UCLA
UCLA Electronic Theses and Dissertations

Title
Algorithms for optimal transport and their applications to PDEs

Permalink
https://escholarship.org/uc/item/59j4x848

Author
Lee, Wonjun

Publication Date
2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
Los Angeles

Algorithms for optimal transport and their applications to PDEs

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Wonjun Lee

2022

© Copyright by
Wonjun Lee
2022

ABSTRACT OF THE DISSERTATION

Algorithms for optimal transport and their applications to PDEs

by

Wonjun Lee
Doctor of Philosophy in Mathematics
University of California, Los Angeles, 2022
Professor Stanley J. Osher, Chair

Optimal transport theory provides a distance between two probability distributions. It finds the cheapest transport map that moves one measure to the other measure with respect to some ground cost. With its deep theoretical properties, the optimal transport distance has been used in diverse areas such as partial differential equations (PDEs), economics, image processing, and machine learning. However, computing the optimal transport distances and maps is difficult, which has been a significant challenge in applications. In this dissertation, we present new numerical methods using optimal transport distance and their applications in solving challenging convex and nonconvex optimization problems involving non-linear PDEs. We demonstrate the suggested methods' efficiency through numerous numerical results.

The dissertation of Wonjun Lee is approved.

Lieven Vandenberghe

Wotao Yin

Wilfrid Dossou Gangbo

Stanley J. Osher, Committee Chair

University of California, Los Angeles

2022

To Yuri and Riwon

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I wish to express my sincerest appreciation and gratitude to my advisor, Professor Stanley Osher, for taking me as his mentee and inspiring me to be a good mathematician. I found myself to be incredibly fortunate to have done research under his guidance. His support and guidance made my Ph.D. experience satisfying and exceptionally productive. I am also deeply grateful to my friend and a mentor, Matt Jacobs, who devoted countless mentorship hours to me and taught me the value of passion and what it takes to be a good researcher. I am forever indebted to him for his patience, encouragement, and support.

I would like to thank Flavien Leger, Wilfrid Gangbo, and Inwon Kim, who have always been supportive and generously spent their time giving me great advice on how to be a good researcher. I would like to extend my sincere thanks to Stanley Osher, Wilfrid Gangbo, Wotao Yin, and Lieven Vandenberghe for serving on my dissertation committee. Furthermore, many thanks to my most frequent collaborator, Wuchen Li, and other collaborators and my friends, Siting Liu, Levon Nurbekyan, Samy Wu Fung, Kyung Ha, Bohyun Kim, Bumsu Kim, Dohyun Kwon, Younghak Kwon, Jaehoon Lee, Sangchul Lee, and Sangjin Lee, to name only a few among many.

I am grateful to my brother, Thomas, my sister, Seungmin, my mother, Soonyoung, and my father, Wan-ho, without whose love and support I could not have done this dissertation. Last and most importantly, I thank my loving wife, Yuri, who has believed in me since we first met, encouraged me to pursue my dream, supported me to do research productively, and gave birth to our lovely daughter, Riwon. Her love, devotion and support made this journey successful. Without her, none of this would have been possible.

VITA

2015 B.S. in Mathematics, George Mason University.

2017–2021 Teaching Assistant, Mathematics Department, UCLA.

2019–Present Graduate Student Researcher, Mathematics Department, UCLA.

PUBLICATIONS

S. Agrawal, **W. Lee**, S. W. Fung, L. Nurbekyan. "Random Features for High-Dimensional Nonlocal Mean-Field Games." *Journal of Computational Physics*, 2022

A. Vepa, A. Choi, N. Nakhaei, **W. Lee**, et al. "Weakly-Supervised Convolutional Neural Networks for Vessel Segmentation in Cerebral Angiography." *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022

H. Gao, **W. Lee**, Y. Kang, W. Li, Z. Han, S. Osher, H. V. Poor, "Energy-Efficient Velocity Control for Massive Numbers of UAVs: A Mean Field Game Approach." *IEEE Transactions on Vehicular Technology*, 2022.

W. Li, **W. Lee**, S. Osher, "Computation Mean-Field Information Dynamics Associated With Reaction Diffusion Equations." *arXiv preprint arXiv:2107.11501*, 2021

W. Lee, S. Liu, W. Li, S. Osher, "Mean Field Control Problems For Vaccine Distribution." *arXiv preprint arXiv:2104.11887*, 2021.

W. Lee, W. Li, B. Lin, A. Monod, "Tropical Optimal Transport and Wasserstein Distances in Phylogenetic Tree Space." *Information Geometry*, 2021

M. Jacobs, **W. Lee**, F. Léger, "The back-and-forth method for Wasserstein gradient flows." *ESAIM: COCV*, 27:28, 2021.

W. Lee, S. Liu, H. Tembine, W. Li, S. Osher, "Controlling propagation of epidemics via mean-field games." *SIAM Journal on Applied Math*, 2020.

W. Lee, R.J. Lai, W. Li, S. Osher, "Generalized unnormalized optimal transport and its fast algorithms." *Journal of Computational Physics*, 2020.

H. Gao, **W. Lee**, W. Li, Z. Han, S. Osher, H. V. Poor, "Energy-efficient Velocity Control for Massive Rotary-Wing UAVs: A Mean Field Game Approach." *IEEE Globecom*, 2020.

Y. Kang, S. Liu, **W. Lee**, W. Li, H. Zhang, and Z. Han, "EJoint Task Assignment and Trajectory Optimization for a Mobile Robot Swarm by Mean-Field Game." *IEEE Globecom*, 2020.

CHAPTER 1

Introduction

1.1 Background

Optimal transport theory provides a distance between two probability measures μ and ν by computing the cheapest way to transport μ to ν with respect to a cost function c. Consider a domain $\Omega \subset \mathbb{R}^d$ is an open bounded set in d-dimensional Euclidean space (the domain can be more general but we will focus on a simple case). The problem can be written mathematically by

$$\inf_T \left\{ \int_\Omega c(T(x), x)\, d\mu(x) : T_\# \mu = \nu \right\} \tag{1.1}$$

where $T : \Omega \to \Omega$ is a measure preserving transport map and $c : \Omega \times \Omega \to \mathbb{R}$ is a cost function that measures the cost of transporting a mass from $x \in \Omega$ to $y \in \Omega$. In the constraint set, the pushforward measure $T_\# \mu$ is defined through

$$(T_\# \mu)(B) = \mu(T^{-1}(B))$$

for all measurable set $B \subset \Omega$. The minimization problem solves for the optimal map T that minimizes the cost to transport μ to ν. This problem is also known as the *Monge problem* [Mon81]. When c is a quadratic function ($c(x, y) = |x - y|^2$), the distance is often referred to as the *2-Wasserstein distance*. With its deep theoretical properties, the optimal transport has been studied extensively, both analytically and numerically. It has found use in diverse areas such as partial differential equations (PDEs) [BB00, GM18, CCY19, JKT20a], fluid dynamics [BB00, DWH15, GM18], economics [Gal16], image processing [ZYH07, WOS10, WSB13], and machine learning [PW08, ACB17b, TPK17a]. Nonetheless, computing optimal

transport distances and maps numerically have been a major challenge. This thesis focuses on fast and accurate algorithms for optimal transport and the implementations of these algorithms to solve PDEs.

Chapter 2, a collaboration with Prof. Rongjie Lai, Prof. Wuchen Li, and Prof. Stanley Osher, is adapted from [LLL21a]. We proposed a new model for unnormalized optimal transport. The classical optimal transport considers the distance between two densities with the same mass. This model can compute the distance between two measures with different masses, and, thus, generalizes the optimal transport distance. We also presented new numerical methods for solving optimal transport distance using a primal-dual algorithm and an accelerated gradient ascent algorithm.

Chapter 3, a collaboration with Siting Liu, Prof. Hamidou Tembine, Prof. Wuchen Li, and Prof. Stanley Osher, is adapted from [LLT21] with the algorithm section (Section 3.3) adapted from [LLL21b]. In response to the COVID-19 pandemic, we proposed a new mean-field game model of spatial epidemiological dynamics. We added spatial dynamics to the classical epidemiological models, such as SIR (Susceptible, Infectious, or Recovered) model. The new model formulates the nonconvex optimization problem and we provided an efficient algorithm to solve the problem.

Chapter 4, a collaboration with Prof. Matt Jacobs and Prof. Flavien Léger, is adapted from [JLL21]. In this project, we presented a new method to efficiently compute Wasserstein gradient flows. Our approach is based on a generalization of the back-and-forth method (BFM) introduced in [JL20] to solve optimal transport problems. The proposed method is a state-of-the-art algorithm to solve large-scale gradient flows simulations for a large class of internal energies including singular and non-convex energies.

In what follows, we provide background information on alternative formulations of optimal transport which will be used throughout the thesis.

1.2 Optimal transport

In this section, we briefly review two alternate formulations of the Monge problem (1.1): the Kantorovich dual formulation [Kan06] and the Benamou-Brenier formulation [BB00]. Chapter 2 and Chapter 3 use the Benamou-Brenier formulation to formulate new models in unnormalized optimal transport and epidemiology. Chapter 4 uses the dual formulation to derive the dual problem of the *minimizing movement scheme* [JKO98] (often also called the JKO scheme) and presents the efficient algorithm to compute it.

The Kantorovich formulation is a relaxed version of the Monge problem. Suppose μ and ν are discrete probability measures such that $\mu = \delta_x$ and $\nu = \frac{1}{2}\delta_y + \frac{1}{2}\delta_z$ where δ_x is a Dirac measure with a mass at x. Then (1.1) does not have a solution since maps cannot split the mass. One can resolve this issue by considering a joint probability measure instead of a map. Consider a set of joint probability measures $\Pi(\mu, \nu) \subset \mathcal{P}(\Omega \times \Omega)$ such that the first marginal is μ and the second marginal is ν. In other words, if $\gamma \in \Pi(\mu, \nu)$ then

$$\gamma(B \times \Omega) = \mu(B), \quad \gamma(\Omega \times B) = \nu(B) \tag{1.2}$$

for all measurable sets $B \subset \Omega$. The relaxed minimization problem takes the form of

$$\inf_{\gamma} \left\{ \int_{\Omega \times \Omega} c(x, y)\, d\gamma(x, y) : \gamma \in \Pi(\mu, \nu) \right\} \tag{1.3}$$

where $d\gamma(x, y)$ is the amount of mass transported from x to y, which allows the mass to be transported to multiple locations from x. The set $\Pi(\mu, \nu)$ is called a set of *transport plans* between μ and ν and the minimizer γ is called the *optimal transport plan*.

The Kantorovich formulation (1.3) has a dual formulation which is important in Chapter 4. First, we convert the minimization problem (1.3) into a saddle point problem by introducing Lagrangian multipliers ϕ and ψ.

$$\inf_{\gamma} \sup_{\phi, \psi} \int_{\Omega \times \Omega} c(x, y)\, d\gamma(x, y) + \int_{\Omega} \psi(x)\, d\mu(x) - \int_{\Omega \times \Omega} \psi(x)\, d\gamma(x, y) \\ - \int_{\Omega} \phi(y)\, d\nu(y) + \int_{\Omega \times \Omega} \phi(y)\, d\gamma(x, y).$$

Note that the above equation satisfies the constraint set (1.2) and is equivalent to (1.3). When γ is not in the constraint set, by the sup over ϕ and ψ, the value can be $+\infty$ by having arbitrarily large values of ϕ and ψ. By rearranging the terms, we have

$$\begin{aligned}
&= \inf_{\gamma} \sup_{\phi,\psi} \int_{\Omega\times\Omega} (c(x,y) + \phi(y) - \psi(x))\, d\gamma(x,y) + \int_{\Omega} \psi(x)\, d\mu(x) - \int_{\Omega} \phi(y)\, d\nu(y) \\
&= \sup_{\phi,\psi} \int_{\Omega} \psi(x)\, d\mu(x) - \int_{\Omega} \phi(y)\, d\nu(y) + \inf_{\gamma} \int_{\Omega\times\Omega} \Big(c(x,y) + \phi(y) - \psi(x)\Big)\, d\gamma(x,y)
\end{aligned}$$

where the second equality interchanges sup and inf. The interchange is valid under the assumption that the cost function c is lower semi continuous [San15]. The equation can be rewritten as

$$= \sup_{\phi,\psi} \left\{ \int_{\Omega} \psi(x)\, d\mu(x) - \int_{\Omega} \phi(y)\, d\nu(y) :\ \psi(x) - \phi(y) \le c(x,y) \text{ for all } x, y \in \Omega \right\}. \tag{1.4}$$

This maximization problem is the dual formulation of the optimal transport and called the *Kantorovich dual formulation*.

The other alternative formulation, the Benamou-Brenier formulation, is a dynamical formulation of the optimal transport problem. Given probability densities $\mu \in \mathcal{P}(\Omega)$ and $\nu \in \mathcal{P}(\Omega)$ and a cost function $c(x,y) = |x-y|^p$ $(p \ge 1)$, one can calculate the optimal transport distance by solving the following minimization problem:

$$\begin{aligned}
&\inf_{\rho,v} \int_0^1 \int_{\Omega} \rho(t,x)|v(t,x)|^p\, dx\, dt \\
\text{subject to} \quad &\partial_t \rho(t,x) + \nabla \cdot (\rho(t,x)v(t,x)) = 0, \\
&\rho(0,\cdot) = \mu,\ \rho(1,\cdot) = \nu.
\end{aligned} \tag{1.5}$$

The infimum is taken over continuous density function $\rho : [0,1] \times \Omega \to \mathbb{R}_+$ and velocity fields $v : [0,1] \times \Omega \to \mathbb{R}^d$ with zero flux condition on $\partial\Omega$. The formulation solves for a nonnegative density $\rho(t,x)$ that flows from μ at $t = 0$ to ν at $t = 1$ through the continuity equation $\partial_t \rho + \nabla \cdot (\rho v) = 0$ while minimizing the cost of the flow defined as $\rho|v|^p$. The formulation computes the same optimal transport distance as the Monge problem (1.1). The optimal

velocity fields $v(t,x)$ satisfies the following ODE:

$$\begin{cases} y_x(t)' = v(t, y_x(t)) \\ y_x(0) = x. \end{cases}$$

The solution of the ODE defines the map through $T_t(x) = y_x(t)$ and the map satisfies $\rho(t,\cdot) = (T_t)_{\#}\rho(0,\cdot)$ [San15]. Thus, we have $(T_1)_{\#}\mu = \nu$ and the map corresponds to the optimal transport map T in the Monge problem (1.1). By modifying the objective function and constraints, the formulation can describe various optimal control problems and certain cases of the mean-field games model. For example, in Chapter 2, we consider the following formulation

$$\begin{aligned} &\inf_{\rho,v} \int_0^1 \int_\Omega \rho(t,x)|v(t,x)|^p \, dx \, dt + \frac{1}{\alpha} \int_0^1 \int_\Omega |f(t,x)|^p \, dx \, dt : \\ \text{subject to} \quad &\partial_t \rho(t,x) + \nabla \cdot (\rho(t,x)v(t,x)) = f(t,x), \\ &\rho(0,\cdot) = \mu, \ \rho(1,\cdot) = \nu. \end{aligned}$$

The objective function and a constraint are modified by adding new terms involving a function $f : [0,1] \times \Omega \to \mathbb{R}$. This modified formulation computes the optimal transport distance between two densities with different masses. Chapter 3 also utilizes the Benamou-Brenier formulation to propose the new epidemiological model.

CHAPTER 2

Generalized Unnormalized Optimal Transport And Its Fast Algorithms

We introduce fast algorithms for generalized unnormalized optimal transport. To handle densities with different total mass, we consider a dynamic model, which mixes the L^p optimal transport with L^p distance. For $p = 1$, we derive the corresponding L^1 generalized unnormalized Kantorovich formula. We further show that the problem becomes a simple L^1 minimization which is solved efficiently by a primal-dual algorithm. For $p = 2$, we derive the L^2 generalized unnormalized Kantorovich formula, a new unnormalized Monge problem and the corresponding Monge-Ampère equation. Furthermore, we introduce a new unconstrained optimization formulation of the problem. The associated gradient flow is essentially related to an elliptic equation which can be solved efficiently. Here the proposed gradient descent procedure together with the Nesterov acceleration involves the Hamilton-Jacobi equation ariseing from the KKT conditions. Several numerical examples are presented to illustrate the effectiveness of the proposed algorithms.

2.1 Introduction

Optimal transport describes transport plans and metrics between two densities with equal total mass [Vil09]. It has wide applications in various fields such as physics [LL19, LYO18], mean field games [CLO18], image processing [PC18], economics [BT09], inverse problem [EY18, YES18], Kalman filter [GHL19] as well as machine learning [ACB17a, LLO18]. In

practice, it is also natural to consider transport and metrics between two densities with different total mass. For example, in image processing, it is very common that we need to compare and process images with unequal total intensities [RLY17].

Recently, there has been increasing interests in studying the optimal transport between two densities with different total mass. Based on the linear programming formulation, generalized versions for unnormalized optimal transport have been considered in [PR16, TPK17b]. In this paper, our discussion is based on the fluid-dynamic formulation following [BB00], which has significantly fewer variables than the linear programming formulation. We consider a source function to provide dynamical behaviors of a source term during transportation. Adding a source term for handling densities with unequal total mass has been considered in [CL18, CGT19, CPS15, CPS18, LMS18, MRS15, PR14]. These methods consider density-dependent source terms and lead to a dynamical mixture of Wasserstein-2 distance and Fisher-Rao distance. The corresponding minimization of the source term is weighted with the density. More recently, a spatially independent source function was considered in [GLO19] to transport densities with unequal mass. This model results in creating or removing masses in the space uniformly during transportation when moving one density to another. Here, we further extend the model [GLO19] using a spatially dependent source function. As a result, the transportation map between two densities with different masses has the flexibility to create or remove masses locally. In all our models, the source term does not depend on the current density. This property keeps the Hamilton-Jacobi equation arising in the original (normalized) optimal transport problem. We further explore the Kantorovich duality and derive the corresponding unnormalized Monge problems and Monge-Ampère equations. Besides these model derivations, the other main contribution of this paper is to propose fast algorithms for all related dynamical optimal transport problems with source terms.

More specifically, the proposed model is a minimal flux problem mixing both L^p metric

and Wasserstein-p metric, following Benamou-Brenier formula [BB00]:

$$\inf_{v,\mu,f} \int_0^1 \int_\Omega \|v(t,x)\|^p \mu(t,x)\, dx\, dt + \frac{1}{\alpha} \int_0^1 \int_\Omega |f(t,x)|^p\, dx\, dt,$$

such that

$$\partial_t \mu(t,x) + \nabla \cdot (\mu(t,x) v(t,x)) = f(t,x), \quad \mu(0,x) = \mu_0(x), \quad \mu(1,x) = \mu_1(x).$$

The minimization problem solves for the optimal map between two nonnegative densities μ_0 and μ_1, given a source function f (see the details in definition 2.2.1 from section 2.2). The optimal map shows how the masses are added or removed by the source function during the transportation. In this paper, in particular, we focus on the cases $p = 1$ and $p = 2$, and design corresponding fast algorithms. For the L^1 case, we propose a primal-dual algorithm [CP11a]. The method updates variables at each iteration with explicit formulas, which only involve low computational cost shrink operators, such as those used in [LRO18]. For the L^2 case, we formulate the minimal flux problem into a novel unconstrained minimization problem as follows

$$\begin{aligned} \inf_{\mu} \Bigg\{ \int_0^1 \int_\Omega \partial_t \mu(t,x) (-\nabla \cdot (\mu(t,x)\nabla) + \alpha \mathrm{Id})^{-1} \partial_t \mu(t,x) dx dt : \\ \mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x), x \in \Omega \Bigg\}, \end{aligned} \tag{2.1}$$

where α is a given positive scalar, Id is the identity operator, and the infimum is taken among all density paths $\mu(t,x)$ with fixed terminal densities μ_0, μ_1. From the associated Euler-Lagrange equation, we derive a Nesterov accelerated gradient descent method to solve the unnormalized optimal transport problem. It turns out that our method only needs to solve an elliptic equation involving the density at each iteration. Thus, fast solvers for elliptic equations can be directly used. Interestingly, the Euler-Lagrange equation of this formulation introduces the Hamilton-Jacobi equation, which characterizes the Lagrange multiplier (see related studies in [Li18]). We, in fact, construct the gradient descent method in the density path space to solve this equation:

$$\partial_\tau \mu(\tau,t,x) = \partial_t \Phi(\tau,t,x) + \frac{1}{2} \|\nabla \Phi(\tau,t,x)\|^2,$$

with

$$\Phi(\tau,t,x)=(-\nabla\cdot(\mu(\tau,t,x)\nabla)+\alpha Id)^{-1}\partial_t\mu(\tau,t,x).$$

Here τ is an artificial time variable in optimization. The minimizer path $\mu^*(t,x)$ is obtained by solving $\mu^*(t,x)=\lim_{\tau\to\infty}\mu(\tau,t,x)$ numerically.

The outline of this paper is as follows. In section 2.2, we propose a formulation for the generalized unnormalized optimal transport. We then derive the Kantorovich duality for both cases. We also formulate the generalized unnormalized Monge problem and the corresponding Monge-Ampère equation. In section 2.3, we propose a fast algorithm for L^1-generalized unnormalized optimal transport using a primal-dual based method. We also propose a new method for L^2-generalized unnormalized optimal transport based on the Nesterov accelerated gradient descent method. In addition, we discuss detailed numerical discretization of the two problems. In section 2.4, we present several numerical experiments to demonstrate the effectiveness of our algorithms. We conclude the paper in section 2.5.

2.2 Generalized unnormalized optimal transport

In this section, we study a formulation of generalized unnormalized optimal transport problem as a natural extension of the exploration studied in [GLO19]. We specifically discuss the L^1 and L^2 versions of the generalized unnormalized optimal transport and their associated Kantorovich dualities. Furthermore, we derive a new generalized unnormalized Monge problem and the corresponding Monge-Ampère equation.

Let $\Omega\subset\mathbb{R}^d$ be a compact convex domain. Denote the space of unnormalized densities $\mathcal{M}(\Omega)$ by

$$\mathcal{M}(\Omega):=\{\mu\in L^1(\Omega):\mu(x)\geq 0\}.$$

Given two densities $\mu_0,\mu_1\in\mathcal{M}(\Omega)$, we define the generalized unnormalized optimal transport as follows:

Definition 2.2.1 (Generalized Unnormalized Optimal Transport)**.** Define the L^p generalized unnormalized Wasserstein distance $UW_p : \mathcal{M}(\Omega) \times \mathcal{M}(\Omega) \to \mathbb{R}$ by

$$UW_p(\mu_0, \mu_1)^p = \inf_{v,\mu,f} \int_0^1 \int_\Omega \|v(t,x)\|^p \mu(t,x)\, dx\, dt + \frac{1}{\alpha} \int_0^1 \int_\Omega |f(t,x)|^p\, dx\, dt,$$

such that the dynamical constraint, i.e. the unnormalized continuity equation, holds

$$\partial_t \mu(t,x) + \nabla \cdot (\mu(t,x) v(t,x)) = f(t,x), \quad \mu(0,x) = \mu_0(x), \quad \mu(1,x) = \mu_1(x).$$

The infimum is taken over continuous unnormalized density functions $\mu : [0,1] \times \Omega \to \mathbb{R}$, and Borel vector fields $v : [0,1] \times \Omega \to \mathbb{R}^d$ with zero flux condition on $[0,1] \times \partial\Omega$, and Borel spatially dependent source functions $f : [0,1] \times \Omega \to \mathbb{R}$. A positive constant $\alpha \in (0, \infty)$ is a fixed parameter.

This is a generalized definition of unnormalized optimal transport from [GLO19]. Here, we consider a spatially dependent source function $f(t,x)$. In this paper, we will focus on the cases with $p = 1$ and $p = 2$.

Remark 2.2.1. We note that [CGT19] has proposed the model for $p = 2$ without any discussion about numerical methods. In this paper, we mainly study Kantorovich duality and design fast algorithms.

Remark 2.2.2. In literature, [CPS15] studied the other dynamical formulations of unbalanced optimal transport problems. In their approach, the optimal source term is expressed as a product of a density function and a scalar field function. In our approach, the optimal source term only depends on a scalar field function. This fact shows that our approach is different from [CPS15] in variational problems and dual (Kantorovich) problems.

2.2.1 L^1 Generalized Unnormalized Wasserstein metric.

When $p = 1$, the problem (2.2.1) becomes

$$UW_1(\mu_0, \mu_1) = \inf_{v,\mu,f} \Bigg\{ \int_0^1 \int_\Omega \|v(t,x)\| \mu(t,x)\, dx\, dt + \frac{1}{\alpha} \int_0^1 \int_\Omega |f(t,x)|\, dx\, dt : \\ \partial_t \mu(t,x) + \nabla \cdot (\mu(t,x) v(t,x)) = f(t,x) \\ \mu(0,x) = \mu_0(x), \quad \mu(1,x) = \mu_1(x) \Bigg\}. \tag{2.2}$$

Here $\|\cdot\|$ can be any homogeneous of degree one norm, i.e. l_q norm $\|u\|_q = (\sum_{i=1}^d |u_i|^q)^{\frac{1}{q}}$. In particular, we consider $q = 1, 2$ with

$$\|u\|_1 = |u_1| + \cdots + |u_d| \qquad \text{for } u \in \mathbb{R}^d,$$

or

$$\|u\|_2 = \sqrt{|u_1|^2 + \cdots + |u_d|^2} \qquad \text{for } u \in \mathbb{R}^d.$$

Proposition 2.2.1. *The L^1 unnormalized Wasserstein metric is given by*

$$UW_1(\mu_0, \mu_1) = \inf_{\mathbf{m},c} \Bigg\{ \int_\Omega \|\mathbf{m}(x)\| dx + \frac{1}{\alpha} \int_\Omega |c(x)| dx \ : \\ \mu_1(x) - \mu_0(x) + \nabla \cdot \mathbf{m}(x) - c(x) = 0 \Bigg\}. \tag{2.3}$$

There exists $\Phi(x)$, such that the minimizer (m, c) for the problem (2.3) satisfies

$$\nabla \Phi(x) \in \partial \|\mathbf{m}(x)\| \quad \textit{and} \quad \alpha \Phi(x) \in \partial |c(x)|$$

where $\partial\|\mathbf{m}(x)\|$ and $\partial|c(x)|$ denote their sub-differentials.

Proof. Denote

$$\mathbf{m}(x) = \int_0^1 v(t,x) \mu(t,x) dt,$$

Using Jensen's inequality and integration by parts, we can reformulate (2.2).

$$\begin{aligned}&\int_0^1 \int_\Omega \|v(t,x)\|\mu(t,x)dxdt + \frac{1}{\alpha}\int_0^1 \int_\Omega |f(t,x)|dxdt \\ &\geq \int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\int_\Omega \left|\int_0^1 f(t,x)dt\right| dx.\end{aligned} \tag{2.4}$$

Define $c(x) = \int_0^1 f(t,x)dt$. Integrating on the constraint of problem (2.2) with the zero flux condition of v yields,

$$\int_\Omega c(x)dx = \int_0^1 \int_\Omega f(t,x)dxdt = \int_\Omega \mu_1(x)dx - \int_\Omega \mu_0(x)dx.$$

Plug $c(x)$ into the equation (2.4), we obtain a new formulation.

$$\inf_{\mathbf{m},c}\left\{\int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\int_\Omega \|c(x)\|dx \ : \mu_1(x) - \mu_0(x) + \nabla\cdot\mathbf{m}(x) - c(x) = 0\right\}.$$

Note that the minimization path can be attained in the inequality (2.4) by choosing $\mu(t,x) = t\mu_0(x) + (1-t)\mu_1(x)$, $\mathbf{m}(x) = \mu(t,x)v(t,x)$ and $f(t,x) = c(x)$. Then $\{\mu(t,x), v(t,x), f(t,x)\}$ is a feasible solution to (2.2) and (2.3) , hence the two minimization problems have the same optimal value.

Consider the Lagrangian of this minimization problem.

$$\mathcal{L}(\mathbf{m}, c, \Phi) = \int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\int_\Omega |c(x)|dx + \int_\Omega \Phi(x)\Big(\mu_1(x) - \mu_0(x) + \nabla\cdot\mathbf{m}(x) - c(x)\Big), \tag{2.5}$$

where $\Phi(x)$ is a Lagrange multiplier. From the Karush–Kuhn–Tucker (KKT) conditions, we derive the following properties of the minimizer

$$\begin{aligned}&0 \in \partial_{\mathbf{m}}\mathcal{L} \Rightarrow \nabla\Phi(x) \in \partial\|\mathbf{m}(x)\| \\ &0 \in \partial_c\mathcal{L} \Rightarrow \alpha\Phi(x) \in \partial|c(x)| \\ &\delta_\Phi\mathcal{L} = 0 \Rightarrow \mu_1(x) - \mu_0(x) + \nabla\cdot\mathbf{m}(x) - c(x) = 0.\end{aligned}$$

□

Remark 2.2.3. In the case that L^1 unnormalized Wasserstein metric with a spatially independent function $f(t)$, c is defined to be $c = \int_0^1 f(t)dt$, which is a constant. Integrating on a spatial domain for continuity equation,

$$c = \frac{1}{|\Omega|}\bigg(\int_\Omega \mu_0(x)dx - \int_\Omega \mu_0(x)dx\bigg).$$

As a result, the minimization problem becomes

$$UW_1(\mu_0,\mu_1) = \inf_{\mathbf{m}}\bigg\{ \int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\bigg|\int_\Omega \mu_1(x)dx - \int_\Omega \mu_0(x)dx\bigg| : \\ \mu_1(x) - \mu_0(x) + \nabla\cdot\mathbf{m}(x) = \frac{1}{|\Omega|}\bigg(\int_\Omega \mu_1(x)dx - \int_\Omega \mu_0(x)dx\bigg)\bigg\}.$$

This is compatible with the result obtained in [GLO19]. In this case, we note that $\mathbf{m}(x)$ does not depend on α.

Proposition 2.2.2 (L^1 Generalized Unnormalized Kantorovich formulation)**.** *The Kantorovich formulation of L^1 unnormalized Wasserstein metric is the following:*

$$UW_1(\mu_0,\mu_1) = \sup_{\Phi}\bigg\{\int_\Omega \Phi(x)(\mu_1(x)-\mu_0(x))dx : \|\nabla\Phi\| \le 1, |\Phi| \le \frac{1}{\alpha}\bigg\} \tag{2.6}$$

Remark 2.2.4. The Kantorovich formulation of the generalized unnormalized Wasserstein-1 metric has also been stated in [CGN17] for the $\|\cdot\|_2$ norm.

