - \frac{1}{4}u\partial_{xxx}u = 0\,. \tag{1.3}$$

This equation was shown to model shallow water waves [10], and in this context, the Lagrangian in (1.2) yields the appropriate generalization to a higher dimensional domain in $\mathbb{R}^d$ [25]. The CH equation has been intensively studied in literature, mostly because it represents an example of bi-Hamiltonian and integrable equation [16], and its smooth solution blow up in finite time in a process known as wave breaking. Furthermore, even weak solutions cannot be defined globally [32], their blow up being related to the emergence of non-injective Lagrangian maps.

Date: June 22, 2022.

Problem 1.1 was recentently reinterpreted as an L^2 geodesic problem on the cone over M [17], establishing therefore a link with the incompressible Euler equations which share a similar structure, as shown in the pioneering work of Arnold [4].

1.1. **Contributions.** In this paper we construct a relaxation of problem 1.1 inspired by Brenier's relaxation of the incompressible Euler equation. We call the minimizers of such a relaxation *generalized solutions*. This approach allows us to obtain several results on the $H(\mathrm{div})$ geodesic problem. In particular, we show that:

- if M is convex, smooth $H(\mathrm{div})$ geodesics are globally length-minimizing for short times and in any dimension (theorem 6.4). This result generalizes the one in [17], which was only valid on the circle of unit radius S^1_1 and it was local otherwise;
- on the torus $S^1 \times S^1$, there exists $h \in \mathrm{Diff}(S^1 \times S^1)$ such that the infimum of the action in problem 1.1 cannot be attained by any smooth flow (theorem 7.11); on the contrary, for the same h there exists a generalized solution that arises as the limit of a minimizing sequence of smooth flows (theorem 7.12);
- there exists a unique pressure field in the sense of distribution associated with generalized solutions (theorem 5.3). To the best of the authors' knowledge, the pressure field we consider is a variable that has not been studied before in the literature of the CH equation or the $H(\mathrm{div})$ geodesic problem (see remark 3.1). It appears however as a natural variable in the generalized setting.

1.2. **The a-b-c metric.** The Lagrangian in (1.2) is a particular instance of a class of right-invariant Lagrangians on the diffeomorphism group of M considered in [24], which for $d = 3$ can be written as

$$l(u) = a \int_M |u|^2 \,\mathrm{d}\rho_0 + b \int_M |\mathrm{div}\, u|^2 \,\mathrm{d}\rho_0 + c \int_M |\mathrm{curl}\, u|^2 \,\mathrm{d}\rho_0 \,, \tag{1.4}$$

where a, b, c are positive constants. Such Lagrangians give rise to several important nonlinear evolution equations, including the EPDiff equation for the H^1 Sobolev norm of vector fields and the Euler-α model [19, 20], both of which have also been regarded as possible multi-dimensional versions of the CH equation.

1.3. **The $\dot{H}^1$ metric and the Hunter-Saxton equation.** The Hunter-Saxton equation [22, 24] corresponds to choosing $a = c = 0$ in (1.4), in which case the metric is denoted by $\dot{H}^1$, and in one dimension. Lenells provided a simple description of the solutions to this equation as geodesic flows on the infinite-dimensional sphere of L^2 functions with constant norm [27]. This was established by constructing an explicit isometry between the group of orientation-preserving diffeomorphism of the circle S^1 (modulo rotations) and a subset of the sphere, given by the map

$$\varphi \mapsto \sqrt{\partial_x \varphi} \,. \tag{1.5}$$

This geometric point of view was particularly fruitful and led to a number of important results, namely a bound on the diameter of the diffeomeorphism group endowed with the $\dot{H}^1$ metric; that its curvature is positive and constant; that geodesics are gobally length-minimizing. Lenell's interpretation still holds when the domain is a higher dimensional manifold, as showed in [24], which allowed the authors to prove complete integrability of the geodesic equations and that all solutions blow up in finite time. The simplifications that arise for the Hunter-Saxton equation are related to the fact that the $\dot{H}^1$ descends to a non-degenerate metric on the space of densities via the isometry (1.5). This however does not apply to the full H^1 metric or the $H(\mathrm{div})$ metric because of the presence of the transport term, given by the L^2 norm of the velocity.

1.4. **The L^2 metric and the incompressible Euler equations.** The L^2 metric was used by Arnold [4] to interpret the solutions to the incompressible Euler equations as geodesic curves on the group of volume-perserving diffeomorphisms $\mathrm{Diff}_{\rho_0}(M)$. As for the $H(\mathrm{div})$ problem, the existence of length-minimizing geodesics is guaranteed a priori only in a sufficiently strong topology [15]. In fact, Shnirelman proved that the infimum is generally not attained when $d \geq 3$ and that even when $d = 2$ there exist final configurations h which cannot be connected to the identity map with finite action [35]. This motivated Brenier to adopt an extrinsic approach, viewing

$$\mathrm{Diff}_{\rho_0}(M) \subset \{\varphi \in L^2(M; M) \,;\, \varphi_{\#}\rho_0 = \rho_0\} \tag{1.6}$$

and reinterpreting incompressible flows as probability measures $\boldsymbol{\mu}$ on $\Omega(M)$, the space of continuous curves on the domain $\mathrm{x} : t \in [0,T] \to x_t \in M$, satisfying

$$(e_t)_{\#}\boldsymbol{\mu} = \rho_0\,, \tag{1.7}$$

where $e_t : \Omega(M) \to M$ is the evaluation map at time t defined by $e_t(\mathrm{x}) = x_t$, and ρ_0 is normalized so that $\rho_0(M) = 1$. In this interpretation, the marginals $(e_0, e_t)_{\#}\boldsymbol{\mu}$ are probability measures on $M \times M$ and describe how particles move and spread their mass across the domain. Classical deterministic solutions, i.e. curves of volume preserving diffeomorphisms $t \mapsto \varphi_t$, correspond to the case where the marginals $(e_0, e_t)_{\#}\boldsymbol{\mu}$ are concentrated on the graph of φ_t. Then, equation (1.7) is equivalent to the incompressibility constraint $\varphi_{\#}\rho_0 = \rho_0$. Note also that formulating the incompressibility via the push-forward of φ, we allow changes in orientation. The minimization problem in terms of generalized flows consists in minimizing the action

$$\int_{\Omega(M)} \int_0^T \frac{1}{2}|\dot{x}_t|^2 \,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{x}) \tag{1.8}$$

among generalized incompressible flows, with the constraint $(e_0, e_T)_{\#}\boldsymbol{\mu} = (\mathrm{Id}, h)_{\#}\rho_0$. Brenier proved that smooth solutions correspond to the unique minimizers of the generalized problem for sufficiently small times, and are therefore globally length-minimizing [7]. On the other hand, for any coupling there exists a unique pressure, defined as a distribution (but that can actually be defined as a function [1]), associated with generalized solutions.

1.5. **The $H(\mathrm{div})$ metric as an L^2 cone metric.** The link between the incompressible Euler equation and the $H(\mathrm{div})$ geodesic problem was established in [17], where it was proven that problem 1.1 can be reformulated as a geodesic problem for the L^2 cone metric (see equation (3.3); see also section 2.2 for the cone metric structure) on a subgroup of the diffeomorphism group of $M \times \mathbb{R}_{>0}$. More precisely, Lagrangian flows are represented by time dependent automorphisms on $M \times \mathbb{R}_{>0}$, i.e. maps in the form

$$(x, r) \in M \times \mathbb{R}_{>0} \mapsto (\varphi(x), \lambda(x)r) \in M \times \mathbb{R}_{>0}\,, \tag{1.9}$$

where $\varphi : M \to M$ and $\lambda : M \to \mathbb{R}_{>0}$, satisfying

$$\varphi_{\#}(\lambda^2\rho_0) = \rho_0\,. \tag{1.10}$$

This condition relates φ and λ by requiring $\lambda = \sqrt{|\mathrm{Jac}(\varphi)|}$. Importantly, in this picture we cannot capture the blow up of solutions as induced by peakon collisions, as in this case the Jacobian would locally vanish. In addition, the metric space $M \times \mathbb{R}_{>0}$ equipped with the cone metric is not complete. We are then led to work with the cone $\mathcal{C} = (M \times \mathbb{R}_{\geq 0})/(M \times \{0\})$, which allows us to represent solutions with vanishing Jacobian by paths on the cone reaching the apex.

Interestingly, the decoupling between the Lagrangian flow map and its Jacobian has also been used in [26] to construct global weak solutions of the CH equation. However, in their case, one continues solutions after the blowup by allowing the square root of the Jacobian to become negative, which does not occur in the formulation described above.

1.6. **Generalized compressible flows and unbalanced optimal transport.** By analogy with the incompressible Euler case, we reformulate the $H(\mathrm{div})$ geodesic problem using generalized flows interpreted as probability measures $\boldsymbol{\mu}$ on the space $\Omega(\mathcal{C})$ of continuous paths on the cone $\mathrm{z} : t \in [0,T] \to z_t = [x_t, r_t] \in \mathcal{C}$. Our relaxed formulation consists in minimizing the action

$$\int_{\Omega(\mathcal{C})} \int_0^T |\dot{z}_t|^2_{g_{\mathcal{C}}} \,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \tag{1.11}$$

among generalized flows satisfying appropriate constraints enforcing a generalized version of (1.10) and the coupling between initial and final times. Choosing the correct form for such constraints is not trivial. It is the first contribution of the paper to define a formulation that allows to prove existence of minimizers while retaining uniqueness for short time in the smooth setting.

It should be noted that the cone construction has been developed and used extensively in [28, 12] in order to characterize the metric side of the Wasserstein-Fisher-Rao (WFR) distance (which is also called Hellinger-Kantorovich distance) on the space of positive Radon measures. In fact, as noted in [17] this has the same relation to the CH equation as the Wasserstein L^2 distance does to the incompressible Euler equations. In the geodesic problem associated the

WFR distance a relation similar to (1.10) is used to prescribe the initial and final density. The resulting problem coincides with the so-called optimal entropy-transport problem, a widespread form of unbalanced optimal transport based on the Kullback-Leibler divergence [12, 11, 28].

Taking advantage of the optimal transport point of view, we propose a numerical scheme based on multi-marginal optimal transport and entropic regularization [13, 5, 6] to simulate the solutions of problem 1.1.

1.7. **Structure of the paper.** In section 2, we introduce the notations and the needed background. In section 3, we recall the L^2 variational formulation of the $H(\mathrm{div})$ geodesic problem.

In section 4 we introduce our relaxation, for which we prove existence of solutions as generalized compressible flows. We also show that our generalized solutions can be decomposed into two parts, one of which involves directly the cone singularity. When this latter is not trivial, it implies the appearance and disappearance of mass in the domain; we refer to such minimizers as singular solutions.

In section 5 we prove that for any boundary conditions, there always exists a unique pressure field defined as a distribution on $(0,T)\times M$ associated with any given generalized solution.

In section 6 we prove that smooth solutions of the $H(\mathrm{div})$ geodesic equations are the unique minimizers of our generalized model for sufficiently short times. This proves that such solutions are also globally length-minimizing on $\mathrm{Diff}(M)$.

In section 7 we show that for $d \geq 2$, singular solutions emerge naturally from the continuous formulation for appropriate (smooth) boundary conditions. This proves that the infimum of the action in problem 1.1 may not be attained. We construct approximations for such minimizers using a particular form of peakon collision which arises from the Hunter-Saxton equation.

Finally, in section 8 we construct a numerical scheme based on entropic regularization and Sinkhorn algorithm to compute generalized $H(\mathrm{div})$ geodesics.

2. Notation and preliminaries

In this section, we describe the notation and some basic results used throughout the paper. Because of the similarities between our setting and the one of [28], we will adopt a similar notation for the cone construction and the measure theory objects we will employ.

2.1. **Function spaces.** Given two metric spaces X and Y, we denote by $C^0(X;Y)$ the space of continuous functions $f : X \to Y$, by $C^0(X)$ the space of real-valued continuous functions $f : X \to \mathbb{R}$, and by $C^0_b(X)$ the subset of bounded functions $f \in C^0(X)$. If X is compact $C^0(X)$ is a Banach space with respect to the sup norm $\|\cdot\|_{C^0}$. The set of Lipschitz continuous function on X is denoted by $C^{0,1}(X)$ and the associated seminorm and norm are given respectively by

$$|f|_{C^{0,1}} \coloneqq \sup_{x,y\in X, x\neq y} \frac{|f(x)-f(y)|}{d_X(x,y)}, \quad \|f\|_{C^{0,1}} \coloneqq \|f\|_{C^0} + |f|_{C^{0,1}}, \tag{2.1}$$

where d_X denotes the distance function on X.

If X is a subset of $\mathbb{R}^d$, we use standard notation for Sobolev spaces on X. In particular, $H(\mathrm{div};X)$ or simply $H(\mathrm{div})$ denotes the space of L^2 vector fields $f : X \to \mathbb{R}^d$ whose divergence $\mathrm{div}(f)$ is in L^2, with squared norm given by $\|f\|^2_{L^2} + \|\mathrm{div}\, f\|^2_{L^2}$ (which is equivalent to (1.2)). Moreover, we denote by $\mathrm{Diff}(X)$ the group of smooth diffeomorphisms of X.

2.2. **The cone and metric structures.** Throughout the paper, M will denote the closure of an open bounded set in $\mathbb{R}^d$ with Lipschitz boundary. Occasionally, we will also consider the case $M = S^1_R \coloneqq \mathbb{R}/2\pi R\mathbb{Z}$ the circle of radius R, or $M = T^2_{R_1,R_2} \coloneqq S^1_{R_1} \times S^1_{R_2}$ the torus with radii $R_1, R_2 > 0$. We will denote by g the Euclidean metric tensor on M and with $|\cdot|$ the Euclidean norm. We denote by $\mathcal{C} \coloneqq (M \times \mathbb{R}_{\geq 0})/(M \times \{0\})$ the cone over M. A point on the cone is an equivalence class $p = [x, r]$, where the equivalence relation is given by

$$(x_1, r_1) \sim (x_2, r_2) \Leftrightarrow (x_1, r_1) = (x_2, r_2) \text{ or } r_1 = r_2 = 0\,. \tag{2.2}$$

The distinguished point of the cone $[x, 0]$ is the apex of $\mathcal{C}$ and it is denoted by o. Every point on the cone different from the apex can be identified with a couple (x, r) where $x \in M$ and $r \in \mathbb{R}_{>0}$.

Moreover, we fix a point $\bar{x} \in M$ and we introduce the projections $\pi_x : \mathcal{C} \to M$ and $\pi_r : \mathcal{C} \to \mathbb{R}_{\geq 0}$ defined by

$$\pi_x([x,r]) = \begin{cases} x & \text{if } r > 0\,, \\ \bar{x} & \text{if } r = 0\,, \end{cases} \qquad \pi_r([x,r]) = r\,. \tag{2.3}$$

We endow the cone with the metric tensor $g_{\mathcal{C}} = r^2 g + \mathrm{d}r^2$, defined on $M \times \mathbb{R}_{>0}$. We denote the associated norm by $|\cdot|_{g_{\mathcal{C}}}$. All differential operators, e.g., ∇, div and so on, are computed with respect to the Euclidean metric on M; we will use the superscript $g_{\mathcal{C}}$ to indicate when they are computed with respect to the cone metric. The distance on the cone $d_{\mathcal{C}} : \mathcal{C} \times \mathcal{C} \to \mathbb{R}_{\geq 0}$ is given by

$$d_{\mathcal{C}}([x_1,r_1],[x_2,r_2])^2 = r_1^2 + r_2^2 - 2r_1 r_2 \cos(\min(|x_1 - x_2|, \pi)) \tag{2.4}$$

(see, for example, definition 3.6.16 in [9]). The closed subset of the cone composed of points below a given radius $R > 0$ is denoted by $\mathcal{C}_R$, or more precisely

$$\mathcal{C}_R \coloneqq \{[x,r] \in \mathcal{C}\,;\, r \leq R\}\,. \tag{2.5}$$

Given an interval $I \subset \mathbb{R}$, we denote by $C^0(I;\mathcal{C})$ and $AC(I;\mathcal{C})$ the spaces of, respectively, continuous and absolutely continuous curves $\mathrm{z} : t \in I \to z_t \in \mathcal{C}$. We will generally use the notation

$$\mathrm{x} : t \in I \to x_t = \pi_x(z_t) \in M\,, \qquad \mathrm{r} : t \in I \to r_t = \pi_r(z_t) \in [0,+\infty)\,, \tag{2.6}$$

so that $\mathrm{z} = [\mathrm{x},\mathrm{r}]$ and $z_t = [x_t, r_t]$. Note that if z is continuous (resp. absolutely continuous), then so is the path r but not x. However, x is continuous (resp. locally absolutely continuous) when restricted to the open set $\{t \in I; r_t > 0\}$. Then, if we define $\dot{\mathrm{z}} : t \in I \to \dot{z}_t \in \mathbb{R}^{d+1}$ by

$$\dot{z}_t = \begin{cases} (\dot{x}_t, \dot{r}_t) & \text{if } r_t > 0 \text{ and the derivatives exist,} \\ (0,0) & \text{otherwise,} \end{cases} \tag{2.7}$$

we have that $|\dot{z}_t|_{g_{\mathcal{C}}}$ coincides for a.e. $t \in I$ with the metric derivative of z with respect to the distance $d_{\mathcal{C}}$ [28]. We denote by $AC^p(I;\mathcal{C})$ the space of absolutely continuous curves such that $|\dot{z}|_{g_{\mathcal{C}}} \in L^p(I)$. Then, the following variational formula for the distance function holds

$$d_{\mathcal{C}}(p,q)^2 = \inf\left\{ \int_0^1 |\dot{z}_t|^2_{g_{\mathcal{C}}}\,\mathrm{d}t\,;\, \mathrm{z} \in AC^2([0,1];\mathcal{C})\,,\, z_0 = p\,,\, z_1 = q \right\}. \tag{2.8}$$

We will extensively use the class of homogeneous functions on the cone defined as follows. A function $f : \mathcal{C}^n \to \mathbb{R}$ is p-homogeneous (in the radial direction) if for any constant $\lambda > 0$ and for all n-tuples $([x_1,r_1],\dots,[x_n,r_n]) \in \mathcal{C}^n$,

$$f([x_1,\lambda r_1],\dots,[x_n,\lambda r_n]) = \lambda^p f([x_1,r_1],\dots,[x_n,r_n])\,. \tag{2.9}$$

In particular, a p-homogeneous function $f : \mathcal{C} \to \mathbb{R}$ satisfies $f([x,\lambda r]) = \lambda^p f([x,r])$. Similarly, a functional $\sigma : C^0(I;\mathcal{C}) \to \mathbb{R}$ is p-homogeneous if for any constant $\lambda > 0$ and for any path $\mathrm{z} \in C^0(I;\mathcal{C})$,

$$\sigma(t \mapsto [x_t, \lambda r_t]) = \lambda^p \sigma(\mathrm{z})\,, \tag{2.10}$$

where $\mathrm{z} : t \in I \to [x_t, r_t] \in \mathcal{C}$.

2.3. **Measure theoretic background.** Let X be a Polish space, i.e. a complete and separable metric space. We denote by $\mathcal{M}(X)$ the set of non-negative and finite Borel measures on X. The set of probability measures on X is denoted by $\mathcal{P}(X)$. Let Y be another Polish space and $F : X \to Y$ a Borel map. Given a measure $\mu \in \mathcal{M}(X)$ we denote by $F_\#\mu \in \mathcal{M}(Y)$ the push-forward measure defined by $(F_\#\mu)(A) \coloneqq \mu(F^{-1}(A))$ for any Borel set $A \subset Y$. Given a Borel set $B \subset X$ we let $\mu \llcorner B$ the restriction of μ to B defined by $\mu \llcorner B(C) \coloneqq \mu(B \cap C)$ for any Borel set $C \subseteq X$. Note that we will generally use bold symbols to denote measures on product spaces, e.g., $\boldsymbol{\mu} \in \mathcal{M}(X \times \ldots \times X)$.

We endow $\mathcal{P}(X)$ with the topology induced by narrow convergence, which is the convergence in duality with the space of real-valued continuous bounded functions $C^0_b(X)$. In other words, a sequence $\mu_n \in \mathcal{P}(X)$, $n \in \mathbb{N}$, is said to converge narrowly to $\mu \in \mathcal{P}(X)$ if for any $f \in C^0_b(X)$

$$\lim_{n \to +\infty} \int_X f\,\mathrm{d}\mu_n = \int_X f\,\mathrm{d}\mu\,. \tag{2.11}$$

In practice, however, to check for narrow convergence it is sufficient to verify equation (2.11) for all bounded Lipschitz continuous functions. With such a topology, $\mathcal{P}(X)$ can be identified with a subset of $[C_b^0(X)]^*$ with the weak-* topology (see Remark 5.1.2 in [3]). In addition, given a lower semi-continuous function $f : X \to \mathbb{R} \cup \{+\infty\}$, bounded from below, the functional $\mathcal{F} : \mathcal{P}(X) \to \mathbb{R} \cup \{+\infty\}$ defined by

$$\mathcal{F}(\mu) := \int_X f \, \mathrm{d}\mu \tag{2.12}$$

is also lower-semicontinuous (see Lemma 1.6 in [34]) .

As usual in this setting, we will use Prokhorov's theorem for a characterization of compact subsets of $\mathcal{P}(X)$ endowed with the narrow topology.

Theorem 2.1 (Prokhorov's theorem)**.** *A set $\mathcal{K} \subset \mathcal{P}(X)$ is relatively sequentially compact in $\mathcal{P}(X)$ if and only if it is tight, i.e. for any $\epsilon > 0$ there exists a compact set $K_\epsilon \subset X$ such that $\mu(X \setminus K_\epsilon) < \epsilon$ for any $\mu \in \mathcal{K}$.*

We also need a criterion to pass to the limit when computing integrals of unbounded functions: for this will use the concept of uniform integrability. Given a set $\mathcal{K} \subset \mathcal{P}(X)$, we say that a Borel function $f : X \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ is uniformly integrable with respect to $\mathcal{K}$ if for any $\epsilon > 0$ there exists a $k > 0$ such that, for any $\mu \in \mathcal{K}$,

$$\int_{f(x)>k} f(x) \, \mathrm{d}\mu(x) < \epsilon \, . \tag{2.13}$$

Lemma 2.2 (Lemma 5.1.7 in [3])**.** *Let $\{\mu_n\}_{n\in\mathbb{N}}$ be a sequence in $\mathcal{P}(X)$ narrowly convergent to $\mu \in \mathcal{P}(X)$ and let $f \in C^0(X)$. If $|f|$ is uniformly integrable with respect to the set $\{\mu_n\}_{n\in\mathbb{N}}$ then*

$$\lim_{n\to+\infty} \int_X f \, \mathrm{d}\mu_n = \int_X f \, \mathrm{d}\mu \, . \tag{2.14}$$

For a fixed $T > 0$, we will denote by $\Omega(X) := C^0([0, T]; X)$ the space of continuous paths on X. This is a Polish space so that we can use the tools introduced in this section also for probability measures $\boldsymbol{\mu} \in \mathcal{P}(\Omega(X))$. We call such probability measures *generalized flows* or also *dynamic plans*. When $X = \mathcal{C}$, where $\mathcal{C}$ is the cone over $M \subset \mathbb{R}^d$, we will often use Ω to denote $\Omega(\mathcal{C})$.

Since we will work with homogeneous functions on the cone, we also introduce the space of probability measures $\mathcal{P}_p(X)$, for $p > 0$, defined by

$$\mathcal{P}_p(X) := \left\{ \mu \in \mathcal{P}(X) \, ; \, \int_X d_X(x, \bar{x})^p \, \mathrm{d}\mu(x) < +\infty \text{ for some } \bar{x} \in X \right\} . \tag{2.15}$$

Then, if $\boldsymbol{\mu} \in \mathcal{P}_p(\mathcal{C}^n)$ it is easy to verify that any locally-bounded p-homogeneous function on $\mathcal{C}^n$ is $\boldsymbol{\mu}$-integrable.

Finally, we will denote by ρ_0 the Lebesgue measure on M normalized so that $\rho_0(M) = 1$.

3. The variational formulation on the cone

In this section we describe the geometric structure underlying problem 1.1 using the group of automorphisms of the cone. Such a formulation was introduced in [17] and it was used to interpret the CH equation as an incompressible Euler equations on the cone. In this section we will only focus on smooth solutions, but we will later use the variational interpretation presented here to guide the construction of generalized $H(\mathrm{div})$ geodesics. We will keep the discussion formal at this stage and we will use some standard geometric tools and notation commonly adopted in similar contexts.