Proof. From the Lagrangian (2.5),

$$\begin{aligned}
&\inf_{\mathbf{m},c}\sup_{\Phi}\mathcal{L}(\mathbf{m},c,\Phi)\\
&\ge \sup_{\Phi}\inf_{\mathbf{m},c}\mathcal{L}(\mathbf{m},c,\Phi)\\
&= \sup_{\Phi}\inf_{\mathbf{m},c}\bigg\{\int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\int_\Omega |c(x)|dx + \int_\Omega \Phi(x)(\mu_1(x)-\mu_0(x)+\nabla\cdot\mathbf{m}(x)-c(x))dx\bigg\}\\
&= \sup_{\Phi}\inf_{\mathbf{m},c}\bigg\{\int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\int_\Omega |c(x)|dx + \int_\Omega \Phi(x)(\mu_1(x)-\mu_0(x)-c(x))dx\\
&\quad - \int_\Omega \nabla\Phi(x)\cdot\mathbf{m}(x)dx + \int_{\partial\Omega}\Phi(x)\mathbf{m}(x)\cdot n(x)ds(x)\bigg\}\\
&= \sup_{\Phi}\bigg\{\int_\Omega \Phi(x)(\mu_1(x)-\mu_0(x)) + \inf_{\mathbf{m},c}\int_\Omega \|\mathbf{m}(x)\| - \nabla\Phi(x)\cdot\mathbf{m}(x)dx + \int_\Omega \frac{1}{\alpha}|c(x)| - \Phi(x)c(x)dx\bigg\}\\
&= \sup_{\Phi}\bigg\{\int_\Omega \Phi(x)(\mu_1(x)-\mu_0(x))dx : \|\nabla\Phi\| \le 1, |\Phi| \le \frac{1}{\alpha}\bigg\}.
\end{aligned}$$

From the calculation, the optimizer Φ satisfies the following:

$$\nabla\Phi \in \partial\|\mathbf{m}(x)\|, \qquad \alpha\Phi \in \partial|c(x)|.$$

We show the duality gap is zero using the proposition 2.2.1.

$$\begin{aligned}
&\int_\Omega \|\mathbf{m}(x)\|dx + \frac{1}{\alpha}\int_\Omega |c(x)|dx + \int_\Omega \Phi(x)(\mu_1(x) - \mu_0(x) + \nabla\cdot\mathbf{m}(x) - c(x))dx \\
&= \int_\Omega \|\mathbf{m}(x)\| - \nabla\Phi\cdot\mathbf{m}(x)dx + \int_\Omega \frac{1}{\alpha}|c(x)| - \Phi(x)c(x)dx + \int_\Omega \Phi(x)(\mu_1(x) - \mu_0(x))dx \\
&= \int_\Omega \Phi(x)(\mu_1(x) - \mu_0(x))dx
\end{aligned}$$

This concludes the proof. □

2.2.2 L^2 Generalized Unnormalized Wasserstein metric.

Let $p = 2$. From the definition (2.2.1), we now consider

$$\begin{aligned}
UW_2(\mu_0,\mu_1)^2 = \inf_{v,\mu,f}\Bigg\{ &\int_0^1\int_\Omega \|v(t,x)\|^2\mu(t,x)\,dx\,dt + \frac{1}{\alpha}\int_0^1\int_\Omega \|f(t,x)\|^2\,dx\,dt : \\
&\partial_t\mu(t,x) + \nabla\cdot(\mu(t,x)v(t,x)) = f(t,x), t\in[0,1], x\in\Omega, \qquad (2.7)\\
&\mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x)\Bigg\}.
\end{aligned}$$

Proposition 2.2.3. *The L^2 generalized unnormalized Wasserstein metric is a well-defined metric function in $M(\Omega)$. In addition, the minimizer $(v(t,x),\mu(t,x),f(t,x))$ for (2.7) satisfies*

$$v(t,x) = \nabla\Phi(t,x), \qquad f(t,x) = \alpha\Phi(t,x),$$

and

$$\begin{aligned}
&\partial_t\mu(t,x) + \nabla\cdot(\mu(t,x)\nabla\Phi(t,x)) = \alpha\Phi(t,x)\\
&\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2 \le 0.
\end{aligned}$$

In particular, if $\mu(t,x) > 0$, then

$$\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2 = 0.$$

Proof. Denote $\mathbf{m}(t,x) = \mu(t,x)v(t,x)$. Then the problem becomes

$$\frac{1}{2}UW_2(\mu_0,\mu_1)^2 = \inf_{\mathbf{m},\mu,f}\Big\{ \int_0^1\int_\Omega \frac{\|\mathbf{m}(t,x)\|^2}{2\mu(t,x)}\,dx\,dt + \frac{1}{2\alpha}\int_0^1\int_\Omega |f(t,x)|^2 dxdt : \\ \partial_t\mu(t,x) + \nabla\cdot\mathbf{m}(t,x) = f(t,x), \\ \mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x), x\in\Omega, 0\le t\le 1\Big\}. \tag{2.8}$$

Denote $\Phi(t,x)$ as a Lagrange multiplier. Consider the Lagrangian

$$\mathcal{L}(\mathbf{m},\mu,f,\Phi) = \int_0^1\int_\Omega \frac{\|\mathbf{m}(t,x)\|^2}{2\mu(t,x)}\,dx\,dt + \frac{1}{2\alpha}\int_0^1\int_\Omega |f(t,x)|^2 dxdt \\ + \int_0^1\int_\Omega \Phi(t,x)\Big(\partial_t\mu(t,x) + \nabla\cdot\mathbf{m}(t,x) - f(t,x)\Big)dxdt.$$

From KKT condition $\delta_{\mathbf{m}}\mathcal{L} = 0, \delta_\mu\mathcal{L}\ge 0, \delta_f\mathcal{L} = 0, \delta_\Phi\mathcal{L} = 0$, the minimizer satisfies the following properties:

$$\frac{\mathbf{m}(t,x)}{\mu(t,x)} = \nabla\Phi(t,x) \tag{2.9}$$

$$-\frac{\|\mathbf{m}(t,x)\|^2}{2\mu(t,x)^2} - \partial_t\Phi(t,x) \ge 0 \tag{2.10}$$

$$f(t,x) = \alpha\Phi(t,x)$$

$$\partial_t\mu(t,x) + \nabla\cdot\mathbf{m}(t,x) - f(t,x) = 0.$$

Combining (2.9) and (2.10) yields: $\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2 \le 0$. □

We next derive the corresponding Monge problem for unnormalized optimal transport with a spatially dependent source function. We note that the following derivations are formal in Eulerian coordinates of fluid dynamics. We are following the proof of Proposition 4 in [GLO19].

Proposition 2.2.4 (Generalized Unnormalized Monge problem)**.**

$$UW_2(\mu_0,\mu_1)^2 = \inf_{M,f(t,x)}\int_\Omega \|M(x)-x\|^2\mu_0(x)dx + \alpha\int_0^1\int_\Omega |f(t,x)|^2 dxdt \\ + \int_\Omega\int_0^1\int_0^t f\Big(s, sM(x)+(1-s)x\Big)\|M(x)-x\|^2 Det\Big(s\nabla M(x) + (1-s)\mathbb{I}\Big)\mathbf{d}sdtdx \tag{2.11}$$

where $M : \Omega \to \Omega$ is an invertible mapping function and $f : \Omega \times [0,1] \to \mathbb{R}$ is a spatially dependent source function. The unnormlized push forward relation holds

$$\begin{aligned}&\mu(1, M(x))Det(\nabla M(x)) \\ &= \mu(0,x) + \int_0^1 f\Big(t, tM(x) + (1-t)\mathbb{I}\Big)Det\Big(t\nabla M(x) + (1-t)\mathbb{I}\Big)dt.\end{aligned} \tag{2.12}$$

Proof. We derive the Lagrange formulation of the unnormalized optimal transport with $p = 2$. Consider a mapping function $X_t(x)$ with vector field $v(t, X_t(x))$ satisfying

$$\frac{\mathbf{d}}{dt}X_t(x) = v(t, X_t(x)), \qquad X_0(x) = x. \tag{2.13}$$

Then

$$\begin{aligned}\int_\Omega \int_0^1 \|v(t,x)\|^2 \mu(t,x)dtdx &= \int_\Omega \int_0^1 \|v(t, X_t(x))\|^2 \mu(t, X_t(x))Det(\nabla X_t(x))dxdt \\ &= \int_\Omega \int_0^1 \|\frac{\mathbf{d}}{dt}X_t(x)\|^2 \mu(t, X_t(x))Det(\nabla X_t(x))dxdt.\end{aligned} \tag{2.14}$$

Define $J(t,x) := \mu(t, X_t(x))Det\big(\nabla X_t(x)\big)$. Differentiate $J(t,x)$ with respect to t,

$$\begin{aligned}\frac{\mathbf{d}}{dt}J(t,x) &= \frac{\mathbf{d}}{dt}\Big\{\mu(t, X_t(x))Det(\nabla X_t(x))\Big\} \\ &= \partial_t \mu(t, X_t(x))Det(\nabla X_t(x)) + \nabla_X \mu(t, X_t(x)) \cdot \frac{\mathbf{d}}{dt}X_t(x)Det(\nabla X_t(x)) \\ &\quad + \mu(t, X_t(x))\partial_t Det(\nabla X_t(x)) \\ &= \partial_t \mu(t, X_t(x))Det(\nabla X_t(x)) + \nabla_X \mu(t, X_t(x)) \cdot \frac{\mathbf{d}}{dt}X_t(x)Det(\nabla X_t(x)) \\ &\quad + \mu(t, X_t(x))\nabla \cdot v(t, X_t(x))Det(\nabla X_t(x)) \\ &= \Big(\partial_t \mu + v \cdot \nabla \mu + \mu \nabla \cdot v\Big)(t, X_t(x))Det(\nabla X_t(x)) \\ &= \Big(\partial_t \mu + \nabla \cdot (\mu v)\Big)(t, X_t(x))Det(\nabla X_t(x)) \\ &= f\big(t, X_t(x)\big)Det(\nabla X_t(x)).\end{aligned}$$

Denote

$$J(t,x) = J(0,x) + \int_0^t \frac{\mathbf{d}}{\mathbf{d}s} J(s,x)\mathbf{d}s.$$

Since $X_0(x) = x$ and $\nabla X_0(x) = \mathbb{I}$, then $J(0,x) = \mu(0,x)$. This yields

$$\mu(t, X_t(x))Det(\nabla X_t(x)) = \mu(0,x) + \int_0^t f\big(s, X_s(x)\big)Det(\nabla X_s(x))\mathbf{d}s.$$

Since the minimizer in Eulerian coordinates satisfies the Hamilton-Jacobi equation:

$$\partial_t \Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2 = 0,$$

and $\frac{\mathbf{d}}{dt}X_t(x) = \nabla\Phi(t, X_t(x))$, then we have $\frac{\mathbf{d}^2}{dt^2}X_t(x) = 0$. This implies

$$\frac{\mathbf{d}}{dt}X_t(x) = v(t, X_t(x)) = M(x) - x,$$

thus $X_t(x) = (1-t)x + tM(x)$ and $Det(\nabla X_t(x)) = Det((1-t)\mathbb{I} + t\nabla M(x))$. Substitute all the above into (2.14):

$$\begin{aligned}
(2.14) &= \int_0^1 \int_\Omega \|\frac{\mathbf{d}}{dt}X_t(x)\|^2 J(t,x)dxdt \\
&= \int_0^1 \int_\Omega \|M(x) - x\|^2 \bigg(J(0,x) + \int_0^t \frac{\mathbf{d}}{\mathbf{d}s} J(s,x)ds\bigg)dxdt \\
&= \int_0^1 \int_\Omega \|M(x) - x\|^2 \mu(0,x)dxdt \\
&+ \int_0^1 \int_\Omega \|M(x) - x\|^2 \int_0^t f\big(s, X_s(x)\big)Det(\nabla X_s(x))\mathbf{d}s dxdt \\
&= \int_\Omega \|M(x) - x\|^2 \mu(0,x)dx \\
&+ \int_0^1 \int_0^t \int_\Omega \|M(x) - x\|^2 f\bigg(s, sM(x) + (1-s)x\bigg)Det\bigg((1-s)\mathbb{I} + s\nabla M(x)\bigg)dx\mathbf{d}s dt.
\end{aligned}$$

This concludes the derivation. □

We next find the relation between the spatially dependent source function $f(t,x)$ and the mapping function $M(x)$. For the simplicity of presentation, here we assume the periodic boundary conditions on Ω. We are following the proof of Proposition 5 in [GLO19].

Proposition 2.2.5 (Generalized Unnormalized Monge-Ampère equation)**.** *The optimal mapping function $M(x) = \nabla\Psi(x)$ satisfies the following unnormalized Monge-Ampère equation*

$$\begin{aligned}&\mu(1, \nabla\Psi(x))Det(\nabla^2\Psi(x)) - \mu(0, x)\\ &= \alpha\int_0^1 \left(\Psi(x) - \frac{\|x\|^2}{2} + \frac{t\|\nabla\Psi(x) - x\|^2}{2}\right) Det\Big(t\nabla^2\Psi(x) + (1-t)\mathbb{I}\Big)dt.\end{aligned} \tag{2.15}$$

Proof. From the Hopf-Lax formula for the Hamilton-Jacobi equation,

$$\Phi(1, y) = \sup_x \Phi(0, x) + \frac{\|y - x\|^2}{2}.$$

Since M is the optimal mapping function, $x = M^{-1}(y)$ is a maximizer of the supremum for each y. Thus, the maximizer satisfies

$$\nabla\Phi(0, x) + x - M(x) = 0,$$

and we can rewrite the formula as

$$\Phi(1, M(x)) = \Phi(0, x) + \frac{\|M(x) - x\|^2}{2}.$$

We further denote $\Psi(x) = \Phi(0, x) + \frac{\|x\|^2}{2}$, then $M(x) = \nabla\Psi(x)$. From $X_t(x) = (1-t)x + tM(x)$,

$$\begin{aligned}\Phi(t, X_t(x)) &= \Phi(0, x) + \frac{\|X_t(x) - x\|^2}{2t}\\ &= \Phi(0, x) + \frac{t\|M(x) - x\|^2}{2}\\ &= \Psi(x) - \frac{\|x\|^2}{2} + \frac{t\|\nabla\Psi(x) - x\|^2}{2}\end{aligned}$$

and

$$\nabla X_t(x) = (1-t)\mathbb{I} + t\nabla^2\Psi(x).$$

Substituting $f(t, x) = \alpha\Phi(t, x)$ and $M(x) = \nabla\Psi(x)$ into (2.12), we get

$$\begin{aligned}&\mu(1, \nabla\Psi(x))Det(\nabla^2\Psi(x)) - \mu(0, x)\\ &= \int_0^1 \alpha\left(\Psi(x) - \frac{\|x\|^2}{2} + \frac{t\|\nabla\Psi(x) - x\|^2}{2}\right) Det\Big(t\nabla^2\Psi(x) + (1-t)\mathbb{I}\Big)dt.\end{aligned}$$

□

Now, we show the Kantorovich formulation of the problem (2.7).

Proposition 2.2.6 (L^2 Generalized Unnormalized Kantorovich formulation)**.** *The unnormalized Kantorovich formulation with $f(t,x)$ satisfies*

$$\frac{1}{2}UW_2(\mu_0,\mu_1)^2 = \sup_{\Phi}\left\{\int_\Omega \Big(\Phi(1,x)\mu_1(x) - \Phi(0,x)\mu_0(x)\Big)dx - \frac{\alpha}{2}\int_0^1\int_\Omega \Phi(t,x)^2\,dx\,dt\right\},$$

where the supremum is taken among all $\Phi : [0,1]\times\Omega \to \mathbb{R}$ satisfying

$$\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2 \le 0.$$

Proof. We introduce a Lagrange multiplier $\Phi(t,x)$ to reformulate the equation (2.8).

$$\begin{aligned}
&\frac{1}{2}UW_2(\mu_0,\mu_1)^2\\
&= \inf_{\mathbf{m},\mu,f}\sup_{\Phi}\left\{\int_0^1\int_\Omega \frac{\|\mathbf{m}(t,x)\|^2}{2\mu(t,x)} + \frac{1}{2\alpha}f(x,t)^2 + \Phi(t,x)\big(\partial_t\mu(t,x) + \nabla\cdot\mathbf{m}(t,x) - f(t,x)\big)dxdt\right\}\\
&\ge \sup_{\Phi}\inf_{\mathbf{m},\mu,f}\left\{\int_0^1\int_\Omega \frac{\|\mathbf{m}(t,x)\|^2}{2\mu(t,x)} + \frac{1}{2\alpha}f(x,t)^2 + \Phi(t,x)\big(\partial_t\mu(t,x) + \nabla\cdot\mathbf{m}(t,x) - f(t,x)\big)dxdt\right\}\\
&= \sup_{\Phi}\inf_{\mathbf{m},\mu,f}\left\{\int_0^1\int_\Omega \frac{\|\mathbf{m}(t,x)\|^2}{2\mu(t,x)} - \nabla\Phi(t,x)\cdot\mathbf{m}(t,x) + \frac{1}{2\alpha}f(x,t)^2 + \Phi(t,x)\cdot\big(\partial_t\mu(t,x) - f(t,x)\big)dxdt\right\}\\
&= \sup_{\Phi}\inf_{\mathbf{m},\mu,f}\left\{\int_0^1\int_\Omega \frac{1}{2}\left\|\frac{\mathbf{m}(t,x)}{\mu(t,x)} - \nabla\Phi(t,x)\right\|^2\mu(t,x) - \frac{1}{2}\|\nabla\Phi(t,x)\|^2\mu(t,x)\,dx\,dt\right.\\
&\qquad + \int_\Omega \Phi(1,x)\mu_1(x) - \Phi(0,x)\mu_0(x)dx\\
&\qquad \left. + \int_0^1\int_\Omega -\mu(t,x)\partial_t\Phi(t,x) + \frac{1}{2\alpha}f(t,x)^2 - \Phi(t,x)f(t,x))\,dx\,dt\right\}.
\end{aligned}$$

By the Proposition 2.2.3, the minimizer $\mathbf{m}$ satisfies $\dfrac{\mathbf{m}(t,x)}{\mu(t,x)} = \nabla\Phi(t,x)$. Thus,

$$
\begin{aligned}
&= \sup_{\Phi}\Bigg\{\int_\Omega\Big(\Phi(1,x)\mu_1(x) - \Phi(0,x)\mu_0(x)\Big)dx \\
&\qquad + \inf_{\mu}\int_0^1\int_\Omega -\mu(t,x)\Big(\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2\Big)\,dx\,dt \\
&\qquad + \inf_{f}\int_0^1\int_\Omega \frac{1}{2\alpha}f(t,x)^2 - \Phi(t,x)f(t,x)\,dx\,dt\Bigg\} \\
&= \sup_{\Phi}\Bigg\{\int_\Omega\Big(\Phi(1,x)\mu_1(x) - \Phi(0,x)\mu_0(x)\Big)dx \\
&\qquad + \inf_{\mu}\int_0^1\int_\Omega -\mu(t,x)\Big(\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2\Big)\,dx\,dt \\
&\qquad + \inf_{f}\int_0^1\int_\Omega \frac{1}{2\alpha}\Big(f(t,x) - \alpha\Phi(t,x)\Big)^2\,dx\,dt - \frac{\alpha}{2}\int_0^1\int_\Omega \Phi(t,x)^2\,dx\,dt\Bigg\}.
\end{aligned}
$$

Again from Proposition 2.2.3, the minimizer satisfies $f(t,x) = \alpha\Phi(t,x)$. With the assumption $\mu(t,x) \geq 0$ for all $t \in [0,1]$ and $x \in \Omega$, the problem can be written with a constraint.

$$
\begin{aligned}
\frac{1}{2}UW_2(\mu_0,\mu_1)^2 = \sup_{\Phi}\Bigg\{\int_\Omega\Big(\Phi(1,x)\mu_1(x) - \Phi(0,x)\mu_0(x)\Big)dx - \frac{\alpha}{2}\int_0^1\int_\Omega \Phi(t,x)^2\,dx\,dt : \\
\partial_t\Phi(t,x) + \frac{1}{2}\|\nabla\Phi(t,x)\|^2 \leq 0\Bigg\}.
\end{aligned}
$$

We next show that the primal-dual gap is zero.

$$
\begin{aligned}
&\int_0^1 \int_\Omega \frac{\mathbf{m}(t,x)^2}{2\mu(t,x)} + \frac{1}{2\alpha} f(t,x)^2 \, dx \, dt \\
&= \int_0^1 \int_\Omega \frac{1}{2} \|\nabla \Phi\|^2 \mu(t,x) \, dx \, dt + \frac{\alpha}{2} \int_0^1 \int_\Omega \Phi(t,x)^2 \, dx \, dt \\
&= \int_0^1 \int_\Omega \left(-\frac{1}{2} \|\nabla \Phi(t,x)\|^2 \mu(t,x) + \|\nabla \Phi(t,x)\|^2 \mu(t,x) + \frac{\alpha}{2} \Phi(t,x)^2 \right) dx \, dt \\
&= \int_0^1 \int_\Omega \partial_t \Phi(t,x) \mu(t,x) + \Phi(t,x) \Big(-\nabla \cdot \big(\mu(t,x) \nabla \Phi(t,x) \big) \Big) + \frac{\alpha}{2} \Phi(t,x)^2 \, dx \, dt \\
&= \int_\Omega \Phi(1,x) \mu_1(x) - \Phi(0,x) \mu_0(x) dx \\
&\qquad - \int_0^1 \int_\Omega \Phi(t,x) \Big(\partial_t \mu(t,x) + \nabla \cdot \big(\mu(t,x) \nabla \Phi(t,x) \big) \Big) \, dx \, dt + \frac{\alpha}{2} \int_0^1 \int_\Omega \Phi(t,x)^2 dx dt \\
&= \int_\Omega \Phi(1,x) \mu_1(x) - \Phi(0,x) \mu_0(x) dx \\
&\qquad - \int_0^1 \int_\Omega \Phi(t,x) f(t,x) \, dx \, dt + \frac{\alpha}{2} \int_0^1 \int_\Omega \Phi(t,x)^2 dx dt.
\end{aligned}
$$

Using $f(t,x) = \alpha \Phi(t,x)$, we get

$$
\frac{1}{2} UW_2(\mu_0, \mu_1)^2 = \int_\Omega \Phi(1,x) \mu(1,x) - \Phi(0,x) \mu(0,x) dx - \frac{\alpha}{2} \int_0^1 \int_\Omega \Phi(t,x)^2 dx dt.
$$

This concludes the proof. □

Remark 2.2.5. We note that our results and proofs follow directly from the those used in [GLO19]. The major difference between [GLO19] and our paper is that in the case of spatial independent source function, $f(t) = \frac{\alpha}{|\Omega|} \int_\Omega \Phi(t,x) dx$, while in the case of spatial dependent source function, $f(t,x) = \alpha \Phi(t,x)$. This difference remains in the corresponding Monge problem and Kantorvich problem. In particular, we obtain a new spatial dependent unnormalized Monge-Ampère equation (2.15).

2.3 Numerical methods

In this section, we propose a Nesterov accelerated gradient descent method to solve L^2 unnormalized OT. In addition, we design a primal-dual hybrid gradient method to solve L^1

unnormalized OT.

2.3.1 L^2 Generalized Unnormalized Wasserstein metric

In this section, we present a new numerical implementation for L^2 unnormalized Wasserstein metric. We obtain a unconstrained version of the problem by plugging the PDE constraint into the objective function. Then the accelerated Nesterov gradient descent method is applied to solve the problem. We show that each iteration involves a simple elliptic equation where fast solvers can be applied. This novel numerical method can also be used in normalized optimal transport and unnormalized optimal transport with a spatially independent source function $f(t)$.

Using Proposition 2.2.3, we can rewrite the equation (2.8) as follows:

$$\begin{aligned} UW_2(\mu_0,\mu_1)^2 = \inf_{\Phi,\mu}\Bigg\{ \int_0^1\int_\Omega \|\nabla\Phi(t,x)\|_2^2\mu(t,x)dxdt + \alpha\int_0^1\int_\Omega |\Phi(t,x)|^2\,dx\,dt : \\ \partial_t\mu(t,x) + \nabla\cdot(\mu(t,x)\nabla\Phi(t,x)) = \alpha\Phi(t,x), \\ \mu(0,x)=\mu_0(x), \mu(1,x)=\mu_1(x)\Bigg\}. \end{aligned}$$

Define an operator $L_\mu = -\nabla\cdot(\mu\nabla)$. The constraint $\partial_t\mu - L_\mu\Phi = \alpha\Phi$ leads to

$$\Phi = (L_\mu + \alpha Id)^{-1}\partial_t\mu. \tag{2.16}$$

With (2.16), the minimization problem can be reformulated as

$$\begin{aligned} UW_2(\mu_0,\mu_1)^2 = \inf_{\mu}\Bigg\{ \int_0^1\int_\Omega \mu\|\nabla(L_\mu+\alpha Id)^{-1}\partial_t\mu\|_2^2\,dx\,dt + \alpha\int_0^1\int_\Omega |(L_\mu+\alpha Id)^{-1}\partial_t\mu|^2\,dx\,dt : \\ \mu(0,x)=\mu_0(x), \mu(1,x)=\mu_1(x)\Bigg\}. \end{aligned} \tag{2.17}$$

Using integration by parts,

$$\begin{aligned}
&\int_0^1 \int_\Omega \mu \|\nabla (L_\mu + \alpha Id)^{-1} \partial_t \mu\|^2 \, dx \, dt + \alpha \int_0^1 \int_\Omega |(L_\mu + \alpha Id)^{-1} \partial_t \mu|^2 \, dx \, dt \\
&= \int_0^1 \int_\Omega - \Big(\nabla \mu \nabla (L_\mu + \alpha Id)^{-1} \partial_t \mu \Big) \Big((L_\mu + \alpha Id)^{-1} \partial_t \mu \Big) \, dx \, dt \\
&\quad + \alpha \int_0^1 \int_\Omega |(L_\mu + \alpha Id)^{-1} \partial_t \mu|^2 \, dx \, dt \\
&= \int_0^1 \int_\Omega \Big(L_\mu (L_\mu + \alpha Id)^{-1} \partial_t \mu \Big) \Big((L_\mu + \alpha Id)^{-1} \partial_t \mu \Big) \, dx \, dt \\
&\quad + \alpha \int_0^1 \int_\Omega |(L_\mu + \alpha Id)^{-1} \partial_t \mu|^2 \, dx \, dt \\
&= \int_0^1 \int_\Omega \Big((L_\mu + \alpha Id)(L_\mu + \alpha Id)^{-1} \partial_t \mu \Big) \Big((L_\mu + \alpha Id)^{-1} \partial_t \mu \Big) \, dx \, dt \\
&= \int_0^1 \int_\Omega \partial_t \mu(t,x) (L_\mu + \alpha Id)^{-1} \partial_t \mu(t,x) \, dx \, dt.
\end{aligned}$$

Thus, the unnormalized Wasserstein-2 distance forms

$$\begin{aligned}
UW_2(\mu_0, \mu_1)^2 = \inf_\mu \Bigg\{ \int_0^1 \int_\Omega \partial_t \mu(t,x) (L_\mu + \alpha Id)^{-1} \partial_t \mu(t,x) \, dx \, dt : \\
\mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x) \Bigg\}.
\end{aligned} \tag{2.18}$$

Proposition 2.3.1. *If $\mu(t,x) > 0$, then the Euler-Lagrange equation of problem (2.18) satisfies the Hamilton-Jacobi equation, i.e.*

$$\partial_t \Phi(t,x) + \frac{1}{2} \|\nabla \Phi(t,x)\|^2 = 0, \qquad x \in \Omega, t \in [0,1]$$

where $\Phi(t,x) = (L_\mu + \alpha Id)^{-1} \partial_t \mu(t,x)$.

Remark 2.3.1. For unnormalized optimal transport with a spatially independent source function $f(t)$, the formula uses $(L_\mu + \frac{\alpha}{|\Omega|} \int_\Omega)^{-1}$ instead of $(L_\mu + \alpha Id)^{-1}$, i.e.

$$\begin{aligned}
UW_2(\mu_0, \mu_1)^2 = \inf_\mu \Bigg\{ \int_0^1 \int_\Omega \partial_t \mu(t,x) \Big(L_\mu + \frac{\alpha}{|\Omega|} \int_\Omega \Big)^{-1} \partial_t \mu(t,x) \, dx \, dt : \\
\mu(0,x) = \mu_0(x), \mu(1,x) = \mu_1(x) \Bigg\}.
\end{aligned}$$

The Euler-Lagrange equation satisfies the following:

$$\partial_t \Phi(t,x) + \frac{1}{2}\|\nabla \Phi(t,x)\|^2 = 0, \qquad x \in \Omega, t \in [0,1]$$

where $\Phi(t,x) = \left(L_\mu + \frac{\alpha}{|\Omega|}\int_\Omega\right)^{-1}\partial_t \mu(t,x)$.

Remark 2.3.2. If $\mu(t,x) = 0$, one can show that the Euler-Lagrange equation of problem (2.18) satisfies

$$\partial_t \Phi(t,x) + \frac{1}{2}\|\nabla \Phi(t,x)\|^2 \leq 0.$$

Proof. Define

$$\mathcal{I}(\mu) = \int_0^1 \int_\Omega \partial_t \mu(t,x)(L_\mu + \alpha Id)^{-1}\partial_t \mu(t,x) dx dt.$$

We now calculate the first variation of $\mathcal{I}(\mu)$ with a perturbation $\eta(t,x) \in C^\infty(\Omega \times [0,1])$.

$$\begin{aligned}
0 &= \lim_{h\to 0} \frac{\mathcal{I}(\mu + h\eta) - \mathcal{I}(\mu)}{h} \\
&= \lim_{h\to 0} \frac{1}{h}\int_0^1 \int_\Omega \left((\partial_t \mu + h\partial_t \eta)(L_{\mu+h\eta} + \alpha Id)^{-1}(\partial_t \mu + h\partial_t \eta) - \partial_t \mu(t,x)(L_\mu + \alpha Id)^{-1}\partial_t \mu(t,x)\right) dx dt \\
&= \lim_{h\to 0}\Bigg[\int_0^1 \int_\Omega \partial_t \mu\left(\frac{(L_{\mu+h\eta} + \alpha Id)^{-1} - (L_\mu + \alpha Id)^{-1}}{h}\right)\partial_t \mu \\
&\qquad\qquad + 2\partial_t \eta (L_{\mu+h\eta} + \alpha Id)^{-1}\partial_t \mu dx dt + O(h)\Bigg] \\
&= \int_0^1 \int_\Omega -\partial_t \mu (L_\mu + \alpha Id)^{-1} L_\eta (L_\mu + \alpha Id)^{-1}\partial_t \mu + 2\partial_t \eta (L_\mu + \alpha Id)^{-1}\partial_t \mu dx dt \\
&= \int_0^1 \int_\Omega -\Phi L_\eta \Phi + 2\Phi \partial_t \eta dx dt \\
&= \int_0^1 \int_\Omega -\eta\left(\|\nabla \Phi\|^2 + 2\partial_t \Phi\right) dx dt.
\end{aligned}$$

This has to be true for all $\eta \in C^\infty(\Omega \times [0,1])$. Thus, we get

$$\partial_t \Phi + \frac{1}{2}\|\nabla \Phi\|^2 = 0, \qquad x \in \Omega, t \in [0,1].$$

This concludes the proof. □

Using Proposition 2.3.1, we can formulate a Nesterov accelerated gradient descent method [Nes83] to solve the minimization problem (2.18).