For any $\varphi \in \mathrm{Diff}(M)$ and $\lambda \in C^\infty(M; \mathbb{R}_{>0})$, we let $(\varphi, \lambda) : \mathcal{C} \to \mathcal{C}$ be the map defined by $(\varphi, \lambda)([x, r]) = [\varphi(x), \lambda(x)r]$. The automorphism group $\mathrm{Aut}(\mathcal{C})$ is the collection of such maps, i.e.

$$\mathrm{Aut}(\mathcal{C}) = \{(\varphi, \lambda) : \mathcal{C} \to \mathcal{C}; \, \varphi \in \mathrm{Diff}(M), \, \lambda \in C^\infty(M; \mathbb{R}_{>0})\} \, . \tag{3.1}$$

The group composition law is given by

$$(\varphi, \lambda) \cdot (\psi, \mu) = (\varphi \circ \psi, (\lambda \circ \psi)\mu) \, , \tag{3.2}$$

the identity element is $(\mathrm{Id}, 1)$, where Id is the identity map on M, and the inverse is given by $(\varphi,\lambda)^{-1} = (\varphi^{-1}, \lambda^{-1}\circ\varphi^{-1})$. The tangent space of $\mathrm{Aut}(\mathcal{C})$ at (φ,λ) is denoted by $T_{(\varphi,\lambda)}\mathrm{Aut}(\mathcal{C})$ and it can be identified with the space of vector fields $C^\infty(M;\mathbb{R}^{d+1})$. The collection all the tangent spaces is the tangent bundle $T\mathrm{Aut}(\mathcal{C})$. We endow $T\mathrm{Aut}(\mathcal{C})$ with the $L^2(M;T\mathcal{C})$ metric inherited from $g_{\mathcal{C}}$. This is defined as follows: given $(\dot\varphi,\dot\lambda)\in T_{(\varphi,\lambda)}\mathrm{Aut}(\mathcal{C})$,

$$\|(\dot\varphi,\dot\lambda)\|^2_{L^2(M;T\mathcal{C})} \coloneqq \int_M (\lambda^2|\dot\varphi|^2 + \dot\lambda^2)\,\mathrm{d}\rho_0\,, \tag{3.3}$$

where $|\cdot|$ is the Euclidean norm and ρ_0 is the Lebesgue measure on M normalized so that $\rho_0(M)=1$.

In [17] the authors found that the $H(\mathrm{div})$ geodesic equations on M coincide with the geodesic equation on the subgroup $\mathrm{Aut}_{\rho_0}(\mathcal{C})\subset\mathrm{Aut}(\mathcal{C})$ defined as follows:

$$\mathrm{Aut}_{\rho_0}(\mathcal{C}) \coloneqq \{(\varphi,\lambda)\in\mathrm{Aut}(\mathcal{C})\,;\,\varphi_\#(\lambda^2\rho_0)=\rho_0\}\,. \tag{3.4}$$

In other words, the group $\mathrm{Aut}_{\rho_0}(\mathcal{C})$ can be regarded as the configuration space for the $H(\mathrm{div})$ geodesic problem in the same way as the $\mathrm{Diff}_{\rho_0}(M)$ is the configuration space for the incompressible Euler equations, with

$$\mathrm{Diff}_{\rho_0}(M) \coloneqq \{\varphi\in\mathrm{Diff}(M)\,;\,\varphi_\#\rho_0=\rho_0\}\,. \tag{3.5}$$

In order to see this, we first observe that the $L^2(M;T\mathcal{C})$ metric is right invariant when restricted to $\mathrm{Aut}_{\rho_0}(\mathcal{C})$.In particular, for any $(\psi,\vartheta)\in\mathrm{Aut}_{\rho_0}(\mathcal{C})$, consider the right translation map $R_{(\psi,\vartheta)}:\mathrm{Aut}_{\rho_0}(\mathcal{C})\to\mathrm{Aut}_{\rho_0}(\mathcal{C})$ defined by $R_{(\psi,\vartheta)}(\varphi,\lambda)=(\varphi,\lambda)\cdot(\psi,\vartheta)$. Its tangent map at (φ,λ) is given by

$$TR_{(\psi,\mu)}(\dot\varphi,\dot\lambda) = (\dot\varphi\circ\psi, (\dot\lambda\circ\psi)\vartheta). \tag{3.6}$$

Then, it is easy to check that $\|TR_{(\psi,\vartheta)}(\dot\varphi,\dot\lambda)\|^2_{L^2(M;T\mathcal{C})} = \|(\dot\varphi,\dot\lambda)\|^2_{L^2(M;T\mathcal{C})}$. Geodesics on $\mathrm{Aut}_{\rho_0}(\mathcal{C})$ correspond to stationary paths on $T\mathrm{Aut}_{\rho_0}(\mathcal{C})$ for the action functional

$$\int_0^T L((\varphi,\lambda),(\dot\varphi,\dot\lambda))\,\mathrm{d}t \tag{3.7}$$

for a given $T>0$, where the Lagrangian $L((\varphi,\lambda),(\dot\varphi,\dot\lambda)) = \|(\dot\varphi,\dot\lambda)\|^2_{L^2(M;T\mathcal{C})}$. Define the Eulerian velocities $(u,\alpha)\in T_{(\mathrm{Id},1)}\mathrm{Aut}(\mathcal{C})$ by

$$(u,\alpha) = TR_{(\varphi,\lambda)^{-1}}(\dot\varphi,\dot\lambda) = (\dot\varphi\circ\varphi^{-1}, (\dot\lambda\lambda^{-1})\circ\varphi^{-1})\,. \tag{3.8}$$

In terms of these variables the constraint $\varphi_\#(\lambda^2\rho_0)=\rho_0$ becomes $2\alpha=\operatorname{div}u$, since for any $f\in C^\infty(M)$,

$$0 = \frac{\mathrm{d}}{\mathrm{d}t}\int_M f\,\mathrm{d}\varphi_\#(\lambda^2\rho_0) = \int_M(-\operatorname{div}u + 2\alpha)f\,\mathrm{d}\rho_0\,. \tag{3.9}$$

Moreover, by right invariance,

$$L((\varphi,\lambda),(\dot\varphi,\dot\lambda)) = L((\mathrm{Id},1),(u,\alpha)) = \int_M |u|^2 + \frac{1}{4}|\operatorname{div}u|^2\,\mathrm{d}\rho_0\,, \tag{3.10}$$

which is the $H(\mathrm{div})$ norm. Note that the coefficient $1/4$ is directly related to the choice of $g_{\mathcal{C}}$ as cone metric. Using different coefficients in $g_{\mathcal{C}}$ we can obtain the general form of the Lagrangian in equation (1.4) with $c=0$. Introducing P as the Lagrange multiplier for the constraint $\varphi_\#(\lambda^2\rho_0)=\rho_0$, the Euler-Lagrange equations associated with L read as follows

$$\begin{cases} \lambda\ddot\varphi + 2\dot\lambda\dot\varphi + \frac{1}{2}\lambda\nabla P\circ\varphi = 0\,,\\ \ddot\lambda - \lambda|\dot\varphi|^2 + \lambda P\circ\varphi = 0\,, \end{cases} \tag{3.11}$$

which can be expressed in terms of (u,α) by composing both equation with φ^{-1}, yielding

$$\begin{cases} \dot u + \nabla_u u + 2u\alpha = -\frac{1}{2}\nabla P\,,\\ \dot\alpha + u\cdot\nabla\alpha + \alpha^2 - |u|^2 = -P\,. \end{cases} \tag{3.12}$$

In one dimension, using the relation $\alpha=\operatorname{div}u/2$, this finally gives the CH equation for u, i.e. equation (1.3).

Remark 3.1. *Note that in the literature for the CH equation the "pressure field" is sometimes defined in a different way so that, when M is one-dimensional, the first equation in (3.12) can be written as*

$$\partial_t u + u\partial_x u = -\partial_x p\,, \tag{3.13}$$

for an appropriate function $p = (\mathrm{Id} - \frac{1}{4}\partial_{xx})^{-1}(u^2 + \frac{1}{8}(\partial_x u)^2)$ (see, e.g., [21]). Throughout the paper we will instead intend by pressure the Lagrange multiplier P considered above, which is related to p by

$$P = 2p - u^2\,. \tag{3.14}$$

In section 5 we will prove that the pressure P is uniquely defined for minimizers of the $H(\mathrm{div})$ geodesic problem. On the other hand, we cannot prove the same result for p, since the flow φ and as a consequence the velocity field u may not be well-defined for generalized solutions (see the explicit examples of generalized solutions in section 7).

4. The generalized $H(\mathrm{div})$ geodesic formulation

In section 3 we recalled the interpretation of the $H(\mathrm{div})$ metric as an L^2 metric on $\mathrm{Aut}_{\rho_0}(\mathcal{C})$, which is defined in (3.4). Then, we can reformulate problem 1.1 as follows:

Problem 4.1 ($H(\mathrm{div})$ geodesic problem on the cone)**.** *Given a diffemorphism $h \in \mathrm{Diff}(M)$, find a smooth curve $t \in [0,T] \mapsto (\varphi_t, \lambda_t) \in \mathrm{Aut}_{\rho_0}(\mathcal{C})$ satisfying*

$$(\varphi_0, \lambda_0) = (\mathrm{Id}, 1)\,, \qquad (\varphi_T, \lambda_T) = (h, \sqrt{|\mathrm{Jac}(h)|})\,, \tag{4.1}$$

and minimizing the action in equation (3.7).

As already pointed out in [17], there is a remarkable analogy between this problem and Arnold's geometric interpretation of the incompressible Euler equations [4]. This suggests that adapting to this problem Brenier's concept of generalized flow could be a successful strategy to characterize its minimizers. In this section we follow this path and in particular we formulate the generalized $H(\mathrm{div})$ geodesic problem and prove existence of solutions.

By generalized flow or dynamic plan we mean a probability measure on the space of continuous paths of the cone $\boldsymbol{\mu} \in \mathcal{P}(\Omega)$. This is a generalization for curves on the automorphism group since for any smooth curve $(\varphi, \lambda) : t \in [0,T] \to (\varphi_t, \lambda_t) \in \mathrm{Aut}_{\rho_0}(\mathcal{C})$, we can associate the generalized flow $\boldsymbol{\mu}$ defined by

$$\boldsymbol{\mu} = (\varphi, \lambda)_{\#}\rho_0\,, \tag{4.2}$$

where we recall that the Lebesgue measure ρ_0 is normalized in such a way that $\rho_0(M) = 1$. More explicitly, for any Borel functional $\mathcal{F} : \Omega \to \mathbb{R}$,

$$\int_\Omega \mathcal{F}(\mathrm{z})\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \int_M \mathcal{F}([\varphi(x), \lambda(x)])\mathrm{d}\rho_0(x)\,, \tag{4.3}$$

where $[\varphi(x), \lambda(x)] : t \in [0,T] \to [\varphi_t(x), \lambda_t(x)] \in \mathcal{C}$.

The condition $(\varphi_t)_{\#}\lambda_t^2\rho_0 = \rho_0$ is equivalent to requiring $\lambda_t = \sqrt{|\mathrm{Jac}(\varphi_t)|}$. We want to generalize this condition for arbitrary $\boldsymbol{\mu} \in \mathcal{P}(\Omega)$. Let $e_t : \Omega \to \mathcal{C}$ be the evaluation map at time $t \in [0,T]$. Then, if $\boldsymbol{\mu}$ is defined as in (4.2), we have

$$\mathfrak{h}_t^2(\boldsymbol{\mu}) \coloneqq (\pi_x)_{\#}[r^2(e_t)_{\#}\boldsymbol{\mu}] = \rho_0\,. \tag{4.4}$$

In fact, for any $f \in C^0(M)$,

$$\begin{aligned}
\int_M f\,\mathrm{d}\mathfrak{h}_t^2(\boldsymbol{\mu}) &= \int_\Omega f(x_t) r_t^2\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\
&= \int_\Omega f(x_t) r_t^2\,\mathrm{d}(\varphi,\lambda)_{\#}\rho_0 \\
&= \int_M f\circ\varphi_t\lambda_t^2\,\mathrm{d}\rho_0 \\
&= \int_M f\,\mathrm{d}(\varphi_t)_{\#}\lambda_t^2\rho_0 \\
&= \int_M f\,\mathrm{d}\rho_0\,,
\end{aligned} \tag{4.5}$$

where for any path z and any time t, $x_t \coloneqq \pi_x(z_t)$ and $r_t \coloneqq \pi_r(z_t)$. By similar calculations, we also obtain

$$(e_0, e_T)_{\#}\boldsymbol{\mu} = \boldsymbol{\gamma} \coloneqq [(\varphi_0, \lambda_0), (\varphi_T, \lambda_T)]_{\#}\rho_0 \,. \tag{4.6}$$

In other words, enforcing the boundary conditions in the generalized setting boils down to constraining a certain marginal of $\boldsymbol{\mu}$ to coincide with a given *coupling plan* $\boldsymbol{\gamma}$ on the cone, i.e. a probability measure in $\mathcal{P}(\mathcal{C} \times \mathcal{C})$.

Consider now the energy functional $\mathcal{A} : \Omega \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ defined by

$$\mathcal{A}(\mathrm{z}) \coloneqq \begin{cases} \int_0^T |\dot{z}_t|^2_{gc} \,\mathrm{d}t & \text{if } \mathrm{z} \in AC^2([0,T];\mathcal{C})\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{4.7}$$

Setting $\mathcal{F}(\mathrm{z}) = \mathcal{A}(\mathrm{z})$ in (4.3) we obtain the $H(\mathrm{div})$ action expressed in Lagrangian coordinates. This motivates the following definition for the generalized $H(\mathrm{div})$ geodesic problem.

Problem 4.2 (Generalized $H(\mathrm{div})$ geodesic problem)**.** *Given a coupling plan on the cone $\boldsymbol{\gamma} \in \mathcal{P}_2(\mathcal{C}^2)$, find the dynamic plan $\boldsymbol{\mu} \in \mathcal{P}(\Omega)$ satisfying: the* homogeneous coupling constraint

$$\int_\Omega f(z_0, z_T)\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \int_{\mathcal{C}^2} f\,\mathrm{d}\boldsymbol{\gamma}\,, \tag{4.8}$$

for all 2-homogeneous continuous functions $f : \mathcal{C}^2 \to \mathbb{R}$; the homogeneous marginal constraint

$$\int_\Omega \int_0^T f(t, x_t) r_t^2 \,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \int_M \int_0^T f(t,x)\,\mathrm{d}t\,\mathrm{d}\rho_0(x) \quad \forall\, f \in C^0([0,T]\times M)\,; \tag{4.9}$$

and minimizing the action

$$\mathcal{A}(\boldsymbol{\mu}) \coloneqq \int_\Omega \mathcal{A}(\mathrm{z})\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z})\,. \tag{4.10}$$

We remark three basic facts on this formulation:

- we substituted the constraint in (4.4) by its time-integrated version in equation (4.9) as this form will be easier to manipulate in the following. However, the two formulations are equivalent when restricting to generalized flows with finite action (see lemma 4.3);
- we replaced the strong coupling constraint (4.6) by a weaker version, which is always implied by the former as long as $\boldsymbol{\gamma} \in \mathcal{P}_2(\mathcal{C}^2)$ and in particular when $\boldsymbol{\gamma}$ is deterministic, i.e. when it is induced by a diffeomorphism as in equation (4.6);
- we allow for general coupling plans in $\mathcal{P}_2(\mathcal{C}^2)$ so that the integral on the right-hand side of equation (4.8) is finite. However, we will mostly be interested in the case where the coupling is deterministic.

The first of the points above is made explicit in the following lemma, whose proof is postponed to the appendix.

Lemma 4.3. *For any generalized flow $\boldsymbol{\mu}$ with $\mathcal{A}(\boldsymbol{\mu}) < +\infty$ and satisfying the homogeneous coupling constraint in equation (4.8), the homogeneous marginal constraint in equation (4.9) is equivalent to the constraint*

$$\mathfrak{h}^2_t(\boldsymbol{\mu}) = \rho_0 \tag{4.11}$$

for all $t \in [0,T]$.

The main result of this section is contained in the following proposition, which states that generalized $H(\mathrm{div})$ geodesics are well-defined as solutions of problem (4.2).

Proposition 4.4 (Existence of minimizers)**.** *Provided that there exists a dynamic plan $\boldsymbol{\mu}^*$ such that $\mathcal{A}(\boldsymbol{\mu}^*) < +\infty$, the minimum of the action in problem 4.2 is attained.*

Before providing the proof of proposition 4.4, we introduce a useful rescaling operation which will allow us to preserve the homogenous constraint when passing to the limit using sequences of narrowly convergent dynamic plans. Such an operation was introduced in [28] in order to deal with the analogous problem arising from the formulation of optimal entropy-transport (i.e. unbalanced transport) on the cone. Adapting the notation in [28] to our setting, we define for a functional $\theta : \Omega \to \mathbb{R}$,

$$\mathrm{prod}_\theta(\mathrm{z}) \coloneqq (t \in [0,T] \mapsto [x_t, r_t/\theta(\mathrm{z})])\,. \tag{4.12}$$

Then, given a dynamic plan $\boldsymbol{\mu}$, if $\theta(\mathrm{z}) > 0$ for $\boldsymbol{\mu}$-almost any path z, we can define the dilation map

$$\mathrm{dil}_{\theta,2}(\boldsymbol{\mu}) \coloneqq \mathrm{prod}_{\theta\#}(\theta^2\boldsymbol{\mu})\,. \tag{4.13}$$

Since the constraints in equations (4.8) and (4.9) are 2-homogeneous in the radial coordinate r, they are invariant under the dilation map, meaning that if $\boldsymbol{\mu}$ satisfies (4.8) and (4.9), also $\mathrm{dil}_{\theta,2}(\boldsymbol{\mu})$ does. For the same reason, we also have

$$\mathcal{A}(\mathrm{dil}_{\theta,2}(\boldsymbol{\mu})) = \mathcal{A}(\boldsymbol{\mu})\,. \tag{4.14}$$

The map $\mathrm{dil}_{\theta,2}$ performs a *rescaling* on the measure $\boldsymbol{\mu}$ in the sense specified by the following lemma.

Lemma 4.5. *Given a measure $\boldsymbol{\mu} \in \mathcal{M}(\Omega)$ and a 1-homogeneous functional $\sigma : \Omega \to \mathbb{R}$ such that $\sigma(\mathrm{z}) > 0$ for $\boldsymbol{\mu}$-almost every path* z*, suppose that*

$$C \coloneqq \left(\int_\Omega (\sigma(\mathrm{z}))^2\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z})\right)^{1/2} < +\infty\,; \tag{4.15}$$

if $\tilde{\boldsymbol{\mu}} = \mathrm{dil}_{\sigma/C,2}(\boldsymbol{\mu})$ then $\tilde{\boldsymbol{\mu}}(\Omega) = 1$ and

$$\tilde{\boldsymbol{\mu}}(\{\mathrm{z} \in \Omega\,;\, \sigma(\mathrm{z}) = C\}) = 1\,. \tag{4.16}$$

Proof. We prove this by direct calculation. Let $\theta \coloneqq \sigma/C$. By 1-homogeneity of σ, for $\boldsymbol{\mu}$-almost every path z

$$\sigma(\mathrm{prod}_\theta(\mathrm{z})) = \frac{\sigma(\mathrm{z})}{|\theta(\mathrm{z})|} = C\,. \tag{4.17}$$

Then,

$$\begin{aligned}\int_{\{\mathrm{z}\in\Omega\,;\,\sigma(\mathrm{z})=C\}} \mathrm{d}\tilde{\boldsymbol{\mu}}(\mathrm{z}) &= \int_{\{\mathrm{z}\in\Omega\,;\,\sigma(\mathrm{z})=C\}} \mathrm{dprod}_{\theta\#}(\theta^2\boldsymbol{\mu})(\mathrm{z}) \\ &= \int_{\{\mathrm{z}\in\Omega\,;\,\sigma(\mathrm{prod}_\theta(\mathrm{z}))=C\}} \theta^2 \mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\ &= \frac{1}{C^2}\int_\Omega (\sigma(\mathrm{z}))^2 \mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = 1\,.\end{aligned} \tag{4.18}$$

By similar calculations we also have $\tilde{\boldsymbol{\mu}}(\Omega) = 1$. □

Besides the rescaling operator and lemma 4.5, we will also need the following result which will allow us to construct suitable minimizers of the action in problem 4.2.

Lemma 4.6. *The set of measures with uniformly bounded action $\mathcal{A}(\boldsymbol{\mu}) \leq C$ and satisfying the homogeneous constraint in equation (4.9) is relatively sequentially compact for the narrow topology.*

Proof. Due to Therorem 2.1, it is sufficient to prove that sequences of admissible measures are tight. For a given path z with $\mathcal{A}(\mathrm{z}) \leq Q$, for all $0 \leq s \leq t \leq T$,

$$d_{\mathcal{C}}(z_s, z_t) \leq \int_s^t |\dot{z}_{t^*}|_{g_{\mathcal{C}}}\,\mathrm{d}t^* \leq Q^{1/2}|t-s|^{1/2}\,, \tag{4.19}$$

which implies that level sets of $\mathcal{A}(\mathrm{z})$ are equicontinuous. Consider now the set

$$\Omega_R \coloneqq \Omega(\mathcal{C}_R) = \{\mathrm{z} \in \Omega\,;\, \forall\, t \in [0,T]\,,\, r_t \leq R\}\,; \tag{4.20}$$

For any $Q > 0$, the set $\{\mathrm{z} \in \Omega_R\,;\, \mathcal{A}(\mathrm{z}) \leq Q\}$ is also equicontinuous; moreover, since paths in this set are bounded at any time, it is contained in a compact subset of Ω, by the Ascoli-Arzelà theorem.

In order to use such sets to prove tightness we need to be able to control the measure of $\Omega \setminus \Omega_R$. In particular, we now show that there exists a constant $C' > 0$ such that

$$\boldsymbol{\mu}(\Omega \setminus \Omega_R) \leq \frac{C'}{R^2}\,. \tag{4.21}$$

In order to show this, consider first the following set of paths

$$\{\mathrm{z} \in \Omega\,;\, \forall\, t \in [0,T]\,,\, r_t > R\}\,. \tag{4.22}$$

Integrating the constraint in equation (4.9) over such a set with $f = 1$, we obtain

$$\boldsymbol{\mu}(\{\mathrm{z} \in \Omega\,;\, \forall\, t \in [0,T]\,,\, r_t > R\}) \leq \frac{1}{R^2}\,. \tag{4.23}$$

Now, consider the set

$$\{\mathrm{z} \in \Omega \setminus \Omega_R\,;\, \mathcal{A}(\mathrm{z}) < Q\}\,. \tag{4.24}$$

For any z in this set, there exists $t^* \in [0,T]$ such that $r_{t^*} > R$. Moreover, since $\mathcal{A}(\mathrm{z}) < Q$, by equation (4.19),

$$|r_t - r_{t^*}| \leq d_{\mathcal{C}}(z_t, z_{t^*}) \leq Q^{1/2}|t - t^*|^{1/2}\,, \tag{4.25}$$

for all $t \in [0,T]$, which implies

$$r_t \geq r_{t^*} - Q^{1/2}T^{1/2} > R - Q^{1/2}T^{1/2}. \tag{4.26}$$

In particular, if $Q \leq R^2/(4T)$, then $r_t > R/2$, or also

$$\{\mathrm{z} \in \Omega \setminus \Omega_R\,;\mathcal{A}(\mathrm{z}) < Q\} \subseteq \{\mathrm{z} \in \Omega\,;\forall\, t \in [0,T]\,,\, r_t > R/2\}\,. \tag{4.27}$$

Therefore, if $Q \leq R^2/(4T)$,

$$\begin{aligned}\boldsymbol{\mu}(\Omega \setminus \Omega_R) &\leq \boldsymbol{\mu}((\Omega \setminus \Omega_R) \cap \{\mathrm{z}\,;\, \mathcal{A}(\mathrm{z}) < Q\}) + \boldsymbol{\mu}(\{\mathrm{z}\,;\, \mathcal{A}(\mathrm{z}) \geq Q\}) \\ &\leq \boldsymbol{\mu}(\{\mathrm{z} \in \Omega\,;\forall\, t \in [0,T]\,,\, r_t > R/2\}) + \frac{C}{Q} \\ &\leq \frac{4}{R^2} + \frac{C}{Q}\,.\end{aligned} \tag{4.28}$$

Taking $Q = R^2/(4T)$, we deduce that

$$\boldsymbol{\mu}(\Omega \setminus \Omega_R) \leq \frac{4(CT+1)}{R^2}\,, \tag{4.29}$$

which proves equation (4.21).

Recall that $\{\mathrm{z} \in \Omega_R\,;\, \mathcal{A}(\mathrm{z}) \leq Q\}$ is contained in a compact set for any $Q > 0$ and $R > 0$. For any $\epsilon > 0$, set $R = (8(CT+1)/\epsilon)^{1/2}$. For any admissible $\boldsymbol{\mu}$, we have

$$\begin{aligned}\boldsymbol{\mu}(\Omega\setminus\{\mathrm{z} \in \Omega_R\,;\, \mathcal{A}(\mathrm{z}) \leq 2C\epsilon^{-1}\}) &\leq \boldsymbol{\mu}(\Omega\setminus\{\mathrm{z}\,;\mathcal{A}(\mathrm{z}) \leq 2C\epsilon^{-1}\}) + \boldsymbol{\mu}(\Omega\setminus\Omega_R) \\ &\leq \frac{\epsilon}{2C}\int_\Omega \mathcal{A}(\mathrm{z})\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) + \frac{\epsilon}{2} \leq \epsilon\,,\end{aligned} \tag{4.30}$$

which proves tightness. □

We are now ready to prove existence of optimal solutions for the generalized $H(\mathrm{div})$ geodesic problem. Note that due to lemma 4.6, we can always extract a converging subsequence from any minimizing sequence of problem 4.2. However, this approach fails to produce a minimizer, since convergence in the narrow topology is not sufficient to pass the constraints to the limit. Note in particular that this is also true if we enforce the strong coupling constraint (4.6) instead of its homogeneous version in (4.8). On the other hand, by choosing this latter as coupling constraint, we can use lemma 4.5 to construct an appropriate minimizing sequence for which all constraints pass to the limit. We follow this strategy in the proof of proposition 4.4 below.