Algorithm 1 Nesterov Gradient descent method for UW_2 with $f(t,x)$

While not converged

$$\mu^{k+\frac{1}{2}} = \mu^k - \tau\nabla\mathcal{I}(\mu^k) = \mu^k + \frac{\tau}{2}\big(\partial_t\Phi^k + \frac{1}{2}\|\nabla\Phi^k\|^2\big) \text{ where } \Phi^k = (L_{\mu^k} + \alpha Id)^{-1}\partial_t\mu^k$$

$$\mu^{k+\frac{1}{2}}(t,x) = \max\{\mu^{k+\frac{1}{2}}(t,x), 0\} \quad \text{for all } (t,x) \in [0,1]\times\Omega$$

$$\mu^{k+1} = (1-\gamma^k)\mu^{k+\frac{1}{2}} + \gamma^k\mu^k$$

Here, τ and γ^k are step sizes of the algorithm.

$$\gamma^k = \frac{1-\lambda^k}{\lambda^{k+1}}, \quad \lambda^0 = 0, \quad \lambda^k = \frac{1+\sqrt{1+4(\lambda^{k-1})^2}}{2}.$$

Remark 2.3.3. The Nesterov accelerated gradient descent method can be used for a spatially independent source function $f(t)$. We simply replace the operator $L_\mu + \alpha Id$ with $L_\mu + \alpha\int_\Omega$ from Algorithm 1.

Remark 2.3.4. Here we apply an iterative method, such as conjugate gradient, to solve $(L_{\mu^k} + \alpha Id)^{-1}\partial_t\mu^k$.

Remark 2.3.5. We remark that variational problem (2.18) is convex w.r.t. $\mu(t,x)$. This fact holds following the second variational formula derived in Lemma 2 of [Li18]. In other words, our gradient descent algorithm is applied to a convex optimization problem (2.18). For the completeness of this paper, we present the formal derivation here. Denote

$$J(\mu) = \frac{1}{2}\int_0^1\int_\Omega \partial_t\mu(t,x)(L_\mu + \alpha Id)^{-1}\partial_t\mu(t,x)\,dx\,dt.$$

Consider a test function $h \in C^\infty([0,1]\times\Omega)$, such that $h(0,x) = h(1,x) = 0$. Given $\epsilon \in \mathbb{R}^1$, we claim

$$\frac{d^2}{d\epsilon^2}J(\mu + \epsilon h)|_{\epsilon=0} \geq 0.$$

If the above statement is true, we know that the variational problem (2.18) is convex w.r.t $\mu(t,x)$. In fact, by routine computations, we observe that

$$\begin{aligned}&\frac{d^2}{d\epsilon^2}J(\mu+\epsilon h)|_{\epsilon=0}\\ =&\int_\Omega \Big([\partial_t h - L(h)(L_\mu+\alpha Id)^{-1}\partial_t\mu], (L_\mu+\alpha Id)^{-1}[\partial_t h - L(h)(L_\mu+\alpha Id)^{-1}\partial_t\mu]\Big)\,dx\,dt,\end{aligned}$$

which finishes the proof.

We next present the discretization of density path in both time and spatial domains, where the spatial domain is given by $1D$ or $2D$. Here we formulate the operator L_μ and derive its inverse into matrix forms; see similar approaches in [Li18].

2.3.1.1 1D Discretization

Consider the following one dimensional discretization:

$$\begin{aligned}&\boldsymbol{\mu} = (\boldsymbol{\mu}^0, \cdots, \boldsymbol{\mu}^{N_t}) \in \mathbb{R}^{(N_t+1)\times N_x}\\ &\boldsymbol{\mu}^n = (\mu_0^n, \cdots, \mu_{N_x-1}^n) \in \mathbb{R}^{N_x} \qquad (n=0,\cdots,N_t)\\ &\mu_i^n \in \mathbb{R} \qquad (i=0,\cdots,N_x-1, n=0,\cdots,N_t)\\ &\mu_i^0 = \mu_0(i\Delta x), \qquad \mu_i^{N_t} = \mu_1(i\Delta x), \qquad (i=0,\cdots,N_x-1)\\ &\Delta x = \frac{|\Omega|}{N_x-1} \qquad \Delta t = \frac{1}{N_t}.\end{aligned}$$

Using the finite volume method, the weighted Laplacian operator $\tilde{L}_{\boldsymbol{\mu}^n,\alpha} := L_{\boldsymbol{\mu}^n} + \alpha Id$ can be represented as the following matrix:

$$\tilde{L}_{\boldsymbol{\mu}^n,\alpha} = \begin{pmatrix} \frac{\mu_0^n+\mu_1^n}{2\Delta x^2} & -\frac{\mu_0^n+\mu_1^n}{2\Delta x^2} & 0 & \cdots & 0\\ -\frac{\mu_0^n+\mu_1^n}{2\Delta x^2} & \frac{\mu_0^n+\mu_1^n}{2\Delta x^2}+\frac{\mu_1^n+\mu_2^n}{2\Delta x^2} & -\frac{\mu_1^n+\mu_2^n}{2\Delta x^2} & \cdots & 0\\ 0 & -\frac{\mu_1^n+\mu_2^n}{2\Delta x^2} & \frac{\mu_1^n+\mu_2^n}{2\Delta x^2}+\frac{\mu_2^n+\mu_3^n}{2\Delta x^2} & \cdots & 0\\ vdots & \ddots & \ddots & \ddots & vdots\\ 0 & \cdots & \cdots & \cdots & -\frac{\mu_{N_x-2}^n+\mu_{N_x-1}^n}{2\Delta x^2} \end{pmatrix} + \alpha Id$$

Further using the forward Euler method in time, formula (2.18) can be discretized as

$$\begin{aligned}
&\int_0^1 \int_\Omega \partial_t \mu(t,x)(L_\mu + \alpha Id)^{-1} \partial_t \mu(t,x) dx dt \\
&\approx \Delta t \Delta x \sum_{n=0}^{N_t-1} \left\langle \frac{\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n}{\Delta t}, (L_{\boldsymbol{\mu}^n} + \alpha Id)^{-1} \frac{\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n}{\Delta t} \right\rangle_{L^2} \\
&= \frac{\Delta x}{\Delta t} \sum_{n=0}^{N_t-1} \left\langle \boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n, (L_{\boldsymbol{\mu}^n} + \alpha Id)^{-1} (\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n) \right\rangle_{L^2}
\end{aligned}$$

with $\boldsymbol{\mu}^0$ and $\boldsymbol{\mu}^{N_t}$ are given. $\langle \cdot, \cdot \rangle_{L^2}$ is L^2 norm in $\mathbb{R}^{N_x}$ such that

$$\langle \boldsymbol{a}, \boldsymbol{b} \rangle_{L^2} = \sum_{i=0}^{N_x-1} a_i b_i \qquad \text{for } \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^{N_x}.$$

We are now ready to present the derivative of $E(\boldsymbol{\mu})$, and formulate the discrete Hamilton-Jacobi equation as in Algorithm 1.

Proposition 2.3.2. *Denote $\tilde{L}_{\boldsymbol{\mu}^n,\alpha} := L_{\boldsymbol{\mu}^n} + \alpha Id$. Let*

$$E(\boldsymbol{\mu}) := \frac{\Delta x}{\Delta t} \sum_{n=0}^{N_t-1} \left\langle \boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n, \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha} (\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n) \right\rangle_{L^2}.$$

Suppose $x \in \Omega$. The derivative of $E(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}^n$ ($n = 1, \cdots, N_t - 1$) is

$$\begin{aligned}
\frac{\delta E(\boldsymbol{\mu})}{\delta \boldsymbol{\mu}^n} = \frac{\Delta x}{\Delta t} \Bigg(& -2\tilde{L}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n) + 2\tilde{L}_{\boldsymbol{\mu}^{n-1},\alpha}(\boldsymbol{\mu}^n - \boldsymbol{\mu}^{n-1}) \\
& - \left(\left\langle \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n), L_{\boldsymbol{e}_i} \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n) \right\rangle_{L^2} \right)_{i=0}^{N_x-1} \Bigg)
\end{aligned}$$

where $\boldsymbol{e}_i \in \mathbb{R}^{N_x}$ is an index vector defined as

$$\boldsymbol{e}_i = \begin{cases} 1 & i^{th} \text{ index} \\ 0 & \text{else.} \end{cases}$$

Proof. Differentiating $E(\boldsymbol{\mu})$ with respect to $\boldsymbol{\mu}^n$ for $n = 1, \cdots, N_t - 1$, we get

$$\begin{aligned}\frac{\delta E(\boldsymbol{\mu})}{\delta \boldsymbol{\mu}^n} &= \frac{\delta}{\delta \boldsymbol{\mu}^n}\left(\Delta t \Delta x \sum_{m=0}^{N_t-1}(\boldsymbol{\mu}^{m+1}-\boldsymbol{\mu}^m)\tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{m+1}-\boldsymbol{\mu}^m)\right)\\ &= \Delta t \Delta x\Bigg(-2\tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)+2\tilde{L}^{-1}_{\boldsymbol{\mu}^{n-1},\alpha}(\boldsymbol{\mu}^n-\boldsymbol{\mu}^{n-1})\\ &\qquad +(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\frac{\partial \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}}{\partial \boldsymbol{\mu}^n}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\Bigg),\end{aligned}$$

and

$$\begin{aligned}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\frac{\delta \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}}{\delta \boldsymbol{\mu}^n_n}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n) &= -\left\langle \boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n, \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}L_{\boldsymbol{e_i}}\tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\right\rangle_{L^2}\\ &= -\left\langle \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n), L_{\boldsymbol{e}_i}\tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\right\rangle_{L^2}.\end{aligned}$$

This concludes the proof. □

Consider $\boldsymbol{u} = (u_0, \cdots, u_{N_x-1})^T \in \mathbb{R}^{N_x}$, then $\langle \boldsymbol{u}, L_{\boldsymbol{e}_i}\boldsymbol{u}\rangle_{L_2}$ forms the R.H.S. of the discrete Hamilton-Jacobi equation as follows

$$\langle \boldsymbol{u}, L_{\boldsymbol{e}_i}\boldsymbol{u}\rangle_{L^2} = \begin{cases} \frac{1}{2}\left(\frac{u_{i+1}-u_i}{\Delta x}\right)^2 + \frac{1}{2}\left(\frac{u_i-u_{i-1}}{\Delta x}\right)^2, & i = 1, \cdots, N_x - 2\\ \frac{1}{2}\left(\frac{u_{i+1}-u_i}{\Delta x}\right)^2, & i = 0\\ \frac{1}{2}\left(\frac{u_i-u_{i-1}}{\Delta x}\right)^2, & i = N_x - 1.\end{cases}$$

2.3.1.2 2D Discretization

Now, consider the two dimensional discretization. Assume $\Omega = [0,1]\times[0,1]$ and $t \in [0,1]$.

$\boldsymbol{\mu} = (\boldsymbol{\mu}^0, \cdots, \boldsymbol{\mu}^{N_t}) \in \mathbb{R}^{(N_t+1)\times N_x \times N_y}$

$\boldsymbol{\mu}^n = (\mu^n_{ij})_{i=0\ j=0}^{N_x-1\ N_y-1} \in \mathbb{R}^{N_x\times N_y} \qquad (n = 0, \cdots, N_t)$

$\mu^n_{ij} \in \mathbb{R} \qquad (i = 0, \cdots, N_x - 1, j = 0, \cdots, N_y - 1, n = 0, \cdots, N_t)$

$\mu^0_{ij} = \mu_0(i\Delta x, j\Delta y), \qquad \mu^{N_t}_{ij} = \mu_1(i\Delta x, j\Delta y), \qquad (i = 0, \cdots, N_x - 1, j = 0, \cdots, N_y - 1)$

$\Delta x = \dfrac{1}{N_x - 1}, \qquad \Delta y = \dfrac{1}{N_y - 1}, \qquad \Delta t = \dfrac{1}{N_t}.$

Similar to 1D case, using the finite volume method, formula (2.18) can be discretized as

$$\begin{aligned}&\int_0^1\int_0^1\int_0^1 \partial_t\mu(t,x,y)(L_\mu+\alpha Id)^{-1}\partial_t\mu(t,x,y)dx\,dy\,dt\\ &\approx \frac{\Delta x\Delta y}{\Delta t}\sum_{n=0}^{N_t-1}\left\langle \boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n,(L_{\boldsymbol{\mu}^n}+\alpha Id)^{-1}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\right\rangle_{L^2}\end{aligned}$$

with $\boldsymbol{\mu}^0$ and $\boldsymbol{\mu}^{N_t}$ are given and $\langle\cdot,\cdot\rangle_{L^2}$ is defined as

$$\langle \boldsymbol{a},\boldsymbol{b}\rangle_{L^2}=\sum_{i=0}^{N_x-1}\sum_{j=0}^{N_y-1}a_{ij}b_{ij}\qquad \text{for } \boldsymbol{a},\boldsymbol{b}\in\mathbb{R}^{N_x\times N_y}.$$

The major difference between 1D discretization and 2D discretization arises from the weighted Laplacian operator $\tilde{L}_{\boldsymbol{\mu}^n,\alpha}$. Consider $\boldsymbol{w}=(w_{i,j})_{i=0\ j=0}^{N_x-1\ N_y-1}\in\mathbb{R}^{N_x\times N_y}$. For $i=0,\cdots,N_x-1$ and $j=0,\cdots,N_y-1$, the operator can be described as follows:

$$\begin{aligned}&(\tilde{L}_{\boldsymbol{\mu}^n,\alpha}\boldsymbol{w})_{ij}\\ =&-\frac{1}{\Delta x^2}\left(\frac{\mu^n_{i+1,j}+\mu^n_{i,j}}{2}w_{i+1,j}-2\left(\frac{\mu^n_{i+1,j}+\mu^n_{i,j}}{2}+\frac{\mu^n_{i,j}+\mu^n_{i-1,j}}{2}\right)w_{i,j}+\frac{\mu^n_{i,j}+\mu^n_{i-1,j}}{2}w_{i-1,j}\right)\\ &-\frac{1}{\Delta y^2}\left(\frac{\mu^n_{i,j+1}+\mu^n_{i,j}}{2}w_{i,j+1}-2\left(\frac{\mu^n_{i,j+1}+\mu^n_{i,j}}{2}+\frac{\mu^n_{i,j}+\mu^n_{i,j-1}}{2}\right)w_{i,j}+\frac{\mu^n_{i,j}+\mu^n_{i,j-1}}{2}w_{i,j-1}\right)\\ &+\alpha w_{i,j}.\end{aligned}$$

Here, we assume the Neumann boundary on the spatial domain Ω. Thus,

$$\begin{aligned}w_{-1,j}=w_{0,j},\qquad w_{N_x,j}=w_{N_x-1,j},\qquad j=0,\cdots,N_y-1\\ w_{i,-1}=w_{i,0},\qquad w_{i,N_y}=w_{i,N_y-1},\qquad i=0,\cdots,N_x-1\\ \mu^n_{-1,j}=\mu^n_{0,j},\qquad \mu^n_{N_x,j}=\mu^n_{N_x-1,j},\qquad j=0,\cdots,N_y-1\\ \mu^n_{i,-1}=\mu^n_{i,0},\qquad \mu^n_{i,N_y}=\mu^n_{i,N_y-1},\qquad i=0,\cdots,N_x-1.\end{aligned}$$

Proposition 2.3.3. *Denote $\tilde{L}_{\boldsymbol{\mu}^n,\alpha}:=L_{\boldsymbol{\mu}^n}+\alpha Id$. Let*

$$E(\boldsymbol{\mu}):=\frac{\Delta x\Delta y}{\Delta t}\sum_{n=0}^{N_t-1}\left\langle \boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n,\tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1}-\boldsymbol{\mu}^n)\right\rangle_{L^2}.$$

Suppose $x \in \Omega = [0,1] \times [0,1]$. *The derivative of* $E(\boldsymbol{\mu})$ *with respect to* $\boldsymbol{\mu}^n$ $(n = 1, \cdots, N_t - 1)$ *is*

$$\frac{\delta E(\boldsymbol{\mu})}{\delta \boldsymbol{\mu}^n} = \frac{\Delta x \Delta y}{\Delta t} \Bigg(-2\tilde{L}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n) + 2\tilde{L}_{\boldsymbol{\mu}^{n-1},\alpha}(\boldsymbol{\mu}^n - \boldsymbol{\mu}^{n-1}) \\ - \left(\left\langle \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n), L_{\boldsymbol{e}_{ij}} \tilde{L}^{-1}_{\boldsymbol{\mu}^n,\alpha}(\boldsymbol{\mu}^{n+1} - \boldsymbol{\mu}^n) \right\rangle_{L^2} \right)_{i=0,j=0}^{N_x-1,N_y-1} \Bigg).$$

where $\boldsymbol{e}_{ij}$ *is an index vector such that* $\boldsymbol{e}_{k,l} = 1$ *if* $k = i$ *and* $l = j$ *and* 0 *otherwise.*

Proof. The proof follows exactly the one in proposition 2.3.2. □

Consider a vector $\boldsymbol{u} = (u_{ij})_{i=0}^{N_x-1} {}_{j=0}^{N_y-1} \in \mathbb{R}^{N_x \times N_y}$ that satisfies the Neumann boundary condition. Similar to 1D case, $\left\langle \boldsymbol{u}, L_{\boldsymbol{e}_{i,j}} \boldsymbol{u} \right\rangle_{L^2}$ can be computed easily based on the operator and it forms the R.H.S. of the discrete Hamilton-Jacobi equation. For $i = 0, \cdots, N_x - 1$ and $j = 0, \cdots, N_y - 1$,

$$\begin{aligned} \left\langle \boldsymbol{u}, L_{\boldsymbol{e}_{i,j}} \boldsymbol{u} \right\rangle_{L^2} =& \frac{1}{2} \left(\frac{u_{i+1,j} - u_{i,j}}{\Delta x} \right)^2 + \frac{1}{2} \left(\frac{u_{i,j} - u_{i-1,j}}{\Delta x} \right)^2 \\ &+ \frac{1}{2} \left(\frac{u_{i,j+1} - u_{i,j}}{\Delta y} \right)^2 + \frac{1}{2} \left(\frac{u_{i,j} - u_{i,j-1}}{\Delta y} \right)^2. \end{aligned}$$

2.3.2 L^1 Generalized Unnormalized Wasserstein metric

Our discussion here mainly focuses on $\|u\|_1 = \sum_i |u_i|$. The algorithm can be simply extended to $\|u\|_2 = \sqrt{\sum_i u_i^2}$ using the corresponding shrinkage operator. With the Lagrangian (2.5), we consider a saddle point problem.

$$\inf_{\mathbf{m},c} \sup_{\Phi} \mathcal{L}(m, c, \Phi).$$

We can use PDHG [CP11a] to solve the saddle point problem by minimizing $\mathcal{L}(\mathbf{m}, c, \Phi)$ over m and c and maximizing over Φ.

$$\mathbf{m}^{k+1} = \underset{\mathbf{m}}{\operatorname{argmin}} \left(\|\mathbf{m}\|_1 + \frac{\epsilon}{2}\|\mathbf{m}\|_2^2 + \left\langle \Phi^k, \nabla \cdot \mathbf{m} \right\rangle_{L^2} + \frac{1}{2\lambda}\|\mathbf{m} - \mathbf{m}^k\|_2^2 \right) \tag{2.19}$$

$$c^{k+1} = \underset{c}{\operatorname{argmin}} \left(\frac{1}{\alpha}\|c\|_1 + \frac{\epsilon}{2}\|c\|_2^2 - \left\langle \Phi^k, c \right\rangle_{L^2} + \frac{1}{2\lambda}\|c - c^k\|_2^2 \right) \tag{2.20}$$

$$\Phi^{k+1} = \underset{\Phi}{\operatorname{argmax}} \left(\left\langle \Phi, \nabla \cdot (2\mathbf{m}^{k+1} - \mathbf{m}^k) - (2c^{k+1} - c^k) + \mu_1 - \mu_0 \right\rangle_{L^2} - \frac{1}{2\tau}\|\Phi - \Phi^k\|_2^2 \right) \tag{2.21}$$

where λ and τ are step sizes of the algorithm. Note that we add a small $\|\cdot\|_2^2$ perturbation in (2.19) and (2.20) to strictly convexify the problem. This adjustment can overcome the possible non-uniqueness of the optimal transport problem. This trick is also related to so called the elastic net regularization [PB14], whose proximal operator is essentially the same as the proximal operator of L^1 norm shrink operator.

Algorithm 2 PDHG for UW_1 with $f(t, x)$

$$\begin{aligned}
\mathbf{m}^{k+1} &= 1/(1+\epsilon\lambda) shrink\left(\mathbf{m}^k + \lambda \nabla \Phi^k, \lambda \right) \\
c^{k+1} &= 1/(1+\epsilon\lambda) shrink\left(c^k + \lambda \Phi^k, \frac{\lambda}{\alpha} \right) \\
\Phi^{k+1} &= \Phi^k + \tau \left(\nabla \cdot (2\mathbf{m}^{k+1} - \mathbf{m}^k) - (2c^{k+1} - c^k) + \mu_1 - \mu_0 \right)
\end{aligned}$$

where the *shrink* operator is defined as following:

$$(shrink(u,t))_i = \begin{cases} (1 - t/|u_i|)u_i, & \text{for } \|u_i\|_1 \geq t; \\ 0, & \text{for } \|u_i\|_1 < t. \end{cases} \qquad i = 1, \cdots, d.$$

Remark 2.3.6. This algorithm can also be extended to $\|\cdot\|_2$ by simply replacing the above

shrink operator as

$$shrink(u,t) = \begin{cases} (1 - t/\|u\|_2)u, & \text{for } \|u\|_2 \geq t; \\ 0, & \text{for } \|u\|_2 < t. \end{cases}$$

2.3.2.1 Discretization

Consider the following two dimensional discretization on a domain $\Omega = [0,1] \times [0,1]$ based on the finite volume method.

$$\begin{aligned}
&\Delta x = \frac{1}{N_x}, \Delta y = \frac{1}{N_y} \\
&\mu^0_{ij} = \mu_0(i\Delta x, j\Delta y), \qquad \mu^1_{ij} = \mu_1(i\Delta x, j\Delta y) \\
&V = \{(i,j) : i = 0, \cdots, N_x, j = 0, \cdots, N_y\} \\
&E_x = \{(i \pm \frac{1}{2}, j) : i = 1, \cdots, N_x - 1, j = 0, \cdots, N_y)\} \\
&E_y = \{(i, j \pm \frac{1}{2}) : i = 0, \cdots, N_x, j = 1, \cdots, N_y - 1)\} \\
&\boldsymbol{\Phi} = (\Phi_{ij})_{ij \in V} \in \mathbb{R}^{(N_x+1)\times(N_y+1)}, \qquad \boldsymbol{c} = (c_{ij})_{ij \in V} \in \mathbb{R}^{(N_x+1)\times(N_y+1)} \\
&\boldsymbol{mx} = (mx_e)_{e \in E_x} \in \mathbb{R}^{N_x \times (N_y+1)}, \qquad \boldsymbol{my} = (my_e)_{e \in E_y} \in \mathbb{R}^{(N_x+1)\times N_y} \\
&mx_{i+\frac{1}{2},j} \approx \int_{i\Delta x}^{(i+1)\Delta x} \int_{(j-1/2)\Delta y}^{(j+1/2)\Delta y} m_x(x,y) dy dx \\
&my_{i,j+\frac{1}{2}} \approx \int_{(i-1/2)\Delta x}^{(i+1/2)\Delta x} \int_{j\Delta y}^{(j+1)\Delta y} m_y(x,y) dy dx.
\end{aligned}$$

Here m satisfies the zero flux condition. Thus, $\boldsymbol{mx}$ and $\boldsymbol{my}$ satisfy the following boundary conditions on m:

$$\begin{aligned}
&mx_{-\frac{1}{2},j} = mx_{N_x+\frac{1}{2},j} = 0, \qquad j = 0, \cdots, N_y \\
&my_{i,-\frac{1}{2}} = my_{i,N_y+\frac{1}{2}} = 0, \qquad i = 0, \cdots, N_x.
\end{aligned}$$

The discretization of Algorithm 2 can be written as the following:

$$
\begin{aligned}
mx_{i+\frac{1}{2},j}^{k+\frac{1}{2}} =& \frac{1}{1+\epsilon\lambda}\left(mx_{i+\frac{1}{2},j}^{k} + \frac{\lambda}{\Delta x}(\Phi_{i+1,j} - \Phi_{i,j})\right)\\
my_{i,j+\frac{1}{2}}^{k+\frac{1}{2}} =& \frac{1}{1+\epsilon\lambda}\left(my_{i,j+\frac{1}{2}}^{k} + \frac{\lambda}{\Delta y}(\Phi_{i,j+1} - \Phi_{i,j})\right)\\
c_{ij}^{k+\frac{1}{2}} =& \frac{1}{1+\epsilon\lambda} shrink\left(c^k + \lambda\Phi_{ij}^k, \frac{\lambda}{\alpha}\right)\\
\boldsymbol{mx}^{k+1} =& 2\boldsymbol{mx}^{k+\frac{1}{2}} - \boldsymbol{mx}^{k}\\
\boldsymbol{my}^{k+1} =& 2\boldsymbol{my}^{k+\frac{1}{2}} - \boldsymbol{my}^{k}\\
\boldsymbol{c}^{k+1} =& 2\boldsymbol{c}^{k+\frac{1}{2}} - \boldsymbol{c}^{k}\\
\Phi_{ij}^{k+1} =& \Phi_{ij}^{k} + \tau\left(\frac{1}{\Delta x}(mx_{i+\frac{1}{2},j}^{k+1} - mx_{i-\frac{1}{2},j}^{k+1}) + \frac{1}{\Delta y}(my_{i,j+\frac{1}{2}}^{k+1} - my_{i,j-\frac{1}{2}}^{k+1}) - c_{ij}^{k+1} + \mu_{ij}^{1} - \mu_{ij}^{0}\right).
\end{aligned}
$$

2.4 Numerical experiments

In this section, we show the numerical results with various examples for L^1 and L^2 unnormalized optimal transport with the spatially dependent source function. The computations were conducted on 2019 MacBook Pro with 2.6 GHz 6-Core and 16GB RAM.

2.4.1 Nesterov Accelerated Gradient Descent for UW_2

We present four numerical experiments with different μ_0 and μ_1 using Algorithm 1.

2.4.1.1 Experiment 1

Consider a one dimensional problem on $\Omega = [0, 1]$ with μ_0 and μ_1 in $\mathcal{M}(\Omega)$ as

$$
\begin{aligned}
\mu_0 &= N(x; \frac{1}{5}, 0.0001)\\
\mu_1 &= N(x; \frac{4}{5}, 0.0001)\cdot 1.4
\end{aligned}
$$

Here we choose $N(x, \mu, \sigma^2) = C\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$ with an appropriate choice of C satisfying $\int_\Omega N(x; \mu, \sigma^2)dx = 1$. Note that $\int_\Omega \mu_0 dx = 1$ and $\int_\Omega \mu_1 dx = 1.4$. We use the Algorithm 1 to

compute the minimizer $\mu(t,x)$ of $UW_2(\mu_0,\mu_1)$. The parameters chosen for the experiment are

$$N_x = 40, N_t = 30, \tau = 0.1, \text{maximum iterations} = 200,000.$$

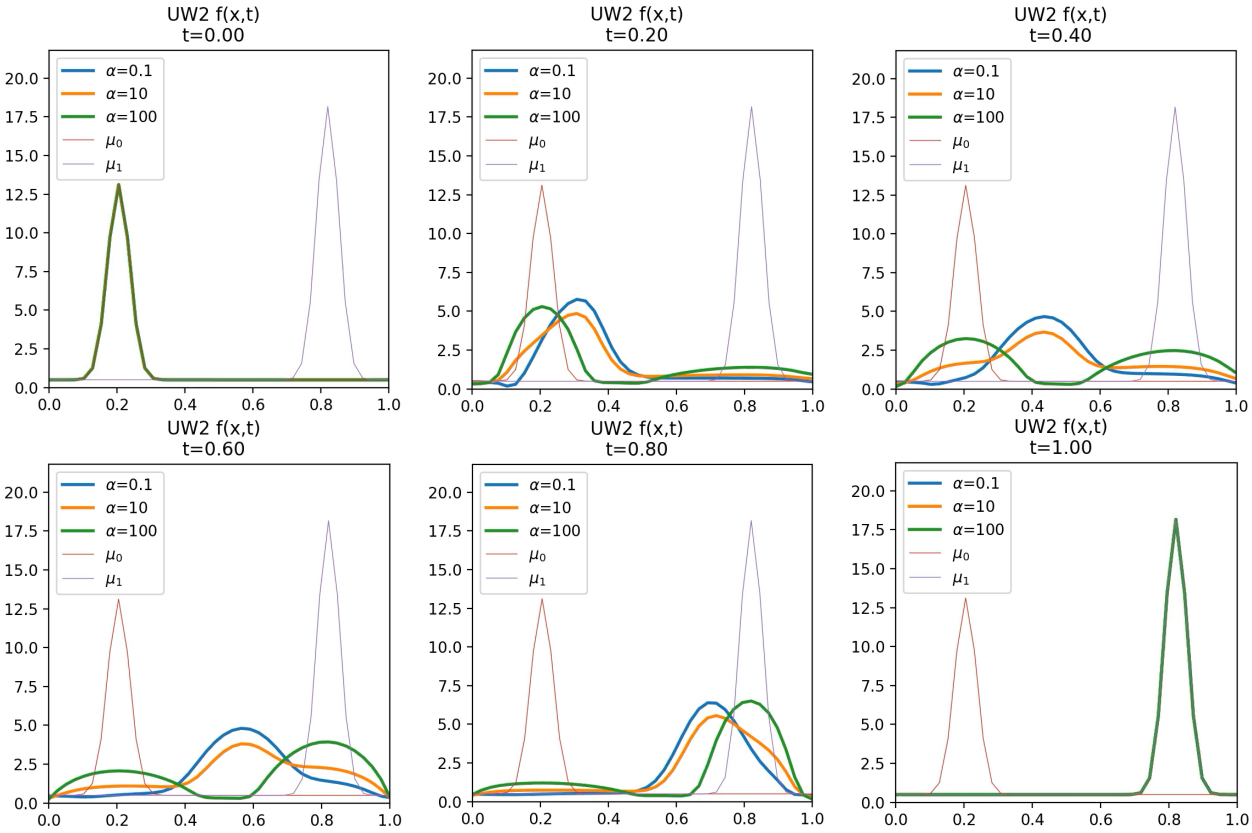


Figure 2.1: *Experiment 1.* L^2 Unnormalized optimal transportation with a spatially dependent source function $f(t,x)$. The figures show the transportation of the densities from $t=0$ (top left) to $t=1$ (bottom right). Blue line shows $\alpha = 0.1$, orange line shows $\alpha = 10$, and green line shows $\alpha = 100$.

Figure 2.1 shows the L^2 unnormalized optimal transport with a spatially dependent source function $f(t,x)$ with different α values. The parameter α determines the ratio between transportation and linear interpolation for μ_0 and μ_1. If α is small, the geodesic of generalized unnormalized optimal transport is similar to the normalized (classical) optimal transport geodesics. As the parameter α increases, the generalized unnormalized optimal transport geodesic behaves closer to the Euclidean geodesics.