Proof of proposition 4.4. The functional $\mathcal{A}(\mathrm{z})$ is lower semi-continuous; hence so is $\mathcal{A}(\boldsymbol{\mu})$. Consider a minimizing sequence $\boldsymbol{\mu}_n$ with $n \in \mathbb{N}$. By assumption we can take $\mathcal{A}(\boldsymbol{\mu}_n) \leq C$ for all $n \in \mathbb{N}$. Let $\mathrm{o} : t \in [0,T] \to o \in \mathcal{C}$ the path on the cone assigning to every time the apex of the cone o. Let $\boldsymbol{\mu}_n^{\mathrm{o}} \coloneqq \boldsymbol{\mu}_n \llcorner \Omega^{\mathrm{o}} \in \mathcal{M}(\Omega)$ the restriction of $\boldsymbol{\mu}_n$ to $\Omega^{\mathrm{o}} \coloneqq \Omega \setminus \{\mathrm{o}\}$. Such an operation preserves both the action and the constraints.

Let $\sigma : \Omega \to \mathbb{R}$ be the 1-homogeneous functional defined by

$$\sigma(\mathrm{z}) \coloneqq \left(r_0^2 + r_T^2 + \int_0^T r_t^2\,\mathrm{d}t\right)^{1/2}. \tag{4.31}$$

For any $\boldsymbol{\mu}_n^{\mathrm{o}}$ in the sequence, we obviously have that $\sigma(\mathrm{z}) > 0$ for $\boldsymbol{\mu}_n^{\mathrm{o}}$-almost every path. Moreover, since $\boldsymbol{\mu}_n^{\mathrm{o}}$ satisfies both the homogeneous marginal and coupling constraint, for all $n \in \mathbb{N}$,

$$\int_\Omega \sigma(\mathrm{z})^2\,\mathrm{d}\boldsymbol{\mu}_n(\mathrm{z}) = T + 2\,. \tag{4.32}$$

Hence we can apply lemma 4.5 and define a sequence $\tilde{\boldsymbol{\mu}}_n \in \mathcal{P}(\Omega)$ by $\tilde{\boldsymbol{\mu}}_n := \mathrm{dil}_{\sigma/\sqrt{T+2},2}\boldsymbol{\mu}_n^{\mathrm{o}}$. In particular, for all $n \in \mathbb{N}$, $\tilde{\boldsymbol{\mu}}_n$ is concentrated on the set of paths such that $\sigma(\mathrm{z}) = \sqrt{T+2}$, i.e.

$$\tilde{\boldsymbol{\mu}}_n\left(\left\{\mathrm{z} \in \Omega\,;\, r_0^2 + r_T^2 + \int_0^T r_t^2\,\mathrm{d}t = T+2\right\}\right) = 1\,. \tag{4.33}$$

Moreover, $\tilde{\boldsymbol{\mu}}_n$ satisfies the homogeneous constraint and the coupling constraint, since these are both 2-homogeneous in the radial direction, and for the same reason $\mathcal{A}(\tilde{\boldsymbol{\mu}}_n) = \mathcal{A}(\boldsymbol{\mu}_n) \leq C$. This is enough to apply lemma 4.6; thus, we can extract a subsequence $(\tilde{\boldsymbol{\mu}}_n)_n \rightharpoonup \tilde{\boldsymbol{\mu}}_\infty \in \mathcal{P}(\Omega)$.

We now show that for any $f \in C^0([0,T]\times M)$ the functional

$$\mathcal{F}(\mathrm{z}) := \int_0^T |f(t,x_t)| r_t^2\,\mathrm{d}t \tag{4.34}$$

is uniformly integrable with respect to the sequence $(\tilde{\boldsymbol{\mu}}_n)_n$, that is, for any $\epsilon > 0$ there exists a constant $K > 0$ such that for all $n \in \mathbb{N}$

$$\int_{\Omega,\mathcal{F}(\mathrm{z})>K} \mathcal{F}(\mathrm{z})\,\mathrm{d}(\tilde{\boldsymbol{\mu}}_n)_n(\mathrm{z}) < \epsilon\,. \tag{4.35}$$

It is sufficient to consider the case $\|f\|_{C^0} = 1$, because the case $\|f\|_{C^0} = 0$ is trivial and otherwise we can always rescale the functional by dividing it by $\|f\|_{C^0}$. Recall the definition of the functional σ in equation (4.31); we have

$$\int_{\Omega,\mathcal{F}(\mathrm{z})>K} \mathcal{F}(\mathrm{z})\,\mathrm{d}(\tilde{\boldsymbol{\mu}}_n)_n(\mathrm{z}) \leq \int_{\Omega,\sigma(\mathrm{z})^2>K} \sigma(\mathrm{z})^2\mathrm{d}(\tilde{\boldsymbol{\mu}}_n)_n(\mathrm{z})\,. \tag{4.36}$$

However, by equation (4.33) the right-hand side is zero if $K > T+2$, which proves uniform integrability. Hence, using lemma 2.2, we deduce that $\tilde{\boldsymbol{\mu}}_\infty$ satisfies the homogeneous marginal constraint. Similarly, we can deduce that $\tilde{\boldsymbol{\mu}}_\infty$ also satisfies the homogeneous coupling constraint since $(e_0,e_T)_\#(\tilde{\boldsymbol{\mu}}_n)_n$ is concentrated on $\mathcal{C}_R^2$ with $R = \sqrt{T+2}$; hence it is an optimal solution of problem 4.2. □

Remark 4.7. *Given $h \in \mathrm{Diff}(M)$, set $\boldsymbol{\gamma} = [(\mathrm{Id},1),(h,\sqrt{|\mathrm{Jac}(h)|})]_\#\rho_0$. For such a coupling, there always exists a dynamic plan $\boldsymbol{\mu}^*$ such that $\mathcal{A}(\boldsymbol{\mu}^*) < +\infty$. This is constructed explicitly in lemma 7.1. Therefore, the minimum of the action in problem 4.2 is attained.*

In general, we cannot ensure that there exists a minimizer $\boldsymbol{\mu}$ of problem 4.2 satisfying the *strong coupling constraint*:

$$(e_0,e_T)_\#\boldsymbol{\mu} = \boldsymbol{\gamma}\,. \tag{4.37}$$

However, we can easily obtain a characterization for the existence of such minimizers when $\boldsymbol{\gamma}$ is deterministic. This relies on the following crucial result which allows us to isolate the part of the solution involving the cone singularity.

Proposition 4.8. *Suppose that $\boldsymbol{\gamma} = [(\mathrm{Id},1),(h,\sqrt{|\mathrm{Jac}(h)|})]_\#\rho_0$. Any measure $\boldsymbol{\mu} \in \mathcal{M}(\Omega)$ satisfying the homogeneous coupling constraint admits the decomposition*

$$\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}^0\,, \tag{4.38}$$

where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega\,;\, r_0 \neq 0\,,\, r_T \neq 0\}$ and $\tilde{\boldsymbol{\mu}}^0 = \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega\,;\, r_0 = r_T = 0\}$. Moreover $\tilde{\boldsymbol{\mu}}^1 := \mathrm{dil}_{r_0,2}\tilde{\boldsymbol{\mu}}$ satisfies the strong coupling constraint, i.e. $(e_0,e_T)_\#\tilde{\boldsymbol{\mu}}^1 = \boldsymbol{\gamma}$.

Proof. Let $\boldsymbol{\mu} \in \mathcal{M}(\Omega)$ be any dynamic plan satisfying the homogeneous coupling constraint. We decompose $\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}^0$ where

$$\tilde{\boldsymbol{\mu}} := \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega\,;\, r_0 \neq 0\}\,, \qquad \tilde{\boldsymbol{\mu}}^0 := \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega\,;\, r_0 = 0\}\,. \tag{4.39}$$

Consider the 1-homogeneous functional $\tilde{\sigma}(\mathrm{z}) : \Omega \to \mathbb{R}$ defined by $\tilde{\sigma}(\mathrm{z}) = r_0$. Clearly $\tilde{\sigma}(\mathrm{z}) > 0$ for $\tilde{\boldsymbol{\mu}}$-almost every path z. Moreover, we have

$$\int_\Omega (\tilde{\sigma}(\mathrm{z}))^2\,\mathrm{d}\tilde{\boldsymbol{\mu}}(\mathrm{z}) = \int_\Omega r_0^2\,\mathrm{d}\tilde{\boldsymbol{\mu}}(\mathrm{z}) = 1\,. \tag{4.40}$$

Hence, by lemma 4.5, the measure $\tilde{\boldsymbol{\mu}}^1 := \mathrm{dil}_{r_0,2}\tilde{\boldsymbol{\mu}} \in \mathcal{P}(\Omega)$ is concentrated on paths such that $r_0 = 1$. Moreover, $\tilde{\boldsymbol{\mu}}^0 + \tilde{\boldsymbol{\mu}}^1$ still satisfies the homogeneous coupling constraint and in particular, for any $\alpha \in [0, 2)$,

$$\begin{aligned}\int_\Omega r_T^\alpha \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) &= \int_\Omega r_0^{2-\alpha} r_T^\alpha \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) \\ &= \int_\Omega r_0^{2-\alpha} r_T^\alpha \,\mathrm{d}(\tilde{\boldsymbol{\mu}}^0 + \tilde{\boldsymbol{\mu}}^1)(\mathrm{z}) \\ &= \int_M \zeta^\alpha \,\mathrm{d}\rho_0 \,.\end{aligned} \tag{4.41}$$

Taking the limit for $\alpha \to 2$, by the dominated convergence theorem,

$$\int_\Omega r_T^2 \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) = \int_M \zeta^2 \,\mathrm{d}\rho_0 = 1 \,. \tag{4.42}$$

In turn, this implies that

$$\int_\Omega r_T^2 \,\mathrm{d}\tilde{\boldsymbol{\mu}}^0(\mathrm{z}) = 0 \,, \tag{4.43}$$

which means that $\tilde{\boldsymbol{\mu}}^0$-almost every path z has $r_T = 0$. This proves that $\tilde{\boldsymbol{\mu}}^0 = \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega \,;\, r_0 = r_T = 0\}$ and that $\tilde{\boldsymbol{\mu}}$ satisfies the homogeneous coupling constraint.

Next, we prove that $(e_0, e_T)_\# \tilde{\boldsymbol{\mu}}^1 = \boldsymbol{\gamma}$. For any $g \in C^0(M^2)$ we can take $f = g r_0^2$ in equation (4.8) yielding

$$\int_\Omega g(x_0, x_T) \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) = \int_M g(x, h(x)) \,\mathrm{d}\rho_0(x) \,. \tag{4.44}$$

Similarly, letting $\zeta := \sqrt{|\mathrm{Jac}(h)|}$,

$$\begin{aligned}\int_\Omega (r_T - \zeta(x_0))^2 \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) &= \int_\Omega (r_T^2 + \zeta(x_0)^2 - 2\zeta(x_0) r_T) \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) \\ &= \int_\Omega (r_T^2 + r_0^2 \zeta(x_0)^2 - 2\zeta(x_0) r_0 r_T) \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) \\ &= 2\int_M \zeta(x)^2 \mathrm{d}\rho_0(x) - 2\int_M \zeta(x)^2 \mathrm{d}\rho_0(x) = 0 \,,\end{aligned} \tag{4.45}$$

which means that for $\tilde{\boldsymbol{\mu}}^1$-almost every path $r_T = \zeta(x_0)$. Then, for any continuous bounded function $f : \mathcal{C}^2 \to \mathbb{R}$, we have

$$\begin{aligned}\int_\Omega f(z_0, z_T) \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) &= \int_\Omega f([x_0, 1], [x_T, \zeta(x_0)]) \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) \\ &= \int_M f([x, 1], [\varphi(x), \zeta(x)]) \,\mathrm{d}\rho_0(x) \,,\end{aligned} \tag{4.46}$$

which proves the second part of the proposition. Finally, we must also have $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega \,;\, r_0 \neq 0 \,,\, r_T \neq 0\}$, since by definition of the dilation map

$$\int_{\{\mathrm{z}\in\Omega \,;\, r_T = 0\}} r_0^2 \,\mathrm{d}\tilde{\boldsymbol{\mu}} = \int_{\{\mathrm{z}\in\Omega \,;\, r_T = 0\}} r_0^2 \,\mathrm{d}\tilde{\boldsymbol{\mu}}^1 = \tilde{\boldsymbol{\mu}}^1(\{\mathrm{z} \in \Omega \,;\, r_T = 0\}) = 0 \,. \tag{4.47}$$

□

Remark 4.9. *It should be noted that proposition 4.8 can be proved also if the coupling constraint in equation (4.8) is enforced only for homogeneous functions $f \in C^0(\mathcal{C}^2)$ in the form $f(z_0, z_1) = g(x_0, x_1) r_0^{2-\alpha} r_1^\alpha$ and $\alpha \in [0, 2]$, for example. Nonetheless, if we defined the constraint in this way, given the fact that $\tilde{\boldsymbol{\mu}}^1$ satisfies the strong coupling constraint, we would still retrieve that (when the coupling is deterministic) $\boldsymbol{\mu}$ satisfies the coupling constraint with respect to any homogeneous function.*

Corollary 4.10 (Existence of minimizers satisfying the strong coupling constraint)**.** *Suppose that $\boldsymbol{\gamma} = [(\mathrm{Id}, 1), (h, \sqrt{|\mathrm{Jac}(h)|})]_\# \rho_0$ and let $\boldsymbol{\mu} \in \mathcal{M}(\Omega)$ (not necessarily a probability measure) be a minimizer of problem 4.2. Then, if*

$$\boldsymbol{\mu}(\{\mathrm{z} \in \Omega \,;\, r_0 = r_T = 0\}) = 0 \,, \tag{4.48}$$

the measure $\boldsymbol{\mu}$ can be rescaled (in the sense of lemma 4.5) to a minimizer satsifying the strong coupling constraint.

The proofs of proposition 4.4 and 4.8 give us several insights on the nature of the generalized solutions of the $H(\mathrm{div})$ geodesic problem. First of all, it is evident that such solutions can only be unique up to rescaling. In fact, since all constraints are homogeneous and preserved by rescaling, given one minimizer one can generate others using the dilation map as in lemma 4.5. In addition, if the coupling is deterministic, even using rescaling, in principle one might not be able to find a minimizer satisfying the coupling constraint in the classical sense. By proposition 4.8, this happens if all minimizers charge paths which start and end at the apex of the cone and are not trivial. In this case the optimal solutions use the the apex to enforce the homogeneous marginal constraint on some time interval contained in $(0,T)$. We will refer to such minimizers as *singular solutions* since they involve the cone singularity. More precisely:

Definition 4.11 (Singular generalized $H(\mathrm{div})$ geodesics)**.** A singular solution of the generalized $H(\mathrm{div})$ geodesic problem is a minimizer $\boldsymbol{\mu} \in \mathcal{P}(\Omega)$ such that

$$\boldsymbol{\mu}(\{\mathrm{z} \in \Omega \setminus \{\mathrm{o}\}\,;\, r_0 = r_T = 0\}) > 0\,, \tag{4.49}$$

where $\mathrm{o} : t \in [0,T] \to o \in \mathcal{C}$.

Proposition 4.8 can also help us visualize such solutions. In fact, for deterministic boundary conditions, to any singular minimizer $\boldsymbol{\mu}$ we can still associate a measure $\tilde{\boldsymbol{\mu}}^1 = \mathrm{dil}_{r_0,2}\tilde{\boldsymbol{\mu}}$ which satisfies the strong coupling constraint but not necessarily the homogeneous marginal constraint. In section 7 we will construct some specific examples of singular minimizers, which will provide some intuition on their meaning.

5. Existence and uniqueness of the pressure

In the previous section, we proved existence of minimizers of the generalized $H(\mathrm{div})$ geodesic problem. In general, given that all constraints are homogeneous, such minimizers are only defined up to rescaling. However, even using rescaling, it might not always be possible to find a minimizer that satisfies the strong coupling constraint. Here, we show that independently of this, the pressure field P in (3.12) is uniquely defined as a distribution for any given deterministic coupling constraint. This reproduces a similar result proved by Brenier for the incompressible Euler case [8].

The idea is to extend the set of admissible generalized flows in order to define appropriate variations of the action. By analogy to the Euler case, we consider dynamic plans whose homogeneous marginals are not the Lebesgue measure ρ_0, but are sufficiently close to it. Given a dynamic plan $\boldsymbol{\nu} \in \mathcal{P}(\Omega)$ we denote by $\rho^{\boldsymbol{\nu}} : [0,T] \times M \to \mathbb{R}$ the function defined by

$$\rho^{\boldsymbol{\nu}}(t,\cdot) \coloneqq \frac{\mathrm{d}\mathfrak{h}_t^2\boldsymbol{\nu}}{\mathrm{d}\rho_0}\,, \tag{5.1}$$

for any $t \in [0,T]$. For an admissible generalized flow $\boldsymbol{\nu}$, $\rho^{\boldsymbol{\nu}} = 1$. Dynamic plans $\boldsymbol{\nu}$ with $\rho^{\boldsymbol{\nu}} \neq 1$ correspond to generalized automorphisms of the cone with a mismatch between the radial variable and the Jacobian of the flow map on the base space.

Definition 5.1 (Almost diffeomorphisms)**.** A generalized almost diffeomorphism is a probability measure $\boldsymbol{\nu} \in \mathcal{P}(\Omega)$ such that $\rho^{\boldsymbol{\nu}} \in C^{0,1}([0,T] \times M)$ and

$$\|\rho^{\boldsymbol{\nu}} - 1\|_{C^{0,1}([0,T]\times M)} \leq \frac{1}{2}\,. \tag{5.2}$$

For any $\rho \in C^{0,1}([0,T] \times M)$ with $\rho > 0$, let $\Phi^\rho : \Omega \to \Omega$ be the map defined by

$$\Phi^\rho(\mathrm{z}) \coloneqq (t \in [0,T] \mapsto [x_t, r_t\sqrt{\rho(t,x_t)}] \in \mathcal{C})\,. \tag{5.3}$$

We use this map in the following proposition, which is the equivalent of proposition 2.1 in [8] and justifies our choice for the space of densities in definition 5.1.

Proposition 5.2. *Fix a $\rho \in C^{0,1}([0,T] \times M)$ such that*

$$\|\rho - 1\|_{C^{0,1}} \leq \frac{1}{2}\,, \qquad \rho(0,\cdot) = \rho(1,\cdot) = 1\,. \tag{5.4}$$

Then, given any dynamic plan $\boldsymbol{\mu} \in \mathcal{P}(\Omega)$ with finite action $\mathcal{A}(\boldsymbol{\mu}) < +\infty$, satisfying the homogeneous constraint in equation (4.9), i.e. $\rho^{\boldsymbol{\mu}} = \rho_0$, and the coupling constraint (4.8), the dynamic plan $\boldsymbol{\nu} := \Phi^{\rho}_{\#}\boldsymbol{\mu} \in \mathcal{P}(\Omega)$ still satisfies the coupling constraint and we have $\rho^{\boldsymbol{\nu}} = \rho$; moreover,

$$\mathcal{A}(\boldsymbol{\nu}) \leq \mathcal{A}(\boldsymbol{\mu}) + \|\rho - 1\|_{C^{0,1}} \left(\frac{T}{2} + \mathcal{A}(\boldsymbol{\mu}) \right) + |\rho - 1|^2_{C^{0,1}} (T + \mathcal{A}(\boldsymbol{\mu}))\,. \tag{5.5}$$

Proof. The fact that $\rho^{\boldsymbol{\nu}} = \rho$ and that $\boldsymbol{\nu}$ satisfies the coupling constraint follows from direct computation. As for equation (5.5), observe that $\boldsymbol{\mu}$-almost every path is absolutely continuous and that the map $([x, r], t) \in \mathcal{C} \times [0, T] \mapsto r\sqrt{\rho(t, x)} \in \mathbb{R}_{\geq 0}$ is Lipschitz. Then, for $\boldsymbol{\mu}$-almost every path z the curve $t \in [0, T] \mapsto r_t\sqrt{\rho(t, x_t)} \in \mathbb{R}_{\geq 0}$ is also absolutely continuous and we have

$$\begin{aligned} \mathcal{A}(\boldsymbol{\nu}) &= \int_\Omega \int_0^T \mathcal{A}(\Phi^\rho(\mathrm{z}))\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\ &= \int_\Omega \int_0^T \rho(t, x_t)|\dot{z}_t|^2_{g_{\mathcal{C}}} + r_t\dot{r}_t\partial_t(\rho(t, x_t)) + r_t^2(\partial_t\sqrt{\rho(t, x_t)})^2\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\ &\leq \|\rho\|_{C^0}\mathcal{A}(\boldsymbol{\mu}) + \int_\Omega \int_0^T r_t\dot{r}_t\partial_t(\rho(t, x_t)) + r_t^2(\partial_t\sqrt{\rho(t, x_t)})^2\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z})\,. \end{aligned} \tag{5.6}$$

Moreover,

$$\begin{aligned} \int_\Omega \int_0^T r_t\dot{r}_t\partial_t(\rho(t, x_t))\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) &\leq |\rho - 1|_{C^{0,1}} \int_\Omega \int_0^T r_t|\dot{r}_t|(1 + |\dot{x}_t|)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\ &\leq |\rho - 1|_{C^{0,1}} \left(\frac{T}{2} + \mathcal{A}(\boldsymbol{\mu}) \right), \end{aligned} \tag{5.7}$$

and similarly, since $\rho \geq 1/2$,

$$\begin{aligned} \int_\Omega \int_0^T r_t^2(\partial_t\sqrt{\rho(t, x_t)})^2\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) &\leq \frac{1}{2}\int_\Omega \int_0^T r_t^2(\partial_t(\rho(t, x_t)))^2\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\ &\leq \frac{1}{2}|\rho - 1|^2_{C^{0,1}} \int_\Omega \int_0^T r_t^2(1 + |\dot{x}_t|)^2\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) \\ &\leq |\rho - 1|^2_{C^{0,1}}(T + \mathcal{A}(\boldsymbol{\mu}))\,. \end{aligned} \tag{5.8}$$

Reinserting these estimates into equation (5.6) we obtain (5.5). □

Consider now the following space

$$B_0 := \{\rho \in C^{0,1}([0, T] \times M)\,;\, \rho(0, \cdot) = \rho(1, \cdot) = 0\}\,, \tag{5.9}$$

which we regard as a Banach space with the $C^{0,1}$ norm. The following theorem shows that we can define the pressure as an element $P \in B_0^*$ and it is the analogue of Theorem 6.2 in [2].

Theorem 5.3. *Let $\boldsymbol{\mu}^*$ be a minimizer for the generalized $H(\mathrm{div})$ geodesic problem such that $\mathcal{A}(\boldsymbol{\mu}^*) < +\infty$. Then, there exists $P \in B_0^*$ such that*

$$\langle P, \rho^{\boldsymbol{\nu}} - 1\rangle \leq \mathcal{A}(\boldsymbol{\nu}) - \mathcal{A}(\boldsymbol{\mu}^*)\,, \tag{5.10}$$

for all generalized almost diffeomorphisms $\boldsymbol{\nu}$ satisfying the coupling constraint (4.8).

Proof. First of all, observe that for any generalized almost diffeomorphism $\boldsymbol{\nu}$ satisfying the coupling constraint,

$$\rho^{\boldsymbol{\nu}}(0, \cdot) = \rho^{\boldsymbol{\nu}}(1, \cdot) = 1\,; \tag{5.11}$$

hence $\rho^{\boldsymbol{\nu}} - 1 \in B_0$ and the pairing in equation (5.10) is well defined. Now, consider the convex set $C := \{\tilde{\rho} \in B_0; \|\tilde{\rho}\|_{C^{0,1}} \leq \frac{1}{2}\}$ and the functional $\phi : B_0 \to \mathbb{R}^+ \cup \{+\infty\}$ defined by

$$\phi(\tilde{\rho}) := \begin{cases} \inf\{\mathcal{A}(\boldsymbol{\nu})\,;\, \rho^{\boldsymbol{\nu}} = \tilde{\rho} + 1 \text{ and (4.8) holds}\} & \text{if } \tilde{\rho} \in C\,, \\ +\infty & \text{otherwise}\,. \end{cases} \tag{5.12}$$

We observe that $\phi(0) = \mathcal{A}(\boldsymbol{\mu}^*) < +\infty$ and so ϕ is a proper convex function. We prove that it is bounded in a neighborhood of $\tilde{\rho} = 0$. By proposition 5.2, for any $\tilde{\rho} \in C$ there exists a $\boldsymbol{\nu} \in \mathcal{P}(\Omega)$ satisfying $\rho^{\boldsymbol{\nu}} = \tilde{\rho} + 1$ and the coupling constraint, such that

$$\mathcal{A}(\boldsymbol{\nu}) \leq \mathcal{A}(\boldsymbol{\mu}^*) + \|\tilde{\rho}\|_{C^{0,1}} \left(\frac{T}{2} + \mathcal{A}(\boldsymbol{\mu}^*) \right) + |\tilde{\rho}|^2_{C^{0,1}}(T + \mathcal{A}(\boldsymbol{\mu}^*))\,, \tag{5.13}$$

which implies

$$\phi(\rho) \leq \phi(0) + \|\tilde{\rho}\|_{C^{0,1}} \left(\frac{T}{2} + \mathcal{A}(\boldsymbol{\mu}^*) \right) + |\tilde{\rho}|^2_{C^{0,1}} (T + \mathcal{A}(\boldsymbol{\mu}^*)) \,. \tag{5.14}$$

Therefore, ϕ is bounded in a neighborhood of $\tilde{\rho} = 0$. As a consequence, by standard convex analysis arguments, ϕ is also locally Lipschitz on the same neighborhood and the subdifferential of ϕ at 0 is not empty, i.e. there exists $P \in B_0^*$ such that

$$\langle P, \tilde{\rho} \rangle \leq \phi(\tilde{\rho}) - \phi(0) \,. \tag{5.15}$$

By the definition of ϕ, this implies

$$\langle P, \tilde{\rho} \rangle \leq \mathcal{A}(\boldsymbol{\nu}) - \mathcal{A}(\boldsymbol{\mu}^*) \,, \tag{5.16}$$

for all generalized almost diffeomorphisms $\boldsymbol{\nu}$ satisfying $\rho^{\boldsymbol{\nu}} = \tilde{\rho} + 1$ and the coupling constraint in (4.8). □

Theorem 5.3 tells us that $\boldsymbol{\mu}^*$ is also a minimizer for the augmented action

$$\mathcal{A}^p(\boldsymbol{\nu}) \coloneqq \mathcal{A}(\boldsymbol{\nu}) - \langle P, \rho^{\boldsymbol{\nu}} - 1 \rangle \,, \tag{5.17}$$

defined on generalized almost diffeomorphisms. Then, for any $\tilde{\rho} \in B_0$, $\boldsymbol{\mu}^*_\epsilon \coloneqq \Phi^{1+\epsilon\tilde{\rho}}_{\#} \boldsymbol{\mu}^*$ is a generalized almost diffeomorphism if ϵ is sufficiently small. Moreover, we must have

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathcal{A}(\boldsymbol{\mu}^*_\epsilon) \Big|_{\epsilon=0} = 0 \,. \tag{5.18}$$

By the same calculation as in the proof of proposition 5.2, this implies

$$\langle P, \tilde{\rho} \rangle = \int_\Omega \int_0^T \tilde{\rho}(t, x_t) |\dot{z}_t|^2_{g_\mathcal{C}} + \partial_t(\tilde{\rho}(t, x_t)) r_t \dot{r}_t \, \mathrm{d}t \, \mathrm{d}\boldsymbol{\mu}^*(\mathrm{z}) \,, \tag{5.19}$$

for any $\tilde{\rho} \in B_0$, which defines P uniquely as a distribution. This also implies that the functional ϕ is actually differentiable at 0 since its subdifferential reduces to a single element.