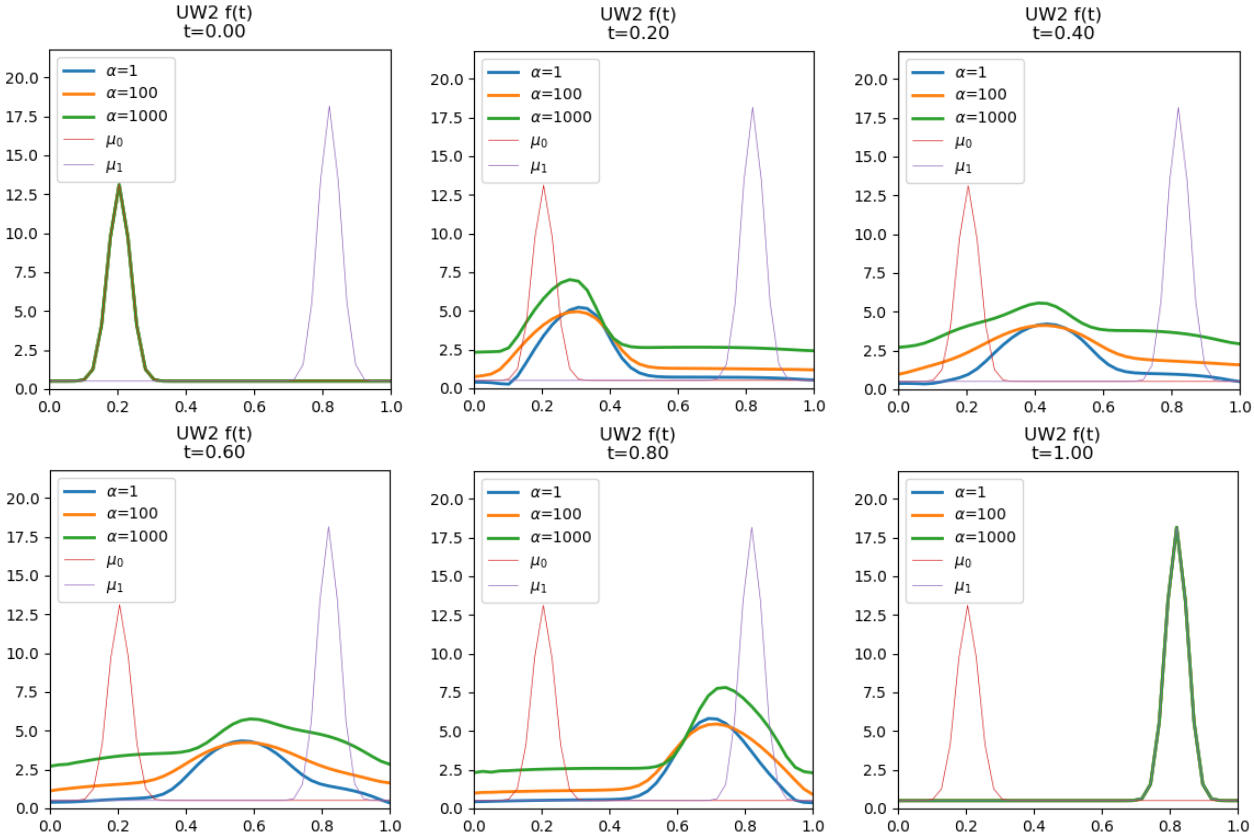
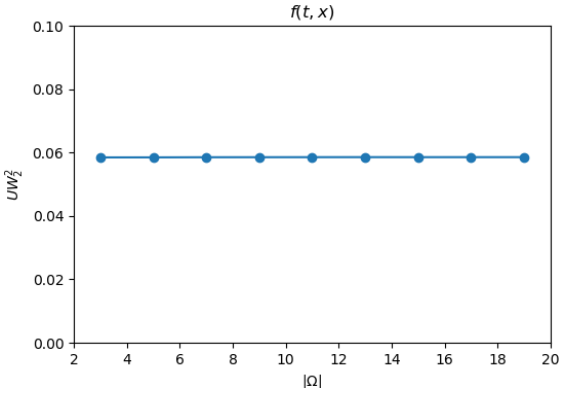


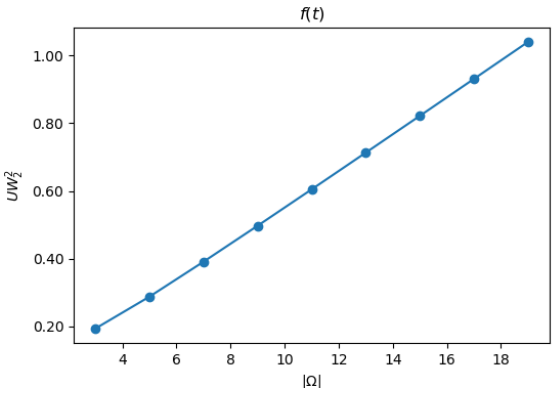
Figure 2.2: *Experiment 1.* L^2 Unnormalized optimal transportation with a spatially independent source function $f(t)$. The figures show the transportation of the densities from $t = 0$ (top left) to $t = 1$ (bottom right). Blue lines show $\alpha = 1$, orange lines show $\alpha = 100$, and green lines show $\alpha = 1000$.

Figure 2.2 shows the transportation with a spatially independent source function $f(t)$. It is clear to see that the masses are created or removed locally for the transportation with $f(t, x)$, while they are created or removed globally for the transportation with $f(t)$.

For each 1-dimensional numerical experiment, the computation took about 5 seconds for $200,000$ iterations.



(a) $|\Omega|$ vs. UW_2^2 with $f(t,x)$



(b) $|\Omega|$ vs. UW_2^2 with $f(t)$

Figure 2.3: *Experiment 2.* The size of the domain $|\Omega|$ vs. L^2 unnormalized Wasserstein metrics for $f(t,x)$ and $f(t)$. x-axis represents $|\Omega|$ and y-axis represents $UW_2(\mu_0,\mu_1)^2$. Both $f(t,x)$ and $f(t)$ use $\alpha = 100$.

2.4.1.2 Experiment 2

In this experiment, we can see how the size of the domain affects the unnormalized Wasserstein distances for both a spatially dependent source function $f(t,x)$ and a spatially independent source function $f(t)$. Consider a one dimensional problem between two densities with different total masses. Figure 2.3 shows plots for the size of the domain $|\Omega|$ vs. the unnormalized Wasserstein distance UW_2. As expected, for the spatially independent source function, the distance increases as $|\Omega|$ increases since the source function affects the transportation globally. Thus, more masses are created or removed as $|\Omega|$ increases. However, the unnormalized Wasserstein distance with the spatially dependent source function does not depend on $|\Omega|$. This actually provides an advantage of using the spatially dependent source function over the spatially independent source function when we need a consistent Wasserstein distance for any size of the domain.

2.4.1.3 Experiment 3

Consider a two dimensional problem with the following input values:

$$\mu_0 = N\left((x,y),(\frac{1}{3},\frac{1}{3}),(\frac{\sqrt{2}}{20},\frac{\sqrt{2}}{20})\right) + N\left((x,y),(\frac{2}{3},\frac{1}{3}),(\frac{\sqrt{2}}{20},\frac{\sqrt{2}}{20})\right)$$
$$\mu_1 = N\left((x,y),(\frac{2}{3},\frac{2}{3}),(\frac{\sqrt{2}}{20},\frac{\sqrt{2}}{20})\right)$$

where $N\left((x,y);(\mu_x,\mu_y),(\sigma_x^2,\sigma_y^2)\right) = C\exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2} - \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)$ and C is a constant such that $\int_\Omega N((x,y);(\mu_x,\mu_y),(\sigma_x^2,\sigma_y^2))dxdy = 1$. Using the Algorithm 1, we calculate the minimizers of $UW_2(\mu_0,\mu_1)$ with a spatially dependent source function $f(t,x)$. The parameters are chosen as

$$N_x = 35, N_y = 35, N_t = 15, \tau = 0.1, \text{maximum iterations} = 6,000.$$

Figure 2.4 shows the transportation with $\alpha = 1$ and $\alpha = 1000$, respectively. The same phenomena can be observed as in 1D case from *Experiment 1.* In other words, the geodesic with the spatially dependent source function with small α in Figure 2.4 behaves closer to the normalized (classical) optimal transport geodesic and the geodesic with large α behaves closer to the Euclidean geodesic. The computation took 402.54 seconds for $\alpha = 1$ and 77.82 seconds for $\alpha = 1000$. Note that when α is small, the condition number of the Laplacian operator gets larger. This results in slower convergence rate of conjugate gradient method for inverting the Laplacian operator.

2.4.1.4 Experiment 4

In this experiment, we are interested in calculating L^2 unnormalized Wasserstein distance between two images. We show two sets of experiments with different initial and terminal densities. First, consider images of two cats with different total masses defined on the domain $\Omega = [0,1]\times[0,1]$. We use Algorithm 1 with the following parameters:

$$N_x = 64, N_y = 64, N_t = 15, \text{maximum iterations} = 4,000.$$

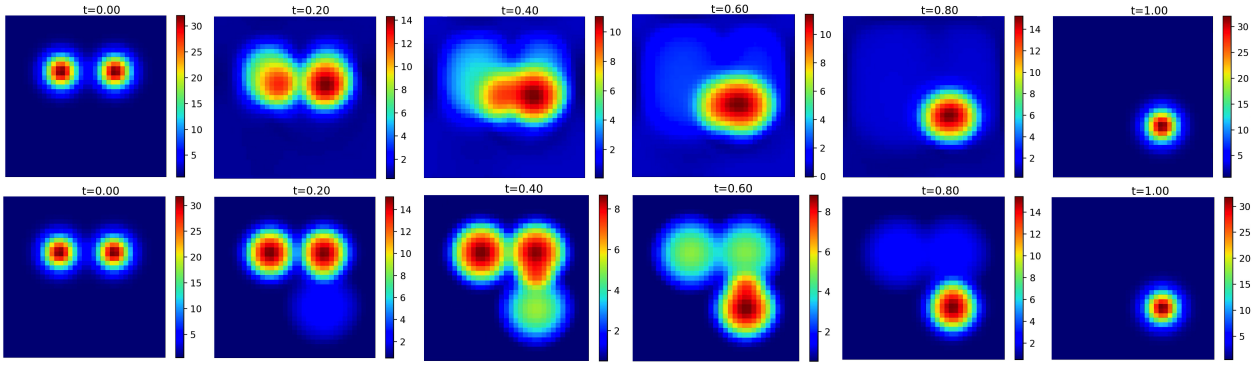


Figure 2.4: *Experiment 3.* L^2 generalized unnormalized optimal transportation: 2D example with a spatially dependent source function $f(t,x)$. The first row is with $\alpha = 1$. The second row is with $\alpha = 1000$.

Figure 2.5 shows transportation between two cats images with $\alpha = 10$ and $\alpha = 1000$, respectively. The computation took 353.85 seconds for $\alpha = 10$ and 93.67 seconds for $\alpha = 1000$.

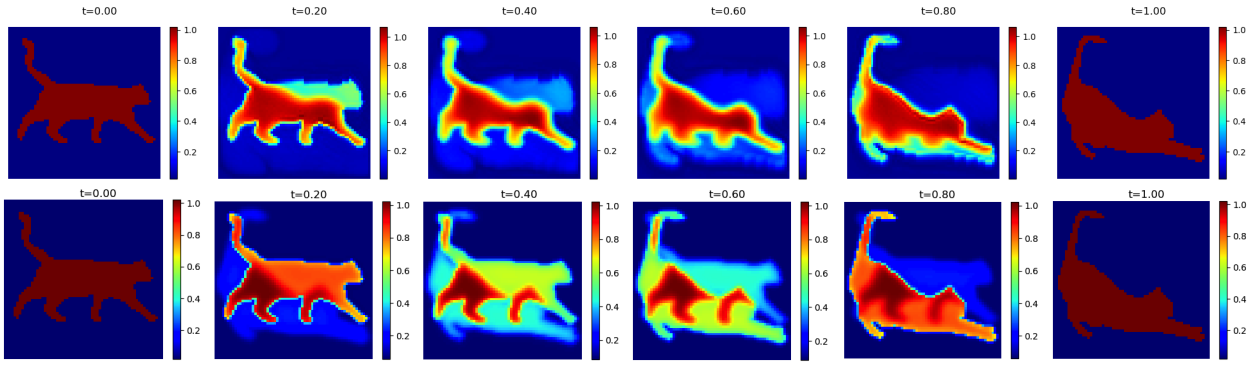


Figure 2.5: *Experiment 4.* L^2 generalized unnormalized optimal transportation between two cats with a spatially dependent source function $f(t,x)$. The first row is with $\alpha = 10$. The second row is with $\alpha = 1000$.

Additionally, we consider images of a pair of scissors and Homer Simpson. We again use Algorithm 1 with the same set of parameters as above. Figure 2.6 shows transportation between two images with $\alpha = 10$ and $\alpha = 1000$, respectively. The computation took 353.00 seconds for $\alpha = 10$ and 93.96 seconds for $\alpha = 1000$.

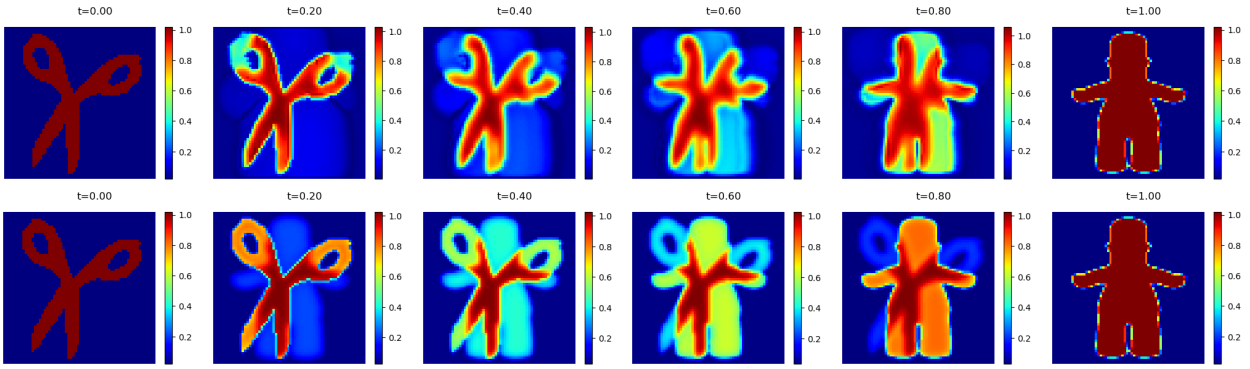


Figure 2.6: *Experiment 4.* L^2 generalized unnormalized optimal transportation between a pair of scissors and Homer Simpson with a spatially dependent source function $f(t,x)$. The first row is with $\alpha = 10$. The second row is with $\alpha = 1000$.

2.4.2 Primal dual algorithm for UW_1

We conduct two numerical examples of L^1 unnormalized optimal transport using Algorithm 2. For UW_1 experiments, we use maximum iterations for the stopping condition.

2.4.2.1 Experiment 5

Assume $\Omega = [0,1] \times [0,1]$. Consider the two dimensional problem with the following initial densities:

$$\mu_0 = N\left((x,y), (\frac{1}{3}, \frac{1}{2}), (\frac{1}{10}, \frac{1}{10})\right)$$
$$\mu_1 = N\left((x,y), (\frac{2}{3}, \frac{1}{2}), (\frac{1}{10}, \frac{1}{10})\right) \cdot 1.4$$

N is the same as the one used in Experiment 3. We chose the parameters as:

$$N_x = N_y = 40, \epsilon = 0.001, \lambda = 0.001, \tau = 0.1, \text{maximum iterations} = 30,000.$$

In Figure 2.7, the initial densities μ_0 and μ_1 are shown on the top two plots and the minimizers $\boldsymbol{m}$'s are plotted for three different α values in the second row. As a comparison, the top right picture in Figure 2.7 shows the result from L^1 transportation with a spatially independent

source function $f(t)$. This experiment shows the clear difference between L^1 unnormalized optimal transport with $f(t,x)$ and with $f(t)$. While the minimizer $\boldsymbol{m}$ from the unnormalized optimal transport with $f(t,x)$ is nonzero only on the area between two densities, the minimizer from the unnormalized optimal transport with $f(t)$ is nonzero everywhere. This is because the spatially dependent source function $f(t,x)$ affects the minimizer locally but the spatially independent source function $f(t)$ affects the minimizer globally. The computation took 2.02 seconds for $\alpha = 1$, 2.74 seconds for $\alpha = 10$, and 2.67 seconds for $\alpha = 100$.

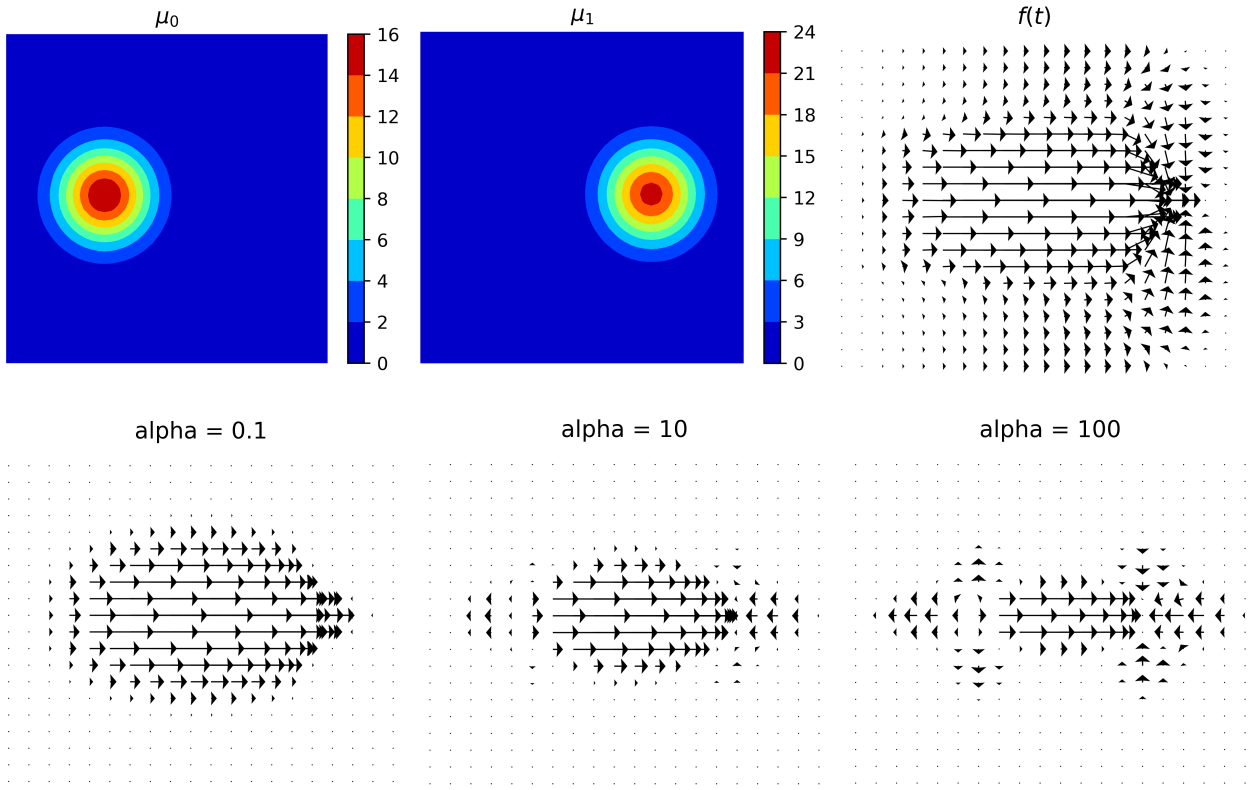


Figure 2.7: *Experiment 5.* Top left: initial density μ_0. Top middle: the terminal density μ_1. Top right: the solution $\mathbf{m}$ of L^1 unnormalized optimal transportation with a spatially independent source $f(t)$. The bottom images show the solution of L^1 unnormalized optimal transportation with $f(t,x)$ using different α values. Bottom left: $\alpha = 0.1$, bottom middle: $\alpha = 10$, bottom right: $\alpha = 100$.

2.4.2.2 Experiment 6

In this experiment, we are interested in UW_1 distance between two images. Consider the same 2D example as in the *Experiment 4*. We use the Algorithm 2 with the following parameters:

$$N_x = N_y = 256, \epsilon = 0.001, \lambda = 0.0001, \tau = 0.01, \text{maximum iterations} = 40,000.$$

Figure 2.8 plots the results of L^1 unnormalized optimal transport with a spatially dependent source function $f(t,x)$ with different α values 0.1, 5, and 10 in the second row. As a comparison, the top right picture in Figure 2.8 shows the result from L^1 transportation with a spatially independent source function $f(t)$. The result is similar to the *Experiment 5*. The minimizer $\boldsymbol{m}$ from L^1 unnormalized optimal transport with $f(t)$ has nonzero values on the surrounding area of the two densities, but the minimizers from unnormalized optimal transport with $f(t,x)$ are zero on that surrounding area. The computation took 153.07 seconds for $\alpha = 1$, 192.82 seconds for $\alpha = 10$, and 186.27 seconds for $\alpha = 100$.

2.4.2.3 Experiment 7

In this last experiment, we demonstrate the spatial convergence of Algorithm 2. Assume $\Omega = [0,1] \times [0,1]$. Consider an initial density to be a circle of radius 0.05 at $(0.2, 0.2)$ with a mass 1 and a terminal density to be a circle of radius 0.05 at $(0.8, 0.8)$ with a mass 1 (Figure 2.9). UW_1 distance between these two densities is 0.8 which equals L^1 distance. We use the following parameters:

$$\alpha = 0.001, \epsilon = 0.01, \lambda = 0.001, \tau = 0.01, \text{maximum iterations} = 50,000.$$

We repeat the experiment with 4 different space discretizations ($N_x = N_y = 16, 32, 64, 128$). The table 2.1 summarizes the result of the experiment which shows the algorithm is accurate to order 1 in space.

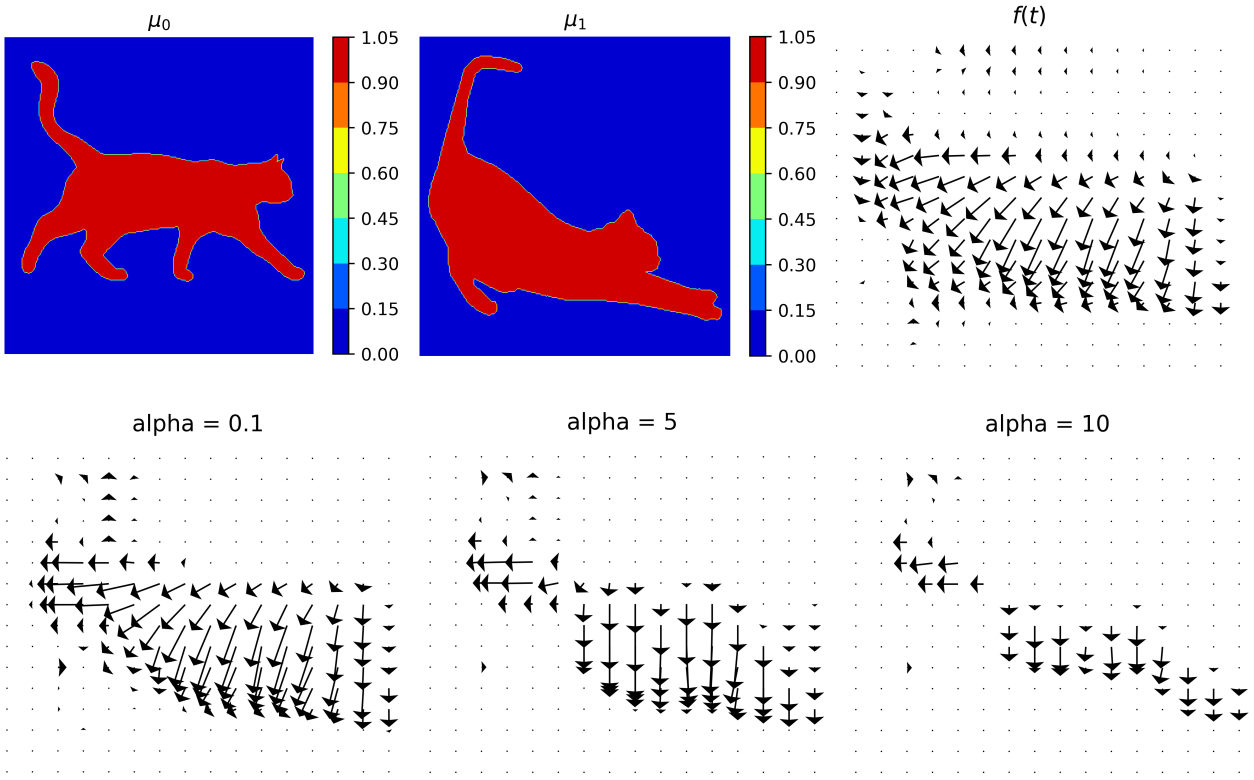


Figure 2.8: *Experiment 6.* Top left: initial density μ_0. Top middle: the terminal density μ_1. Top right: the solution $\mathbf{m}$ of L^1 unnormalized optimal transportation with a spatially independent source $f(t)$. The bottom images show the solution of L^1 unnormalized optimal transportation with $f(t,x)$ using different α values. Bottom left: $\alpha = 0.1$, bottom middle: $\alpha = 5$, bottom right: $\alpha = 10$.

2.5 Discussion

In this paper, we introduced a new class of L^p generalized unnormalized optimal transport distance with a spatially dependent source function. We presented new fast algorithms for L^1 and L^2 generalized unnormalized optimal transport. For L^1 case, we derived the Kantorovich duality and used a primal-dual algorithm which has explicit formulas with low computational costs. For L^2 case, we derived the duality formula, the generalized unnormalized Monge problem and corresponding Monge-Ampère equation. We applied a weighted Laplacian

Table 2.1: The summary of the results of Experiment 7

N_x	N_y	Time (s)	Error
16	16	0.62	5.4×10^{-2}
32	32	2.39	3.8×10^{-2}
64	64	9.87	1.4×10^{-2}
128	128	44.14	9.4×10^{-3}

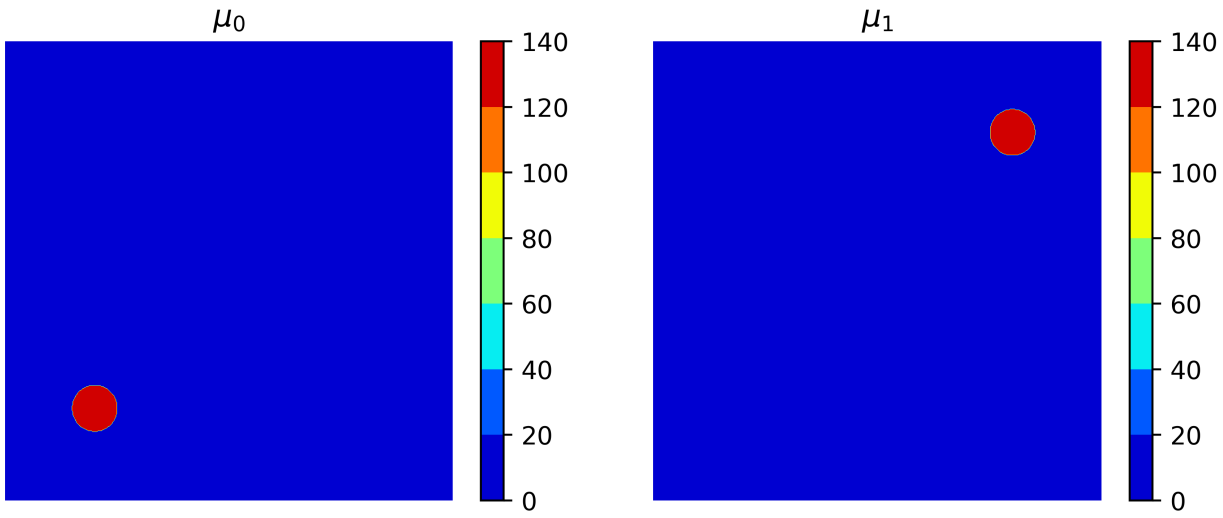


Figure 2.9: Initial and terminal densities for Experiment 7.

operator L_μ to formulate the problem into an unconstrained optimization. The gradient operator of this unconstrained optimization is precisely the Hamilton-Jacobi equation. We apply the Nesterov accelerated gradient descent method to solve this minimization problem.

Our algorithm can be applied to general unnormalized/unbalanced optimal transport problems. It is also suitable for considering general variational mean-field games. In future works, we will derive new formulations for all related L^p unbalanced or unnormalized mean-field games and design fast numerical algorithms to solve them.

2.6 Acknowledgments and Disclosure of Funding

W. Lee, W. Li and S. Osher's research are supported in part by AFOSR MURI FA9550-18-1-0502. R. Lai's reserach is supported in part by an NSF CAREER Award DMS–1752934.

CHAPTER 3

Controlling Propagation of Epidemics via Mean-Field Controls

The coronavirus disease 2019 (COVID-19) pandemic is changing and impacting lives on a global scale. In this paper, we introduce a mean-field control model in controlling the propagation of epidemics on a spatial domain. The control variable, the spatial velocity, is first introduced for the classical disease models, such as the SIR model. For this proposed model, we provide fast numerical algorithms based on proximal primal-dual methods. Numerical experiments demonstrate that the proposed model illustrates how to separate infected patients in a spatial domain effectively.

3.1 Introduction

The outbreak of the COVID-19 epidemic has resulted in millions of confirmed cases and hundreds of thousands of deaths globally. It has a huge impact on the global economy as well as everyone's daily life. There has been a lot of interest in modeling the dynamics and propagation of the epidemic. One of the well-known and basic models in epidemiology is the SIR model proposed by Kermack and McKendrick [KM27] in 1927. Here, S, I, R represent the number of susceptible, infected, and recovered people respectively. They use an ODE system to describe the transmission dynamics of infectious diseases among the population. As the propagation of COVID-19 has a significant spatial characteristic actions such as travel restrictions, physical distancing and self-quarantine are taken to slow down the spread of

the epidemic. It is important to have a spatial-type SIR model to study the spread of the infectious disease and movement of individuals [Ken65, Kal84, HI95].

Since the epidemic has affected society and individuals significantly, mean-field games and mean-field controls (MFG, MFC) provide a perspective to study and understand the underlying population dynamics. Mean-field games were introduced by Jovanovic and Rosenthal [JR88], Huang, Malhamé, and Caines [HMC06], and Lasry and Lions [LL06a, LL06b]. They model a huge population of agents playing dynamic games. There is growing research interest in this direction. For a review of MFG theory, we refer to [LL07, Gom14]. With wide application to various fields [GNP15, BDM13, LLL16, AL19], computational methods are also designed to solve related high dimensional MFG problems [BC15, AKS18, EHL18, LFL20, ROL19, LJL20].