6. Correspondence with deterministic solutions

In this section we study the correspondence between generalized and classical solutions of the $H(\mathrm{div})$ geodesic equations. In particular, we show that for sufficiently short times classical solutions generate dynamic plans which are the unique minimizers of problem 4.2.

We start by proving a modified version of a result presented in [17] stating that smooth solutions of the $H(\mathrm{div})$ geodesic equations are length-minimizing for short times in an L^∞ neighborhood on $\mathrm{Aut}_{\rho_0}(\mathcal{C})$. Let (φ, λ) be a smooth solution of the system (3.11) on the interval $[0, T]$. Let P be the associated pressure and $\Psi_p(t, x, r) \coloneqq P(t, x) r^2$. Following [7] we introduce the following functional on Ω,

$$\mathcal{B}(\mathrm{z}) \coloneqq \begin{cases} \int_0^T |\dot{z}_t|^2_{g_\mathcal{C}} - \Psi_p(t, x_t, r_t) \, \mathrm{d}t & \text{if } \mathrm{z} \in AC^2([0, T]; \mathcal{C}) \,, \\ +\infty & \text{otherwise} \,. \end{cases} \tag{6.1}$$

Moreover, we consider the function $b : \mathcal{C}^2 \to \mathbb{R}$ defined by

$$b(p, q) \coloneqq \inf\{\mathcal{B}(\mathrm{z}) \,;\, z_0 = p \,,\, z_T = q\} \,. \tag{6.2}$$

Lemma 6.1. *Suppose that $M \subset \mathbb{R}^d$ is convex and let (φ, λ) be a smooth solution of (3.11) on $[0, T] \times M$, with P being the associated pressure and $\Psi_p(t, x, r) \coloneqq P(t, x) r^2$. For any fixed $x \in M$, let $\mathrm{z}^* = [\mathrm{x}^*, \mathrm{r}^*] \in \Omega$ be the curve defined by $\mathrm{x}^* : t \to x^*_t \coloneqq \varphi_t(x)$ and $\mathrm{r}^* : t \to r^*_t \coloneqq \lambda_t(x)$. Let $r_{min} \coloneqq \min_{(t,x)\in[0,T]\times M} \lambda_t(x)$, $r_{max} \coloneqq \max_{(t,x)\in[0,T]\times M} \lambda_t(x)$ and $\varrho \coloneqq 2r_{max}/r_{min}$. There exists a constant $C_0 > 0$ such that, if*

- *for all $t \in [0, T]$ and for all $w \in T_{z^*_t}\mathcal{C}$,*

$$|\mathrm{Hess}^{g_\mathcal{C}} \, \Psi_p(w, w)| \leq \frac{C_0 \pi^2}{T^2} |w|^2_{g_\mathcal{C}} \,; \tag{6.3}$$

- *for all $t_0, t_1 \in [0, T]$,*

$$d_\mathcal{C}(z_{t_0}, z_{t_1}) \leq \frac{r_{min}}{4} \,; \tag{6.4}$$

- *the following inequality holds:*

$$\left[\varrho^2 + (\varrho+1)^2\right] \|P\|_{C^0} \leq \frac{3}{2T^2}\,; \tag{6.5}$$

then, $\mathcal{B}(\mathrm{z}^*) = b(z_0^*, z_T^*)$; moreover, for any other $\mathrm{z} \in AC^2([0,T];\mathcal{C})$ such that $z_0 = z_0^*$ and $z_T = z_T^*$, $\mathcal{B}(\mathrm{z}) = \mathcal{B}(\mathrm{z}^*)$ if and only if $\mathrm{z} = \mathrm{z}^*$. When M is the circle of unit radius $S_1^1 := \mathbb{R}/2\pi\mathbb{Z}$ the same holds with $C_0 = 2$ but without the conditions in equations (6.4) and (6.5).

Remark 6.2. *The assumption in (6.3) amounts to requiring that the spectral norm of the matrix*

$$g_{\mathcal{C}}^{-1/2}(\mathrm{Hess}^{g_{\mathcal{C}}}\,\Psi_p)g_{\mathcal{C}}^{-1/2} = \begin{pmatrix} 2P + (\nabla)^2 P & \nabla P \\ (\nabla P)^T & 2P \end{pmatrix} \tag{6.6}$$

be bounded by $C_0\pi^2/T^2$. This is verified for sufficiently small T if, e.g., $P \in L^\infty([0,T];C^2(M))$. Similarly, the assumption in (6.5) is verified for sufficiently small T if $P \in C^0([0,T]\times M)$, since for a given smooth solution φ with $\varphi_0 = \mathrm{Id}$, $\varrho = 2r_{max}/r_{min} \to 2$ as $T \to 0$. In addition, note that when $M = S_1^1$ the cone $\mathcal{C}$ can be identified with $\mathbb{R}^2$ and we do not have to deal with the singularity introduced by the apex. This is the reason why the assumptions in (6.4) and (6.5) are not necessary in this case.

The proof of lemma 6.1 is postponed to the appendix. Lemma 6.1 is the equivalent of lemma 5.2 in [7] on the cone. As in [7], we can use it to prove the optimality of the plan concentrated on the continuous solution. For this, however, we also need the following additional result on the function b, which characterizes the mininimizing paths starting and ending at the apex.

Lemma 6.3. *Suppose $P \in C^0([0,T]\times M)$ and $P_{max} := \max_{(t,x)\in[0,T]\times M} P(t,x) \leq (\pi/T)^2$. Then $b(o,o) = \mathcal{B}(\mathrm{o}) = 0$ where $\mathrm{o} : t \in [0,T] \to o \in \mathcal{C}$. If the inequality is strict then for any other $\mathrm{z} \in AC^2([0,T];\mathcal{C})$ such that $z_0 = o$ and $z_T = o$, $\mathcal{B}(\mathrm{z}) = \mathcal{B}(\mathrm{o})$ if and only if $\mathrm{z} = \mathrm{o}$.*

Proof. For the first part, observe that for any $\mathrm{z} \in AC^2([0,T];\mathcal{C})$ such that $r_0 = r_T = 0$, using Poincaré inequality on $\mathrm{r} : t \in [0,T] \to r_t \in \mathbb{R}_{\geq 0}$

$$\begin{aligned} \mathcal{B}(\mathrm{z}) &\geq \int_0^T |\dot z_t|^2_{g_{\mathcal{C}}} - r_t^2 P_{max}\,\mathrm{d}t \\ &\geq \int_0^T r_t^2|\dot x_t|^2 + \frac{\pi^2}{T^2} r_t^2 - r_t^2 P_{max}\,\mathrm{d}t \\ &\geq \left(\frac{\pi^2}{T^2} - P_{max}\right)\int_0^T r_t^2\,\mathrm{d}t\,. \end{aligned} \tag{6.7}$$

This implies that $b(o,o) \geq 0$. Clearly, $b(o,o) \leq \mathcal{B}(\mathrm{o}) = 0$ and therefore $b(o,o) = 0$. For the second part, if the inequality is strict, $C := \frac{\pi^2}{T^2} - P_{max} > 0$. Then, for any other $\mathrm{z} \in AC^2([0,T];\mathcal{C})$ such that $z_0 = o$ and $z_T = o$, and satisfying $\mathcal{B}(\mathrm{z}) = \mathcal{B}(\mathrm{o})$, we have

$$0 = \mathcal{B}(\mathrm{z}) \geq C\int_0^T r_t^2\,\mathrm{d}t\,, \tag{6.8}$$

which implies $\mathrm{z} = \mathrm{o}$. □

Theorem 6.4. *Under the assumptions of lemma 6.1, the dynamic plan $\boldsymbol{\mu}^* = (\varphi,\lambda)_{\#}\rho_0$ is an optimal solution of problem 4.2 with $\boldsymbol{\gamma} = [(\varphi_0,\lambda_0),(\varphi_T,\lambda_T)]_{\#}\rho_0$. If the inequalities (6.3) and (6.5) are strict, the solution $\boldsymbol{\mu}^*$ is unique in the following sense: for any minimizer $\boldsymbol{\mu}$, the measure $\boldsymbol{\mu}^{\mathrm{o}} := \boldsymbol{\mu}\llcorner\Omega^{\mathrm{o}}$, with $\Omega^{\mathrm{o}} := \Omega\setminus\{\mathrm{o}\}$, is equal to $\boldsymbol{\mu}^*$ up to rescaling (defined in lemma 4.5).*

Proof. Let $\boldsymbol{\mu}$ be any dynamic plan with finite action, i.e. $\mathcal{A}(\boldsymbol{\mu}) < +\infty$, and satisfying the constraints in (4.8) and (4.9). Consider the functional

$$\mathcal{P}(\mathrm{z}) = \int_0^T \Psi_p(t,x_t,r_t)\mathrm{d}t\,. \tag{6.9}$$

Then,

$$\begin{aligned}\int_\Omega \mathcal{P}(\mathrm{z})\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) &= \int_\Omega \int_0^T \Psi_p(t,x_t,r_t)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z})\\ &= \int_\Omega \int_0^T P(t,x_t) r_t^2\,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z})\\ &= \int_0^T \int_M P\,\mathrm{d}\rho_0\mathrm{d}t\,.\end{aligned} \tag{6.10}$$

Hence,

$$\mathcal{B}(\boldsymbol{\mu}) \coloneqq \int_\Omega \mathcal{B}(\mathrm{z})\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \mathcal{A}(\boldsymbol{\mu}) - \int_0^T \int_M P\,\mathrm{d}\rho_0\mathrm{d}t\,, \tag{6.11}$$

and since equation (6.11) also holds replacing $\boldsymbol{\mu}$ with $\boldsymbol{\mu}^*$,

$$\mathcal{B}(\boldsymbol{\mu}) - \mathcal{B}(\boldsymbol{\mu}^*) = \mathcal{A}(\boldsymbol{\mu}) - \mathcal{A}(\boldsymbol{\mu}^*)\,. \tag{6.12}$$

Now, by proposition 4.8 we have the decomposition $\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}^0$ where $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega\,;\, r_0 \neq 0\}$ and $\tilde{\boldsymbol{\mu}}^0 = \boldsymbol{\mu} \llcorner \{\mathrm{z} \in \Omega\,;\, r_0 = r_T = 0\}$. Therefore, integrating the function b defined in (6.2) with respect to $\boldsymbol{\mu}$ we obtain

$$\begin{aligned}\int_\Omega b(z_0,z_T)\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) &= \int_\Omega b(z_0,z_T)\,\mathrm{d}\tilde{\boldsymbol{\mu}}(\mathrm{z}) + \int_\Omega b(o,o)\,\mathrm{d}\tilde{\boldsymbol{\mu}}^0(\mathrm{z})\\ &= \int_\Omega b(z_0,z_T)\,\mathrm{d}\tilde{\boldsymbol{\mu}}(\mathrm{z})\,,\end{aligned} \tag{6.13}$$

where we used the fact that $b(o,o) = 0$ by lemma 6.3. By proposition 4.8, $\tilde{\boldsymbol{\mu}}^1 \coloneqq \mathrm{dil}_{r_0,2}\tilde{\boldsymbol{\mu}}$ satisfies the strong coupling constraint $(e_0,e_T)_\#\tilde{\boldsymbol{\mu}}^1 = \boldsymbol{\gamma}$. Moreover, b is 2-homogeneous (because $\mathcal{B}$ is) and therefore

$$\int_\Omega b(z_0,z_T)\,\mathrm{d}\tilde{\boldsymbol{\mu}}(\mathrm{z}) = \int_\Omega b(z_0,z_T)\,\mathrm{d}\tilde{\boldsymbol{\mu}}^1(\mathrm{z}) = \int_{\mathcal{C}^2} b(p,q)\,\mathrm{d}\boldsymbol{\gamma}(p,q)\,. \tag{6.14}$$

We get the same result integrating b with respect to $\boldsymbol{\mu}^*$. In particular, by lemma 6.1,

$$\int_\Omega b(z_0,z_T)\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \mathcal{B}(\boldsymbol{\mu}^*)\,. \tag{6.15}$$

By definition of b in (6.2), for any path $\mathrm{z} \in \Omega$, $\mathcal{B}(\mathrm{z}) \geq b(z_0,z_T)$ and therefore

$$\mathcal{B}(\boldsymbol{\mu}) \geq \int_\Omega b(z_0,z_T)\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \mathcal{B}(\boldsymbol{\mu}^*)\,, \tag{6.16}$$

which implies the same inequality for $\mathcal{A}$ due to equation (6.12). This proves that $\boldsymbol{\mu}^*$ is an optimal solution.

In order to prove uniqueness, let $\boldsymbol{\mu}$ be a solution of problem 4.2. Without loss of generality we can assume that $\boldsymbol{\mu} = \boldsymbol{\mu}^\circ$. Then, equations (6.12) and (6.14) imply

$$\int_\Omega \mathcal{B}(\mathrm{z}) - b(z_0,z_T)\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \mathcal{B}(\boldsymbol{\mu}) - \mathcal{B}(\boldsymbol{\mu}^*) = \mathcal{A}(\boldsymbol{\mu}) - \mathcal{A}(\boldsymbol{\mu}^*) = 0\,. \tag{6.17}$$

Since for any $\mathrm{z} \in \Omega$ we have $\mathcal{B}(\mathrm{z}) \geq b(z_0,z_T)$, then for $\boldsymbol{\mu}$-almost every path z, $\mathcal{B}(\mathrm{z}) = b(z_0,z_T)$. Clearly, also for $\boldsymbol{\mu}^*$-almost every path z, $\mathcal{B}(\mathrm{z}) = b(z_0,z_T)$. Now, if $\boldsymbol{\mu}$ satisfies the strong coupling constraint, for $\boldsymbol{\mu}$-almost every path z and for $\boldsymbol{\mu}^*$-almost every path z^* such that $z_0 = z_0^*$ and $z_T = z_T^*$, we have $\mathcal{B}(\mathrm{z}) = \mathcal{B}(\mathrm{z}^*) = b(z_0^*,z_T^*)$. This implies $\mathrm{z} = \mathrm{z}^*$ by lemma 6.1. In other words, $\boldsymbol{\mu}$ and $\boldsymbol{\mu}^*$ are concentrated on the same paths and due to the strong coupling constraint they must coincide.

On the other hand, suppose that $\boldsymbol{\mu}$ does not satisfy the strong coupling constraint. Recall that for $\boldsymbol{\mu}$-almost every path z we have $\mathcal{B}(\mathrm{z}) = b(z_0,z_T)$. Then, defining $\tilde{\Omega} \coloneqq \{\mathrm{z} \in \Omega\,;\, r_0 = r_T = 0\}$, we have

$$\int_{\tilde{\Omega}} \mathcal{B}(\mathrm{z})\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \int_{\tilde{\Omega}} b(z_0,z_T)\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = b(o,o)\,\boldsymbol{\mu}(\tilde{\Omega}) = 0\,. \tag{6.18}$$

For any $\mathrm{z} \in \tilde{\Omega}$ we also have $\mathcal{B}(\mathrm{z}) \geq b(o,o) = 0$. Hence, we find that for $\boldsymbol{\mu}$-almost every path z such that $z_0 = z_T = o$, we have $\mathcal{B}(\mathrm{z}) = 0$ which by lemma 6.3 is equivalent to $\mathrm{z} = \mathrm{o}$. This implies

$$\boldsymbol{\mu}(\{\mathrm{z} \in \Omega\,;\, r_0 = r_T = 0\}) = 0\,. \tag{6.19}$$

Then, corollary 4.10 implies that $\boldsymbol{\mu}$ can be rescaled to a measure satisfying the strong coupling. □

The assumptions on the pressure in lemma 6.1 are less strict for the case of the circle. This leads to the following result.

Corollary 6.5. *Let $M = S^1_1$ and let (φ, λ) be a smooth solution of (3.11) on $[0,T] \times M$, with P being the associated pressure and $\Psi_p(t,x,r) \coloneqq P(t,x)r^2$. If for all $t \in [0,T]$ and for all $w \in T_{z^*_t}\mathcal{C}$,*

$$|\mathrm{Hess}^{gc}\, \Psi_p(w,w)| \leq \frac{2\pi^2}{T^2}|w|^2_{gc}\,, \tag{6.20}$$

then the dynamic plan $\boldsymbol{\mu}^* = (\varphi,\lambda)_{\#}\rho_0$ is an optimal solution of problem 4.2 for the coupling $\boldsymbol{\gamma} = [(\varphi_0,\lambda_0),(\varphi_T,\lambda_T)]_{\#}\rho_0$. If the inequality in equation (6.20) is strict, it is unique up to rescaling (in the sense of lemma 4.5).

7. Some examples of generalized $H(\mathrm{div})$ geodesics

In this section we construct some instructive examples of generalized $H(\mathrm{div})$ geodesics which shed some light on the need of the relaxation and its tightness. In particular, we will focus on deterministic boundary conditions and construct singular solutions, i.e. minimizers that charge (non-trivial) paths starting and ending at the apex of the cone. This will allow us to prove two main results. First, that our relaxation is not tight on S^1_R, the circle of radius R, when R is sufficiently large; and second, that on the torus, for specific boundary conditions, problem 1.1 may admit no smooth minimizer, whereas we can construct a singular solution as the limit of a minimizing sequence of smooth flows. This suggests that problem 4.2 is a tight relaxation of problem 1.1 in dimension $d \geq 2$.

We start by considering an important generalized flow which provides an upper bound on the action on any domain and for any deterministic coupling.

Lemma 7.1. *Consider the generalized $H(\mathrm{div})$ geodesic problem with coupling given by $\boldsymbol{\gamma} = (h, \sqrt{|\mathrm{Jac}(h)|})_{\#}\rho_0$ where $h \in \mathrm{Diff}(M)$. Denote by ρ_0 the Lebesgue measure on M, normalized so that $\rho_0(M) = 1$. Then the measure*

$$\boldsymbol{\mu}^* = \frac{1}{2}(\mathrm{Id}, \zeta^0)_{\#}\rho_0 + \frac{1}{2}(\psi^1, \zeta^1)_{\#}\rho_0\,, \tag{7.1}$$

with

$$\zeta^0_t(x) = \sqrt{2}\sin(\sqrt{P^*}t)\,, \quad \zeta^1_t(x) = \begin{cases} \sqrt{2}\cos(\sqrt{P^*}t) & t \leq T/2\,, \\ -\sqrt{2\,|\mathrm{Jac}(h(x))|}\cos(\sqrt{P^*}t) & t > T/2\,, \end{cases} \tag{7.2}$$

$$\psi^1_t(x) = \begin{cases} x & t \leq T/2\,, \\ h(x) & t > T/2\,, \end{cases} \tag{7.3}$$

where $P^* = \pi^2/T^2$, is an admissible generalized flow and the action of the minimizer is bounded from above by $\mathcal{A}(\boldsymbol{\mu}^*) = \pi^2/T$.

Proof. We need to check that $\boldsymbol{\mu}^*$ is a probability measure and that it satisfies the homogeneous marginal and coupling constraints. The fact that $\boldsymbol{\mu}^*(\Omega) = 1$ is immediate from the definition.

As for the marginal constraint, observe that for any $f \in C^0([0,T]\times M)$,

$$\begin{aligned}\int_\Omega \int_0^T f(t,x_t) r_t^2 \,\mathrm{d}t\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) &= \frac{1}{2}\int_M \int_0^T f(t,x) 2\sin^2(\sqrt{P^*}t)\,\mathrm{d}t\,\mathrm{d}\rho_0(x)\\ &\quad+\frac{1}{2}\int_M \int_0^{T/2} f(t,x) 2\cos^2(\sqrt{P^*}t)\,\mathrm{d}t\,\mathrm{d}\rho_0(x)\\ &\quad+\frac{1}{2}\int_M \int_{T/2}^T f(t,h(x)) 2|\mathrm{Jac}(h(x))|\cos^2(\sqrt{P^*}t)\,\mathrm{d}t\,\mathrm{d}\rho_0(x)\\ &= \int_M\int_0^T f(t,x)\,\mathrm{d}t\,\mathrm{d}\rho_0(x)\,.\end{aligned} \tag{7.4}$$

By similar calculations also the homogeneous coupling constraint holds and therefore $\boldsymbol{\mu}^*$ is admissible. Moreover, the action associated with $\boldsymbol{\mu}^*$ is given by

$$\begin{aligned}\mathcal{A}(\boldsymbol{\mu}^*) &= \frac{1}{2}\int_M\int_0^T |\dot\zeta_t^0(x)|^2 + |\dot\zeta_t^1(x)|^2 \mathrm{d}t\,\mathrm{d}\rho_0(x)\\ &= \int_M\int_0^T P^*\mathrm{d}t\,\mathrm{d}\rho_0(x) = \frac{\pi^2}{T}\,.\end{aligned} \tag{7.5}$$

□

The dynamic plan in lemma 7.1 shows that in our generalized formulation we can reach any final configuration only by changes in the Jacobian, although in a non-deterministic sense. In the following we will consider several instances of this flow for different domains and couplings and we will prove that in some cases it also minimizes the generalized $H(\mathrm{div})$ action. In fact, the idea behind the construction of the generalized flow $\boldsymbol{\mu}^*$ is that as for geodesics on the cone, we expect that for a sufficiently large displacement optimal solutions would concentrate on straight lines in the radial direction passing by the apex of the cone. If there is no motion on the base space M, the geodesic equation (3.11) in the radial direction reduces to

$$\ddot\lambda + \lambda P = 0 \tag{7.6}$$

The dynamic plan $\boldsymbol{\mu}^*$ concentrates precisely on solutions to this equation with constant pressure $P = P^*$.

It should also be noted that $\boldsymbol{\mu}^*$ is exactly in the form discussed in proposition 4.8, i.e. it is decomposed in the sum of two measures, $\boldsymbol{\mu}^* = \tilde{\boldsymbol{\mu}} + \tilde{\boldsymbol{\mu}}^0$, where

$$\tilde{\boldsymbol{\mu}}^0 = \frac{1}{2}(\mathrm{Id},\zeta^0)_{\#}\rho_0\,,\quad \tilde{\boldsymbol{\mu}} = \frac{1}{2}(\psi^1,\zeta^1)_{\#}\rho_0\,. \tag{7.7}$$

In particular, $\tilde{\boldsymbol{\mu}}$ does not charge paths starting and ending at the apex, so it can be rescaled to a probability measure satisfying the strong coupling constraint but not the homogeneous marginal constraint. This is given by

$$\tilde{\boldsymbol{\mu}}^1 = \mathrm{dil}_{r_0,2}\tilde{\boldsymbol{\mu}} = (\psi^1,\zeta^1/\sqrt{2})_{\#}\rho_0\,. \tag{7.8}$$

The dynamic plan $\tilde{\boldsymbol{\mu}}^1$ describes a peculiar solution in which particles gradually disappear up to time $T/2$, when the whole domain vanishes, and then gradually reappear in the final configuration. This phenomenon is related to the collision of a peakon and an anti-peakon in one dimension, which is a well-known solution of the CH equation [10]. Such a solution implies that arbitrarily small portions of the domain can be stretched to occupy finite area at finite bounded cost. The generalized solution in (7.1) replicates this behavior in an averaged sense across the domain. This will be made precise by the approximation results in proposition 7.8 and theorem 7.12.