In this paper, we combine the above ideas of the spatial SIR model and MFG. In other words, we introduce a mean-field game (control) model for controlling the virus spreading within a spatial domain. Here the goal is to minimize the number of infectious agents and the amount of movement of the population. In short, we formalize the following constrained optimization problem

$$\inf_{(\rho_i,v_i)_{i\in\{S,I,R\}}} \quad E(\rho_I(T,\cdot)) + \int_0^T \int_\Omega \sum_{i\in\{S,I,R\}} \frac{\alpha_i}{2}\rho_i|v_i|^2 + \frac{c}{2}(\rho_S+\rho_I+\rho_R)^2 dxdt$$

subject to

$$\begin{cases} \partial_t\rho_S + \nabla\cdot(\rho_S v_S) + \beta\rho_S\rho_I - \dfrac{\eta_S^2}{2}\Delta\rho_S = 0 \\ \partial_t\rho_I + \nabla\cdot(\rho_I v_I) - \beta\rho_S\rho_I + \gamma\rho_I - \dfrac{\eta_I^2}{2}\Delta\rho_I = 0 \\ \partial_t\rho_R + \nabla\cdot(\rho_R v_R) - \gamma\rho_I - \dfrac{\eta_R^2}{2}\Delta\rho_R = 0 \\ \rho_S(0,\cdot), \rho_I(0,\cdot), \rho_R(0,\cdot) \text{ are given.} \end{cases}$$

Here ρ_i represents the population density and v_i describes the movement, with $i \in \{S, I, R\}$ corresponding to the susceptible, infected and recovered compartmental state or class. We consider the spatial SIR model with nonlocal spreading modeled by an integration kernel K representing the physical distancing and spatial diffusion of population and set it as dynamic

to our mean-field control problem. This is the constraint to the minimization problem. The minimization objective include both the movement and the congestion of the population. The kinetic energy terms describe the situation that, if the population (the susceptible, infected or recovered) needs to be moved to alleviate the local medical shortage. The congestion term models the fact that the government doesn't want the population to get too concentrated in one place. This might increase the risk of disease outbreaks and their faster and wider spread. Due to the multiplicative nature of the interaction term between susceptible and infectious agents $\beta\rho_S\rho_I$, the mean-field control problem can be a non-convex optimization problem. By using Lagrange multipliers, we formalize the mean-field control problem as an unconstrained optimization problem. Fast numerical algorithms are designed to solve the non-convex optimization problem in $2D$ with the $G - prox$ preconditioning method [JLL19].

In the literature, spatial SIR models in the form of a nonlinear integrodifferential [Aro77, Die79, Thi77] and reaction-diffusion system [Kal84, HI95] have been studied. Traveling waves are studied to understand the propagation of various types of epidemics, such as Lyme disease, measles, etc, and recently, COVID-19 [CGC02, GBK01, WW10, BRR20]. In [BRR20], they introduce a SIRT model to study the effects of the presence of a road on the spatial propagation of the epidemic. For surveys, see [Mur01, Rua07]. As for numerical modeling of epidemic model concerning the spatial effect, finite-difference methods are used to discretize the reaction-diffusion system and solve the spatial SIR model, and its various extensions [CC10, JC14, FH16]. Epidemic models have been treated using optimal control theory, with major control measures on medicare (vaccination) [SS78, LES18, JKL20]. In [JKL20], a feedback control problem of the SIR model is studied to help determine the vaccine policy to minimize the number of infected people. In [LZM19], they introduce a nonlinear SIQS epidemic model on complex networks and study the optimal quarantine control. Compared to previous works, our model is the first to consider an optimal control problem for the SIR model on a spatial domain, combining optimal transport and mean field controls. As SIR model can be interpreted in terms of stochastic processes of agent-based models, it can be

obtained as a motion of the law of a three-state Markov chain with the transition from S to I and I to R [All17]. Here, we formulate velocity fields among S, I, R populations as control variables. And our model applies a pair of PDEs, consisting Fokker-Planck equation and Hamilton-Jacobi equation. These equations describe how different populations (susceptible, infected or recovered) react to the propagation of pandemic on a spatial domain.

Our paper is organized as follows. In section 3.2, we introduce the mean field control model for the propagation of epidemics. We introduce a primal-dual hybrid gradient algorithm for this model in section 3.3. In section 3.4, several numerical examples are demonstrated.

3.2 Model

In this section, we briefly review the classical epidemics models, e.g., SIR dynamics. We then introduce a mean field control model for SIR dynamics on a spatial domain. We derive a system to find the minimizer of the proposed model.

3.2.1 Review

We first review the classical SIR model.

$$\begin{cases} \dfrac{dS(t)}{dt} = -\beta S(t)I(t) \\ \dfrac{dI(t)}{dt} = \beta S(t)I(t) - \gamma I(t) \\ \dfrac{dR(t)}{dt} = \gamma I(t) \end{cases}$$

where $S, I, R : [0, T] \to [0, 1]$ represent the proportion of the susceptible population, infected population, and recovered population, respectively, given time $t \in [0, T]$. Susceptible people become infected with a rate of β, and infected people are recovered with a rate of γ. The SIR model can be derived based on the mean-field assumptions. Thus it can be interpreted as the mean field equations for a three-state Markov chain on S, I, R states.

3.2.2 Spatial SIR variational problem

We consider the spatial dimension of the S, I, R functions. Let $\Omega \subset \mathbb{R}^d$ be a bounded domain. Consider the following functions

$$\rho_S, \rho_I, \rho_R : \ [0,T] \times \Omega \to [0, \infty).$$

Here, ρ_S, ρ_I, and ρ_R represent susceptible, infected and recovered populations, respectively. We assume ρ_i for each $i \in \{S, I, R\}$ moves on a spatial domain Ω with velocities v_i. We can describe these movements by continuity equations.

$$\begin{cases} \partial_t \rho_S + \nabla \cdot (\rho_S v_S) + \beta \rho_S \rho_I - \dfrac{\eta_S^2}{2} \Delta \rho_S = 0 \\ \partial_t \rho_I + \nabla \cdot (\rho_I v_I) - \beta \rho_S \rho_I + \gamma \rho_I - \dfrac{\eta_I^2}{2} \Delta \rho_I = 0 \\ \partial_t \rho_R + \nabla \cdot (\rho_R v_R) - \gamma \rho_I - \dfrac{\eta_R^2}{2} \Delta \rho_R = 0 \\ \rho_S(0, \cdot), \rho_I(0, \cdot), \rho_R(0, \cdot) \text{ are given.} \end{cases} \tag{3.1}$$

Here $v_i : [0,T] \times \Omega \to \mathbb{R}^d$ ($i \in \{S, I, R\}$) are vector fields that represent the velocity fields for ρ_i ($i \in \{S, I, R\}$) and nonnegative constants η_i ($i \in \{S, I, R\}$) are coefficients for viscosity terms. We add these viscosity terms to regularize the systems of continuity equations, thus stabilize our numerical method that will be discussed in later sections. In addition, we assume zero flux conditions by the Neumann boundary conditions, that is no mass can flow in or out of Ω. These systems of continuity equations satisfy the following equality:

$$\frac{\partial}{\partial t} \int_\Omega \rho_S(t,x) + \rho_I(t,x) + \rho_R(t,x) dx = 0.$$

This means that the total mass of the three populations will be conserved for all time.

Lastly, we introduce the proposed mean field control models. Consider the following variational problem:

$$\inf_{(\rho_i, v_i)_{i \in \{S,I,R\}}} \quad E(\rho_I(T, \cdot)) + \int_0^T \int_\Omega \sum_{i \in \{S,I,R\}} \frac{\alpha_i}{2} \rho_i |v_i|^2 + \frac{c}{2} (\rho_S + \rho_I + \rho_R)^2 dx dt \tag{3.2}$$

subject to (3.1) with fixed initial densities.

Here E is a convex functional and α_i ($i \in \{S, I, R\}$) and c are nonnegative constants. The formulation is mainly divided into two parts: a terminal cost and a running cost. The functional E is a terminal cost which increases if there is greater mass of the infected population at the terminal time. For example, we choose $E(\rho(T,\cdot)) = \frac{1}{2}\int_\Omega \rho^2(T,x)dx$ for the experiments (Section 3.4). The rest of the terms besides the functional E are running costs. Kinetic energy terms $\frac{\alpha_i}{2}\rho_i|v_i|^2$ ($i \in \{S, I, R\}$) represent the cost of moving the density ρ_i with velocities v_i over time $0 \le t \le T$. A high value of α_i means that it is expensive to move ρ_i for corresponding $i \in \{S, I, R\}$. In the numerical experiments (Section 3.4), we assume $\alpha_S = \alpha_R = 1$ and $\alpha_I = 10$ to simulate the real life scenario where infected group is harder to move than other groups. The last term in the running cost, $\frac{c}{2}(\rho_S + \rho_I + \rho_R)^2$, penalizes the congestion of the total population. A high value of c means more penalization on the congestion. The minimizers of the variational problem will provide the optimal movements for each population while minimizing the terminal cost functional with respect to the infected population ρ_I.

We note that the function $(\rho_i, v_i) \mapsto \rho_i|v_i|^2$ is not convex. By introducing new variables $m_i := \rho_i v_i$, we convert the cost functional to be convex in term of (ρ_i, m_i). In other words,

$$\min_{\rho_i, v_i} \quad P(\rho_i, m_i)_{i\in\{S,I,R\}} \tag{3.3a}$$

subject to

$$\begin{cases} \partial_t\rho_S + \nabla\cdot m_S + \beta\rho_S\rho_I - \dfrac{\eta_S^2}{2}\Delta\rho_S = 0 \\ \partial_t\rho_I + \nabla\cdot m_I - \beta\rho_S\rho_I + \gamma\rho_I - \dfrac{\eta_I^2}{2}\Delta\rho_I = 0 \\ \partial_t\rho_R + \nabla\cdot m_R - \gamma\rho_I - \dfrac{\eta_R^2}{2}\Delta\rho_R = 0 \\ \rho_S(0,\cdot), \rho_I(0,\cdot), \rho_R(0,\cdot) \text{ are given} \end{cases} \tag{3.3b}$$

where

$$P(\rho_i, m_i)_{i\in\{S,I,R\}} = E(\rho_I(T,\cdot)) + \int_0^T \int_\Omega \sum_{i\in\{S,I,R\}} \frac{\alpha_i|m_i|^2}{2\rho_i} + \frac{c}{2}(\rho_S + \rho_I + \rho_R)^2 dxdt.$$

From an optimization viewpoint, we note that the minimization problem is not a convex problem since the coupling terms, $\beta\rho_S\rho_I$, in constraints make the feasible set nonconvex. We replace the nonconvex coupling term $\beta\rho_S\rho_I$ with convolution. Note that Kendall [Ken65] introduced this kernel for modeling pandemic dynamics and took the nonlocal exposure to infectious agents into consideration. This term also helps regularize the minimization problem.

$$\min_{(\rho_i,v_i)_{i\in\{S,I,R\}}} \quad P(\rho_i, m_i)_{i\in\{S,I,R\}} \tag{3.4a}$$

subject to

$$\begin{cases} \partial_t\rho_S(t,x) + \nabla\cdot m_S(t,x) + \beta\rho_S(t,x)\displaystyle\int_\Omega K(x,y)\rho_I(t,y)dy - \frac{\eta_S^2}{2}\Delta\rho_S(t,x) = 0 \\ \partial_t\rho_I(t,x) + \nabla\cdot m_I(t,x) - \beta\rho_I(t,x)\displaystyle\int_\Omega K(x,y)\rho_S(t,y)dy + \gamma\rho_I(t,x) - \frac{\eta_I^2}{2}\Delta\rho_I(t,x) = 0 \\ \partial_t\rho_R(t,x) + \nabla\cdot m_R(t,x) - \gamma\rho_I(t,x) - \frac{\eta_R^2}{2}\Delta\rho_R(t,x) = 0 \\ \rho_S(0,\cdot), \rho_I(0,\cdot), \rho_R(0,\cdot) \text{ given.} \end{cases} \tag{3.4b}$$

Here, $K(x,y)$ is a symmetric positive definite kernel. In this paper, we focus on a Gaussian kernel

$$K(x,y) = \frac{1}{\sqrt{(2\pi)^d}}\prod_{k=1}^{d}\frac{1}{\sigma_k}\exp\left(-\frac{|x_k-y_k|^2}{2\sigma_k^2}\right).$$

The variance σ_k of Gaussian kernel can be viewed as a parameter for modeling the spatial spreading effect of the virus. Let's consider the convolution term in the first continuity equation, $\rho_S(t,x)\int_\Omega K(x,y)\rho_I(t,y)dy$. Larger values of variance σ_k's in K mean that a susceptible agent located at position x can be affected by infectious agents farther away from x. Note that by letting $\sigma_k \to 0$, we get

$$\rho_S(t,x)\int_\Omega K(x,y)\rho_I(t,y)dy \to \rho_S(t,x)\rho_I(t,x).$$

Thus, when σ_k becomes close to 0, the susceptible agent is only affected by infectious agents nearby. If we let $\sigma_k \to \infty$, then

$$\rho_S(t,x)\int_\Omega K(x,y)\rho_I(t,y)dy \to \rho_S(t,x)\int_\Omega \rho_I(t,y)dy,$$

which means the susceptible group is affected by the total number of the infected population.

Remark 3.2.1. The formulation is not limited to the SIR model we chose in this paper. It can be used to solve any types of spatial epidemiological models. For example, if we use the SEIR model where E stands for the exposed group, we just add one additional variable ρ_E and add one more continuity equation.

3.2.3 Properties

We next derive the mean field control system, i.e. the minimizer system associated with spatial SIR variational problem (3.4). We introduce three dual variables ϕ_i ($i \in \{S, I, R\}$) to convert the minimization problem (3.4) into a saddle problem.

$$\inf_{(\rho_i, v_i)_{i\in\{S,I,R\}}} \left\{ P(\rho_i, m_i)_{i\in\{S,I,R\}} : \text{subject to (3.4b)} \right\}$$

$$\begin{aligned}
= \inf_{(\rho_i, v_i)_{i\in\{S,I,R\}}} \sup_{(\phi_i)_{i\in\{S,I,R\}}} & P(\rho_i, m_i)_{i\in\{S,I,R\}} \\
& - \int_0^T \int_\Omega \phi_S \left(\partial_t \rho_S + \nabla \cdot m_S + \beta \rho_S K * \rho_I - \frac{\eta_S^2}{2} \Delta \rho_S \right) dxdt \\
& - \int_0^T \int_\Omega \phi_I \left(\partial_t \rho_I + \nabla \cdot m_I - \beta \rho_S K * \rho_I + \gamma \rho_I - \frac{\eta_I^2}{2} \Delta \rho_I \right) dxdt \\
& - \int_0^T \int_\Omega \phi_R \left(\partial_t \rho_R + \nabla \cdot m_R - \gamma \rho_I - \frac{\eta_R^2}{2} \Delta \rho_R \right) dxdt.
\end{aligned}$$

Simplifying the above function, we define the Lagrangian functional

$$\begin{aligned}
& \mathcal{L}((\rho_i, m_i, \phi_i)_{i\in\{S,I,R\}}) \\
=& P(\rho_i, m_i)_{i\in\{S,I,R\}} - \int_0^T \int_\Omega \sum_{i\in\{S,I,R\}} \phi_i \left(\partial_t \rho_i + \nabla \cdot m_i - \frac{\eta_i^2}{2} \Delta \rho_i \right) dxdt \\
& + \int_0^T \int_\Omega \beta \rho_S (\phi_I - \phi_S) K * \rho_I + \gamma \rho_I (\phi_R - \phi_I) dxdt.
\end{aligned} \tag{3.5}$$

Thus, we have the following saddle problem:

$$\inf_{(\rho_i, m_i)_{i\in\{S,I,R\}}} \sup_{(\phi_i)_{i\in\{S,I,R\}}} \mathcal{L}((\rho_i, m_i, \phi_i)_{i\in\{S,I,R\}}). \tag{3.6}$$

The existence of the saddle point of this mini-max problem is based on the assumption that the dual gap is zero. In other words, given a primal solution with respect to optimal primal variables $(\rho_i, m_i)_{i\in\{S,I,R\}}$ and a dual solution with respect to optimal dual variables $(\phi_i)_{i\in\{S,I,R\}}$, the difference between these two solutions is zero. However, the dual gap may not be zero for this problem because the nonconvex functional $(\rho_S, \rho_I) \mapsto \rho_S K * \rho_I$ makes the feasible set of the problem nonconvex. Throughout the paper, we assume the dual gap is zero.

The following propositions are the properties of the saddle point problem derived from optimality conditions, known as Karush–Kuhn–Tucker (KKT) conditions.

Proposition 3.2.1 (Mean-field control SIR system)**.** *By KKT conditions, the saddle point* (ρ_i, m_i, ϕ_i) *of* (3.6) *satisfies the following equations.*

$$\begin{cases} \partial_t\phi_S - \dfrac{\alpha_S}{2}|\nabla\phi_S|^2 + \dfrac{\eta_S^2}{2}\Delta\phi_S + c(\rho_S+\rho_I+\rho_R) + \beta(\phi_I-\phi_S)K*\rho_I = 0 \\ \partial_t\phi_I - \dfrac{\alpha_I}{2}|\nabla\phi_I|^2 + \dfrac{\eta_I^2}{2}\Delta\phi_I + c(\rho_S+\rho_I+\rho_R) \\ \qquad\qquad + \beta K*(\rho_S(\phi_I-\phi_S)) + \gamma(\phi_R-\phi_I) = 0 \\ \partial_t\phi_R - \dfrac{\alpha_R}{2}|\nabla\phi_R|^2 + \dfrac{\eta_R^2}{2}\Delta\phi_R + c(\rho_S+\rho_I+\rho_R) = 0 \\ \partial_t\rho_S - \dfrac{1}{\alpha_S}\nabla\cdot(\rho_S\nabla\phi_S) + \beta\rho_S K*\rho_I - \dfrac{\eta_S^2}{2}\Delta\rho_S = 0 \\ \partial_t\rho_I - \dfrac{1}{\alpha_I}\nabla\cdot(\rho_I\nabla\phi_I) - \beta\rho_S K*\rho_I + \gamma\rho_I - \dfrac{\eta_I^2}{2}\Delta\rho_I = 0 \\ \partial_t\rho_R - \dfrac{1}{\alpha_R}\nabla\cdot(\rho_R\nabla\phi_R) - \gamma\rho_I - \dfrac{\eta_R^2}{2}\Delta\rho_R = 0 \\ \phi_I(T,\cdot) = \delta E(\rho_I(T,\cdot)). \end{cases} \tag{3.7}$$

The term δE *is the functional derivative. Given a smooth functional* $F : \mathcal{H} \to \mathbb{R}$ *where* $\mathcal{H}$ *is a separable Hilbert space and* $\rho \in \mathcal{H}$*, we say a map* $\frac{\delta F}{\delta \rho}$ *is the functional derivative of* F *with respect to* ρ *if it satisfies*

$$\lim_{\epsilon\to 0}\frac{F(\rho+\epsilon h)-F(\rho)}{\epsilon} = \int_\Omega \frac{\delta F}{\delta\rho}(\rho(x))h(x)\,dx,$$

for any arbitrary function $h \in \mathcal{H}$*.*

The proof of Poposition 3.2.1 can be found in the appendix. We note that dynamical system (3.7) models the optimal vector field strategies for S, I, R populations. It combines both strategies from mean field controls and SIR models. For this reason, we call (3.7) *Mean-field control SIR system.*

3.3 Algorithm

In this section, we propose an algorithm to solve the proposed SIRV variational problem. We use the primal-dual hybrid gradient (PDHG) algorithm [CP11b, CP16]. The PDHG can solve the following convex optimization problem

$$\min_u \ f(Au) + g(u)$$

where f and g are convex functions and A is a continuous linear operator. Since f is a convex function, by the convex duality relation, we have $f^{**} = f$ where f^* is the Legendre transform such that

$$f^*(p) = \sup_u \langle u, p\rangle_{L^2} - f(u)$$

where $\langle \cdot, \cdot \rangle_{L^2}$ is L^2 inner product. Thus, f can be represented as a Legendre transform of f^*, i.e.,

$$f(Au) = f^{**}(Au) = \sup_p \langle Au, p\rangle_{L^2} - f^*(p).$$

The algorithm solves the problem by converting it into a saddle point problem by introducing a dual variable p and using the convex duality relation

$$\min_u \max_p \ g(u) + \langle Au, p\rangle_{L^2} - f^*(p).$$

The method solves for the saddle point (u_*, p_*) by iterating

$$\begin{aligned} p^{(k+1)} &= \operatorname*{argmax}_p \langle Au^{(k)}, p\rangle_{L^2} - f^*(p) - \frac{1}{2\sigma}\|p - p^{(k)}\|_{L^2}^2 \\ u^{(k+1)} &= \operatorname*{argmin}_u \ g(u) + \langle u, A^T(2p^{(k+1)} - p^{(k)})\rangle_{L^2} + \frac{1}{2\tau}\|u - u^{(k)}\|_{L^2}^2. \end{aligned} \tag{3.8}$$

The scheme converges if the step sizes τ and σ satisfy

$$\tau\sigma\|A^TA\|_{L^2} < 1, \tag{3.9}$$

where $\|\cdot\|_{L^2}$ is an operator norm in L^2. However, the SIR variational problem has a nonlinear function A for the constraint. Thus, we use the extension of the algorithm from [CV17a] which solves the nonlinear constrained optimization problem.

$$\min_u \max_p g(u) + \langle A(u), p\rangle_{L^2} - f^*(p), \tag{3.10}$$

where A is a nonlinear function. The scheme iterates the algorithm (3.8) with a linear approximation of A at a base point $\bar{u}$

$$A(u) \approx A(\bar{u}) + [\nabla A(\bar{u})](u - \bar{u}).$$

Denote $A_u := \nabla A(u)$. We have a linearized saddle point problem

$$\min_u \max_p g(u) + \langle A(\bar{u}) + A_{\bar{u}}(u - \bar{u}), p\rangle_{L^2} - f^*(p) \tag{3.11}$$

and the scheme iterates

$$\begin{aligned}
u^{(k+1)} &= \operatorname*{argmin}_u \; g(u) + \langle u, A^T_{u^{(k)}} p^{(k)}\rangle_{L^2} + \frac{1}{2\tau^{(k)}}\|u - u^{(k)}\|^2_{L^2} \\
\tilde{u}^{(k+1)} &= 2u^{(k+1)} - u^{(k)} \\
p^{(k+1)} &= \operatorname*{argmax}_p \langle A(u^{(k)}) + A_{u^{(k)}}(\tilde{u}^{(k+1)} - u^{(k)}), p\rangle_{L^2} - f^*(p) - \frac{1}{2\sigma^{(k)}}\|p - p^{(k)}\|^2_{L^2}
\end{aligned} \tag{3.12}$$

The paper [CV17a] proves that the sequence $\{u^{(k)}, p^{(k)}\}_{k=0}^{\infty}$ of the algorithm converges to some saddle point (u_*, p_*) that satisfies the following KKT conditions of (3.12):

$$\begin{aligned}
[\nabla A(u_*)]^T p_* &= -\partial g(u_*) \\
A(u_*) &= \partial f^*(p_*).
\end{aligned} \tag{3.13}$$

However, the scheme converges if the step sizes satisfy

$$\sigma^{(k)}\tau^{(k)}\|\nabla A(u^{(k)})\|^2_{L^2} < 1, \quad k = 1, 2, \cdots.$$

Suppose we use an unbounded operator that depends on the grid size, for example, $A = \nabla$. Then the operator norm $\|\nabla A(u^{(k)})\|_{L^2}^2$ increases as the grid sizes increase. Thus, the scheme can result in a very slow convergence if we use a fine grid resolution. To circumvent the problem, we use the General-proximal Primal-Dual Hybrid Gradient (G-prox PDHG) method from [JLL19] which is another variation of the PDHG algorithm. This variant provides an appropriate choice of norms for the algorithm, and the authors prove that choosing the proper norms allows the algorithm to have larger step sizes than the vanilla PDHG algorithm. The G-prox PDHG iterates

$$\begin{aligned} u^{(k+\frac{1}{2})} &= \underset{u}{\operatorname{argmin}}\ g(u) + \langle u, A_{u^{(k)}}^T p^{(k)} \rangle_{L^2} + \frac{1}{2\tau^{(k)}} \|u - u^{(k)}\|_{L^2}^2 \\ u^{(k+1)} &= 2u^{(k+\frac{1}{2})} - u^{(k)} \\ p^{(k+1)} &= \underset{p}{\operatorname{argmax}}\ \langle A(u^{(k)}) + A_{u^{(k)}}(u^{(k+1)} - u^{(k)}), p \rangle_{L^2} - f^*(p) - \frac{1}{2\sigma^{(k)}} \|p - p^{(k)}\|_{\mathcal{H}^{(k)}}^2. \end{aligned} \tag{3.14}$$

where the norm $\|\cdot\|_{\mathcal{H}^{(k)}}$ is defined as

$$\|p\|_{\mathcal{H}^{(k)}}^2 = \|A_{u^{(k)}}^T p\|_{L^2}^2.$$

By choosing the proper norms, the step sizes only need to satisfy

$$\sigma^{(k)} \tau^{(k)} < 1, \quad k = 1, 2, \cdots$$

which are clearly independent of the grid size.

3.3.1 Local convergence of the algorithm

In this section, we show the iterations from the algorithm (3.14) locally converges to the saddle point. The local convergence theorem in this paper is mainly based on the Theorem 2.11 from [CV17a]. However, we add a preconditioning operator from the G-prox PDHG method. We show that the method converges locally to the saddle point with the step sizes independent of the nonlinear operator A.

From the algorithm (3.14), $(u^{(k+1)}, p^{(k+1)})$ satisfies the following first-order optimality conditions

$$\begin{aligned} 0 &\in \partial g(u^{(k+1)}) + A_{u^{(k)}}^T p^{(k)} + \frac{1}{\tau^{(k)}}(u^{(k+1)} - u^{(k)}) \\ 0 &\in A(u^{(k)}) + 2A_{u^{(k)}}(u^{(k+1)} - u^{(k)}) - \partial f^*(p^{(k+1)}) - \frac{1}{\sigma^{(k)}} A_{u^{(k)}} A_{u^{(k)}}^T (p^{(k+1)} - p^{(k)}) \end{aligned} \tag{3.15}$$

which can be rewritten as

$$0 \in H_{u^{(k)}}(q^{(k+1)}) + M^{(k)}(q^{(k+1)} - q^{(k)}) \tag{3.16}$$

with $q = (u, p)$. Here, the monotone operator $H_{\bar{u}}$ is defined as

$$H_{\bar{u}}(q) := \begin{pmatrix} \partial g(u) + A_{\bar{u}}^T p \\ \partial f^*(p) - A(\bar{u}) - A_{\bar{u}}(u - \bar{u}) \end{pmatrix}$$

and

$$M^{(k)} := \begin{pmatrix} \frac{1}{\tau^{(k)}} Id & -A_{u^{(k)}}^T \\ -A_{u^{(k)}} & \frac{1}{\sigma^{(k)}} A_{u^{(k)}} A_{u^{(k)}}^T \end{pmatrix}$$

where Id is an identity operator.

Recall that from (3.13), the saddle point $q_* = (u_*, p_*)$ has to satisfy

$$0 \in H_{u_*}(u_*, p_*).$$

Throughout, we assume that

$$\|\nabla A(u_*)\| > 0 \tag{3.17}$$

and $u \mapsto \nabla A(u)$ is continuous.

Lemma 3.3.1. *There exists constants $0 < c < C$ and $R > 0$ such that*

$$c \leq \|\nabla A(u)\| \leq C, \quad (\|u - u_*\|_{L^2} \leq R)$$

where $\|\cdot\|$ is an operator norm.

Proof. This follows immediately from (3.17) and the fact that the derivative $\nabla A(u)$ is continuous with respect to u. □

Lemma 3.3.2. *Suppose* (3.17) *holds and let* $\tau^{(k)}\sigma^{(k)} < 1$. *Then there exist constants* $0 < \theta < \Theta$ *such that*

$$\theta^2 \|q\|_{L^2}^2 \leq \langle q, M^{(k)} q \rangle \leq \Theta^2 \|q\|_{L^2}^2$$

where

$$\|q\|_{L^2}^2 = \|u\|_{L^2}^2 + \|p\|_{L^2}^2.$$

A proof of Lemma 3.3.2 is provided in the appendix.

With the above Lemmas, we can use the Theorem 2.11 from [CV17a] to show the local convergence of the algorithm.

Theorem 3.3.3. *Let* $(u_*, p_*) \in L^2 \times \mathcal{H}^{(*)}$ *be a solution to* (3.13) *where* $\|p\|_{\mathcal{H}^{(*)}}^2 = \|A_{u_*}^T p\|_{L^2}^2$. *Let the step sizes* $\tau^{(k)}$ *and* $\sigma^{(k)}$ *satisfy* $\tau^{(k)}\sigma^{(k)} < 1$ *for all* k. *Then there exists* $\delta > 0$ *such that for any initial point* $(u^{(0)}, p^{(0)}) \in L^2 \times \mathcal{H}^{(0)}$ *satisfying*

$$\|u^{(0)} - u_*\|_{L^2}^2 + \|p^{(0)} - p_*\|_{L^2}^2 < \delta^2,$$

the iterates $(u^{(k)}, p^{(k)})$ *from* (3.14) *converges to the saddle point* (u_*, p_*).

Proof. By Lemma 3.3.1, Lemma 3.3.2, and strong convexity of the functional P, we can use [CV17a, Theorem 2.11], which proves the theorem. □

Remark 3.3.1. [CV17a, Theorem 2.11] requires H_{u_*} to satisfy the condition called metric regularity. In our formulation, the constraint $A(u) = 0$ makes H_{u_*} metrically regular by [CV17b, Section 5.3]. We refer readers to [CV17a, CV17b, RW09] for further details about metric regularity.

3.3.2 Implementation of the algorithm

To implement the algorithm to the minimization problem (3.4), we set

$$\begin{aligned} u &= ((\rho_i, m_i)_{i\in\{S,I,R\}}) \\ p &= (\phi_i)_{i\in\{S,I,R\}} \\ g(u) &= P(u) \\ f(A(u)) &= \begin{cases} 0 & \text{if } A(u) = 0 \\ \infty & \text{otherwise} \end{cases} \\ f^*(p) &= 0 \end{aligned}$$

where A is a nonlinear operator defined as

$$\begin{aligned} A(u) = \big(A_S(u), A_I(u), A_R(u)\big) = (&\partial_t\rho_S + \nabla\cdot m_S - \frac{\eta^2}{2}\Delta\rho_S + \beta\rho_S K * \rho_I, \\ &\partial_t\rho_I + \nabla\cdot m_I - \frac{\eta^2}{2}\Delta\rho_I - \beta\rho_I K * \rho_S + \gamma\rho_I, \\ &\partial_t\rho_R + \nabla\cdot m_R - \frac{\eta^2}{2}\Delta\rho_R). \end{aligned}$$

Define the Lagrangian functional as

$$\mathcal{L}(u,p) := P(u) - \langle A(u), p\rangle_{L^2}$$

where $\langle\cdot,\cdot\rangle_{L^2}$ is an inner product defined as

$$\langle p, q\rangle_{L^2} = \sum_{i=S,I,R}\int_0^1\int_\Omega p_i(t,x)\, q_i(t,x)\, dx\, dt$$

for $p = (p_S, p_I, p_R)$ and $q = (q_S, q_I, q_R)$. Thus, using definitions of the inner product and the operator A, $\langle A(u), p\rangle_{L^2}$ can be written as

$$\begin{aligned} \langle A(u), p\rangle_{L^2} = &\int_0^1\int_\Omega \phi_S\left(\partial_t\rho_S + \nabla\cdot m_S - \frac{\eta^2}{2}\Delta\rho_S + \beta\rho_S K * \rho_I\right) dx\, dt \\ &+ \int_0^1\int_\Omega \phi_I\left(\partial_t\rho_I + \nabla\cdot m_I - \frac{\eta^2}{2}\Delta\rho_I - \beta\rho_I K * \rho_S + \gamma\rho_I\right) dx\, dt \\ &+ \int_0^1\int_\Omega \phi_R\left(\partial_t\rho_R + \nabla\cdot m_R - \frac{\eta^2}{2}\Delta\rho_R\right) dx\, dt. \end{aligned}$$

Now we are ready to implement the algorithm (3.14) to solve the SIR variataional problem.

Algorithm 3 G-proximal PDHG for mean-field control SIR system

Input: $\rho_i(0,\cdot)$ $(i \in \{S, I, R\})$

Output: ρ_i, m_i, ϕ_i $(i \in \{S, I, R\})$ for $x \in \Omega$, $t \in [0, T]$

While relative error $>$ tolerance **For** $i \in \{S, I, R\}$

$$\phi_i^{(k+1)} = \text{argmax}_\phi \, \mathcal{L}(\rho^{(k)}, m_i^{(k)}, \phi) - \tfrac{1}{2\sigma_i}\|\phi - \phi_i^{(k)}\|^2_{H_i^{(k)}}$$
$$\rho_i^{(k+1)} = \text{argmin}_\rho \, \mathcal{L}(\rho, m_i^{(k)}, 2\phi_i^{(k+1)} - \phi_i^{(k)}) + \tfrac{1}{2\tau_i}\|\rho - \rho_i^{(k)}\|^2_{L^2}$$
$$m_i^{(k+1)} = \text{argmin}_m \, \mathcal{L}(\rho^{(k+1)}, m, 2\phi_i^{(k+1)} - \phi_i^{(k)}) + \tfrac{1}{2\tau_i}\|m - m_i^{(k)}\|^2_{L^2}$$

Here, with abuse of notations, L^2 norms are defined as

$$\|\rho_i\|^2_{L^2} = \int_0^T \int_\Omega \rho_i^2(t,x) dx\, dt$$
$$\|m_i\|^2_{L^2} = \int_0^T \int_\Omega |m_i(t,x)|^2 dx\, dt, \quad (i = S, I, R)$$

and $H_i^{(k)}$ are defined as

$$\|\phi_i\|^2_{H_i^{(k)}} = \|[\nabla A_i(u^{(k)})]^T \phi_i\|^2_{L^2}, \quad (i = S, I, R)$$
$$\|\phi_S\|^2_{H_S^{(k)}} = \int_0^T \int_\Omega (\partial_t \phi_S)^2 + |\nabla \phi_S|^2 + \frac{\eta^4}{4}(\Delta \phi_S)^2 + \beta^2 (K * \rho_I^{(k)} \phi_S)^2 \, dx\, dt$$
$$\|\phi_I\|^2_{H_I^{(k)}} = \int_0^T \int_\Omega (\partial_t \phi_I)^2 + |\nabla \phi_I|^2 + \frac{\eta^4}{4}(\Delta \phi_I)^2 + \beta^2 (K * \rho_S^{(k)} \phi_I)^2 + \gamma^2 (\phi_I)^2 \, dx\, dt$$
$$\|\phi_R\|^2_{H_R^{(k)}} = \int_0^T \int_\Omega (\partial_t \phi_R)^2 + |\nabla \phi_R|^2 + \frac{\eta^4}{4}(\Delta \phi_R)^2 \, dx\, dt.$$

Moreover, the relative error is defined as

$$\text{relative error} = \frac{|P(\rho_i^{(k+1)}, m_i^{(k+1)}) - P(\rho_i^{(k)}, m_i^{(k)})|}{|P(\rho_i^{(k)}, m_i^{(k)})|}.$$

By formulating these optimality conditions, we can find explicit formulas for each variable.

Proposition 3.3.4. *The variables $\rho_i^{(k+1)}, m_i^{(k+1)}, \phi_i^{(k+1)}$ $(i \in \{S, I, R\})$ from the Algorithm 4 satisfy the following explicit formulas:*

$$\rho_S^{(k+1)} = root_+\Bigg(\frac{\tau_S}{1+c\tau_S}\Big(\partial_t\phi_S^{(k)} + \frac{\eta_S^2}{2}\Delta\phi_S^{(k)} - \frac{1}{\tau_S}\rho_S^{(k)} + \beta\left(K*(\phi_I^{(k)}\rho_I^{(k)}) - \phi_S^{(k)}K*\rho_I^{(k)}\right) + c(\rho_I+\rho_R)\Big), 0, -\frac{\tau_S\alpha_S(m_S^{(k)})^2}{2(1+c\tau_S)}\Bigg)$$

$$\rho_I^{(k+1)} = root_+\Bigg(\frac{\tau_I}{1+c\tau_I}\Big(\partial_t\phi_I^{(k)} + \frac{\eta_I^2}{2}\Delta\phi_I^{(k)} - \frac{1}{\tau_I}\rho_I^{(k)} + \beta\left(\phi_I^{(k)}K*\rho_S^{(k)} - K*(\phi_S^{(k)}\rho_S^{(k)})\right) + \gamma(\phi_R-\phi_I) + c(\rho_S+\rho_R)\Big), 0, -\frac{\tau_I\alpha_I(m_I^{(k)})^2}{2(1+c\tau_I)}\Bigg)$$

$$\rho_R^{(k+1)} = root_+\Bigg(\frac{\tau_R}{1+c\tau_R}\Big(\partial_t\phi_R^{(k)} + \frac{\eta_R^2}{2}\Delta\phi_R^{(k)} - \frac{1}{\tau_R}\rho_R^{(k)} + c(\rho_S+\rho_I)\Big), 0, -\frac{\tau_R\alpha_R(m_R^{(k)})^2}{2(1+c\tau_R)}\Bigg)$$

$$m_i^{(k+1)} = \frac{\rho_i^{(k+1)}}{\tau\alpha_i + \rho_i^{(k+1)}}\left(m_i^{(k)} - \tau\nabla\phi_i^{(k)}\right), \qquad (i\in\{S,I,R\})$$

$$\phi_S^{(k+1)} = \phi_S^{(k)} + \sigma_S(A_SA_S^T)^{-1}\left(-\partial_t\rho_S^{(k+1)} - \nabla\cdot m_S^{(k+1)} - \beta\rho_S^{(k+1)}K*\rho_I^{(k+1)} + \frac{\eta_S^2}{2}\Delta\rho_S^{(k+1)}\right)$$

$$\phi_I^{(k+\frac{1}{2})} = \phi_I^{(k)} + \sigma_I(A_IA_I^T)^{-1}\Big(-\partial_t\rho_I^{(k+1)} - \nabla\cdot m_I^{(k+1)} + \beta\rho_I^{(k+1)}K*\rho_S^{(k+1)} - \gamma\rho_I^{(k+1)} + \frac{\eta_I^2}{2}\Delta\rho_I^{(k+1)}\Big)$$

$$\phi_R^{(k+\frac{1}{2})} = \phi_R^{(k)} + \sigma_R(A_RA_R^T)^{-1}\left(-\partial_t\rho_R^{(k+1)} - \nabla\cdot m_R^{(k+1)} + \gamma\rho_I^{(k+1)} + \frac{\eta_R^2}{2}\Delta\rho_R^{(k+1)}\right)$$

where $root_+(a,b,c)$ *is a positive root of a cubic polynomial* $x^3+ax^2+bx+c=0$ *and*

$$A_SA_S^T = -\partial_{tt} + \frac{\eta_S^4}{4}\Delta^2 - (1+2\beta\eta_S)\Delta + \beta^2$$
$$A_IA_I^T = -\partial_{tt} + \frac{\eta_I^4}{4}\Delta^2 - (1+2(\gamma+\beta)\eta_S)\Delta + (\gamma+\beta)^2$$
$$A_RA_R^T = -\partial_{tt} + \frac{\eta_R^4}{4}\Delta^2 - \Delta.$$

We use FFTW library to compute $(A_i A_i^T)^{-1}$ $(i \in \{S, I, R\})$ and convolution terms by Fast Fourier Transform (FFT), which is $O(n \log n)$ operations per iteration with n being the number of points. Thus, the algorithm takes just $O(n \log n)$ operations per iteration.

In this section, we implement optimization methods to solve the proposed SIR variational problems. Specifically, we use G-Prox Primal-Dual Hybrid Gradient (G-Prox PDHG) method [JLL19]. This is a variation of Chambolle-Pock primal-dual algorithm [CP11b, CP16]. G-Prox PDHG proposes a way of choosing proper norms for the optimization based on the given minimization problem whereas the Chambolle-Pock primal-dual algorithm just uses L^2 norms. Choosing appropriate norms results in faster and more robust convergence of the algorithm.

3.3.3 Review of primal-dual algorithms

The PDHG method solves the minimization problem

$$\min_x f(Ax) + g(x)$$

by converting it into a saddle point problem

$$\min_x \sup_y \{L(x, y) := \langle Ax, y\rangle + g(x) - f^*(y)\}.$$

Here, f and g are convex functions, A is a continuous linear operator, and

$$f^*(y) = \sup_x \langle x\, y\rangle - f(x)$$

is a Legendre transform of f. For each iteration, the algorithm finds the minimizer x_* by gradient descent method and the maximizer y_* by gradient ascent method. Thus, the minimizer and maximizer are calculated by iterating

$$\begin{cases} x^{k+1} & = \text{argmin}_x\, L(x, y^k) + \frac{1}{2\tau}\|x - x^k\|^2 \\ y^{k+\frac{1}{2}} & = \text{argmax}_y\, L(x^{k+1}, y) + \frac{1}{2\sigma}\|y - y^k\|^2 \\ y^{k+1} & = 2y^{k+\frac{1}{2}} - y^k \end{cases}$$

where τ and σ are step sizes for the algorithm.

G-Prox PDHG is a modified version of PDHG that solves the minimization problem by choosing the most appropriate norms for updating x and y. Choosing the appropriate norms allows us to choose larger step sizes. Hence, we get a faster convergence rate. In details,

$$\begin{cases} x^{k+1} & = \text{argmin}_x \, L(x, y^k) + \frac{1}{2\tau}\|x - x^k\|_{\mathcal{H}}^2 \\ y^{k+\frac{1}{2}} & = \text{argmax}_y \, L(x^{k+1}, y) + \frac{1}{2\sigma}\|y - y^k\|_{\mathcal{G}}^2 \\ y^{k+1} & = 2y^{k+\frac{1}{2}} - y^k \end{cases}$$

where $\mathcal{H}$ and $\mathcal{G}$ are two Hilbert spaces with the inner product

$$(u_1, u_2)_{\mathcal{G}} = (Au_1, Au_2)_{\mathcal{H}}.$$

In particular, we use G-Prox PDHG to solve the minimization problem (3.4) by setting $\mathcal{H} = L^2$ and $\mathcal{G} = H^2$. Furthermore,

$$x = (\rho_S, \rho_I, \rho_R, m_S, m_I, m_R), \quad g(x) = P(\rho_i, m_i)_{i\in\{S,I,R\}}, \quad f(Ax) = \begin{cases} 0 & \text{if } Ax = (0, 0, \gamma\rho_I) \\ \infty & \text{otherwise.} \end{cases}$$

$$\begin{aligned} Ax = (&\partial_t\rho_S + \nabla\cdot m_S - \frac{\eta^2}{2}\Delta\rho_S + \beta\rho_S K * \rho_I, \\ &\partial_t\rho_I + \nabla\cdot m_I - \frac{\eta^2}{2}\Delta\rho_I - \beta\rho_I K * \rho_S + \gamma\rho_I, \\ &\partial_t\rho_R + \nabla\cdot m_R - \frac{\eta^2}{2}\Delta\rho_R). \end{aligned}$$

Thus, we have the following inner products

$$(u_1, u_2)_{L^2} = \int_0^T \int_\Omega u_1(t,x)u_2(t,x)dxdt, \quad (u_1, u_2)_{H^2} = \int_0^T \int_\Omega Au_1(t,x)Au_2(t,x)dxdt.$$

Note that the operator A is nonlinear. In the implementation, we approximate the operator with the following linear operator

$$\begin{aligned} Ax \approx (&\partial_t\rho_S + \nabla\cdot m_S - \frac{\eta^2}{2}\Delta\rho_S + \beta\rho_S, \\ &\partial_t\rho_I + \nabla\cdot m_I - \frac{\eta^2}{2}\Delta\rho_I + (\gamma+\beta)\rho_I, \\ &\partial_t\rho_R + \nabla\cdot m_R - \frac{\eta^2}{2}\Delta\rho_R). \end{aligned}$$

3.3.4 G-Prox PDHG on SIR variational problem

In this section, we implement G-Prox PDHG to solve the saddle problem (3.6). For $i \in \{S, I, R\}$,

$$\begin{aligned}
\rho_i^{(k+1)} &= \operatorname*{argmin}_{\rho} \mathcal{L}(\rho, m_i^{(k)}, \phi_i^{(k)}) + \frac{1}{2\tau_i}\|\rho - \rho_i^{(k)}\|_{L^2}^2 \\
m_i^{(k+1)} &= \operatorname*{argmin}_{m} \mathcal{L}(\rho_i^{(k+1)}, m, \phi_i^{(k)}) + \frac{1}{2\tau_i}\|m - m_i^{(k)}\|_{L^2}^2 \\
\phi_i^{(k+\frac{1}{2})} &= \operatorname*{argmax}_{\phi} \mathcal{L}(\rho_i^{(k+1)}, m_i^{(k+1)}, \phi) - \frac{1}{2\sigma_i}\|\phi - \phi_i^{(k)}\|_{H^2}^2 \\
\phi_i^{(k+1)} &= 2\phi_i^{(k+\frac{1}{2})} - \phi_i^{(k)}
\end{aligned}$$

where τ_i, σ_i ($i \in \{S, I, R\}$) are step sizes for the algorithm and by G-Prox PDHG, L^2 norm and H^2 norm are defined as

$$\|u\|_{L^2}^2 = \int_0^T \int_\Omega u^2 dxdt, \quad \|u\|_{H^2}^2 = \int_0^T \int_\Omega (\partial_t u)^2 + |\nabla u|^2 + \frac{\eta^4}{4}(\Delta u)^2 dxdt$$

for any $u : [0, T] \times \Omega \to [0, \infty)$.

By formulating these optimality conditions, we can find explicit formulas for each variable.

$$\begin{aligned}
\rho_S^{(k+1)} = root_+\Bigg(& \frac{\tau_S}{1 + c\tau_S}\Bigg(\partial_t \phi_S^{(k)} + \frac{\eta_S^2}{2}\Delta\phi_S^{(k)} - \frac{1}{\tau_S}\rho_S^{(k)} + \beta\left(K * (\phi_I^{(k)}\rho_I^{(k)}) - \phi_S^{(k)} K * \rho_I^{(k)}\right) \\
& + c(\rho_I + \rho_R)\Bigg), 0, -\frac{\tau_S \alpha_S (m_S^{(k)})^2}{2(1 + c\tau_S)}\Bigg) \\
\rho_I^{(k+1)} = root_+\Bigg(& \frac{\tau_I}{1 + c\tau_I}\Bigg(\partial_t \phi_I^{(k)} + \frac{\eta_I^2}{2}\Delta\phi_I^{(k)} - \frac{1}{\tau_I}\rho_I^{(k)} + \beta\left(\phi_I^{(k)} K * \rho_S^{(k)} - K * (\phi_S^{(k)}\rho_S^{(k)})\right) \\
& + \gamma(\phi_R - \phi_I) + c(\rho_S + \rho_R)\Bigg), 0, -\frac{\tau_I \alpha_I (m_I^{(k)})^2}{2(1 + c\tau_I)}\Bigg) \\
\rho_R^{(k+1)} = root_+\Bigg(& \frac{\tau_R}{1 + c\tau_R}\Bigg(\partial_t \phi_R^{(k)} + \frac{\eta_R^2}{2}\Delta\phi_R^{(k)} - \frac{1}{\tau_R}\rho_R^{(k)} + c(\rho_S + \rho_I)\Bigg), 0, -\frac{\tau_R \alpha_R (m_R^{(k)})^2}{2(1 + c\tau_R)}\Bigg)
\end{aligned}$$

$$m_i^{(k+1)} = \frac{\rho_i^{(k+1)}}{\tau\alpha_i + \rho_i^{(k+1)}}\left(m_i^{(k)} - \tau\nabla\phi_i^{(k)}\right), \qquad (i \in \{S, I, R\})$$

$$\phi_S^{(k+1)} = \phi_S^{(k)} + \sigma_S (A_S^T A_S)^{-1} \left(-\partial_t \rho_S^{(k+1)} - \nabla \cdot m_S^{(k+1)} - \beta \rho_S^{(k+1)} K * \rho_I^{(k+1)} + \frac{\eta_S^2}{2} \Delta \rho_S^{(k+1)} \right)$$

$$\phi_I^{(k+\frac{1}{2})} = \phi_I^{(k)} + \sigma_I (A_I^T A_I)^{-1} \Big(-\partial_t \rho_I^{(k+1)} - \nabla \cdot m_I^{(k+1)} + \beta \rho_I^{(k+1)} K * \rho_S^{(k+1)} - \gamma \rho_I^{(k+1)} + \frac{\eta_I^2}{2} \Delta \rho_I^{(k+1)} \Big)$$

$$\phi_R^{(k+\frac{1}{2})} = \phi_R^{(k)} + \sigma_R (A_R^T A_R)^{-1} \left(-\partial_t \rho_R^{(k+1)} - \nabla \cdot m_R^{(k+1)} + \gamma \rho_I^{(k+1)} + \frac{\eta_R^2}{2} \Delta \rho_R^{(k+1)} \right)$$

where $root_+(a, b, c)$ is a positive root of a cubic polynomial $x^3 + ax^2 + bx + c = 0$ and

$$\begin{aligned} A_S^T A_S &= -\partial_{tt} + \frac{\eta_S^4}{4} \Delta^2 - (1 + 2\beta\eta_S)\Delta + \beta^2 \\ A_I^T A_I &= -\partial_{tt} + \frac{\eta_I^4}{4} \Delta^2 - (1 + 2(\gamma + \beta)\eta_S)\Delta + (\gamma + \beta)^2 \\ A_R^T A_R &= -\partial_{tt} + \frac{\eta_R^4}{4} \Delta^2 - \Delta. \end{aligned}$$

We use FFTW library to compute $(A_i^T A_i)^{-1}$ $(i \in \{S, I, R\})$ and convolution terms by Fast Fourier Transform (FFT), which is $O(n \log n)$ operations per iteration with n being the number of points. Thus, the algorithm takes just $O(n \log n)$ operations per iteration.

In all, we summarize the algorithm as follows.

Algorithm 4 G-proximal PDHG for mean-field control SIR system

Input: $\rho_i(0, \cdot)$ $(i \in \{S, I, R\})$

Output: ρ_i, m_i, ϕ_i $(i \in \{S, I, R\})$ for $x \in \Omega$, $t \in [0, T]$

While relative error $>$ tolerance **For** $i \in \{S, I, R\}$

$\phi_i^{(k+1)} = \text{argmax}_\phi \, \mathcal{L}(\rho^{(k)}, m_i^{(k)}, \phi) - \frac{1}{2\sigma_i} \|\phi - \phi_i^{(k)}\|_{H^2}^2$

$\rho_i^{(k+1)} = \text{argmin}_\rho \, \mathcal{L}(\rho, m_i^{(k)}, 2\phi_i^{(k+1)} - \phi_i^{(k)}) + \frac{1}{2\tau_i} \|\rho - \rho_i^{(k)}\|_{L^2}^2$

$m_i^{(k+1)} = \text{argmin}_m \, \mathcal{L}(\rho^{(k+1)}, m, 2\phi_i^{(k+1)} - \phi_i^{(k)}) + \frac{1}{2\tau_i} \|m - m_i^{(k)}\|_{L^2}^2$

Here, the relative error is defined as

$$\text{relative error} = \frac{|P(\rho_i^{(k+1)}, m_i^{(k+1)}) - P(\rho_i^{(k)}, m_i^{(k)})|}{|P(\rho_i^{(k)}, m_i^{(k)})|}.$$

3.4 Experiments

In this section, we present several sets of numerical experiments using the algorithm with various parameters. We wrote C++ codes to run the numerical experiments. Let $\Omega = [0,1]^2$ be a unit cube in $\mathbb{R}^2$ and $T = 1$. The domain Ω is discretized with the regular rectangular mesh

$$\Delta x = \frac{1}{N_x}, \quad \Delta y = \frac{1}{N_y}, \quad \Delta t = \frac{1}{N_t - 1}$$

$$x_{kl} = ((k+0.5)\Delta x, (l+0.5)\Delta y), \quad k = 0, \cdots, N_x - 1, \quad l = 0, \cdots, N_y - 1$$

$$t_n = n\Delta t, \qquad n = 0, \cdots, N_t - 1$$

where N_x, N_y are the number of data points in space and N_t is the number of data points in time. For all the experiments, we use the same set of parameters,

$$N_x = 128, \quad N_y = 128, \quad N_t = 32$$

$$\sigma = 0.02, \quad c = 0.01, \quad \eta_i = 0.01 \quad (i \in \{S, I, R\})$$

$$\alpha_S = 1, \quad \alpha_I = 10, \quad \alpha_R = 1$$

and choose the same terminal cost functional

$$E(\rho_I(1, \cdot)) = \frac{1}{2} \int_\Omega \rho_I^2(1, x) dx.$$

By setting a higher value for α_I, we penalize the infected population's movement more than other populations. Considering the immobility of the infected individuals, this is a reasonable choice in terms of real-world applications.

We would like to minimize the terminal cost functional $E(\rho_I(T, \cdot))$. A solution needs to reduce the number of the infected population. There are mainly two ways of reducing the

number of infected. One way is to recover infected to recovered population. However, it may not be feasible if the rate of recovery γ is low. Another way to reduce the number of infected is by separating the susceptible population from the infected population. The number of infected doesn't increase if there are no susceptible people near infected. However, the total cost increases when densities move due to the kinetic energy term $\rho_i|v_i|^2$ ($i \in \{S, I, R\}$) in the running cost. A solution needs to find the optimal balance between the terminal cost and the running cost. Experiment 1 shows the effectiveness of controlling populations' movements. We compute two solutions of the model: with and without control of movements. The comparison between these solutions shows that the number of infected people can be reduced effectively with the control at the terminal time. Experiment 2 shows that the algorithm finds the proper solutions based on different recovery rates given nonsymmetric initial densities. In Experiment 3, we consider a more complicated terminal energy functional $E(\rho_I(T, \cdot))$, and compute the solutions based on different infection rates.

3.4.1 Experiment 1

In this experiment, we compare the solutions of the SIR model with and without control. We set initial densities for susceptible, infected and recovered populations as

$$\rho_S(0, x = (x_1, x_2)) = 0.6 \exp\Big(-10\big((x_1 - 0.5)^2 + (x_2 - 0.5)^2\big)\Big)$$
$$\rho_I(0, x = (x_1, x_2)) = 0.6 \exp\Big(-35\big((x_1 - 0.6)^2 + (x_2 - 0.6)^2\big)\Big)$$
$$\rho_R(0, x = (x_1, x_2)) = 0$$

Susceptible population and infected population are Gaussian distributions centered at $(0.5, 0.5)$ and $(0.6, 0.6)$, respectively. We set $\beta = 0.7$ and $\gamma = 0.1$.

We show two different numerical results: one with control and one without control. The

formulation without control has the following system of equations,

$$\begin{aligned}
\frac{\partial \rho_S(t,x)}{dt} &= -\beta \rho_S(t,x)\rho_I(t,x) \\
\frac{\partial \rho_I(t,x)}{dt} &= \beta \rho_S(t,x)\rho_I(t,x) - \gamma \rho_I(t,x) \\
\frac{\partial \rho_R(t,x)}{dt} &= \gamma \rho_I(t,x).
\end{aligned}$$

By removing the velocity terms, we assume no movements of population. We solve these equations by using Euler's method. Thus, the solution can be computed by iterating $n = 0, \cdots, N_t - 2$,

$$\begin{aligned}
\rho_S(t_{n+1}, x_{kl}) &= \rho_S(t_n, x_{kl}) - \Delta t \beta \rho_S(t_n, x_{kl})\rho_I(t_n, x_{kl}) \\
\rho_I(t_{n+1}, x_{kl}) &= \rho_I(t_n, x_{kl}) + \Delta t \left(\beta \rho_S(t_n, x_{kl})\rho_I(t_n, x_{kl}) - \gamma \rho_I(t_n, x_{kl})\right) \\
\rho_R(t_{n+1}, x_{kl}) &= \rho_R(t_n, x_{kl}) + \Delta t \gamma \rho_I(t_n, x_{kl}),
\end{aligned}$$

for $k = 0, \cdots, N_x - 1, l = 0, \cdots, N_y - 1$. The results can be seen in Figure 3.1 and Figure 3.2. Figure 3.1 shows snapshots of the initial and terminal densities. The first row shows the initial densities of susceptible, infected and recovered (from left to right) based on the equations above. The second row and the third row show the terminal densities without control and with control, respectively. Figure 3.2 shows a quantitative comparison between these two solutions. The graphs indicate the total sum of each group over time. More specifically, they show $\int_\Omega \rho_i(t,x)dx$ for $i \in \{S, I, R\}$ from $t = 0$ to $t = 1$.

In Figure 3.1, when we compare the susceptible groups from second and third rows, the susceptible group with control moves more than the susceptible group without control. If there is no control (the second row in Figure 3.1), the groups don't move, and the susceptible group is exposed to the infected group, leading to a high chance of susceptible being infected over time. If the population is in control (the third row in Figure 3.1), we see a clear separation between susceptible and infected at the terminal time. This separation decreases the exposure of susceptible to infected effectively and, as a result, we see less number of the infected and more number of susceptible at the terminal time from the solution with control.

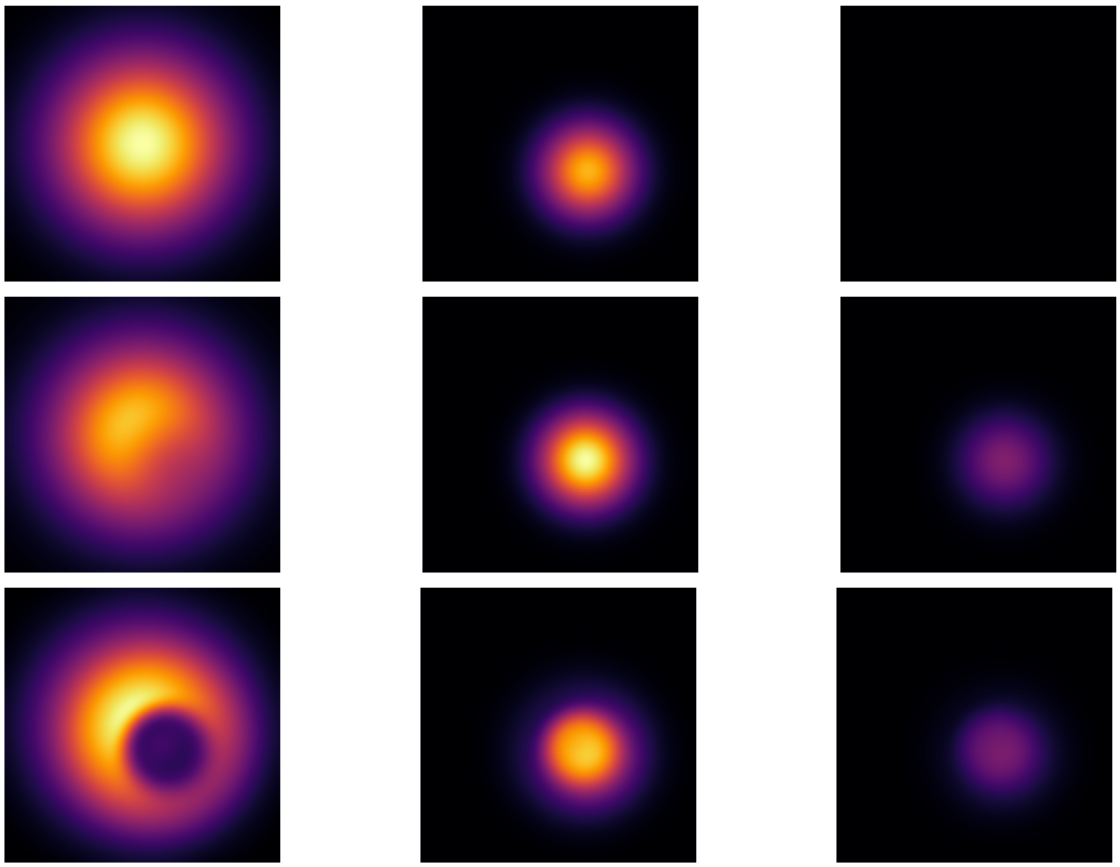
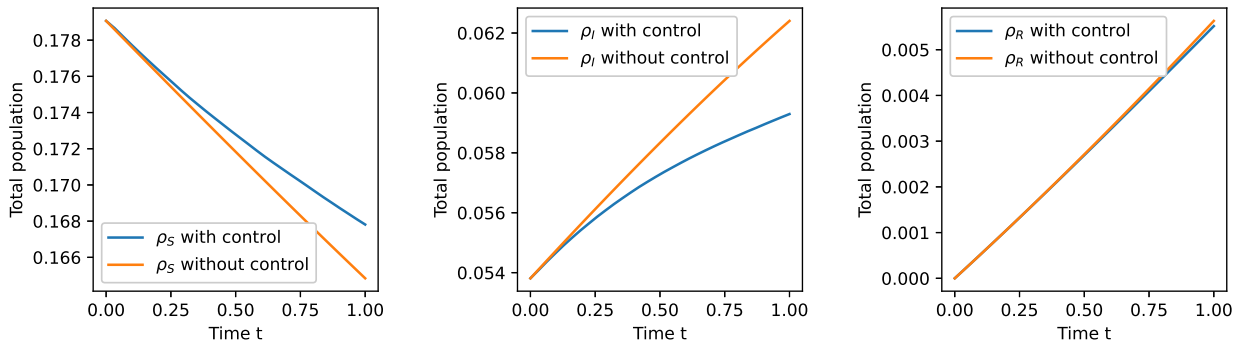


Figure 3.1: Snapshots of susceptible (column 1), infected (column 2) and recovered populations (column 3). The first row shows the initial densities, the second row shows the solution without control at the terminal time and the third row shows the solution with control at the terminal time.



(a) ρ_S (b) ρ_I (c) ρ_R

Figure 3.2: The comparison between solutions with and without control. The graphs show the total population of each group $\int_\Omega \rho_i(t,x)dx$ for $0 \le t \le 1$ and $i \in \{S, I, R\}$.