7.1. **Construction of a generalized solution on the circle.** We now consider the generalized $H(\mathrm{div})$ geodesic problem on S^1_R, the circle of radius R. For specific boundary conditions given by uniform rotation and when $R = 1$, we show that the generalized flow in lemma 7.1 is a minimizer although not unique, having the same cost as the deterministic solution. When $R > 1$, the constant speed rotation is not a minimizer since its action is strictly larger than π^2/T. Moreover, if R is sufficiently large, there is no minimizing sequence of deterministic smooth flows whose action tend to π^2/T. This implies that in this case, the relaxed problem 4.2

is not tight. In order to make this precise, we start by proving the following lower bound on the action of problem 1.1 on S^1_R and for boundary conditions given by uniform rotation.

Lemma 7.2. *Let $\varphi^* : [0,T]\times S^1_R \to S^1_R$ be a smooth flow satisfying $\varphi^*_0 = \mathrm{Id}$ and $\varphi^*_T = h$, where h prescribes uniform rotation by half of the circle length, i.e. $h : x \in \mathbb{R}/2\pi R\mathbb{Z} \to x + \pi R \in \mathbb{R}/2\pi\mathbb{Z}$. Then,*

$$\frac{1}{2\pi R}\int_0^{2\pi R} \mathcal{A}([\varphi^*(x), \sqrt{\mathrm{Jac}(\varphi^*(x))}])\,\mathrm{d}x \geq \frac{\tanh(2\pi R)}{2T}\pi R\,. \tag{7.9}$$

Proof. Consider the following problem

$$\inf\left\{\frac{1}{2\pi R}\int_0^{2\pi R} \mathcal{A}([\varphi(x), \sqrt{\mathrm{Jac}(\varphi(x))}])\,\mathrm{d}x\,;\, \varphi_0(0) = 0\,, \varphi_T(0) = \pi R\right\}. \tag{7.10}$$

where the infimum is taken over smooth curves $t \in [0,T] \mapsto \varphi_t \in \mathrm{Diff}(S^1_R)$. The quantity in equation (7.10) provides a lower bound for the action associated with φ^*. Fix a smooth flow φ. For any $t \in (0,1)$, let $u_t \in H^1(S^1_R)$ be the velocity field minimizing

$$\frac{1}{2\pi R}\int_0^{2\pi R} |u|^2 + \frac{1}{4}|\partial_x u|^2 \mathrm{d}x\,, \tag{7.11}$$

over all $u \in H^1(S^1_R)$ such that $u(\varphi_t(0)) = \partial_t\varphi_t(0)$. In particular, we have $u_t = G * m_t$ where m is in the form

$$m_t(x) = p_t\,\delta(x - \varphi_t(0))\,, \tag{7.12}$$

with $p_t \in \mathbb{R}$ depends on the boundary conditions, and G is the Green's function for the operator $\mathrm{Id} - \frac{1}{4}\partial_{xx}$, which is given by

$$G(x,y) = \frac{\cosh(2|x-y| - 2\pi R)}{\sinh(2\pi R)} \tag{7.13}$$

(note that u_t has the same form of a peakon on S^1_R, see section 7.2). Then, by direct calculation,

$$\begin{aligned}\frac{1}{2\pi R}\int_0^{2\pi R} \mathcal{A}([\varphi(x), \sqrt{\mathrm{Jac}(\varphi(x))}])\,\mathrm{d}x &\geq \int_0^T\int_0^{2\pi R} |u_t|^2 + \frac{1}{4}|\partial_x u_t|^2 \mathrm{d}x\,\mathrm{d}t \\ &= \frac{\tanh(2\pi R)}{2\pi R}\int_0^T |\partial_t\varphi_t(0)|^2\mathrm{d}t\,.\end{aligned} \tag{7.14}$$

Using the boundary conditions on φ from equation (7.10) gives the result. □

In view of lemma 7.1, lemma 7.2 implies that our relaxation (problem 4.2) is not tight on S^1_R for sufficiently large R. This is made precise in the following theorem.

Theorem 7.3. *Consider the generalized $H(\mathrm{div})$ geodesic problem on S^1_R with coupling constraint given by uniform rotation by half of the circle length, i.e. in polar coordinates $h : \theta \in \mathbb{R}/2\pi\mathbb{Z} \to \theta + \pi \in \mathbb{R}/2\pi\mathbb{Z}$. Denote by $\rho_0 = (2\pi)^{-1}\mathrm{d}\theta$ the normalized Lebesgue measure on the circle. The following holds:*

1. *when $R = 1$ the dynamic plan $\boldsymbol{\mu}^*$ in lemma 7.1, i.e. equation (7.1) with*

$$\zeta^0_t(\theta) = \sqrt{2}\sin(\sqrt{P^*}t)\,,\quad \zeta^1_t(\theta) = \sqrt{2}|\cos(\sqrt{P^*}t)|\,,\quad \psi^1_t(\theta) = \begin{cases}\theta & t \leq T/2\,,\\ \theta+\pi & t > T/2\,,\end{cases} \tag{7.15}$$

   as well as the dynamic plan induced by constant speed rotation are minimizers corresponding to the constant pressure $P^* = (\pi/T)^2$;
2. *when $R > 1$ the constant speed rotation is not a minimizer; moreover, if R is sufficiently large, the infimum of the deterministic $H(\mathrm{div})$ geodesic problem 1.1 is strictly larger than that of the generalized geodesic problem 4.2.*

Proof. For the first point, observe that from the Euler-Lagrange equations (3.12) the pressure relative to constant speed rotation on S^1_1 is given by

$$P^{rot} = |u|^2 = \frac{\pi^2}{T^2}. \tag{7.16}$$

This satisfies the hypotheses of corollary 6.5 (see remark 6.2) and therefore the constant rotation is a minimizer. Since the Jacobian stays constant during the rotation, the associated action is given by

$$\mathcal{A}^{rot} = \frac{1}{2\pi}\int_0^{2\pi}\int_0^T |u|^2 \,\mathrm{d}t\,\mathrm{d}\theta = \frac{\pi^2}{T}\,. \tag{7.17}$$

On the other hand, by lemma 7.1, $\boldsymbol{\mu}^* \in \mathcal{P}(\Omega)$ is admissible and its action is equal to $\mathcal{A}(\boldsymbol{\mu}^*) = \pi^2/T$, independently of R. Hence $\boldsymbol{\mu}^*$ is also a minimizer and it must share the same pressure with the constant speed rotation, $P^{rot} = P^*$. For the second point, observe that the action for constant speed rotation on S^1_R is $\mathcal{A}^{rot}_R = R^2\mathcal{A}^{rot} > \mathcal{A}(\boldsymbol{\mu}^*)$ whenever $R > 1$. Similarly, for R sufficiently large we have

$$\frac{\tanh(2\pi R)}{2T}\pi R > \mathcal{A}(\boldsymbol{\mu}^*)\,. \tag{7.18}$$

We conclude applying lemma 7.2. □

Remark 7.4. *For $d = 1$ one can produce a tight relaxation of problem 1.1 using different techniques than those used in the present paper. This approach is developed in* [14] *and is specific to dimension one. However note that one still needs to rely on theorem 6.4 in order to conclude that smooth geodesics are the unique global length-minimizers for this tight one-dimensional relaxation.*

7.2. **Collision of peakons and an approximation result.** Before going further with the construction of a generalized solutions on a two-dimensional domain, we need to clarify the connection between the solution presented in theorem 7.3 and diffeomorphisms of the circle. In particular, here we show that if no rotation occurs, the generalized flow in theorem 7.3 can be approximated using linear peakon/anti-peakon collisions. This will serve as a basis to construct a sequence of deterministic flows converging to a non-deterministic minimizer in two dimensions.

Consider the CH equation on the circle S^1_1 with Lagrangian $\int_0^{2\pi} |u|^2 + \frac{1}{4}|\partial_\theta u|^2 \,\mathrm{d}\theta$, where $u : [0,T]\times S^1_1 \to \mathbb{R}$ is the Eulerian velocity field. Peakon solutions can be described in terms of momentum $m = u - \frac{1}{4}\partial^2_\theta u$ as a linear combination of Dirac delta functions, i.e.

$$m(t,\theta) = \sum_{i=1}^N p_i(t)\delta(\theta - \theta_i(t))\,, \tag{7.19}$$

where $p_i : [0,T] \to \mathbb{R}$ and $\theta_i : [0,T] \to S^1_1$ are appropriate functions specifying the momentum carried by the ith peakon and its location, respectively. The associated velocity field is given by $u = G * m$ where G is the Green's function for the operator $\mathrm{Id} - \frac{1}{4}\partial^2_\theta$ (see equation (7.13)).

The collision of a peakon and an anti-peakon corresponds to the case $N = 2$, $p_2 = -p_1$, $\theta_2 = 2\pi - \theta_1$, in which case there exists a finite time T^* such that as $t \to T^*$ collision occurs, i.e. $\theta_1 = \theta_2$. A similar behavior occurs for the Lagrangian $\int_0^{2\pi} \frac{1}{4}|\partial_\theta u|^2 \,\mathrm{d}\theta$, which corresponds to the Hunter-Saxton equation. In this case, the velocity field is simply given by the linear interpolation of the velocity at θ_1 and θ_2 (see figure 1) and the Jacobian of the flow map is piecewise constant. Hence specifying the trajectory $\theta_1(t)$ uniquely defines the flow. We refer to such a solution as *linear peakon/anti-peakon* collision. The associated flow on a circle of arbitrary radius R is described in the following lemma.

Lemma 7.5. *For a given $\epsilon > 0$, let $\varphi^\epsilon : [0,T]\times S^1_R \to S^1_R$ be the flow map defined in polar coordinates by*

$$\varphi^\epsilon_t(0) = 0\,, \qquad \partial_\theta\varphi^\epsilon_t(\theta) = \begin{cases} 1 - \sin\left(\frac{\pi t}{2(T+\epsilon)}\right) & \text{if } \frac{\pi}{2} < \theta < \frac{3\pi}{2}\,, \\ 1 + \sin\left(\frac{\pi t}{2(T+\epsilon)}\right) & \text{otherwise}\,, \end{cases} \tag{7.20}$$

Then the associated action is uniformly bounded and

$$\lim_{R\to 0}\lim_{\epsilon\to 0}\frac{1}{2\pi}\int_0^{2\pi}\mathcal{A}([R\varphi^\epsilon(\theta),\lambda^\epsilon(\theta)])\,\mathrm{d}\theta = \frac{\pi^2}{16T}\,, \tag{7.21}$$

where $\lambda^\epsilon = \sqrt{\mathrm{Jac}(\varphi^\epsilon)}$ and

$$\frac{1}{2\pi}\int_0^{2\pi}\mathcal{A}([R\varphi^\epsilon(\theta),\lambda^\epsilon(\theta)])\,\mathrm{d}\theta = \frac{1}{2\pi}\int_0^{2\pi}\int_0^T R^2(\lambda^\epsilon_t(\theta))^2|\dot\varphi^\epsilon_t(\theta)|^2 + |\dot\lambda^\epsilon_t(\theta)|^2\,\mathrm{d}t\,\mathrm{d}\theta \tag{7.22}$$

is the action expressed in polar coordinates.

Proof. The result follows by direct computation and by definition of the functional $\mathcal{A}$ in equation (4.7). Note that the expression for the action in equation (7.22) can be justified by an appropriate change of variables. Specifically, denoting by φ^ϵ_R the flow map in arc length coordinates $x \in \mathbb{R}/2\pi R\mathbb{Z}$, we have $\theta = x/R$ and

$$\varphi^\epsilon_R(x) = R\varphi^\epsilon\left(\frac{x}{R}\right), \quad \partial_x\varphi^\epsilon_R(x) = \partial_\theta\varphi^\epsilon\left(\frac{x}{R}\right). \tag{7.23}$$

Denoting $\lambda^\epsilon_R = \sqrt{\partial_x\varphi^\epsilon_R}$, since $\rho_0 = (2\pi R)^{-1}\mathrm{d}x$ we obtain that the action is given by

$$\begin{aligned}\int_{S^1_R} \mathcal{A}([\varphi^\epsilon_R(x), \lambda^\epsilon_R(x)])\,\mathrm{d}\rho_0(x) &= \frac{1}{2\pi R}\int_0^{2\pi R}\int_0^T ((\lambda^\epsilon_R)_t(x))^2|(\dot\varphi^\epsilon_R)_t(x)|^2 + |(\dot\lambda^\epsilon_R)_t(x)|^2\,\mathrm{d}t\,\mathrm{d}x \\ &= \frac{1}{2\pi}\int_0^{2\pi}\int_0^T R^2(\lambda^\epsilon_t(\theta))^2|\dot\varphi^\epsilon_t(\theta)|^2 + |\dot\lambda^\epsilon_t(\theta)|^2\,\mathrm{d}t\,\mathrm{d}\theta.\end{aligned} \tag{7.24}$$

□

Remark 7.6. *The flow described in lemma 7.5 coincides with a linear peakon/anti-peakon solution of the Hunter-Saxton equation where the momentum is in the form of equation (7.19) and the two peak trajectories are given by*

$$\theta^\epsilon_1(t) = \frac{\pi}{2}\left(1 + \sin\left(\frac{\pi t}{2(T+\epsilon)}\right)\right), \quad \theta^\epsilon_2(t) = \frac{\pi}{2}\left(3 - \sin\left(\frac{\pi t}{2(T+\epsilon)}\right)\right). \tag{7.25}$$

The reason why we consider solutions Hunter-Saxton rather than CH peakons is due to the fact that as $R \to 0$ the action in (7.22) tends to the $\dot H^1$ action.

In figure 2, we give an illustration of the flow defined in equation (7.20) for fixed ϵ both in terms of particle trajectories and as a measure on the cone for $R = 1$ (in which case the cone can be identified with $\mathbb{R}^2$). Note that at collision time the trajectories of particles between the peaks reach the apex of the cone.

In the next lemma we construct a flow using n linear peakon/anti-peakon collisions that converges as $n \to +\infty$ to a measure in the same form as the one in lemma 7.1.

Lemma 7.7. *Let $\varphi^\epsilon : [0,T]\times S^1_R \to S^1_R$ the flow in lemma 7.5 and for each $n \in \mathbb{N}$ let $\hat\varphi^n : [0,T]\times S^1_R \to S^1_R$ be defined by*

$$\hat\varphi^n(\theta) \coloneqq \frac{2\pi}{n}\left\lfloor\frac{\theta n}{2\pi}\right\rfloor + \frac{1}{n}\varphi^{\epsilon_n}\left(n\theta - 2\pi\left\lfloor\frac{\theta n}{2\pi}\right\rfloor\right) \tag{7.26}$$

with ϵ_n being any positive sequence such that $\epsilon_n \to 0$. Then $\hat{\boldsymbol{\mu}}_n \coloneqq (\hat\varphi^n, \sqrt{\mathrm{Jac}(\hat\varphi^n)})_{\#}\rho_0 \rightharpoonup \hat{\boldsymbol{\mu}}^*$, where $\hat{\boldsymbol{\mu}}^*$ is defined by

$$\hat{\boldsymbol{\mu}}^* = \frac{1}{2}(\mathrm{Id}, \zeta^0)_{\#}\rho_0 + \frac{1}{2}(\mathrm{Id}, \zeta^1)_{\#}\rho_0, \tag{7.27}$$

with

$$\zeta^0_t(\theta) = \sqrt{2}\sin\left(\frac{\pi t}{4T} + \frac{\pi}{4}\right), \quad \zeta^1_t(\theta) = \sqrt{2}\cos\left(\frac{\pi t}{4T} + \frac{\pi}{4}\right). \tag{7.28}$$

Moreover, $\mathcal{A}(\hat{\boldsymbol{\mu}}_n) \to \mathcal{A}(\hat{\boldsymbol{\mu}}^*) = \pi^2/(16T)$.

Proof. For simplicity, we prove the result for $R = 1$ but the argument presented here applies for any $R > 0$. Let $\mathcal{F}$ be any bounded Lipschitz functional on Ω with Lipschitz constant L. We need to check that

$$\lim_{n\to+\infty}\int_\Omega \mathcal{F}(\mathrm{z})\,\mathrm{d}\hat{\boldsymbol{\mu}}_n(\mathrm{z}) = \int_\Omega \mathcal{F}(\mathrm{z})\,\mathrm{d}\hat{\boldsymbol{\mu}}^*(\mathrm{z}). \tag{7.29}$$

Denoting by $\hat\lambda^n = \sqrt{\mathrm{Jac}(\hat\varphi^n)}$ and by $\hat\lambda^{\epsilon_n} = \sqrt{\mathrm{Jac}(\hat\varphi^{\epsilon_n})}$, we observe that

$$\begin{aligned}\int_\Omega \mathcal{F}(\mathrm{z})\,\mathrm{d}\hat{\boldsymbol{\mu}}_n(\mathrm{z}) &= \frac{1}{2\pi}\int_0^{2\pi}\mathcal{F}([\hat\varphi^n(\theta), \hat\lambda^n(\theta)])\,\mathrm{d}\theta \\ &= \frac{1}{2\pi}\sum_{i=0}^{n-1}\int_0^{2\pi/n}\mathcal{F}\left(\left[\frac{2\pi i}{n} + \frac{1}{n}\varphi^{\epsilon_n}(n\theta), \lambda^{\epsilon_n}(n\theta)\right]\right)\mathrm{d}\theta,\end{aligned} \tag{7.30}$$

and similarly,

$$\int_\Omega \mathcal{F}(\mathrm{z})\,\mathrm{d}\hat{\boldsymbol{\mu}}^*(\mathrm{z}) = \frac{1}{4\pi}\sum_{i=0}^{n-1}\int_0^{2\pi/n} \mathcal{F}\left(\left[\frac{2\pi i}{n}+\theta,\zeta^0(\theta)\right]\right)+\mathcal{F}\left(\left[\frac{2\pi i}{n}+\theta,\zeta^1(\theta)\right]\right)\mathrm{d}\theta\,. \tag{7.31}$$

We consider separately each integral in the sums in equation (7.30) and (7.31). Rescaling the integrals in θ and using Lipschitz continuity of $\mathcal{F}$, we observe that the result is proven if

$$\lim_{n\to+\infty}\frac{1}{n}\sum_{i=0}^{n-1}\left|\frac{1}{2\pi}\int_0^{2\pi}\mathcal{F}\left(\left[\frac{2\pi i}{n}+\frac{1}{n}\,\varphi^{\epsilon_n}(\theta),\lambda^{\epsilon_n}(\theta)\right]\right)\mathrm{d}\theta - I_i^n\right| = 0\,, \tag{7.32}$$

where

$$I_i^n = \frac{1}{2}\mathcal{F}\left(\left[\frac{2\pi i}{n},\zeta^0\left(\frac{2\pi i}{n}\right)\right]\right)+\frac{1}{2}\mathcal{F}\left(\left[\frac{2\pi i}{n},\zeta^1\left(\frac{2\pi i}{n}\right)\right]\right)\,. \tag{7.33}$$

For any fixed sufficiently large n, we need to provide an appropriate bound for each term in the sum in equation (7.32). For any integer i with $0\le i\le n-1$, we have

$$\begin{aligned} E^{i,n} &\coloneqq \left|\frac{1}{2\pi}\int_0^{2\pi}\mathcal{F}\left(\left[\frac{2\pi i}{n}+\frac{1}{n}\,\varphi^{\epsilon_n}(\theta),\lambda^{\epsilon_n}(\theta)\right]\right)\mathrm{d}\theta - I_i^n\right| \\ &\le \frac{1}{2\pi}\left|\int_0^{2\pi}\mathcal{F}\left(\left[\frac{2\pi i}{n}+\frac{1}{n}\,\varphi^{\epsilon_n}(\theta),\lambda^{\epsilon_n}(\theta)\right]\right)\mathrm{d}\theta - \int_0^{2\pi}\mathcal{F}\left(\left[\frac{2\pi i}{n},\lambda^{\epsilon_n}(\theta)\right]\right)\mathrm{d}\theta\right| \\ &\quad + \left|\frac{1}{2\pi}\int_0^{2\pi}\mathcal{F}\left(\left[\frac{2\pi i}{n},\lambda^{\epsilon_n}(\theta)\right]\right)\mathrm{d}\theta - I_0^n\right| \coloneqq E_0^n + E_1^n\,. \end{aligned} \tag{7.34}$$

Observe that for $\alpha\in[0,\pi/2]$, $\sqrt{1-\sin(\alpha)} = \sqrt{2}\cos(\alpha/2+\pi/4)$ and $\sqrt{1+\sin(\alpha)} = \sqrt{2}\sin(\alpha/2+\pi/4)$ therefore

$$\lambda^{\epsilon_n}(\theta) = \sqrt{\partial_\theta\varphi_t^{\epsilon_n}(\theta)} = \begin{cases} \sqrt{2}\cos\left(\frac{\pi t}{4(T+\epsilon_n)}+\frac{\pi}{4}\right) & \text{if } \frac{\pi}{2}<\theta<\frac{3\pi}{2}\,,\\ \sqrt{2}\sin\left(\frac{\pi t}{4(T+\epsilon_n)}+\frac{\pi}{4}\right) & \text{otherwise}\,. \end{cases} \tag{7.35}$$

Since λ^{ϵ_n} is piecewise constant in θ we can write

$$\frac{1}{2\pi}\int_0^{2\pi}\mathcal{F}\left(\left[\frac{2\pi i}{n},\lambda^{\epsilon_n}(\theta)\right]\right)\mathrm{d}\theta = \frac{1}{2}\mathcal{F}\left(\left[\frac{2\pi i}{n},\lambda^{\epsilon_n}(0)\right]\right)+\frac{1}{2}\mathcal{F}\left(\left[\frac{2\pi i}{n},\lambda^{\epsilon_n}(\pi)\right]\right)\,. \tag{7.36}$$

Comparing the expression for ζ^0 and ζ^1 with that of λ^{ϵ_n} and using the fact that $\mathcal{F}$ is Lipschitz we obtain $E_1^n\le C(\epsilon_n)$, where $C(\epsilon_n)>0$ is a constant depending on ϵ_n and L such that $C(\epsilon_n)\to 0$ as $n\to+\infty$. A similar argument holds for E_0^n and therefore we can find a constant C_n independent of i such that $E^{i,n}\le C_n$ and $C_n\to 0$ as $n\to+\infty$. This implies equation (7.32).

Finally, convergence of the action is a consequence of lemma 7.5. In particular, it is immediate to verify that $\mathcal{A}(\hat{\boldsymbol{\mu}}^*) = \pi^2/(16T)$. Moreover, by the same reasoning as in the proof of lemma 7.5 and the change of variables in equation (7.30) we obtain that the action $\mathcal{A}(\hat{\boldsymbol{\mu}}_n)$ is given by

$$\begin{aligned} \frac{1}{2\pi}\int_0^{2\pi}\mathcal{A}([\hat{\varphi}^n(\theta),\hat{\lambda}^n(\theta)])\,\mathrm{d}\theta &= \frac{1}{2\pi}\sum_{i=0}^{n-1}\int_0^{2\pi/n}\mathcal{A}\left(\left[\frac{1}{n}\,\varphi^{\epsilon_n}(n\theta),\lambda^{\epsilon_n}(n\theta)\right]\right)\mathrm{d}\theta \\ &= \frac{1}{2\pi}\int_0^{2\pi}\mathcal{A}\left(\left[\frac{1}{n}\,\varphi^{\epsilon_n}(n\theta),\lambda^{\epsilon_n}(n\theta)\right]\right)\mathrm{d}\theta\,. \end{aligned} \tag{7.37}$$

Therefore, the limit of $\mathcal{A}(\hat{\boldsymbol{\mu}}_n)$ for $n\to+\infty$ is the same to that in equation (7.21). □

In figure 3, we give an illustration of the flow defined in equation (7.26) for fixed n both in terms of particle trajectories and as a measure on the cone for $R=1$. It can be seen that convergence towards the measure $\boldsymbol{\mu}^*$ defined in lemma 7.7 is due to the appearance of fast oscillations in the Jacobian together with the fact that particles tend to stay still as $n\to+\infty$.

We can use the flows defined in lemma 7.7 to construct a sequence that converges to the generalized flow $\boldsymbol{\mu}^*$ in theorem 7.3 but where no rotation occurs. The construction consists in concatenating in time the flows in lemma 7.7 so that a small portion of the domain stretches and then return to its original size. This is shown in figure 4. The convergence result is stated explicitly in the following proposition.

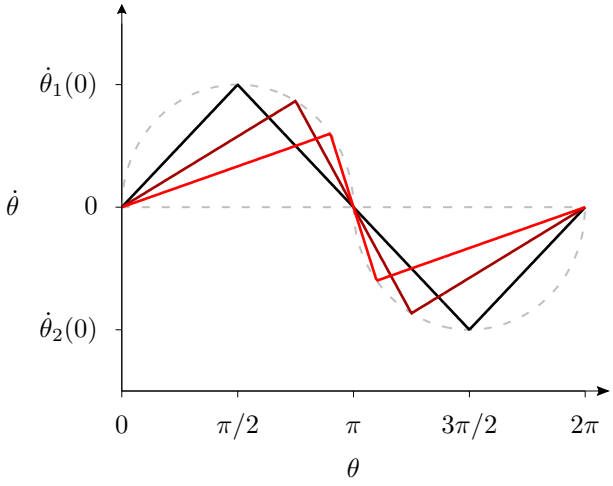


FIGURE 1. Velocity field evolution for the linear peakon/anti-peakon solution of the Hunter-Saxton equation.

Proposition 7.8. *Let $\hat\varphi^n : [0,T]\times S^1_R \to S^1_R$ be the sequence defined in lemma 7.7 and for each $n\in\mathbb{N}$ let $\varphi^n : [0,T]\times S^1_R \to S^1_R$ be defined by $\varphi^n_t = \varphi^n_{T-t}$ and*

$$\varphi^n_t(\theta) = \begin{cases} \hat\varphi^n_{T-4t}((\hat\varphi^n_T)^{-1}(\theta)) & \text{if } t\le \frac{T}{4}\,, \\ \hat\varphi^n_{4t-T}\left((\hat\varphi^n_T)^{-1}(\theta)+\frac{\pi}{n}\right)-\frac{\pi}{n} & \text{if } \frac{T}{4}<t\le\frac{T}{2}\,. \end{cases} \tag{7.38}$$

Then $\boldsymbol{\mu}_n := (\varphi^n, \sqrt{\mathrm{Jac}(\varphi^n)})_\#\rho_0$ can be rescaled to a sequence $\tilde{\boldsymbol{\mu}}_n \rightharpoonup \boldsymbol{\mu}^*$, where $\boldsymbol{\mu}^*$ is defined as in equation (7.27) with

$$\zeta^0_t(\theta) = \sqrt{2}\sin\left(\frac{\pi t}{T}\right)\,, \quad \zeta^1_t(\theta) = \sqrt{2}\left|\cos\left(\frac{\pi t}{T}\right)\right|\,. \tag{7.39}$$

Moreover, $\mathcal{A}(\boldsymbol{\mu}_n)\to\mathcal{A}(\boldsymbol{\mu}^*) = \pi^2/T$.

Proof. The rescaling to be performed in order to obtain the sequence $\tilde{\boldsymbol{\mu}}_n$ is given by

$$\tilde{\boldsymbol{\mu}}_n = \mathrm{dil}_{r_{T/4},2}\boldsymbol{\mu}_n\,. \tag{7.40}$$

In fact, by lemma 4.5, $\tilde{\boldsymbol{\mu}}_n$ is concentrated on paths such that $r_{T/4}=1$. Then, the result can be deduced from lemmas 7.7 and 7.5. □

Remark 7.9. *The maps defined by equation (7.38) are piecewise smooth in space since their Jacobian is piecewise constant with a finite number of discontinuities. However, using a regularization argument, it is not difficult to construct a sequence of smooth diffeomorphisms satisfying proposition 7.8. For this it is sufficient to repeat the construction above using a regularized version of the linear peakon/anti-peakon collision, which can be defined by convolution of the flow map with a sequence of positive symmetric mollifiers.*

7.3. **Construction of a generalized solution on the torus.** We now consider the generalized $H(\mathrm{div})$ geodesic problem on the torus $T^2_{1,R} := S^1_1\times S^1_R$, with one of the two radii set to one. We consider as boundary condition a uniform rotation on the torus in which each particle rotates of half of the length on both circles. For this specific boundary condition we can construct a generalized minimizer using the construction of the previous section which realizes smaller action than any deterministic smooth flow.

In the following lemma we provide a lower bound on the action associated with a deterministic minimizer.

Lemma 7.10. *Suppose that the smooth curve $t\in[0,T]\mapsto\varphi^*_t\in\mathrm{Diff}(T^2_{1,R})$ is a minimizer for the deterministic $H(\mathrm{div})$ geodesic problem 1.1, with $\varphi_0=\mathrm{Id}$, $\varphi_T = h$ and where h is given in polar*

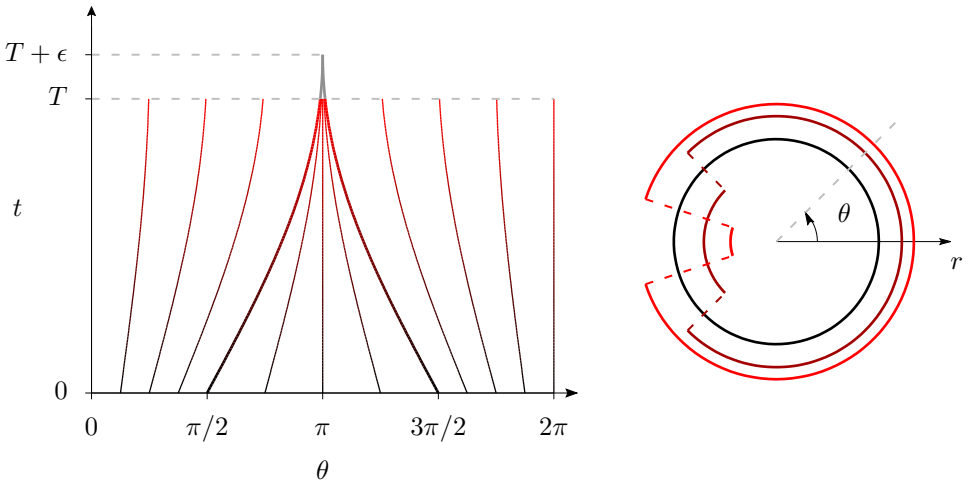


FIGURE 2. Particle trajectories $t \mapsto \varphi_t^\epsilon(\theta)$ for the linear peakon/anti-peakon solution (left) and support of fixed time marginals for the measure $(\varphi^\epsilon, \sqrt{\mathrm{Jac}(\varphi^\epsilon)})_\# \rho_0$ (right).

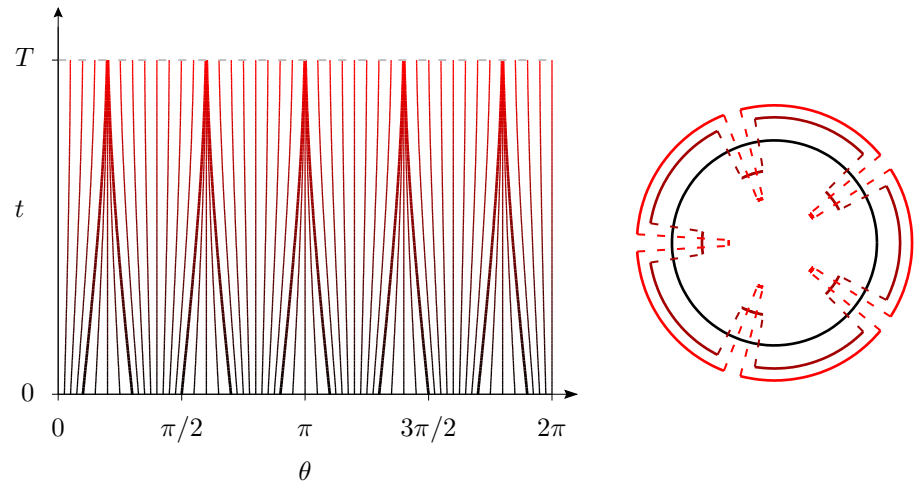


FIGURE 3. Particle trajectories $t \mapsto \hat{\varphi}_t^n(\theta)$ relative to the map constructed in proposition 7.7 (left) and support of fixed time marginals for the measure $(\hat{\varphi}^n, \sqrt{\mathrm{Jac}(\hat{\varphi}^n)})_\# \rho_0$ (right), for $n = 5$.

coordinates by $h : (\theta, \phi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2 \to (\theta+\pi, \phi+\pi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2$. *Denote by* $\rho_0 = (2\pi)^{-2} \mathrm{d}\theta \mathrm{d}\phi$ *the normalized Lebesgue measure on the torus. Then,*

$$\int_{T^2_{1,R}} \mathcal{A}([\varphi^*, \sqrt{\mathrm{Jac}(\varphi^*)}])\, \mathrm{d}\rho_0 \geq \frac{\tanh(2\pi\sqrt{1+R^2})}{2T} \pi^2 \sqrt{1+R^2}\,. \tag{7.41}$$

Proof. By an appropriate change of coordinates, it is sufficient to show the result on $T^2_{R_1,R_2}$ with

$$R_1 = \frac{R}{\sqrt{1+R^2}}\,, \quad R_2 = \sqrt{1+R^2}\,, \tag{7.42}$$

and $h : (\theta, \phi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2 \to (\theta, \phi+\pi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2$. Let $t \in [0,T] \mapsto \varphi_t^* = (\varphi_t^\theta, \varphi_t^\phi) \in \mathrm{Diff}(T^2_{R_1,R_2})$ be a smooth minimizer for these boundary conditions. Define the flow $t \in [0,T] \mapsto \tilde{\varphi}_t \in \mathrm{Diff}(T^2_{R_1,R_2})$ by

$$\tilde{\varphi}_t(\theta,\phi) = \begin{cases} (\varphi_t^\theta(2\theta,\phi)/2, \varphi_t^\phi(2\theta,\phi)) & \text{if } 0 < \theta \leq \pi\,, \\ (\varphi_t^\theta(2\theta,\phi)/2 + \pi, \varphi_t^\phi(2\theta,\phi)) & \text{if } \pi < \theta \leq 2\pi\,. \end{cases} \tag{7.43}$$

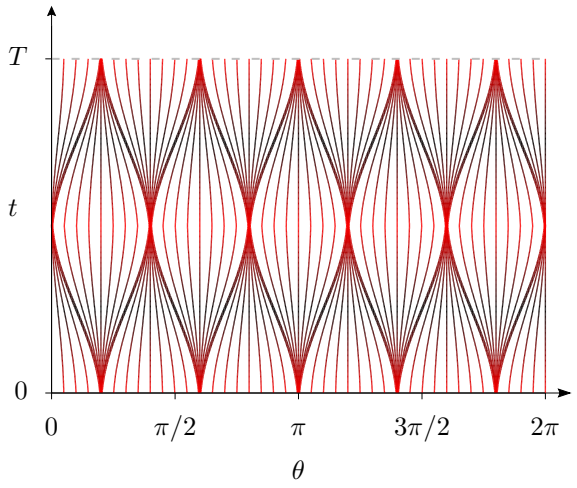


FIGURE 4. Particle trajectories $t \mapsto \varphi_t^n(\theta)$ relative to the map constructed in proposition 7.8 for $n = 5$.

Then, by direct computation,

$$\int_{T^2_{1,R}} \mathcal{A}([\tilde{\varphi}, \sqrt{\mathrm{Jac}(\tilde{\varphi})}])\,\mathrm{d}\rho_0 \leq \int_{T^2_{1,R}} \mathcal{A}([\varphi^*, \sqrt{\mathrm{Jac}(\varphi^*)}])\,\mathrm{d}\rho_0\,; \tag{7.44}$$

moreover, the inequality is strict unless $\dot{\varphi}_t^\theta = 0$ for all $t \in (0,T)$. Since φ^* is a minimizer, we conclude that we must have $\varphi_t^\theta = \theta$ for all $t \in [0,T]$. Then, we obtain the result applying lemma 7.2. □

Theorem 7.11. *Consider the generalized $H(\mathrm{div})$ geodesic problem on $T^2_{1,R}$ with coupling constraint given by uniform rotation on both circles by half of the circles length, i.e. in polar coordinates $h : (\theta,\phi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2 \to (\theta+\pi, \phi+\pi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2$. Denote by $\rho_0 = (2\pi)^{-2}\mathrm{d}\theta\mathrm{d}\phi$ the normalized Lebesgue measure on the torus. Then, the dynamic plan $\boldsymbol{\mu}^*$ in lemma 7.1, i.e. equation (7.1) with*

(7.45)

$$\zeta_t^0(\theta,\phi) = \sqrt{2}\sin(\sqrt{P^*}t)\,,\quad \zeta_t^1(\theta,\phi) = \sqrt{2}|\cos(\sqrt{P^*}t)|\,,\quad \psi_t^1(\theta,\phi) = \begin{cases} (\theta,\phi) & t \leq T/2\,,\\ (\theta+\pi,\phi+\pi) & t > T/2\,,\end{cases}$$

where $P^* = (\pi/T)^2$, is a minimizer, whereas the constant speed rotation is not a minimizer. Moreover, if R is sufficiently large no smooth flow can be a minimizer.

Proof. Consider the functional $\pi_\theta : \Omega(T^2_{1,R}) \to \Omega(S^1_1)$ defined by

$$\pi_\theta(\mathrm{z}) \coloneqq (t \in [0,T] \mapsto [\theta_t, r_t] \in \mathcal{C})\,, \tag{7.46}$$

for any $\mathrm{z} = (t \in [0,T] \mapsto [(\theta_t,\phi_t), r_t] \in \mathcal{C})$. In other words, π_θ applies at each time the canonical projection on the circle of unit radius. We observe that for any admissible dynamic plan $\boldsymbol{\mu} \in \mathcal{P}(\Omega(T^2_{1,R}))$ for the generalized $H(\mathrm{div})$ geodesic problem on the torus,

$$\boldsymbol{\mu}_\theta \coloneqq \pi_{\theta\#}\boldsymbol{\mu} \in \mathcal{P}(\Omega(S^1_1)) \tag{7.47}$$

is admissible for the generalized $H(\mathrm{div})$ geodesic problem on S^1_1 with boundary conditions associated with the map $h_\theta : \theta \in \mathbb{R}/2\pi\mathbb{Z} \to \theta+\pi$. In fact, if for example $\boldsymbol{\mu}$ satisfies the homogeneous marginal constraint with respect to the normalized measure $(2\pi)^{-2}\mathrm{d}\theta\mathrm{d}\phi$, then also $\boldsymbol{\mu}_\theta$ satisfies the same constraint since for any $t \in [0,T]$ and $f \in C^0(S^1_1)$,

$$\int_{\Omega(S^1_1)} f(\theta_t) r_t^2\,\mathrm{d}\boldsymbol{\mu}_\theta(\mathrm{z}) = \int_{\Omega(T^2_{1,R})} f(\theta_t) r_t^2\,\mathrm{d}\boldsymbol{\mu}(\mathrm{z}) = \frac{1}{2\pi}\int_{S^1_1} f(\theta)\,\mathrm{d}\theta\,, \tag{7.48}$$

and similarly for the coupling constraint. The problem on S^1_1 admits a non-deterministic minimizer, which was given in theorem 7.3 and we denote it by $\boldsymbol{\mu}^*_\theta$. Then, we have for any admissible

$\boldsymbol{\mu} \in \mathcal{P}(\Omega(T^2_{1,R}))$,

$$\mathcal{A}(\boldsymbol{\mu}) \geq \mathcal{A}(\boldsymbol{\mu}_\theta) \geq \mathcal{A}(\boldsymbol{\mu}^*_\theta) = \frac{\pi^2}{T}\,. \tag{7.49}$$

However, by lemma 7.1, the dynamic plan $\boldsymbol{\mu}^*$ defined by equation (7.45) satisfies $\mathcal{A}(\boldsymbol{\mu}^*) = \mathcal{A}(\boldsymbol{\mu}^*_\theta)$ and so it must be a minimizer. On the other hand, the action for constant speed rotation is given by $\mathcal{A}^{rot} = \pi^2(R^2+1)/T$ and therefore such a solution cannot be a minimizer since $R > 0$. Similarly, if R is sufficiently large, by lemma 7.10, no smooth minimizer can exist, since otherwise its action would be strictly larger than π^2/T. □

7.4. **Approximation of a generalized minimizer on the torus.** The generalized minimizer in theorem 7.11 is very similar to its one-dimensional counterpart of theorem 7.3. Importantly, however, the extra dimension gives us enough flexibility to produce deterministic approximations, which is the main result of this section. Such approximations will be similar in spirit to those presented in the one-dimensional case. In brief, using again peakon/anti-peakon collisions we will be able to reach the final configuration by moving two complementary subsets of the domain at different times, when they occupy a small volume.

Theorem 7.12. *Let $\boldsymbol{\mu}^*$ and h be the minimizer and the coupling, respectively, defined in theorem 7.11 on the torus $M = T^2_{1,R}$. There exists a sequence of continuous flow maps $\varphi^n : [0,T] \times M \to M$, $n \in \mathbb{N}$, such that for every $t \in [0,T]$, $\varphi^n_t : M \to M$ is smooth almost everywhere, and*

- *for all $n \in \mathbb{N}$, $\varphi^n_0 = \mathrm{Id}$ and $\varphi^n_T = h$;*
- *the sequence $\boldsymbol{\mu}_n := (\varphi^n, \sqrt{\mathrm{Jac}(\varphi^n)})_\# \rho_0$ can be rescaled to a sequence $\tilde{\boldsymbol{\mu}}_n \rightharpoonup \boldsymbol{\mu}^*$;*
- *$\mathcal{A}(\tilde{\boldsymbol{\mu}}_n) \to \mathcal{A}(\boldsymbol{\mu}^*)$.*

Proof. For simplicity, we prove the result for $R = 1$ but the argument presented here applies for any $R > 0$. In addition, performing an appropriate change of variables, one can easily verify that it is sufficient to prove the theorem with $h : (\theta, \phi) \in \mathbb{R}^2/(2\pi\mathbb{Z})^2 \to (\theta, \phi + \pi)$ and $\boldsymbol{\mu}^*$ defined as in equation (7.45), but with ψ^1 defined by

$$\psi^1_t(\theta,\phi) = \begin{cases} (\theta,\phi) & t \leq T/2\,, \\ (\theta,\phi+\pi) & t > T/2\,. \end{cases} \tag{7.50}$$

For each $n \in \mathbb{N}$, the map φ^n will be constructed using two basic flows. The first is defined as follows. Fix a sequence $\epsilon_n = \epsilon_0/n^3$, $n \in \mathbb{N}$, where ϵ_0 is a sufficiently small constant. Moreover, for any $\epsilon > 0$ consider the set $B_\epsilon \subset S^1_1$ defined by

$$B_\epsilon := \bigcup_{i=0}^{n-1} \left[\frac{\pi}{n}(2i+1) - \frac{\epsilon}{2}, \frac{\pi}{n}(2i+1) + \frac{\epsilon}{2}\right]\,, \tag{7.51}$$

and let $\phi^n_{rot} : S^1_1 \to S^1_1$ such that $0 \leq \phi^n_{rot} \leq \pi$, $\phi^n_{rot}(\theta) = \pi$ for all $\theta \in B_{\epsilon_n}$ and $\phi^n_{rot}(\theta) = 0$ for all $\theta \in S^1_1 \setminus B_{2\epsilon_n}$. For $k = 0, 1$, we let $\varphi^{k,n}_{rot} : [0, \sqrt{\epsilon_n}] \times T^2_{1,1} \to T^2_{1,1}$ be the flow defined by

$$(\varphi^{0,n}_{rot})_t(\theta,\phi) := \left(\theta, \phi + \frac{t}{\sqrt{\epsilon_n}}\phi^n_{rot}(\theta)\right)\,,\ (\varphi^{1,n}_{rot})_t(\theta,\phi) := \left(\theta, \phi + \frac{t}{\sqrt{\epsilon_n}}(\pi - \phi^n_{rot}(\theta))\right)\,. \tag{7.52}$$

Consider now the flow $\hat{\varphi}^n$ defined in equation (7.26), with ϵ_n defined as above. With a slight abuse of notation, we will also denote by $\hat{\varphi}^n$ its canonical extension to the torus which leaves the ϕ coordinate fixed. Moreover, for any $\alpha \in \mathbb{R}/2\pi\mathbb{Z}$ denote by $R^\theta_\alpha : T^2_{1,1} \to T^2_{1,1}$ the map $R^\theta_\alpha(\theta,\phi) := (\theta + \alpha, \phi)$. Then, we define the flow $\varphi^{0,n}_{exp} : [\sqrt{\epsilon_n}, T/2] \times T^2_{1,1} \to T^2_{1,1}$ by

$$(\varphi^{0,n}_{exp})_t(\theta,\phi) := \begin{cases} \hat{\varphi}^n_{a_n(t)}((\hat{\varphi}^n_T)^{-1}(\theta,\phi)) & \text{if } t \leq \frac{T}{4}\,, \\ R^\theta_{-\pi/n} \circ \hat{\varphi}^n_{4t-T}\left(R^\theta_{\pi/n} \circ (\hat{\varphi}^n_T)^{-1}(\theta,\phi)\right) & \text{if } \frac{T}{4} < t \leq \frac{T}{2}\,, \end{cases} \tag{7.53}$$

where $a_n(t) := T(T-4t)(T-4\sqrt{\epsilon_n})^{-1}$. Note that setting $\epsilon_n = 0$ this flow coincides with the canonical extension to the torus of the flow in equation (7.38). Similarly,

$$\varphi^{1,n}_{exp} := R^\theta_{-\pi/n} \circ \varphi^{0,n}_{exp} \circ R^\theta_{\pi/n}. \tag{7.54}$$

We construct the sequence φ^n by glueing together the maps $\varphi^{k,n}_{rot}$ and $\varphi^{k,n}_{exp}$ so that for each $n \in \mathbb{N}$ the final flow consists of four stages: in the first, n stripes of the domain rotate while the rest of the domain stays put as prescribed by $\varphi^{0,n}_{rot}$; in the second, the stripes expand up to a symmetric configuration in which the rest of the domain occupies stripes of the same size, as

prescribed by $\varphi^{0,n}_{exp}$; in the third, the rest of the points rotate as prescribed by $\varphi^{1,n}_{rot}$; finally, we use $\varphi^{1,n}_{exp}$ to compress the stripes to their original size. More precisely,

$$\varphi^n_t := \begin{cases} (\hat\varphi^{0,n}_{rot})_t & \text{if } t \le \sqrt{\epsilon_n}\,, \\ (\hat\varphi^{0,n}_{exp})_t \circ (\hat\varphi^{0,n}_{rot})_{\sqrt{\epsilon_n}} & \text{if } \sqrt{\epsilon_n} < t \le \tfrac{T}{2}\,, \\ (\hat\varphi^{1,n}_{rot})_{t-T/2} \circ (\hat\varphi^{0,n}_{exp})_{T/2} \circ (\hat\varphi^{0,n}_{rot})_{\sqrt{\epsilon_n}} & \text{if } \tfrac{T}{2} < t \le \tfrac{T}{2} + \sqrt{\epsilon_n}\,, \\ (\hat\varphi^{1,n}_{exp})_{t-T/2} \circ (\hat\varphi^{1,n}_{rot})_{\sqrt{\epsilon_n}} \circ (\hat\varphi^{0,n}_{exp})_{T/2} \circ (\hat\varphi^{0,n}_{rot})_{\sqrt{\epsilon_n}} & \text{if } \tfrac{T}{2} + \sqrt{\epsilon_n} < t \le T\,. \end{cases} \tag{7.55}$$

A graphical representation of this flow is given in figure 5 for $n = 1$ (so that we have only one stripe) and in the original coordinates (so that the boundary conditions are those associated with double rotation).

Note that the flow defined in equation (7.55) is very similar to the one defined in proposition 7.8, whose canonical extension to the torus will be denoted by $\varphi^{0,n}$. As in proposition 7.8, we define again the rescaled measure $\tilde{\boldsymbol{\mu}}_n$ using equation (7.40). This means that for any Lipschitz continuous bounded functional $\mathcal{F} : \Omega \to \mathbb{R}$,

$$\int_\Omega \mathcal{F}(\mathrm{z})\,\mathrm{d}\tilde{\boldsymbol{\mu}}_n(\mathrm{z}) = \frac{1}{4\pi^2}\int_{T^2_{1,1}} \mathcal{F}\left(\left[\varphi^n \circ (\varphi^n_{T/4})^{-1}(\theta,\phi), \bar\lambda^n(\theta,\phi)\right]\right)\mathrm{d}\theta\,\mathrm{d}\phi\,, \tag{7.56}$$

where

$$\bar\lambda^n_t := \left(\frac{\mathrm{Jac}(\varphi^n_t)}{\mathrm{Jac}(\varphi^n_{T/4})}\right)^{1/2} \circ (\varphi^n_{T/4})^{-1} = \left(\mathrm{Jac}(\varphi^n_t \circ (\varphi^n_{T/4})^{-1})\right)^{1/2}\,. \tag{7.57}$$

Note that equation (7.56) is a direct consequence of the definition of the dilation map and the change of variables formula. Due to proposition 7.8 and the way φ^n is constructed, to prove the convergence $\tilde{\boldsymbol{\mu}}_n \rightharpoonup \boldsymbol{\mu}^*$, it is sufficient to focus on the interval $[0, T/2]$ and check that $I^n \to 0$, where

$$I^n := \int_{T^2_{1,1}} \sup_{t\in[0,T/2]} d_{\mathcal{C}}([\varphi^n_t \circ (\varphi^n_{T/4})^{-1}, \bar\lambda^n_t], [\varphi^{0,n}_t \circ (\varphi^{0,n}_{T/4})^{-1}, \bar\lambda^{0,n}_t])\,\mathrm{d}\theta\,\mathrm{d}\phi \tag{7.58}$$

and $\bar\lambda^{0,n}_t := \left(\mathrm{Jac}(\varphi^n_t \circ (\varphi^{0,n}_{T/4})^{-1})\right)^{1/2}$. Because of the similar structure of the flows φ^n and $\varphi^{0,n}$, I^n reduces to

$$I^n = \int_{T^2_{1,1}} \sup_{t\in[0,T/4]} d_{\mathcal{C}}([\varphi^n_t \circ (\varphi^n_{T/4})^{-1}, \bar\lambda^n_t], [\varphi^{0,n}_t \circ (\varphi^{0,n}_{T/4})^{-1}, \bar\lambda^{0,n}_t])\,\mathrm{d}\theta\,\mathrm{d}\phi\,. \tag{7.59}$$

Let $A_\epsilon := \varphi^{0,n}_{T/4}(B_\epsilon \times S^1_1)$, for any $\epsilon > 0$. We decompose $I^n = I^{0,n} + I^{1,n}$ where $I^{0,n}$ and $I^{1,n}$ are the integrals over $A_{2\epsilon_n}$ and $T^2_{1,1} \setminus A_{2\epsilon_n}$, respectively. Define for $0 \le a < b \le T/4$

$$I^{0,n}_{a,b} := \int_{A_{2\epsilon_n}} \sup_{t\in[a,b]} d_{\mathcal{C}}([\varphi^n_t \circ (\varphi^n_{T/4})^{-1}, \bar\lambda^n_t], [\varphi^{0,n}_t \circ (\varphi^{0,n}_{T/4})^{-1}, \bar\lambda^{0,n}_t])\,\mathrm{d}\theta\,\mathrm{d}\phi\,. \tag{7.60}$$

We have $I^n \le I^{0,n}_{0,\sqrt{\epsilon_n}} + I^{0,n}_{\sqrt{\epsilon_n},T/4} + I^{1,n}$. By continuity of the flow maps it is easy to verify that $I^{0,n}_{\sqrt{\epsilon_n},T/4} \to 0$ and $I^{1,n} \to 0$ as $n \to +\infty$. On the other hand, by construction $0 < \bar\lambda^n, \bar\lambda^{0,n} < \sqrt{2}$. Therefore, by the triangular inequality

$$\begin{aligned} I^{0,n}_{0,\sqrt{\epsilon_n}} &\le \int_{A_{2\epsilon_n}} \sup_{t\in[0,\sqrt{\epsilon_n}]} (\bar\lambda^n_t + \bar\lambda^{0,n}_t)\,\mathrm{d}\theta\,\mathrm{d}\phi\,, \\ &\le 4\pi n\epsilon_n(2\sqrt{2}) + \int_{B_{\pi/n}\times S^1_1} \sup_{t\in[0,\sqrt{\epsilon_n}]} (\bar\lambda^n_t + \bar\lambda^{0,n}_t)\,\mathrm{d}\theta\,\mathrm{d}\phi\,, \end{aligned} \tag{7.61}$$

where on the second line, we decomposed the integral over the part of $A_{2\epsilon_n}$ that gets stretched and the part that gets compressed under $\varphi^n \circ (\varphi^n_{T/4})^{-1}$ for $t \in [0, \sqrt{\epsilon_n}]$. In particular, the integrand in the second line tends to 0 as $n \to +\infty$, which yields $I^n \to 0$. A similar argument can be applied on the interval $[T/2, T]$, which proves that $\tilde{\boldsymbol{\mu}}_n \rightharpoonup \boldsymbol{\mu}^*$.