3.4.2 Experiment 2

In this experiment, we consider nonsymmetric initial densities.

$$\begin{aligned}
\rho_S(0,x) &= 0.45\Big(\exp\big(-15((x-0.3)^2+(y-0.3)^2)\big) \\
&\quad + \exp\big(-25((x-0.5)^2+(y-0.75)^2)\big) \\
&\quad + \exp\big(-30((x-0.8)^2+(y-0.35)^2)\big)\Big) \\
\rho_I(0,x) &= 10\big(0.04-(x-0.2)^2-(y-0.65)^2\big)_+ \\
&\quad + 12\big(0.03-(x-0.5)^2-(y-0.2)^2\big)_+ \\
&\quad + 12\big(0.03-(x-0.8)^2-(y-0.55)^2\big)_+ \\
\rho_R(0,x) &= 0.
\end{aligned}$$

The susceptible population is the sum of three Gaussian distributions, and the infected population is the sum of the positive part of quadratic polynomials. We conduct this experiment to show that the algorithm works well for nonsymmetric initial densities. Moreover, we choose $\beta = 0.34$ (an infection rate) and $\gamma = 0.12$ (a recovery rate) from [BFM20] based on the data in California, U.S. from March to May 2020. Figure 3.3 shows the evolution of densities using these parameters. We repeat the experiment with the same initial densities and

β but with different γ (Figure 3.4). In this experiment, we show the solution to the problem based on $\gamma = 0.36$. This experiment is under the scenario when the vaccine comes to the public. In both figures, evolutions of densities ρ_i ($i \in \{S, I, R\}$) are shown at $t = 0, 0.21, 0.47, 0.74, 1$. The total population of each density is indicated as *sum* in the subtitle of each plot, and it is calculated as $\int_\Omega \rho_i(t, x)dx$ for $0 \leq t \leq T$ and $i \in \{S, I, R\}$.

When $\gamma = 0.12$ (a low recovery rate), the solution separates susceptible population away from infected population. By separating susceptible from infected, the solution prevents susceptible populations from becoming infected, thus reducing the terminal cost at $t = 1$. When $\gamma = 0.36$ (a high recovery rate), recovering the infected is considered a better choice than separating the susceptible population from the infected population. In Figure 3.4, the susceptible population barely moves over time. We also observe that less number of the infected and more number of recovered. The total population of the infected at the terminal time in Figure 3.4 is 0.045, which is smaller than the total population of the infected in Figure 3.3. This experiment tells us that, with a high recovery rate, the optimal way to minimize the number of infected is by focusing on recovering them rather than moving the susceptible population.

3.4.3 Experiment 3

In this experiment, we consider the initial densities

$$\rho_S(0, x) = \begin{cases} 0.4 & \text{if } x \in B_{0.3}(0.5, 0.5) \\ 0 & \text{else} \end{cases}, \quad \rho_I(0, x) = \begin{cases} 0.4 & \text{if } x \in B_{0.2}(0.5, 0.5) \\ 0 & \text{else} \end{cases}, \quad \rho_R(0, \cdot) = 0$$

where $B_R(x_1, x_2)$ is a ball of radius R centered at (x_1, x_2) with value. Furthermore, we consider the following energy functional:

$$E(\rho_I(T, \cdot)) = \int_\Omega \frac{1}{2}\rho_I^2(T, x) + \rho_I(T, x)V(x)dx$$

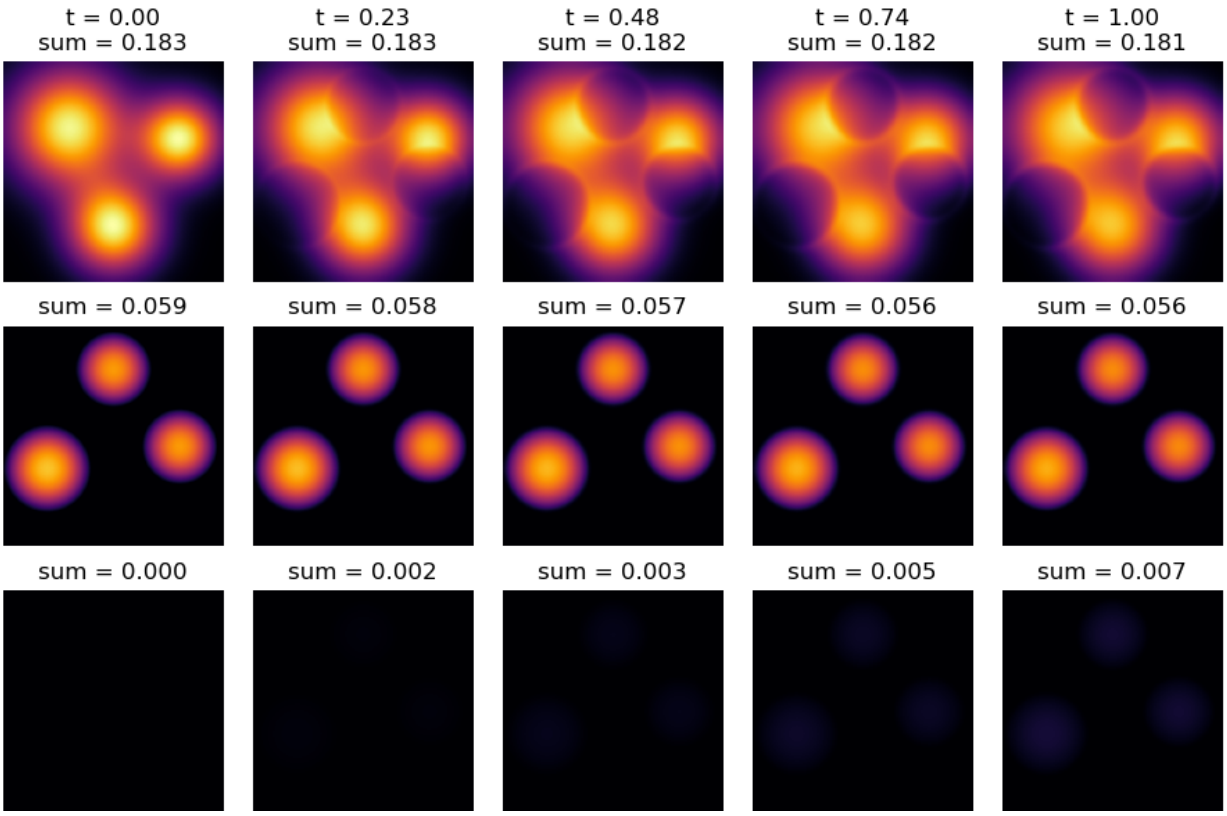


Figure 3.3: Experiment 2. The evolution of populations from $t = 0$ to $t = 1$ with $\beta = 0.34$ and $\gamma = 0.12$. The first row represents the susceptible population, the second row represents the infected population, and the last row represents the recovered population.

where, for $x = (x_1, x_2)$,

$$V(x) = \begin{cases} 1 & \text{if } |x_1 - 0.5| < 0.1 \text{ and } |x_2 - 0.5| < 0.1 \\ 0 & \text{otherwise.} \end{cases}$$

Here $V(x)$ is a step function that equals 1 on a square with a side length 0.2 at the center of the domain and 0 elsewhere. This energy penalizes if there is a positive infected density on the square. Thus, the solution has to move the infected density away from the square region while minimizing the total infected population. In this set of experiments, we show how the solution changes based on an infection rate β. We consider the case with a high infection rate $\beta = 0.96, \gamma = 0.12$ (Figure 3.5) and with $\beta = 0.34, \gamma = 0.12$ (Figure 3.6) same

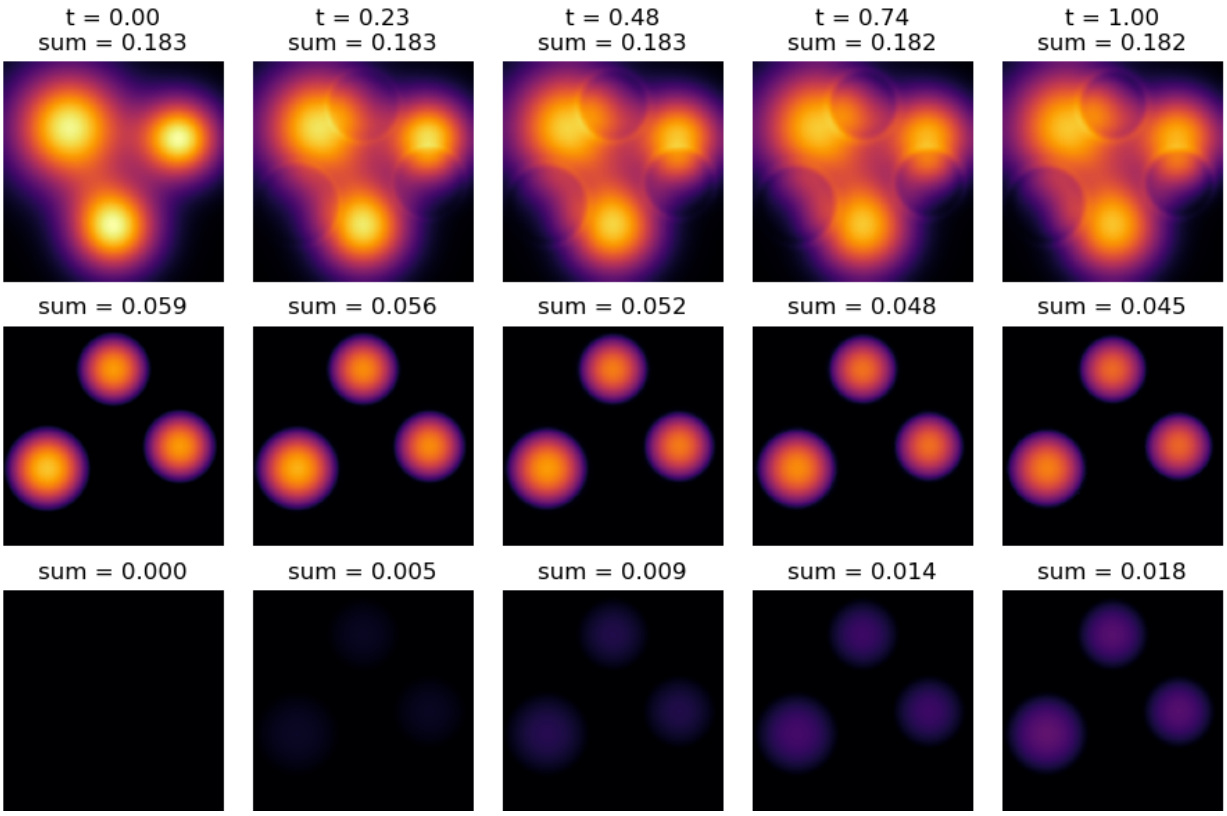


Figure 3.4: Experiment 2. The evolution of populations from $t = 0$ to $t = 1$ with $\beta = 0.34$ and $\gamma = 0.36$. The first row represents the susceptible population, the second row represents the infected population, and the last row represents the recovered population.

as Experiment 2.

When $\beta = 0.96$ (a high infection rate), the solution minimizes the total infected population by separating the susceptible from the infected. Due to the usage of this energy functional, the infected population has to move away from the square region at the center. Since there is going to be no infected population in this square region at the terminal time, the optimal place for the susceptible population is inside this square region. As a result, we can see the concentrated susceptible population inside this square at the terminal time. When $\beta = 0.32$ (a low infection rate), the susceptible population does not move as much as in the case when β is large. There are more overlaps between susceptible and infected groups at the

terminal time when β is small. However, when β is large, there is a complete separation between these groups. Thus, based on β and γ values, our model's solution can find the most cost-effective way of moving susceptible and infected populations while minimizing the total infected population.

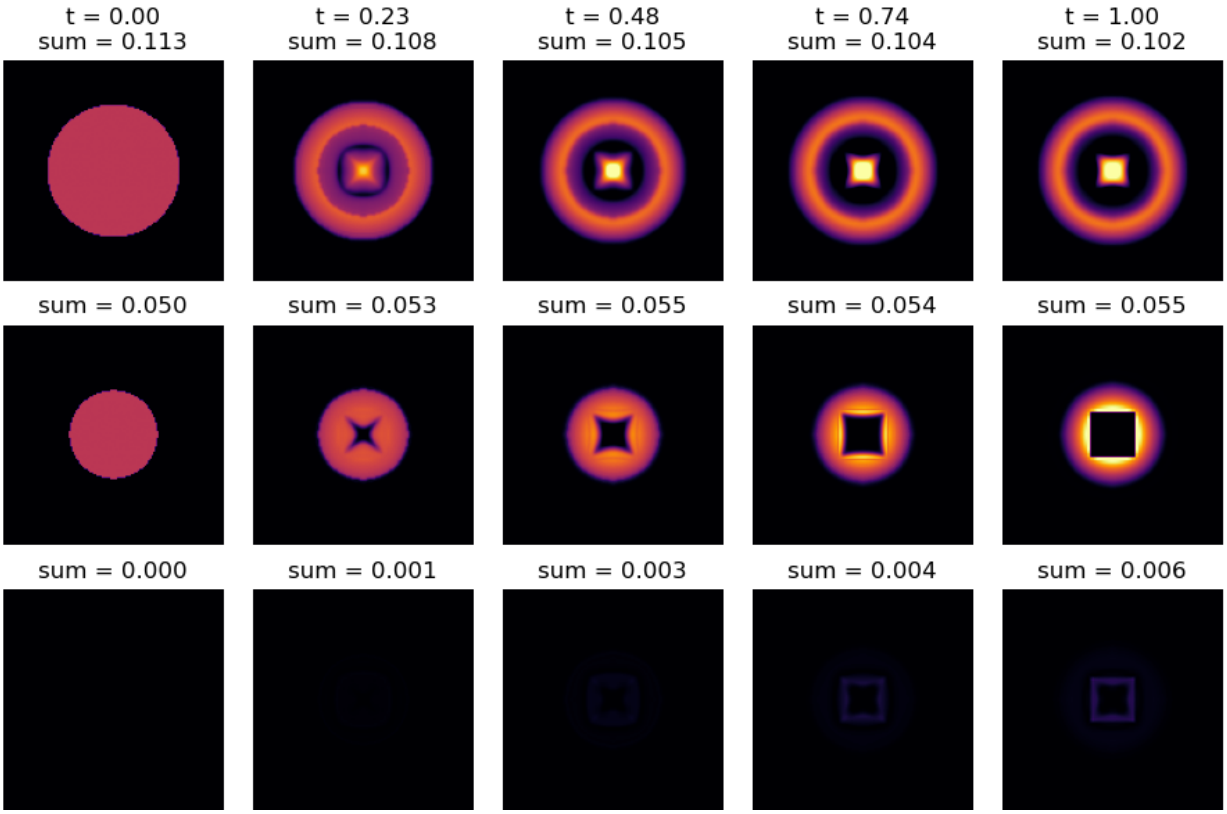


Figure 3.5: Experiment 3. The evolution of populations from $t = 0$ to $t = 1$ with $\beta = 0.96$ and $\gamma = 0.12$. The first row represents the susceptible population, the second row represents the infected population, and the last row represents the recovered population.

3.5 Discussion

In this paper, we introduce a mean-field control model for controlling the virus spreading of a population in a spatial domain, which extends the current SIR model with spatial effect.

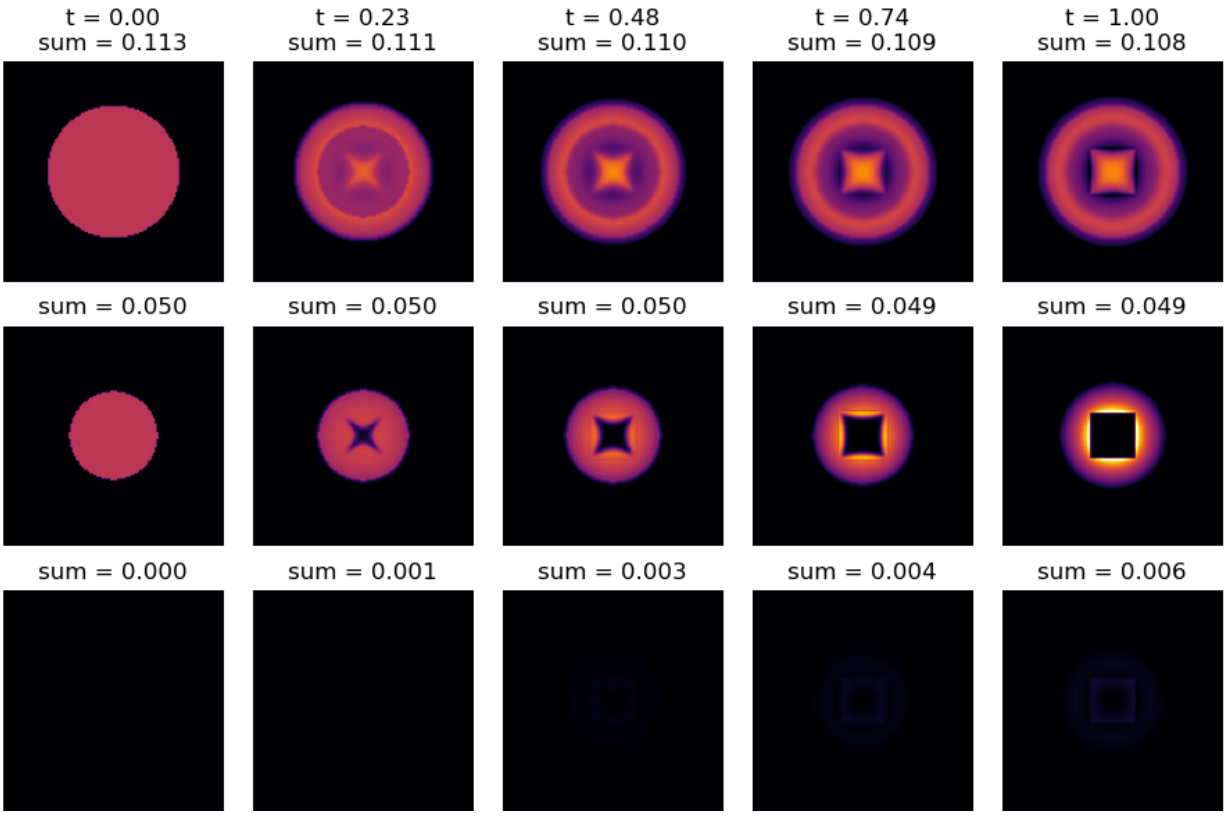


Figure 3.6: Experiment 3. The evolution of populations from $t = 0$ to $t = 1$ with $\beta = 0.34$ and $\gamma = 0.12$. The first row represents the susceptible population, the second row represents the infected population, and the last row represents the recovered population.

Here, the state variable represents the population status, such as S, I, R, with a spatial domain, while the control variable is the population's velocity of motion. The terminal cost forms government's goal, which balances the total infection number and maintains suitable physical movement of essential tasks and goods. Numerical algorithms are derived to solve the proposed model. Several experiments demonstrate that our model can effectively demonstrate how to separate the infected and susceptible population in a spatial domain.

Our model opens the door to many questions in modeling, inverse problems, and computations, especially during this COVID-19 pandemic. On the modeling side, first, we are interested in generalizing the geometry of the spatial domain. Second, our current model

only focuses on the control of population movement. The control of the diffusion operator among populations is also of great interest in future work. Third, the government can also put restrictions on the interaction for a different class of populations, depending on their infection status. Fourth, in real life, the spatial domain is often inhomogeneous, containing airports, schools, subways, etc. We also need to formulate our mean-field control model on a discrete spatial graph (network). Besides, our model focuses on the forward problem of modeling the virus's dynamics. In practice, real-time data is generated as a virus spreading across different regions. To effectively model this dynamic, a suitable inverse mean-field control problem needs to be constructed. On the computational side, our model involves a non-convex optimization problem, which comes from the multiplicative term of the SIR model itself. In future work, we expect to design a fast and reliable algorithm for these advanced models. We will develop and apply AI numerical algorithms to compute models in high dimensions.

3.6 Acknowledgments and Disclosure of Funding

The authors are supported by AFOSR MURI FA9550-18-1-0502.

CHAPTER 4

The Back-And-Forth Method For Wasserstein Gradient Flows

We present a method to efficiently compute Wasserstein gradient flows. Our approach is based on a generalization of the back-and-forth method (BFM) introduced in [JL20] to solve optimal transport problems. We evolve the gradient flow by solving the dual problem to the JKO scheme. In general, the dual problem is much better behaved than the primal problem. This allows us to efficiently run large scale gradient flows simulations for a large class of internal energies including singular and non-convex energies.

4.1 Introduction

In this work, we are interested in simulating the evolution of parabolic equations of the form

$$\begin{aligned} \partial_t \rho - \nabla \cdot (\rho \nabla \phi) &= 0, \\ \phi &= \delta U(\rho). \end{aligned} \tag{4.1}$$

Equation (4.1), often referred to as Darcy's law or the generalized porous medium equation, describes the evolution of a mass density ρ flowing along a pressure gradient $\nabla\phi$ generated by an internal energy functional U. This class of equations models various physical phenomena such as fluid flow, heat transfer, aggregation-diffusion, and crowd motion [Vaz07, San15]. In general, these equations are both stiff and non-linear making them challenging to solve numerically. For example, in the important special case where $U(\rho) = \frac{1}{m-1}\int \rho^m$ $(m > 1)$,

equation (4.1) becomes a non-linear version of the heat equation

$$\partial_t \rho - \Delta(\rho^m) = 0,$$

known as the porous medium equation (PME). When U is non-differentiable or non-convex, simulation of these equations becomes even more difficult. Thus, in this paper, our goal is to design a method that can efficiently and accurately simulate equation (4.1) for a wide variety of internal energies U.

Our approach to simulating Darcy's law is based on the celebrated interpretation of equation (4.1) as a gradient flow with respect to the Wasserstein metric [JKO98, Ott01]. This interpretation can be used to create a discrete-in-time approximation scheme known as the JKO scheme [JKO98]. The scheme constructs approximate solutions by iterating

$$\rho^{(n+1)} := \operatorname*{argmin}_{\rho} U(\rho) + \frac{1}{2\tau} W_2^2(\rho, \rho^{(n)}). \tag{4.2}$$

Here, τ plays the role of the time step in the scheme and $W_2(\cdot, \cdot)$ is the 2-Wasserstein metric from the theory of optimal transportation [San15] (see Section 4.2.1 for a brief overview of optimal transport and the 2-Wasserstein metric). Thanks to the variational structure of the scheme, the iterates are unconditionally energy stable and one can choose the time step τ independently from any spatial discretization. In addition, the JKO scheme retains many desirable properties of the continuum equation, such as comparison and contraction type principles [JKT20b, DMS16, AKY14].

In light of the many favorable properties of the JKO scheme, there have been many works devoted to the computation of minimizers for problem (4.2), see [BCW10, CM10, Pey15, BCM16, BCL16, CDP17, CCW19, CWX20, LMS20] to name just a few. Despite the amount of work on this problem, it remains a challenge to efficiently solve the JKO scheme at a high resolution. The main difficulty in solving problem (4.2) lies in the handling of the Wasserstein distance term. Indeed, there is not a simple formula that gives the variation of the Wasserstein distance with respect to the density ρ. As such, essentially all methods for

solving (4.2) are adaptations of algorithms for computing the Wasserstein distance between two fixed densities.

In this paper, we solve problem (4.2) by adapting the back-and-forth method (BFM) introduced in [JL20]. BFM is a state-of-the-art algorithm for computing optimal transport maps between two fixed densities. Instead of directly solving Monge's optimal transportation problem, BFM finds optimal maps by solving the associated Kantorovich dual problem. Building on this approach, rather than directly solving problem (4.2), we instead compute solutions to its dual problem. The dual problem is a concave maximization problem that produces the pressure variable at the next time step $\phi^{(n+1)}$. The optimal density variable can then easily be recovered from the pressure via the duality relation $\phi^{(n+1)} = \delta U(\rho^{(n+1)})$.

There are several advantages to solving the dual problem rather than the original primal problem. The pressure variable ϕ has better regularity than the density variable ρ. Indeed, at worst, the pressure gradient must be square integrable. As a result, the pressure is better suited to discrete approximation schemes. In addition, there is an explicit formula to compute derivatives of the dual functional, hence one can apply gradient ascent to solve the dual problem (the corresponding gradient descent scheme for the primal problem is much more difficult). Finally, the dual approach is very convenient when U encodes hard constraints (such as incompressibility of the density), as the dual problem will be unconstrained.

Leveraging the advantages of the dual problem to (4.2) and the special gradient ascent structure of BFM, we are able to rapidly and accurately solve the JKO scheme for a large class of internal energies U. We show that the algorithm increases the value of the dual problem at every step. In particular, this analysis holds even in cases where the Hessian of U is singular and our analysis has no dependence on the size of the computational grid. As a result, we are able to simulate equation (4.1) on a much larger scale than previous methods, and we are easily able to handle difficult cases like incompressible crowd motion models with obstacles and aggregation-diffusion equations.

4.1.1 Overall approach

The back-and-forth method for Wasserstein gradient flows is based on solving the dual problem associated to the JKO scheme. The starting point for this analysis is Kantorovich's dual formulation of optimal transport. Given two measures μ and ν, the dual formulation of the 2-Wasserstein distance is given by

$$\frac{1}{2\tau}W_2^2(\mu,\nu) = \sup_{(\phi,\psi)\in\mathcal{C}} \int_\Omega \psi(x)\, d\mu(x) - \int_\Omega \phi(y)\, d\nu(y), \tag{4.3}$$

where we maximize over the constraint

$$\mathcal{C} := \{(\phi,\psi) \in C(\Omega)\times C(\Omega) : \psi(x) - \phi(y) \le \frac{1}{2\tau}|x-y|^2\}.$$

Using the dual formulation of optimal transport, we can rewrite problem (4.2) as

$$\inf_\rho \sup_{(\phi,\psi)\in\mathcal{C}} U(\rho) + \int_\Omega \psi(x)\, d\rho^{(n)}(x) - \int_\Omega \phi(y)\, d\rho(y).$$

When U is convex, we can interchange the inf and sup to get an equivalent dual problem to (4.2):

$$\sup_{(\phi,\psi)\in\mathcal{C}} \int_\Omega \psi(x)\, d\rho^{(n)}(x) - U^*(\phi), \tag{4.4}$$

where U^* is the convex conjugate of U,

$$U^*(\phi) := \sup_\rho \int_\Omega \phi(y)\, d\rho(y) - U(\rho).$$

Problem (4.4) looks difficult due to the constraint encoded by $\mathcal{C}$. Nevertheless, there is a very convenient way to reformulate the problem. Because $\rho^{(n)}$ is a nonnegative measure, it is favorable to choose ψ to be pointwise as large as possible. If we fix ϕ, it then follows that the corresponding largest possible choice for ψ is given by

$$\phi^c(x) := \inf_{y\in\Omega} \phi(y) + \frac{1}{2\tau}|x-y|^2. \tag{4.5}$$

Conversely, U^* is increasing with respect to ϕ (see Section 4.2.1), therefore, we would like to choose ϕ to be pointwise as small as possible. Thus, if we fix ψ, then the corresponding

smallest choice for ϕ is given by

$$\psi^{\bar{c}}(y) := \sup_{x \in \Omega} \psi(x) - \frac{1}{2\tau}|x - y|^2. \tag{4.6}$$

Formulas (4.5) and (4.6) are known as the backward-c-transform and forward-c-transform respectively. These transforms play an essential role in optimal transport and are integral to our method. Crucially, we can use these transforms to eliminate the constraint $\mathcal{C}$ and either one of the variables ϕ or ψ. More explicitly, problem (4.4) is equivalent to maximizing either one of the following two *unconstrained* functionals:

$$J(\phi) := \int_\Omega \phi^c(x)\, d\rho^{(n)}(x) - U^*(\phi), \tag{4.7}$$

$$I(\psi) := \int_\Omega \psi(x)\, d\rho^{(n)}(x) - U^*(\psi^{\bar{c}}). \tag{4.8}$$

Indeed, if ϕ_* is a maximizer of J and ψ_* is a maximizer of I, then we must have the relations

$$\phi_*^c = \psi_*, \quad \psi_*^{\bar{c}} = \phi_*,$$

and (ϕ_*, ψ_*) is a maximizer of (4.4). The reformulations I and J genuinely simplify the task of finding maximizers. On a regular discrete grid, the c-transform can be computed very efficiently [Luc97, JL20]. As a result, it is much more tractable to maximize I and J, rather than trying work with (4.4) directly.

We will find the maximizers ϕ_* and ψ_* by building upon the BFM algorithm introduced in [JL20]. The original BFM gives a very efficient scheme for finding the maximizers in the special case where U^* is a linear functional. Rather than focusing on either I or J, BFM simultaneously maximizes both functionals. The method proceeds by hopping back-and-forth between gradient ascent updates on J in ϕ-space and gradient ascent updates on I in ψ-space (hence the name). In between gradient steps, information in one space (ϕ-space or ψ-space) is propagated back to the other by taking a forward/backward c-transform. As noted in [JL20], the advantage of the back-and-forth approach is that certain features of the optimal solution pair (ϕ_*, ψ_*) may be easier to build in one space compared to the other. As a result, the

back-and-forth method converges far more rapidly than vanilla gradient ascent methods that operate only on ϕ-space or only on ψ-space.

In order to generalize BFM to the Wasserstein gradient flow case, we need to be able to guarantee the stability of gradient ascent steps on (4.7) and (4.8) when U^* is nonlinear. In fact, for many important cases, the Hessian of U^* may have a singular component. To overcome this difficulty, we perform the gradient ascent steps in an appropriately weighted Sobolev space. The Sobolev control allows us to use Stokes' Theorem to convert boundary integrals into integrals over the full space, thus taming the singularities of U^* (see Section 4.3.2). As a result of this continuous analysis, the discretized scheme will have a convergence rate that is independent of the grid size. The back-and-forth method is summarized in Algorithm 5, where H is the aforementioned weighted Sobolev space.

Algorithm 5 The back-and-forth scheme for solving (4.4)

Given $\rho^{(n)}$ and ϕ_0, iterate:

$$\begin{aligned}
\phi_{k+\frac{1}{2}} &= \phi_k + \nabla_H J(\phi_k) \\
\psi_{k+\frac{1}{2}} &= (\phi_{k+\frac{1}{2}})^c \\
\psi_{k+1} &= \psi_{k+\frac{1}{2}} + \nabla_H I(\psi_{k+\frac{1}{2}}) \\
\phi_{k+1} &= (\psi_{k+1})^{\bar{c}}
\end{aligned}$$

Once we have solved the dual problem, we can recover the solution to the original problem (4.2). If U is convex, then the optimal dual variable ϕ_* is related to $\rho^{(n+1)}$ through the duality relation $\rho^{(n+1)} = \delta U^*(\phi_*)$ (see Theorem 4.2.7 in Section 4.2.2). When U is not convex, the connection between (4.2) and the dual problem becomes more tenuous. Luckily, we can circumvent this difficulty using a convexity splitting scheme [Eyr98]. Indeed, if we write $U = U_1 + U_0$ where U_1 is convex and U_0 is concave, then we can replace the JKO scheme (4.2) with the modified scheme

$$\rho^{(n+1)} = \operatorname*{argmin}_{\rho} U_1(\rho) + U_0(\rho^{(n)}) + (\delta U_0(\rho^{(n)}), \rho - \rho^{(n)}) + \frac{1}{2\tau} W_2^2(\rho, \rho^{(n)}). \tag{4.9}$$

It is well-known that convexity splitting retains the energy stability of a fully implicit scheme. Crucially, the energy term $U_1(\rho) + U_0(\rho^{(n)}) + (\delta U_0(\rho^{(n)}), \rho - \rho^{(n)})$ in (4.9) is a convex function of the variable ρ, and thus, we can apply the duality approach. All together, our method gives an extremely rapid way to simulate the PDE (4.1) even when U is non-convex or irregular.