In order to prove convergence of the action, in view of lemma 7.5, it is sufficient to show

$$\int_{T^2_{1,1}} \mathcal{A}([\varphi^{k,n}_{rot}(\theta,\phi), 1])\,\mathrm{d}\theta\,\mathrm{d}\phi \to 0\,, \tag{7.62}$$

for $k = 0, 1$, as $n \to +\infty$, where the action is computed over the time interval $[0, \sqrt{\epsilon_n}]$. For all $n \in \mathbb{N}$, under the flow $\varphi^{k,n}_{rot}$ only points with $\theta \in B_{2\epsilon_n}$ rotate with velocity bounded by $\pi/\sqrt{\epsilon_n}$; hence

$$\int_{T^2_{1,1}} \mathcal{A}([\varphi^{k,n}_{rot}(\theta,\phi),1])\,\mathrm{d}\theta\,\mathrm{d}\phi \leq 2\pi(n2\epsilon_n)\frac{\pi^2}{\sqrt{\epsilon_n}} = \sqrt{\frac{\epsilon_0}{n}}4\,\pi^3\,, \tag{7.63}$$

which concludes the proof. □

Remark 7.13. *As for the one-dimensional case (see remark 7.9), the maps defined by equation (7.55) are piecewise smooth in space since their Jacobian is piecewise constant with a finite number of discontinuities. Also in this case, it is sufficient to repeat the construction above using a regularized version of the linear peakon/anti-peakon collision, to obtain a sequence of smooth diffeomorphisms satisfying theorem 7.12.*

Remark 7.14. *In theorem 7.3 we proved that our relaxation is not tight on S^1_R (for R sufficiently large), whereas theorem 7.11 suggests it is tight for $d \geq 2$. It should be noted that the situation is similar for the incompressible Euler equations. In fact, Shnirelman proved that Brenier's relaxation is not tight for $d = 2$ but it is tight when $d = 3$, as in this case any generalized incompressible flows can be approximated using deterministic maps* [35].

8. Discrete generalized solutions

There are two main obstacles in translating problem 4.2 to the discrete setting. On one hand, we need to make computations on an unbounded domain; on the other, we need to be able to single out a representative for the equivalence class of minimizers with respect to rescaling. However, if one is interested in simulating solutions that are not singular (see definition 4.11), it is appropriate to enforce the strong coupling constraint in (4.6) instead of (4.8). Hence, if we substitute $\mathcal{C}$ by $\mathcal{C}_R$ for a fixed $R > 1$ and use the strong coupling constraint in the generalized $H(\mathrm{div})$ geodesic problem, we obtain a modified formulation that is able to reproduce a particular class of solutions, which includes all deterministic solutions with bounded Jacobian. In this section we describe a numerical algorithm based on entropy regularization and Sinkhorn algorithm that solves such a modified formulation. Our scheme is based on similar methods for the incompressible Euler equations developed in [33, 6, 5]. We also provide some numerical results illustrating the behavior of generalized $H(\mathrm{div})$ geodesics.

8.1. **Discrete formulation.** We set $M = [0,1]^d$ and consider a uniform discretization with points $\{x_i\}_{i=1}^{N_x}$, and a discretization of the interval $(0, R]$ with points $\{r_i\}_{i=1}^{N_r}$ such that $r_j = 1$ for a fixed $j \in \{1, \dots, N_r\}$. These induce a discretization of the cone with points $\{z_i\}_{i=1}^{N}$ where $N = N_x N_r$. Similarly, we also consider a uniform discretization $\{t_i\}_{i=1}^{K}$ of $[0, T]$. Generalized flows are then replaced by a coupling arrays $\boldsymbol{\mu} \in (\mathbb{R}^N_{\geq 0})^K$. Note that we can incorporate the boundary condition $\lambda_0 = 1$ by reducing the dimension of $\boldsymbol{\mu}$. In particular, we now denote by π_x and π_r the canonical projections from $M \times (0, R]$ to M and $(0, R]$ respectively. We use the same notation to indicate the maps $\pi_x : \{1, \dots, N\} \to \{1, \dots, N_x\}$ and $\pi_r : \{1, \dots, N\} \to \{1, \dots, N_r\}$ mapping directly the discretization indices. Then, we set for any $\{j_1, \dots, j_K\} \in \{1, \dots, N\}^K$,

$$\boldsymbol{\mu}_{j_1,\dots,j_K} = \mathbb{1}_{\{\pi_r(z_{j_1})=1\}}\tilde{\boldsymbol{\mu}}_{\pi_x(j_1),j_2,\dots,j_K}\,, \tag{8.1}$$

where $\mathbb{1}$ is the indicator function and $\tilde{\boldsymbol{\mu}} \in \mathbb{R}^{N_x}_{\geq 0} \times (\mathbb{R}^N_{\geq 0})^{K-1}$. We denote by Π_0 the set of couplings satisfying (8.1). The marginal at a given time t_k is a discrete measure on $M \times (0, R]$. We denote this by $S_k(\boldsymbol{\mu}) \in \mathbb{R}^N_{\geq 0}$, and it is defined as follows:

$$[S_k(\boldsymbol{\mu})]_j = \sum_{j_1,\dots,j_{k-1},j_{k+1},\dots,j_K} \boldsymbol{\mu}_{j_1,\dots,j_{k-1},j,j_{k+1},\dots,j_K}\,. \tag{8.2}$$

We denote by $M_n : \mathbb{R}^N_{\geq 0} \to \mathbb{R}^{N_x}_{\geq 0}$ the nth moment taken in the radial direction, i.e.

$$M_n[A]_i = \sum_{j,\pi_x(j)=i} \pi_r(z_j)^n A_j\,. \tag{8.3}$$

Hence the constraint in (4.4) becomes

$$M_2[S_k(\boldsymbol{\mu})]_i = 1/N_x\,. \tag{8.4}$$

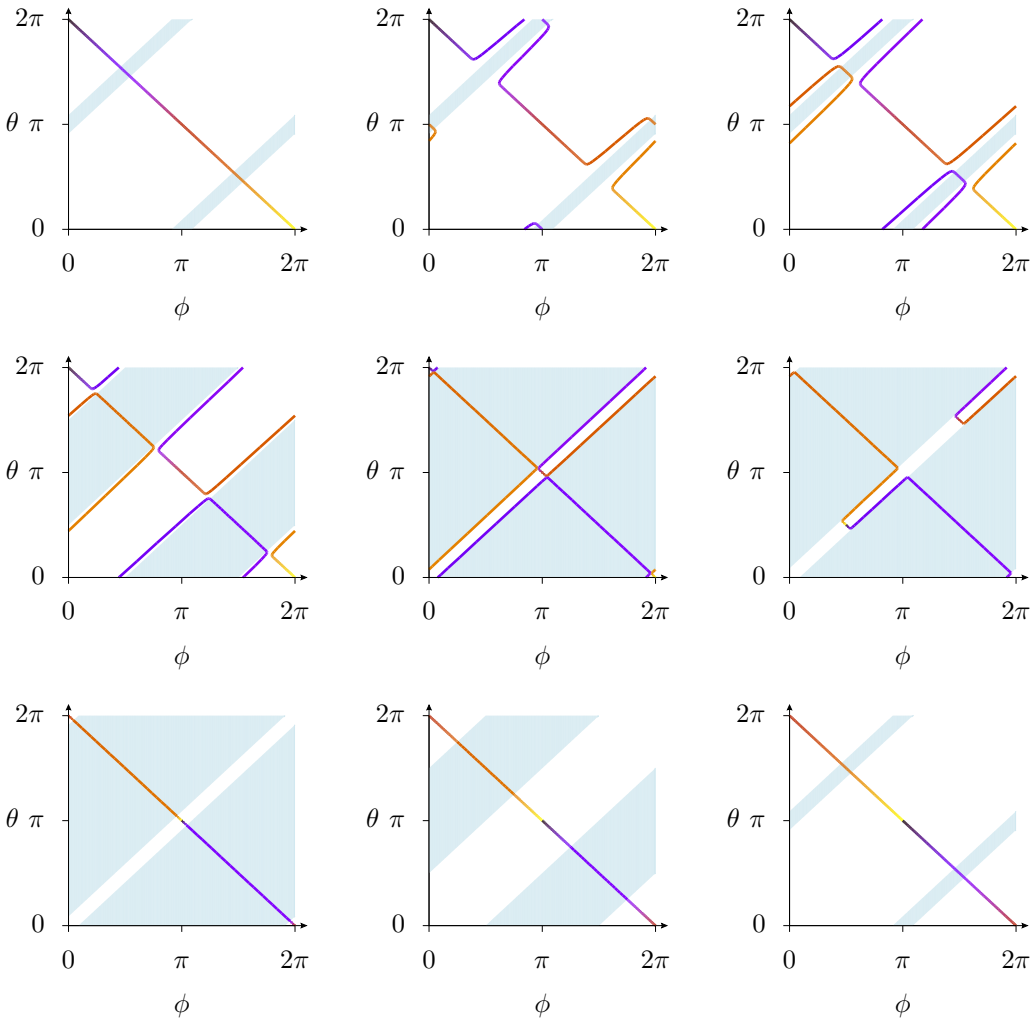


FIGURE 5. Particle trajectories for the flow φ^n in (7.55) for $n = 1$ (in appropriate coordinates to determine double rotation on the torus). We use different colors to label particles and follow their motion. The stripes indicate the particles in between the peakons' peaks. In the limit, the trajectories of such particles lifted to the cone will start and end at the apex.

Moreover, we denote by Π the set of admissible coupling arrays,

$$\Pi = \{\boldsymbol{\mu} \in \Pi_0;\, \forall i,\, M_2[S_k(\boldsymbol{\mu})]_i = 1/N_x\}\,. \tag{8.5}$$

The constraint on the coupling between time 0 and T can be enforced weakly by including it directly in the cost, which is given by the following array

$$C_{j_1,\dots,j_K} = \frac{K-1}{T}\sum_{k=1}^{K-1} d_{\mathcal{C}}(z_{j_k}, z_{j_{k+1}})^2 + \alpha\, d_{\mathcal{C}}(z_{j_K}, (h(\pi_x(z_{j_1})), \sqrt{|\mathrm{Jac}(h)|}))^2\,, \tag{8.6}$$

where $\alpha > 0$ is a parameter. The regularized discrete problem is then,

$$\min_{\boldsymbol{\mu}\in\Pi} \langle C, \boldsymbol{\mu}\rangle - \epsilon E(\boldsymbol{\mu})\,, \tag{8.7}$$

where $\epsilon > 0$ is another parameter and $E(\boldsymbol{\mu})$ is the entropy of the coupling defined by

$$E(\boldsymbol{\mu}) = -\langle \boldsymbol{\mu}, \log(\boldsymbol{\mu}) - 1\rangle\,. \tag{8.8}$$

Problem (8.7) can be solved by means of alternating projections which consist in enforcing recursively the marginal constraints at the different time levels. In particular, we consider the

following augmented functional

$$\min_{\boldsymbol{\mu}} \langle C, \boldsymbol{\mu}\rangle - \epsilon E(\boldsymbol{\mu}) - \sum_{i,k} p_i^k (M_2[S_k(\boldsymbol{\mu})]_i - 1/N_x)\,, \tag{8.9}$$

where $p^k \in \mathbb{R}^{N_x}$ for all $k \in \{1,\dots,K\}$. From (8.9) we obtain

$$\boldsymbol{\mu}_{j_1,\dots,j_K} = e^{-\frac{C_{j_1,\dots,j_K}}{\epsilon}} e^{\sum_k p^k_{\pi_x(j_k)} r^2_{\pi_r(j_k)}}\,. \tag{8.10}$$

Enforcing the constraint at time level n allows us to solve for p^n given the set $\{p^k\}_{k\neq n}$. This amounts to solving the following nonlinear equation for all $i \in \{1,\dots,N_x\}$,

$$\sum_j B_{i,j} e^{p_i^n r_j^2} r_j^2 = 1/N_x\,, \tag{8.11}$$

where

$$B = S_n\left[e^{-\frac{C_{j_1,\dots,j_K}}{\epsilon}} e^{\sum_{k,k\neq n} p^k_{\pi_x(j_k)} r^2_{\pi_r(j_k)}}\right]\,. \tag{8.12}$$

Due to the structure of the cost, we only need to store two arrays $D^0, D^1 \in \mathbb{R}^N \times \mathbb{R}^N$, given by

$$D^0_{i,j} = d_{\mathcal{C}}(z_i, z_j)^2\,, \qquad D^1_{i,j} = d_{\mathcal{C}}(z_i, (h(\pi_x(z_j)), \sqrt{|\mathrm{Jac}(h)|}))^2\,. \tag{8.13}$$

8.2. **Numerical results: from CH to Euler.** We now present some numerical results illustrating the behavior of generalized solutions of the $H(\mathrm{div})$ geodesic problem and their relation to generalized incompressible Euler solutions. We consider two types of couplings to define the boundary conditions: a classical deterministic coupling, which we use to illustrate the emergence of discontinuities in the flow map, and a generalized coupling that obliges particles to cross each other so that the solution is not deterministic. For both cases, the domain will be the one-dimensional interval $M = [0,1]$ and $T = 1$.

A peakon-like solution. Consider the continuous map $h : [0,1] \to [0,1]$, defined by

$$h(x) = \begin{cases} 1.4\,x & \text{if } x \leq 0.5\,, \\ 0.6\,x + 0.4 & \text{if } x > 0.5\,. \end{cases} \tag{8.14}$$

We use this map to define the coupling on the cone as in equation (4.6). We compute the solution using the algorithm presented in the previous section with $N_x = 40$, $N_r = 41$, $0.55 \leq r \leq 1.45$, $K = 35$, $\alpha = 40$, $\epsilon = 5 \cdot 10^{-4}$. In figure 6 we show the evolution of the transport plan on the domain M given by $(e^M_{0,t_k})_{\#}\boldsymbol{\mu} \in \mathcal{P}(M^2)$, where $e^M_{0,t_k}(\mathrm{z}) \coloneqq (x_0, x_{t_k})$, for selected times. In figure 7 we show the evolution of the marginals on the cone given by $(e_{t_k})_{\#}\boldsymbol{\mu} \in \mathcal{P}(\mathcal{C})$ for the same times. We remark that the dynamic plan is approximately deterministic since there is very little diffusion of the mass in the domain, which is at least partially due to the entropic regularization. In addition the discontinuity in the Jacobian of the coupling map propagates to the whole solution, which resembles a peakon with the discontinuity point corresponding to the peak of the peakon.

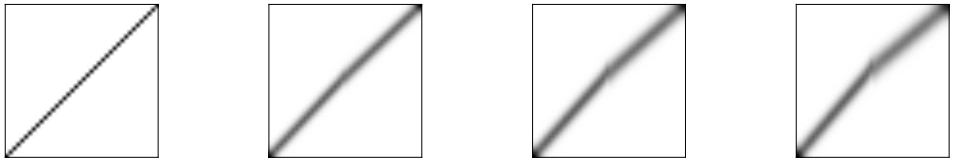
A non-deterministic solution. The homogeneous marginal constraint allows us to consider very general couplings even defined by non-injective maps or maps that do not preserve the local orientation of the domain. Measure-preserving maps provide a special example since these were used by Brenier to define boundary conditions for generalized incompressible Euler flows. In fact if h is measure-preserving, i.e. $h_{\#}\rho_0 = \rho_0$, then we can use as coupling

$$\boldsymbol{\gamma} = [(\mathrm{Id}, 1), (h, 1)]_{\#}\rho_0\,. \tag{8.15}$$

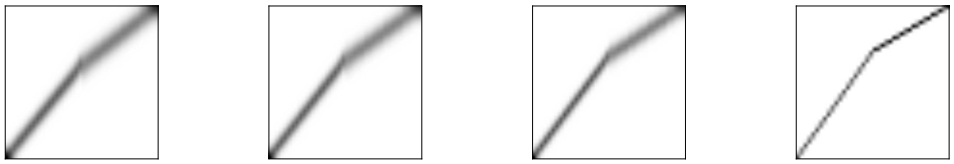
Here, we take $h : [0,1] \to [0,1]$ to be the map

$$h(x) = 1 - x\,, \tag{8.16}$$

which can only be realized by a non-deterministic plan. We compute the discrete solution associated with such boundary conditions with $N_x = 40$, $N_r = 41$, $0.6 \leq r \leq 1.4$, $K = 35$, $\alpha = 40$, $\epsilon = 5 \cdot 10^{-4}$. As before, we show the evolution of the transport plan on the domain M given by $(e^M_{0,t_k})_{\#}\boldsymbol{\mu} \in \mathcal{P}(M^2)$ in figure 8. In figure 9 we show the evolution of the marginals on the cone given by $(e_{t_k})_{\#}\boldsymbol{\mu} \in \mathcal{P}(\mathcal{C})$. The transport plan evolution is remarkably similar to that of the incompressible Euler equation for the same coupling (see, e.g., [6]). However, the two do not

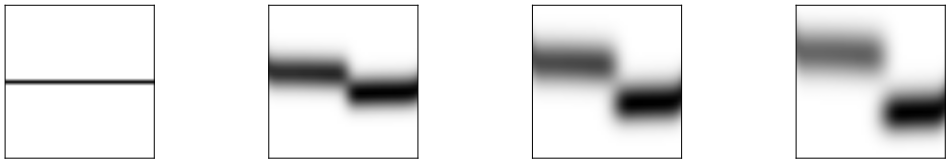


(a) $k = 1$ (b) $k = 6$ (c) $k = 11$ (d) $k = 16$

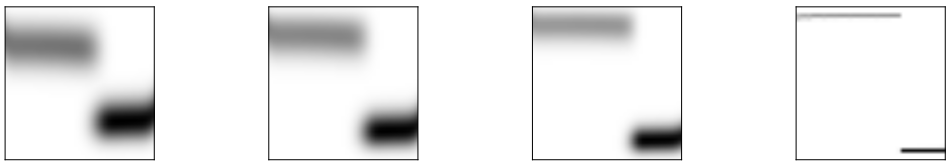


(e) $k = 20$ (f) $k = 25$ (g) $k = 30$ (h) $k = 35$

FIGURE 6. Transport couplings $(e^M_{0,t_k})_{\#}\boldsymbol{\mu}$ on $M \times M$ for the peakon-like solution associated with the boundary conditions specified by the map in equation (8.14).



(a) $k = 1$ (b) $k = 6$ (c) $k = 11$ (d) $k = 16$



(e) $k = 20$ (f) $k = 25$ (g) $k = 30$ (h) $k = 35$

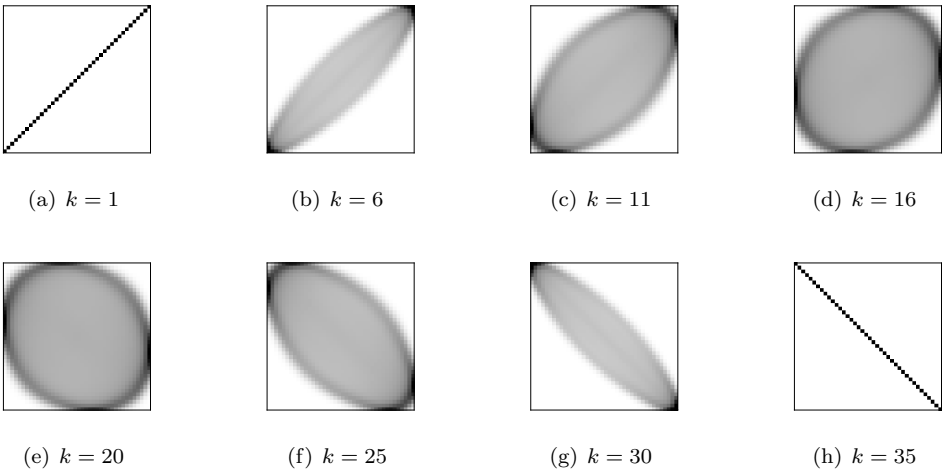
FIGURE 7. Fixed time marginals $(e_{t_k})_{\#}\boldsymbol{\mu}$ on the cone section $M \times [r_{min}, r_{max}]$ ($r_{min} = 0.55$, $r_{max} = 1.45$) for the peakon-like solution associated with the boundary conditions specified by the map in equation (8.14).

coincide as it is evident from the marginals on the cone in figure 9. In the case of incompressible Euler, these marginals are concentrated on $r = 1$ for every time, i.e. the transport plan remains measure-preserving during the evolution. This is clearly not the case for the generalized CH solution, for which also the Jacobian appears to be non-deterministic.

9. Outlook

There are several natural questions that were not addressed in this paper and that we reserve to future work:

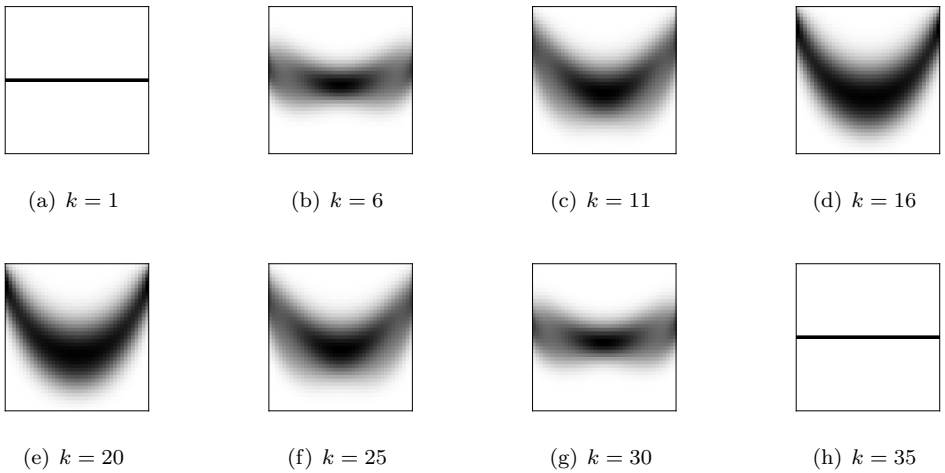
- *Tight relaxation.* Brenier's relaxation of incompressible Euler is not tight in two dimensions but it is in three dimensions due to the work of Shnirelman [35]. It is an open question whether a similar result holds for the generalized problem studied in this paper. The approximation results in section 7 suggest that this is the case. In particular, we conjecture that our formulation is a tight relaxation of the $H(\mathrm{div})$ geodesic problem in dimension $d \geq 2$.



(a) $k = 1$ (b) $k = 6$ (c) $k = 11$ (d) $k = 16$

(e) $k = 20$ (f) $k = 25$ (g) $k = 30$ (h) $k = 35$

FIGURE 8. Transport couplings $(e^M_{0,t_k})_{\#}\boldsymbol{\mu}$ on $M \times M$ for the non-deterministic solution associated to the boundary conditions specified by the map in equation (8.16).



(a) $k = 1$ (b) $k = 6$ (c) $k = 11$ (d) $k = 16$

(e) $k = 20$ (f) $k = 25$ (g) $k = 30$ (h) $k = 35$

FIGURE 9. Fixed time marginals $(e_{t_k})_{\#}\boldsymbol{\mu}$ on the cone section $M \times [r_{min}, r_{max}]$ ($r_{min} = 0.6$, $r_{max} = 1.4$) for the non-deterministic associated with the boundary conditions specified by the map in equation (8.16).

As for the generalized Euler solutions, a better understanding of the structure of minimizing generalized flows is of theoretical interest:

- *Occurrence of singular solutions.* In this paper we did not fully characterize the emergence of singular solutions. Even for the case of rotation on the circle or on the torus, for example, we did not prove that these are the unique minimizers for the problem. In addition, such examples suggest that singular solutions appear whenever particles' displacement is sufficiently large. It would be interesting to give a full characterization in this direction, specifying when solutions are singular in terms of the boundary conditions and the dimension and geometry of the base space M;
- *Regularity of the pressure.* Brenier's result on the existence and uniqueness of the pressure in incompressible Euler was subsequently improved by Ambrosio and Figalli [1] in terms of regularity of the pressure field. It is natural to ask whether such a result can be extended to the generalized $H(\mathrm{div})$ geodesic problem. This question is related to the

previous one, due to the fact that a sufficiently regular pressure field can prevent the occurrence of singular solutions as it can be deduced from the proofs in section 6.

Addressing these theoretical questions will also guide the development of numerical schemes which are better suited to the formulation considered in this paper than methods based on entropic regularization. A viable alternative in this context is given by semi-discrete methods [29] (see also the schemes developed for the incompressible Euler equations in [30, 18]), whose use for the generalized $H(\mathrm{div})$ geodesic problem will also be studied in future work.

Acknowledgments

The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement n. PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France. The authors would also like to acknowledge the support from the project MAGA ANR-16-CE40-0014 (2016-2020).

Appendix A. Proof of lemma 4.3

Proof. Here we prove that the homogeneous marginal constraint can be enforced at each time rather than in integral form as in equation (4.9).

First, we prove that the constraint in equation (4.9) implies the one in equation (4.11). In order to show this, for any fixed $t^* \in [0,T]$ and $f \in C^0(M)$, consider the following functionals

$$\mathcal{F}(\mathrm{z}) \coloneqq r_{t^*}^2 f(x_{t^*})\,, \qquad \mathcal{F}_n(\mathrm{z}) \coloneqq \int_0^T r_t^2 f(x_t)\delta_{n,t^*}(t)\,\mathrm{d}t\,, \tag{A.1}$$

where $\delta_{n,t^*} : [0,T] \to \mathbb{R}$, $n \in \mathbb{N}$, is a Dirac sequence of continuous functions converging to δ_{t^*}. Then for any $\mathrm{z} \in \Omega$, $\mathcal{F}_n(\mathrm{z}) \to \mathcal{F}(\mathrm{z})$ as $n \to +\infty$. Moreover, using Jensen's inequality,

$$\begin{aligned}
\mathcal{F}_n(\mathrm{z}) &\leq \|f\|_{C^0} \int_0^T r_t^2 \delta_{n,t^*}\,\mathrm{d}t \\
&\leq 2\|f\|_{C^0} \left(r_0^2 + \int_0^T (r_t - r_0)^2 \delta_{n,t^*}\,\mathrm{d}t \right) \\
&\leq 2\|f\|_{C^0} \left(r_0^2 + \int_0^T \dot{r}_t^2\,\mathrm{d}t \int_0^T t\,\delta_{n,t^*}\,\mathrm{d}t \right) \\
&\leq 2\|f\|_{C^0} \left(r_0^2 + T\mathcal{A}(\mathrm{z}) \right)\,.
\end{aligned} \tag{A.2}$$

The right-hand side is $\boldsymbol{\mu}$-integrable since $\mathcal{A}(\boldsymbol{\mu}) < +\infty$ and because of the coupling constraint. Hence, we get the result by the dominated convergence theorem.

Similarly, if $f \in C^0([0,T] \times M)$, we take

$$\mathcal{F}(\mathrm{z}) \coloneqq \int_0^T f(t, x_t) r_t^2\,\mathrm{d}t\,, \qquad \mathcal{F}_n(\mathrm{z}) \coloneqq \frac{T}{K} \sum_{k=0}^{K} f(t_k, x_{t_k}) r_{t_k}^2\,, \tag{A.3}$$

where $t_k \coloneqq kT/K$. Then for any $\mathrm{z} \in \Omega$, $\mathcal{F}_n(\mathrm{z}) \to \mathcal{F}(\mathrm{z})$ as $n \to +\infty$. Moreover,

$$\begin{aligned}
\mathcal{F}_n(\mathrm{z}) &\leq 2\|f\|_{C^0} \left(r_0^2 + \frac{T}{K} \sum_{k=1}^{K} (r_{t_k} - r_0)^2 \right) \\
&\leq 2\|f\|_{C^0} \left(r_0^2 + \frac{T}{K} \sum_{k=1}^{K} t_k \int_0^{t_k} \dot{r}_t^2\,\mathrm{d}t \right) \\
&\leq 2\|f\|_{C^0} \left(r_0^2 + T^2 \mathcal{A}(\mathrm{z}) \right)\,,
\end{aligned} \tag{A.4}$$

and we can apply again the dominated convergence theorem to conclude the proof. □

Appendix B. Proof of lemma 6.1

Proof. Throughout this proof, most metric operations are performed with respect to the cone metric $g_{\mathcal{C}}$, so to simplify the notation we will use $|\cdot|$ and $\langle\cdot,\cdot\rangle$ to denote, respectively, the norm and the inner product both on $T\mathcal{C}$ and $\mathbb{R}^d$ according to the context. Moreover, given a vector field u on the cone and a curve $t \mapsto p(t) \in \mathcal{C}$, $\nabla_t u(p(t)) \coloneqq \nabla_{\dot p(t)} u(p(t))$ is the covariant derivative of u at $p(t)$ with respect to the vector $\dot p(t)$.

Given a smooth solution (φ, λ) and a fixed $x \in M$, let $\mathrm{z}^* = [\mathrm{x}^*, \mathrm{r}^*] \in \Omega$ be the curve defined by $\mathrm{x}^* : t \to x_t^* \coloneqq \varphi_t(x)$ and $\mathrm{r}^* : t \to r_t^* \coloneqq \lambda_t(x)$. We want to show that for any curve $\mathrm{z} \in AC^2([0,T];\mathcal{C})$ such that $\mathrm{z} \neq \mathrm{z}^*$, $z_0 = z_0^*$ and $z_T = z_T^*$, we have $\mathcal{B}(\mathrm{z}) > \mathcal{B}(\mathrm{z}^*)$. We proceed in two steps: first we show that the inequality holds when z is smooth and when the geodesics between z_t^* and z_t are smooth for all $t \in [0,T]$; then we derive sufficient conditions for which the inequality holds also for curves z which are farther away from z^*.

Let $s \in [0,1] \mapsto c(t,s) \in \mathcal{C}$ be a family of geodesics parameterized by $t \in [0,T]$ such that $c(t,0) = z_t^*$ and $c(t,1) = z_t$. In order for such geodesics to be smooth we need to assume

$$|x_t^* - x_t| < \pi\,, \qquad \forall\, t \in [0,T]\,. \tag{B.1}$$

Let $J(t,s) \coloneqq \partial_t c(t,s)$, which is a Jacobi field when restricted to any geodesic $c(t,\cdot)$ for any fixed $t \in [0,T]$. Moreover, $J(t,0) = \dot z_t^*$ and $J(t,1) = \dot z_t$. Hence we want to show that

$$\int_0^T |J(t,0)|^2 - \Psi_p(t,c(t,0))\,\mathrm{d}t \le \int_0^T |J(t,1)|^2 - \Psi_p(t,c(t,1))\,\mathrm{d}t\,. \tag{B.2}$$

Let $C \coloneqq \sup_{t\in[0,T]} \sup_{x\in M} |\mathrm{Hess}\,\Psi_p|$. The Taylor expansion of $\Psi_p(t,c(s,t))$ with respect to s at $s = 0$ yields

$$\Psi_p(t,c(t,1)) - \Psi_p(t,c(t,0)) - \langle \nabla\Psi_p(t,c(t,0)), \partial_s c(t,0)\rangle \le \frac{C}{2}\int_0^1 |\partial_s c(t,s)|^2\,\mathrm{d}s\,. \tag{B.3}$$

Since $\partial_s c(t,s) = 0$ at $t = 0$ and $t = T$, by the Poincaré inequality we also have

$$\int_0^T |\partial_s c(t,s)|^2\,\mathrm{d}t \le \frac{T^2}{\pi^2}\int_0^T |\partial_t|\partial_s c(t,s)||^2\,\mathrm{d}t \le \frac{T^2}{\pi^2}\int_0^T |\nabla_t \partial_s c(t,s)|^2\,\mathrm{d}t\,. \tag{B.4}$$

Let $\dot J(t,s) \coloneqq \nabla_s \partial_t c(t,s)$ and exchanging the order of derivatives in the equation above we obtain

$$\int_0^T |\partial_s c(t,s)|^2\,\mathrm{d}t \le \frac{T^2}{\pi^2}\int_0^T |\dot J(t,s)|^2\,\mathrm{d}t\,. \tag{B.5}$$

Integrating over $[0,T]$ equation (B.3) and using equation (B.5) we get

$$\int_0^T \Psi_p(t,c(t,1)) - \Psi_p(t,c(t,0)) - \langle \nabla\Psi_p(t,c(t,0)), \partial_s c(t,0)\rangle\,\mathrm{d}t \le \frac{CT^2}{2\pi^2}\int_0^1 |\dot J(t,s)|^2\,\mathrm{d}s\,. \tag{B.6}$$

Consider the term involving the gradient of Ψ_p. Substituting $\nabla\Psi_p(t,c(t,0)) = -2\nabla_t \dot z_t^* = -2\nabla_t J(t,0)$, integrating by parts in t, and exchanging the order of derivatives for this term yields

$$\int_0^T \Psi_p(t,c(t,1)) - \Psi_p(t,c(t,0)) - 2\langle J(t,0), \dot J(t,0)\rangle\,\mathrm{d}t \le \frac{CT^2}{2\pi^2}\int_0^1 |\dot J(t,s)|^2_{g_c}\,\mathrm{d}s\,. \tag{B.7}$$

Let $f(s) \coloneqq \int_0^T |J(t,s)|^2\,\mathrm{d}t$, then

$$f'(0) = \int_0^T 2\langle J(t,0), \dot J(t,0)\rangle\,\mathrm{d}t\,, \tag{B.8}$$

and

$$\begin{aligned} f(1) - f(0) - f'(0) &= \int_0^1 (1-s) f''(s)\,\mathrm{d}s \\ &= \int_0^1\int_0^T 2(1-s)(|\dot J(t,s)|^2 + \langle J(t,s), \nabla_s \dot J(t,s)\rangle)\,\mathrm{d}t\mathrm{d}s \\ &\ge \int_0^1\int_0^T 2(1-s)|\dot J(t,s)|^2\,\mathrm{d}t\mathrm{d}s\,, \end{aligned} \tag{B.9}$$

where the last inequality is due to the fact that for a Jacobi field $J(t,s)$,

$$\nabla_s \dot{J}(t,s) = -R(J(t,s),\partial_s c(t,s))\partial_s c(t,s)\,, \tag{B.10}$$

where R is the Riemann tensor, which for any tangent vectors X and Y at the same point on the cone over a flat manifold satisfies $\langle X, R(X,Y)Y\rangle \leq 0$. Moreover since the Jacobi fields are finite dimensional and $[0,T]\times M$ is compact, there exists a constant $C_0 > 0$ such that

$$f(1) - f(0) - f'(0) \geq \frac{C_0}{2}\int_0^1\int_0^T |\dot{J}(t,s)|^2\,\mathrm{d}t\mathrm{d}s\,. \tag{B.11}$$

Combining this with (B.7) and rearranging terms we obtain

$$\begin{aligned}\left(\frac{C_0}{2} - \frac{CT^2}{2\pi^2}\right)\int_0^1\int_0^T |\dot{J}(t,s)|^2\,\mathrm{d}t\mathrm{d}s + \int_0^T |J(t,0)|^2 - \Psi_p(t,c(t,0))\,\mathrm{d}t \\ \leq \int_0^T |J(t,1)|^2 - \Psi_p(t,c(t,1))\,\mathrm{d}t\,.\end{aligned} \tag{B.12}$$

Because of the inequality (6.3), shows that z^* is minimizing among all paths $\mathrm{z}\in\Omega$ which satisfy (B.1) and it is unique when the inequality is strict. Note that when $M = S^1_1$, the circle of unit radius, we can identify $\mathcal{C}$ with $\mathbb{R}^2$ and condition (B.1) is not necessary. Furthermore, since geodesics are straight lines with constant speed, from equation (B.9) we find $C_0 = 2$. This concludes the proof for the case $M = S^1_1$.

Now, assume that for all $x\in M$, $d_{\mathcal{C}}(z_{t_0}, z_{t_1}) \leq \epsilon$, for all $t_0, t_1 \in [0,T]$. Let

$$B_\delta \coloneqq \bigcap_{t\in[0,T]} \{q\in\mathcal{C}\,;\, d_{\mathcal{C}}(q, z^*_t) \leq \delta\}\,, \tag{B.13}$$

and take $\epsilon < \delta \coloneqq \frac{r_{min}}{2}$, where $r_{min} \coloneqq \min_{(t,x)\in[0,T]\times M}\lambda_t(x)$. For any $q\in B_\delta$ and any $t\in[0,T]$ the geodesic path between q and z^*_t cannot pass through the apex, since otherwise the distance between the two points should be at least equal to r_{min}. In other words, we must have $d_{\mathcal{C}}(q,z^*_t) < \pi$ and the path z^* is minimizing among all paths $\mathrm{z}\in\Omega$ contained in B_δ. Moreover, the geodesic path from z^*_0 to z^*_T is also included in B_δ. Consider the following quantity

$$E(\delta, q, T^*) \coloneqq \inf_{p\in\partial B_\delta/\mathcal{C}(\partial M)}\left\{\inf_{z\in AC^2([0,T^*];\mathcal{C})}\left\{\int_0^{T^*} |\dot{z}_t|^2 - \Psi_p(t,z_t)\,\mathrm{d}t\,;\, z_0 = q\in B_\delta\,,\, z_T = p\right\}\right\}\,, \tag{B.14}$$

which is the infimum action over the interval $[0,T^*]$ among paths starting at a point $q\in B_\delta$ and reaching its boundary ∂B_δ (but not points on ∂M) at time T^*. Given any path z such that $z_0 = z^*_0$ and $z_T = z^*_T$ not contained in B_δ, we have

$$\mathcal{B}(\mathrm{z}) \geq \inf_{T_1+T_2\leq T}(E(\delta, z^*_0, T_1) + E(\delta, z^*_T, T_2))\,, \tag{B.15}$$

and we want to show that $\mathcal{B}(\mathrm{z}) > \mathcal{B}(\mathrm{z}^*)$. We have

$$\begin{aligned}E(\delta, z^*_0, T_1) &\geq \inf_p\inf_{\mathrm{z}}\int_0^{T_1} |\dot{z}_t|^2\,\mathrm{d}t - (r_{max}+\delta)^2 CT_1 \\ &\geq \frac{(\delta-\epsilon)^2}{T_1} - (r_{max}+\delta)^2 CT_1\,,\end{aligned} \tag{B.16}$$

where $C \coloneqq \sup_{(t,x)\in[0,T]\times M}|P(t,x)|$ and $r_{max} \coloneqq \max_{(t,x)\in[0,T]\times M}\lambda_t(x)$. Hence, by equation (B.15),

$$\mathcal{B}(\mathrm{z}) \geq \frac{4(\delta-\epsilon)^2}{T} - (r_{max}+\delta)^2 CT\,. \tag{B.17}$$

On the other hand, we can deduce an upper bound for $\mathcal{B}(\mathrm{z}^*)$ using the geodesic path z^g between z^*_0 and z^*_T, yielding

$$\mathcal{B}(\mathrm{z}) \leq \int |\dot{z}^g_t|^2\,\mathrm{d}t + r^2_{max}CT \leq \frac{\epsilon^2}{T} + r^2_{max}CT\,. \tag{B.18}$$

Therefore we find the following sufficient condition for optimality of the path z^*:

$$[r^2_{max} + (r_{max}+\delta)^2]CT \leq \frac{4(\delta-\epsilon)^2}{T} - \frac{\epsilon^2}{T}\,. \tag{B.19}$$

The right-hand side is positive if $\epsilon < 2\delta/3$. Hence taking $\epsilon = \delta/2$ and substituting $\delta = \frac{r_{min}}{2}$,

$$\left[r_{max}^2 + \left(r_{max} + \frac{r_{min}}{2}\right)^2\right] CT \leq \frac{3r_{min}^2}{8T}\,. \tag{B.20}$$

This is the same as equation (6.5). For uniqueness we only need to substitute the inequality in (B.20) by a strict one, which concludes the proof. □

References

[1] Luigi Ambrosio and Alessio Figalli. On the regularity of the pressure field of Brenier's weak solutions to incompressible Euler equations. *Calculus of Variations and Partial Differential Equations*, 31(4):497–509, 2008.

[2] Luigi Ambrosio and Alessio Figalli. Geodesics in the space of measure-preserving maps and plans. *Archive for rational mechanics and analysis*, 194(2):421–462, 2009.

[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] Vladimir I Arnold. Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applicationsa l'hydrodynamique des fluides parfaits. *Ann. Inst. Fourier*, 16(1):319–361, 1966.

[5] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[6] Jean-David Benamou, Guillaume Carlier, and Luca Nenna. Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm. *arXiv preprint arXiv:1710.08234*, 2017.

[7] Yann Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

[8] Yann Brenier. The dual least action problem for an ideal, incompressible fluid. *Archive for rational mechanics and analysis*, 122(4):323–351, 1993.

[9] Dmitri Burago, Yuri Burago, and Sergei Ivanov. *A course in metric geometry*, volume 33. American Mathematical Soc., 2001.

[10] Roberto Camassa and Darryl D Holm. An integrable shallow water equation with peaked solitons. *Physical Review Letters*, 71(11):1661, 1993.

[11] 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.

[12] Lénaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Unbalanced optimal transport: Dynamic and Kantorovich formulations. *Journal of Functional Analysis*, 274(11):3090 – 3123, 2018.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

[14] Simone Di Marino, Andrea Natale, Rabah Tahraoui, and François-Xavier Vialard. Metric completion of Diff$([0,1])$ with the H^1 right-invariant metric. HAL preprint https://hal.archives-ouvertes.fr/hal-02161686/file/CHRelaxation.pdf, June 2019.

[15] David G Ebin and Jerrold Marsden. Groups of diffeomorphisms and the motion of an incompressible fluid. *Annals of Mathematics*, pages 102–163, 1970.

[16] Benno Fuchssteiner and Athanassios S Fokas. Symplectic structures, their bäcklund transformations and hereditary symmetries. *Physica D: Nonlinear Phenomena*, 4(1):47–66, 1981.

[17] Thomas Gallouët and François-Xavier Vialard. The Camassa–Holm equation as an incompressible Euler equation: A geometric point of view. *Journal of Differential Equations*, 2017.

[18] Thomas O Gallouët and Quentin Mérigot. A Lagrangian scheme à la Brenier for the incompressible Euler equations. *Foundations of Computational Mathematics*, pages 1–31, 2017.

[19] Darryl D Holm, Jerrold E Marsden, and Tudor S Ratiu. Euler-Poincaré models of ideal fluids with nonlinear dispersion. *Physical Review Letters*, 80(19):4173, 1998.

[20] Darryl D Holm, Jerrold E Marsden, and Tudor S Ratiu. The Euler–Poincaré equations and semidirect products with applications to continuum theories. *Advances in Mathematics*, 137(1):1–81, 1998.

[21] Darryl D Holm and Martin F Staley. Wave structure and nonlinear balances in a family of evolutionary PDEs. *SIAM Journal on Applied Dynamical Systems*, 2(3):323–380, 2003.

[22] John K Hunter and Ralph Saxton. Dynamics of director fields. *SIAM Journal on Applied Mathematics*, 51(6):1498–1521, 1991.

[23] Robert L Jerrard and Cy Maor. Vanishing geodesic distance for right-invariant sobolev metrics on diffeomorphism groups. *arXiv preprint arXiv:1805.01410*, 2018.

[24] B Khesin, Jonatan Lenells, G Misiołek, and SC Preston. Geometry of diffeomorphism groups, complete integrability and geometric statistics. *Geometric and Functional Analysis*, 23(1):334–366, 2013.

[25] H-P Kruse, J Scheurle, and W Du. A two-dimensional version of the Camassa-Holm equation. In *Symmetry and Perturbation Theory: SPT 2001*, pages 120–127. World Scientific, 2001.

[26] Jae Min Lee. Global Lagrangian solutions of the Camassa-Holm equation. *arXiv preprint arXiv:1710.05484*, 2017.

[27] Jonatan Lenells. The Hunter–Saxton equation describes the geodesic flow on a sphere. *Journal of Geometry and Physics*, 57(10):2049–2064, 2007.

[28] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae*, 211(3):969–1117, 2018.
[29] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.
[30] Quentin Mérigot and Jean-Marie Mirebeau. Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport. *SIAM Journal on Numerical Analysis*, 54(6):3465–3492, 2016.
[31] Peter W Michor and David Mumford. Vanishing geodesic distance on spaces of submanifolds and diffeomorphisms. *Doc. Math*, 10:217–245, 2005.
[32] Luc Molinet. On well-posedness results for Camassa-Holm equation on the line: a survey. *Journal of Nonlinear Mathematical Physics*, 11(4):521–533, 2004.
[33] Luca Nenna. *Numerical methods for multi-marginal optimal transportation*. PhD thesis, PSL Research University, 2016.
[34] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pages 99–102, 2015.
[35] Alexander I Shnirelman. Generalized fluid flows, their approximation and applications. *Geometric & Functional Analysis GAFA*, 4(5):586–620, 1994.

RÉSUMÉ

Ce document traite du transport optimal et de son application aux équations aux dérivées partielles telles que des flows de gradient ou d'Euler dans les espaces de Wasserstein. Nous étudions des questions théoriques et numériques. Du côté théorique du transport optimal, nous abordons des questions telles que la construction de splines dans l'espace de Wasserstein, l'extrapolation de géodésiques pour la métrique Wasserstein et certaines questions liées à la régularité du transport optimal non équilibré (projection polaire de Brenier non équilibrée, équations de Monge-Ampère non équilibrées, une classe spéciale de fonctions Cône-convexes). Nous utilisons ensuite la structure de flot de gradient/Euler dans l'espace de Wasserstein pour l'étude de certaines EDP.

D'une part, cette structure spéciale est utilisée pour prouver des résultats théoriques, par exemple pour monter l'existence de solutions au système d'écoulements multiphasiques immiscibles et incompressible en milieu poreux, ou encore introduire la notion de solution généralisées pour les équations de Camassa-Holm, qui s'avère être la contrepartie pour le Transport Optimal Déséquilibré de ce que l'Euler Incompressible est au Transport Optimal classique. D'autre part, la structure géométrique est également utilisée pour concevoir, mettre en œuvre et prouver la convergence de différents schémas numériques. Par exemple, nous introduisons la notion de schémas variationnels volumes finis pour les flots de gradient Wasserstein. Ces schémas sont des schémas de volumes finis définis comme les équations d'Euler-Lagrange pour une discrétisation en espace d'un schéma de mouvement minimisant (JKO), dans les l'esprit des approches "d'abord discrétiser puis optimiser". Nous avons également défini des schémas numériques lagrangiens pour une classe de flots de Gradient/Euler. Ces schémas numériques sont des EDO préservant la structure géométrique sous-jacente avec une énergie approchée définie à l'aide d'un problème de transport optimal semi-discret. Par ailleurs à l'aide d'une méthode d'optimisation alterné et à l'utilisation du transport optimal non équilibré, nous montrons que tous les efforts entrepris pour approcher les flots de gradient Wasserstein peuvent être étendus pour englober des équations de diffusion-réaction plus générales ne préservant pas la masse.

ABSTRACT

This document is about Optimal Transport and its application to partial differential equations such as gradient flows or Euler flows in the Wasserstein spaces. We investigate theoretical as well as numerical questions. On the theoretical side of optimal transport, we address questions such as Wasserstein splines, Wasserstein extrapolation and some questions related to the smoothness of Unbalanced Opti- mal Transport (Unbalanced Brenier polar projection, Unbalanced Monge-Ampère equations, a special class of Cone convex functions). We then apply the Wasserstein Gradient/Euler flow structure to the study of some PDEs.

On the one hand, the flow structure is used to prove theoretical results, notably the existence of solutions to the system of incompressible immiscible multiphase flows in porous media, and the definition of the notion of relaxed solution for the Camassa-Holm equations, which happens to be the counter part for the Unbalanced Optimal Transport of what Incompressible Euler is for the classical Optimal Trans- port. One the other hand, the geometrical structure is also used to design, implement and prove convergence for different numerical schemes. For instance we introduce the notion of variational Finite Volume schemes for Wasserstein Gradient flows. These schemes are finite volume schemes defined as the Euler-Lagrange equations for a space discretization of a minimizing movement (JKO) scheme, a "first discretize then optimize" approach. We also defined Lagrangian numerical schemes for a class of Gradient and Euler flows. These schemes are ODEs preserving the underlying geometrical structure with an approximated energy defined through semi discrete Optimal Transport. Through a splitting procedure and using Unbalanced Optimal Transport, all the effort undertaken for Wasserstein Gradient Flows can by extended to encompass more general and non conservative reaction diffusion equations.