The remainder of the paper is organized as follows. In Section 4.2, we review important background information on optimal transport, convex analysis, and optimization. In Section 4.3, we present the back-and-forth algorithm and explain how to guarantee stability and choose step sizes. Lastly, in Section 4.4, we demonstrate the accuracy, speed, and versatility of the algorithm through a wide suite of numerical experiments. In particular, our experiments include many cases that are well-known to be numerically challenging.

4.2 Background

In this section, we will rigorously establish the connection between the primal and dual formulations of the JKO scheme. Furthermore, we will review key concepts from optimal transport and convex analysis that are needed to compute the gradients $\nabla_H J, \nabla_H I$ and establish stability of Algorithm 5. Note that throughout the paper we shall assume that $\Omega \subset \mathbb{R}^d$ is a bounded open set.

4.2.1 The c-transform and optimal transport

Throughout this section the space of continuous functions over Ω will be denoted by $C(\Omega)$.

Definition 4.2.1. Given $\phi \in C(\Omega)$ its *backward c-transform* is

$$\phi^c(x) := \inf_{y \in \Omega} \phi(y) + \frac{1}{2\tau}|x - y|^2.$$

Given $\psi \in C(\Omega)$ its *forward c-transform* is

$$\psi^{\bar{c}}(y) := \sup_{x \in \Omega} \psi(x) - \frac{1}{2\tau}|x - y|^2.$$

Lemma 4.2.1 ([San15])**.** *Given $\phi, \psi \in C(\Omega)$, we have*

$$\phi^{c\bar{c}} \leq \phi, \quad \psi \leq \psi^{\bar{c}c},$$

and

$$\phi^{c\bar{c}c} = \phi^c, \quad \psi^{\bar{c}c\bar{c}} = \psi^{\bar{c}}.$$

Definition 4.2.2. Given $\phi, \psi \in C(\Omega)$, we say that ϕ is *c-convex* if $\phi^{c\bar{c}} = \phi$ and we say that ψ is *c-concave* if $\psi^{\bar{c}c} = \psi$. Furthermore, if $\phi^c = \psi$ and $\psi^{\bar{c}} = \phi$, then we say the pair (ϕ, ψ) is *c-conjugate*.

The following two propositions establish the fundamental relationship between optimal transport and the c-transform.

Proposition 4.2.2 ([Gan94, Gan95b, GM96])**.** *If $\phi\colon \Omega \to \mathbb{R}$ is c-convex and $\psi\colon \Omega \to \mathbb{R}$ is c-concave, then the maps*

$$T_\phi(x) := \operatorname*{argmin}_{y \in \Omega} \phi(y) + \frac{1}{2\tau}|x - y|^2 \tag{4.10}$$

and

$$S_\psi(y) := \operatorname*{argmax}_{x \in \Omega} \psi(x) - \frac{1}{2\tau}|x - y|^2 \tag{4.11}$$

are well-defined and unique almost everywhere. Furthermore, if $u \in C(\Omega)$, then for almost every $x, y \in \Omega$ we have the following perturbation formulas for the c-transform

$$\lim_{t \to 0^+} \frac{(\phi + tu)^c(x) - \phi^c(x)}{t} = u(T_\phi(x)), \tag{4.12}$$

$$\lim_{t \to 0^+} \frac{(\psi + tu)^{\bar{c}}(y) - \psi^{\bar{c}}(y)}{t} = u(S_\psi(y)). \tag{4.13}$$

Finally, if ϕ and ψ are c-conjugate, then

$$S_\psi(y) = y + \tau \nabla \phi(y),$$

$$T_\phi(x) = x - \tau \nabla \psi(x),$$

and $T_\phi\big(S_\psi(y)\big) = y$, $S_\psi\big(T_\phi(x)\big) = x$ almost everywhere.

Proposition 4.2.3 ([San15]). *If $\mu, \nu \in L^1(\Omega)$ are nonnegative densities with the same mass, then*

$$\frac{1}{2\tau}W_2^2(\mu,\nu) = \sup_{\phi \in C(\Omega)} \int_\Omega \phi^c(x)\,\mu(x)dx - \int_\Omega \phi(y)\,\nu(y)dy,$$
$$\frac{1}{2\tau}W_2^2(\mu,\nu) = \sup_{\psi \in C(\Omega)} \int_\Omega \psi(x)\,\mu(x)dx - \int_\Omega \psi^{\bar{c}}(y)\,\nu(y)dy.$$

Now we can state the fundamental result that guarantees the existence and uniqueness of the optimal transport maps.

Theorem 4.2.4 ([Bre91, Gan95a, GM96]). *If $\mu, \nu \in L^1(\Omega)$ are nonnegative densities with the same mass, then there exists a c-conjugate pair (ϕ_*, ψ_*) such that*

$$\phi_* \in \operatorname*{argmax}_{\phi \in C(\Omega)} \int_\Omega \phi^c(x)\,\mu(x)dx - \int_\Omega \phi(y)\,\nu(y)dy,$$
$$\psi_* \in \operatorname*{argmax}_{\psi \in C(\Omega)} \int_\Omega \psi(x)\,\mu(x)dx - \int_\Omega \psi^{\bar{c}}(y)\,\nu(y)dy,$$
$$\frac{1}{2\tau}W_2^2(\mu,\nu) = \int_\Omega \psi_*(x)\,\mu(x)dx - \int_\Omega \phi_*(y)\,\nu(y)dy,$$

and T_{ϕ_*}, S_{ψ_*} are the unique optimal transport maps sending μ to ν and ν to μ respectively, i.e. $T_{\phi_*}{}_\#\mu = \nu$ and $S_{\psi_*}{}_\#\nu = \mu$.

4.2.2 Convex duality

Now that we have developed the basics of optimal transport, we are ready to return to the JKO scheme. To iterate the JKO scheme, one must be able to solve generalized optimal transport (GOT) problems of the form

$$\rho_* = \operatorname*{argmin}_{\rho \in L^1(\Omega)} U(\rho) + \frac{1}{2\tau}W_2^2(\rho,\mu), \tag{4.14}$$

where $\mu \in L^1(\Omega)$ is a given nonnegative density. Our method solves the GOT problem by appealing to its dual formulation. In the rest of this subsection, we shall derive the dual problem and develop its basic properties. To obtain a well-behaved dual problem, we shall need the following assumptions on the energy U.

Assumption 1. The internal energy is given by a proper, convex, and lower semicontinuous functional $U\colon L^1(\Omega) \to \mathbb{R} \cup \{+\infty\}$ such that $U(\rho) = \infty$ if ρ is negative on a set of positive measure.

Assumption 2. There exists a function $s\colon \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ with superlinear growth such that

$$U(\rho) \geq \int_\Omega s(\rho(y))\, dy.$$

Remark 4.2.1. Assumption 1 encodes the fact that ρ must be a nonnegative density, while Assumption 2 guarantees that for each $B \in \mathbb{R}$ the sets $\{\rho \in L^1(\Omega) : U(\rho) < B\}$ are weakly compact.

Remark 4.2.2. Except for the convexity requirement, Assumptions 1 and 2 are very natural in the context of Wasserstein gradient flows. Note that we will eventually consider non-convex U in Section 4.3.3.

At the heart of duality is the notion of convex conjugation.

Definition 4.2.3. Given a functional $U\colon L^1(\Omega) \to \mathbb{R}$ its convex conjugate $U^*\colon L^\infty(\Omega) \to \mathbb{R}$ is defined by

$$U^*(\phi) := \sup_{\rho \in L^1(\Omega)} \int_\Omega \phi(x)\rho(x)\, dx - U(\rho),$$

Thanks to Assumption 1, U^* possess an important monotonicity property.

Lemma 4.2.5. *U^* is monotonically increasing, i.e. if $\phi_0, \phi_1\colon \Omega \to \mathbb{R}$ are functions such that $\phi_0 \leq \phi_1$ pointwise everywhere, then*

$$U^*(\phi_0) \leq U^*(\phi_1).$$

Proof. By Assumption 1 the internal energy is finite only over nonnegative densities, thus,

$$U^*(\phi) = \sup_{\rho \geq 0} \int_\Omega \phi(x)\, \rho(x) dx - U(\rho).$$

If we take some $\rho \in L^1(\Omega)$, with $\rho(x) \geq 0$ a.e., then we have

$$\int_\Omega \phi_0(x)\,\rho(x)dx - U(\rho) \leq \int_\Omega \phi_1(x)\,\rho(x)dx - U(\rho).$$

Taking a supremum over $\rho \geq 0$ finishes the proof. □

Now we are ready to reintroduce the twin dual functionals I and J.

Proposition 4.2.6. *Fix a nonnegative density $\mu \in L^1(\Omega)$. The functionals I, J given by*

$$J(\phi) := \int_\Omega \phi^c(x)\,\mu(x)dx - U^*(\phi)$$

$$I(\psi) := \int_\Omega \psi(x)\,\mu(x)dx - U^*(\psi^{\bar{c}}),$$

are proper, weakly upper semicontinuous, concave and $\sup_{\phi \in C(\Omega)} J(\phi) = \sup_{\psi \in C(\Omega)} I(\psi)$. Furthermore, if ϕ is c-convex and ψ is c-concave, then J and I have first variations

$$\delta J(\phi) = T_{\phi\#}\mu - \delta U^*(\phi),$$

$$\delta I(\psi) = \mu - S_{\psi\#}\delta U^*(\psi^{\bar{c}}),$$

where δU^* is the first variation of U^*.

Proof. Following the logic in the proof of Lemma 4.2.5, we may write

$$U^*(\psi^{\bar{c}}) = \sup_{\rho \geq 0} \int_\Omega \psi^{\bar{c}}(y)\,\rho(y)dy - U(\rho).$$

Next, let $\mathcal{M}(\Omega \times \Omega)$ denote the space of nonnegative measures on $\Omega \times \Omega$, and for any given density $\rho \geq 0$ define

$$\Pi(\rho) := \left\{ \pi \in \mathcal{M}(\Omega \times \Omega) : \iint_{\Omega\times\Omega} f(y)\,d\pi(x,y) = \int_\Omega f(y)\,\rho(y)dy \text{ for all } f \in C(\Omega) \right\}.$$

Using the definition of the c-transform, we can then write

$$\int_\Omega \psi^{\bar{c}}(y)\,\rho(y)dy = \sup_{\pi \in \Pi(\rho)} \iint_{\Omega\times\Omega} \left(\psi(x) - \frac{1}{2\tau}|x-y|^2\right) d\pi(x,y).$$

Therefore, we have

$$-U^*(\psi^{\bar{c}}) = \inf_{\rho \geq 0} \inf_{\pi \in \Pi(\rho)} U(\rho) - \iint_{\Omega \times \Omega} \left(\psi(x) - \frac{1}{2\tau}|x-y|^2\right) d\pi(x,y).$$

Now it is clear that I can be written as the infimum over a family of linear functionals of ψ. Hence, I must be proper, concave and weakly upper semicontinuous. An essentially identical argument applies to J.

Since U^* is monotonically increasing, Lemma 4.2.5 implies that for any $\phi, \psi \in C(\Omega)$

$$J(\phi) \leq I(\phi^c), \quad I(\psi) \leq J(\psi^{\bar{c}}).$$

Therefore, we must have

$$\sup_{\psi \in C(\Omega)} I(\psi) = \sup_{\phi \in C(\Omega)} J(\phi).$$

When ϕ and ψ are c-convex/concave respectively, the formulas for the first variations follow directly from Proposition 4.2.2. □

Finally, we conclude this subsection by stating the essential result linking the primal and dual generalized optimal transport problems. Crucially, this shows how to recover the solution to (4.14) from the maximizers of I and J.

Theorem 4.2.7 ([JKT20b]). *If $\mu \in L^1(\Omega)$, U satisfies Assumptions 1, 2, and $\delta U(\mu)$ is not a constant function, then there exists a unique density ρ_* and a pair of c-conjugate functions (ϕ_*, ψ_*) such that*

$$\rho_* = \operatorname*{argmin}_{\rho \in L^1(\Omega)} U(\rho) + \frac{1}{2\tau} W_2^2(\rho, \mu), \quad \phi_* \in \operatorname*{argmax}_{\phi \in C(\Omega)} J(\phi), \quad \psi_* \in \operatorname*{argmax}_{\psi \in C(\Omega)} I(\psi),$$

$$U(\rho_*) + \frac{1}{2\tau} W_2^2(\rho_*, \mu) = J(\phi_*) = I(\psi_*),$$

$$\rho_* \in \delta U^*(\phi_*), \quad \phi_* \in \delta U(\rho_*), \quad \rho_* = T_{\phi_*} \# \mu.$$

Remark 4.2.3. Note that if $\delta U(\mu)$ is constant, then $\mu = \operatorname{argmin}_{\rho \in L^1(\Omega)} U(\rho) + \frac{1}{2\tau} W_2^2(\rho, \mu)$. Thus, the excluded case is trivial.

4.2.3 Concave gradient ascent

Now that we see how to link the JKO scheme to the dual functionals I and J, it remains to develop a method to find the maximizers of I and J. To that end, in this subsection, we review classical unconstrained gradient ascent. Let us first recall the notion of *gradient*. This will require the structure of a real Hilbert space $\mathcal{H}$ with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and norm $\|\cdot\|_{\mathcal{H}}$.

Definition 4.2.4. Given a point $\varphi \in \mathcal{H}$, we say that a bounded linear map $\delta F(\varphi) \colon \mathcal{H} \to \mathbb{R}$ is the first variation (Fréchet derivative) of F at φ if

$$\lim_{\|h\|_{\mathcal{H}} \to 0} \frac{\|F(\varphi + h) - F(\varphi) - \delta F(\varphi)(h)\|_{\mathcal{H}}}{\|h\|_{\mathcal{H}}} = 0.$$

Definition 4.2.5. We say that a map $\nabla_{\mathcal{H}} F \colon \mathcal{H} \to \mathcal{H}$ is the $\mathcal{H}$-gradient of F (or simply gradient if there is no ambiguity about the space $\mathcal{H}$) if

$$\langle \nabla_{\mathcal{H}} F(\varphi), h \rangle_{\mathcal{H}} = \delta F(\varphi)(h)$$

for all $(\varphi, h) \in \mathcal{H} \times \mathcal{H}$.

The above identity highlights that gradients are intimately linked to the inner product of the Hilbert space, in contrast to first variations. Indeed, note that one can define the notion of a first variation over any normed vector space, while the notion of a gradient requires an inner product.

Gradient ascent method

Given a concave functional J over $\mathcal{H}$, consider the gradient ascent iterations

$$\phi_{k+1} = \phi_k + \nabla_{\mathcal{H}} J(\phi_k). \tag{4.15}$$

The gradient ascent scheme (4.15) can equivalently be written in the variational form

$$\phi_{k+1} = \operatorname*{argmax}_{\phi} J(\phi_k) + \delta J(\phi_k)(\phi - \phi_k) - \frac{1}{2}\|\phi - \phi_k\|_{\mathcal{H}}^2. \tag{4.16}$$

Note that equations (4.15) and (4.16) typically include a step size parameter that controls how far one travels in the gradient direction. For reasons that will become clear shortly (see equation (4.20) and the subsequent discussion), we prefer to incorporate any parameters into the norm $\|\cdot\|_{\mathcal{H}}$ itself.

In order to obtain convergence of the scheme

$$J(\phi_k) \xrightarrow[k\to\infty]{} \sup_{\phi} J(\phi),$$

with an efficient rate, it is essential to choose the norm $\|\cdot\|_{\mathcal{H}}$ properly. If the norm is too weak, then the algorithm may become unstable and fail to converge. On the other hand, if the norm is too strong, then very little change happens at each step and the algorithm converges slowly. The following theorem, one of the cornerstones of optimization, explains how to balance these competing considerations.

Theorem 4.2.8 ([Nes13])**.** *Let $J\colon \mathcal{H} \to \mathbb{R}$ be a twice Fréchet-differentiable concave functional with maximizer ϕ^*. If*

$$-\delta^2 J(\phi)(h,h) \leq \|h\|_{\mathcal{H}}^2, \tag{4.17}$$

for all $\phi, h \in \mathcal{H}$ (J is said to be "1-smooth"), then the gradient ascent scheme

$$\phi_{k+1} = \phi_k + \nabla_{\mathcal{H}} J(\phi_k)$$

starting at a point ϕ_0 satisfies the ascent property

$$J(\phi_{k+1}) \geq J(\phi_k) + \frac{1}{2}\|\nabla_{\mathcal{H}} J(\phi_k)\|_{\mathcal{H}}^2, \tag{4.18}$$

and has the convergence rate

$$J(\phi^*) - J(\phi_k) \leq \frac{\|\phi^* - \phi_0\|_{\mathcal{H}}^2}{2k}. \tag{4.19}$$

From Theorem 4.2.8, we can again see the competing interests of weakening or strengthening the norm $\|\cdot\|_{\mathcal{H}}$. A stronger norm makes it easier to satisfy equation (4.17), while a weaker norm gives a better convergence rate in (4.19). Putting these considerations together, we see that it is optimal to choose the weakest possible norm such that (4.17) holds.

Sobolev norm

Let Ω be an open bounded convex subset of $\mathbb{R}^d$. Our gradient ascent schemes use a norm H based on the Sobolev space $H^1(\Omega)$. For two constants $\Theta_1 > 0$ and $\Theta_2 > 0$ we define

$$\|h\|_H^2 = \int_\Omega \Theta_2 |\nabla h(x)|^2 + \Theta_1 |h(x)|^2 \, dx. \tag{4.20}$$

The precise value of Θ_1 and Θ_2 will depend on the functional being maximized (see for instance Theorem 4.3.3 in Section 4.3). In many instances, it will be optimal to choose Θ_1 and Θ_2 to have rather different values. For this reason, we do not wish to reduce these parameters to a single step size value. The next lemma describes how to compute gradients with respect to this inner product.

Lemma 4.2.9. *Suppose that $F = F(\phi)$ is a Fréchet-differentiable functional such that for any ϕ the first variation $\delta F(\phi)$ evaluated at any point h can be written as integration against a function f_ϕ, i.e.*

$$\delta F(\phi)(h) = \int_\Omega h(x) f_\phi(x) \, dx.$$

Define $\|\cdot\|_H$ by (4.20). Then the H-gradient of F can be written

$$\nabla_H F(\phi) = (\Theta_1 \operatorname{Id} - \Theta_2 \Delta)^{-1} f_\phi,$$

where Id *is the identity operator and Δ is the Laplacian operator, taken together with zero Neumann boundary conditions.*

Proof. Fix ϕ and consider the unique solution to the elliptic equation

$$\begin{cases} (\Theta_1 \operatorname{Id} - \Theta_2 \Delta) g & = f_\phi \quad \text{in } \Omega, \\ n \cdot \nabla g & = 0 \quad \text{on } \partial\Omega. \end{cases}$$

Then we have the chain of equalities

$$\begin{aligned}\delta F(\phi)(h) &= \int_\Omega h(x) f_\phi(x)\,dx \\ &= \int_\Omega h(x)(\Theta_1 \operatorname{Id} - \Theta_2 \Delta) g(x)\,dx \\ &= \int_\Omega \Theta_1 h(x) g(x) + \Theta_2 \nabla h(x) \cdot \nabla g(x)\,dx \\ &= \langle h, g \rangle_H .\end{aligned}$$

This shows that g is the H-gradient of F. □

The above result can be restated as follows: the H-gradient of F is obtained by "preconditioning" δF with the inverse operator $(\Theta_1 \operatorname{Id} - \Theta_2 \Delta)^{-1}$.

4.3 The back-and-forth method

Our goal is to develop an efficient algorithm for solving the JKO scheme for a large class of interesting energies U. We begin in Section 4.3.1 with the case where U is convex with respect to ρ. In this case, the JKO scheme has an equivalent dual problem that we solve using an adaptation of the back-and-forth method from [JL20]. In Section 4.3.2, we show that the algorithm is gradient stable in a properly weighted H^1 space for convex energies of the form

$$U(\rho) = \int_\Omega u_m(\rho(x)) + V(x)\rho(x)\,dx,$$

where $V \colon \Omega \to [0, +\infty]$ is a fixed function, and

$$u_m(\rho) = \begin{cases} \frac{\gamma}{m-1}(\rho^m - \rho) & \text{if } \rho \geq 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{4.21}$$

for some constants $\gamma > 0$ and $m > 1$. We shall also consider the two limiting cases $m \to 1$ and $m \to \infty$. Let us note that our analysis can be extended to more general functionals, however, we focus on the (important) special case above for clarity of exposition. After we

have developed the method for convex energy functionals U, in Section 4.3.3 we show how to generalize the algorithm for non-convex U.

4.3.1 The back-and-forth method for convex U

To iterate the JKO scheme, we must be able to solve the generalized optimal transport (GOT) problem

$$\rho_* = \operatorname*{argmin}_{\rho \in L^1(\Omega)} U(\rho) + \frac{1}{2\tau} W_2^2(\rho, \mu), \tag{4.22}$$

for any fixed nonnegative density $\mu \in L^1(\Omega)$. As we saw in Section 4.2 (see Theorem 4.2.7), when U is convex, the generalized optimal transport problem is in duality with the twin functionals I and J, i.e.

$$\inf_{\rho \in L^1(\Omega)} U(\rho) + \frac{1}{2\tau} W_2^2(\rho, \mu) = \sup_{\phi} J(\phi) = \sup_{\psi} I(\psi).$$

Recall I and J are given by

$$J(\phi) = \int_\Omega \phi^c(x)\, \mu(x) dx - U^*(\phi), \tag{4.23}$$

$$I(\psi) = \int_\Omega \psi(x)\, \mu(x) dx - U^*(\psi^{\bar{c}}). \tag{4.24}$$

Furthermore, the minimizer ρ_* of problem (4.22) is related to the maximizers ϕ_*, ψ_* through the relations

$$\rho_* = T_{\phi_*} \# \mu, \quad \rho_* \in \delta U^*(\phi_*), \quad \phi_*^c = \psi_*. \tag{4.25}$$

Both I and J are unconstrained concave functionals (see Proposition 4.2.6), therefore, it is now clear that one can find the maximizer of either functional via standard gradient ascent methods. On the other hand, choosing to work with solely I or solely J breaks the symmetry of the problem. Thus, rather than focusing on only one of the functionals, the back-and-forth method performs alternating gradient ascent steps on I and J. Although I and J use different variables, we can switch between ϕ and ψ by using the c-transform. As

noted in [JL20], the alternating steps on I and J substantially accelerate the convergence rate of the method beyond standard gradient ascent.

We are now ready to introduce our approach to find the twin dual maximizers (ϕ_*, ψ_*) to problem (4.22). The method is outlined in Algorithm 6 and is based on two main ideas:

1. A back-and-forth update scheme, alternating between gradient ascent steps on I and J.

2. Gradient ascent steps in an H^1-type norm H, with

$$\nabla_H J(\phi) = (\Theta_1 \operatorname{Id} - \Theta_2 \Delta)^{-1}\Big[T_{\phi\,\#}\mu - \delta U^*(\phi)\Big],$$
$$\nabla_H I(\psi) = (\Theta_1 \operatorname{Id} - \Theta_2 \Delta)^{-1}\Big[\mu - S_{\psi\,\#}(\delta U^*(\psi^{\bar{c}}))\Big].$$

Algorithm 6 The back-and-forth scheme for solving (4.23) and (4.24)

Given μ and ϕ_0, iterate:

$$\phi_{k+\frac{1}{2}} = \phi_k + \nabla_H J(\phi_k)$$
$$\psi_{k+\frac{1}{2}} = (\phi_{k+\frac{1}{2}})^c$$
$$\psi_{k+1} = \psi_{k+\frac{1}{2}} + \nabla_H I(\psi_{k+\frac{1}{2}})$$
$$\phi_{k+1} = (\psi_{k+1})^{\bar{c}}$$

Our ultimate goal is to show that each step of Algorithm 6 increases the value of the functionals J and I. Thanks to Lemmas 4.2.1 and 4.2.5 it is easy to check that

$$J(\phi_{k+\frac{1}{2}}) \leq I((\phi_{k+\frac{1}{2}})^c), \quad I(\psi_{k+1}) \leq J((\psi_{k+1})^{\bar{c}}).$$

Thus, we see that the alternating steps where we switch between the ϕ and ψ variables can only increase the values of the dual problems. To show that the gradient steps $\phi_{k+\frac{1}{2}} = \phi_k + \nabla_H J(\phi_k)$ and $\psi_{k+1} = \psi_{k+\frac{1}{2}} + \nabla_H I(\psi_{k+\frac{1}{2}})$ increase the values of J and I respectively requires a more detailed analysis, which will be the main focus of Section 4.3.2. As we shall see, the enhanced stability provided by the H^1 preconditioner $(\Theta_1 \operatorname{Id} - \Theta_2\Delta)^{-1}$ will be essential to ensure that the gradient steps have the ascent property.

Once the dual problems I and J have been solved to sufficient accuracy, one can recover the optimal density ρ_* in (4.22) through the duality relations in (4.25). In certain examples, such as incompressible flows, the subdifferential δU^* may be multivalued. When this happens, the relation $\rho_* \in \delta U^*(\phi_*)$ does not uniquely define ρ_*. However, in practice, δU^* is typically only multivalued on a single level set of ϕ_* which has zero measure. As a result, for numerical purposes, we can simply identify $\rho_* = \delta U^*(\phi_*)$. Note that it is advantageous to recover ρ_* in this way as opposed to the pushforward relation $\rho_* = T_{\phi_*}{}_{\#}\mu$. Indeed, the formula $\rho_* = T_{\phi_*}{}_{\#}\mu$ requires the computation of numerical derivatives of ϕ_*, while the duality relation $\rho_* \in \delta U^*(\phi_*)$ is derivative free.

Combining our work, we obtain an algorithm for evolving the JKO scheme.

Algorithm 7 Running the JKO scheme

Given initial data $\rho^{(0)}$, initialize $\phi^{(0)} = \delta U(\rho^{(0)})$.

for $n = 0, \dots, N$ **do**

  $\phi^{(n+1)} \leftarrow$Run Algorithm 6 with $\mu = \rho^{(n)}$ and $\phi_0 = \phi^{(n)}$.

  $\rho^{(n+1)} = \delta U^*(\phi^{(n+1)})$.

end

4.3.2 H^1 gradient ascent

In order to ensure stability of the gradient ascent steps, the gradients of I and J are computed in a metric based on the H^1 Sobolev norm. Given two constants $\Theta_1 > 0$, $\Theta_2 > 0$, we define the Hilbert norm H by

$$\|h\|_H^2 = \int_\Omega \Theta_2 |\nabla h(x)|^2 + \Theta_1 |h(x)|^2 \, dx. \tag{4.26}$$

The main steps of the back-and-forth scheme are the gradient ascent steps in the in the H norm

$$\phi_{k+\frac{1}{2}} = \phi_k + \nabla_H J(\phi_k)$$

and

$$\psi_{k+1} = \psi_{k+\frac{1}{2}} + \nabla_H I(\psi_{k+\frac{1}{2}}).$$

In order to obtain convergence of our method, we want these steps to increase the values of the concave functionals J and I respectively. The so-called gradient ascent property

$$\begin{aligned} J(\phi_{k+\frac{1}{2}}) - J(\phi_k) &\geq \frac{1}{2}\|\nabla_H J(\phi_k)\|_H^2, \\ I(\psi_{k+1}) - I(\psi_{k+\frac{1}{2}}) &\geq \frac{1}{2}\|\nabla_H I(\psi_{k+\frac{1}{2}})\|_H^2, \end{aligned}$$

can be obtained when the Hessian bounds

$$\begin{aligned} -\delta^2 J(\phi)(h,h) &\leq \|h\|_H^2, \\ -\delta^2 I(\psi)(h,h) &\leq \|h\|_H^2 \end{aligned} \tag{4.27}$$

are satisfied (c.f. Theorem 4.2.8 in Section 4.2.3). When (4.27) holds, I and J are said to be "1-smooth" with respect to H.

We shall devote the rest of this subsection to obtaining inequalities of the form (4.27). Specifically, we shall show how to choose the constants Θ_1 and Θ_2 in equation (4.26) to ensure that I and J are 1-smooth (under regularity assumptions on ϕ and ψ) when U has the form

$$U(\rho) = \int_\Omega u_m(\rho(x))\,dx + \int_\Omega V(x)\rho(x)\,dx, \tag{4.28}$$

where u_m is defined in (4.21) and $V\colon \Omega \to [0, +\infty]$ is some given function.

Crucially, we will give upper bounds on Θ_1 and Θ_2 that can be efficiently computed from the data. Obtaining tight bounds for Θ_1 and Θ_2 is important as they essentially control the step size of the algorithm (note that small values of Θ_1 and Θ_2 correspond to large gradient steps). As we explained in Section 4.2.3, it is optimal to choose the smallest values of Θ_1 and Θ_2 such that (4.27) holds. This analysis is actually practical, as our numerical experiments confirm that the convergence of BFM can be substantially accelerated by making good choices for Θ_1 and Θ_2.

Those who are interested in the analysis of these bounds can continue reading this section, otherwise, one can immediately jump to the statements of Theorems 4.3.3 and 4.3.4, which give approximately optimal values of Θ_1 and Θ_2 for the functionals I and J.

4.3.2.1 Hessian bound analysis

It turns out that the Hessian bound analysis is nearly identical for I and J. Therefore, we will primarily focus on the analysis for J, and we will later explain how to deal with I in a similar fashion. To obtain Hessian bounds on $J(\phi) = \int_\Omega \phi^c \mu - U^*(\phi)$, we first derive bounds on the c-transform term

$$F(\phi) := \int_\Omega \phi^c(x)\, \mu(x) dx, \tag{4.29}$$

and then on the internal energy term $U^*(\phi)$. Let us begin by providing an expression for $\delta^2 F(\phi)$, the Hessian of F at a point ϕ that is c-convex.

Lemma 4.3.1 (Hessian bounds on the c-transform)**.** *Let F be the functional defined in (4.29). If ϕ is a c-convex function, then the Hessian of F at ϕ can be written as*

$$\delta^2 F(\phi)(h,h) = -\tau \int_\Omega \nabla h(y) \cdot \operatorname{cof}(I_{d\times d} + \tau D^2\phi(y)) \nabla h(y)\, \mu(y + \tau \nabla \phi(y))\, dy,$$

where $\operatorname{cof}(I_{d\times d} + \tau D^2\phi(y))$ denotes the cofactor matrix of $I_{d\times d} + \tau D^2\phi(y)$. Furthermore, if the eigenvalues of $I_{d\times d} + \tau D^2\phi(y)$ are bounded above by some constant Λ for every $y \in \Omega$, then we have the bound

$$-\delta^2 F(\phi)(h,h) \le \tau \|\mu\|_{L^\infty} \Lambda^{d-1} \|\nabla h\|_{L^2}^2. \tag{4.30}$$

The proof of Lemma 4.3.1 can be found in the appendix. To gain some insight into the bound (4.30), note that given a positive definite symmetric matrix $M \in \mathbb{R}^{d\times d}$ with eigenvalues $\{\lambda_1, \ldots, \lambda_d\}$, the eigenvalues of $\operatorname{cof}(M)$ are $\{\frac{\det(M)}{\lambda_1}, \ldots, \frac{\det(M)}{\lambda_d}\}$. This produces the $d-1$ degree scaling of Λ^{d-1}. To understand the meaning of Λ itself better, recall that the optimal primal variable ρ_* is given by $T_{\phi_*}{}_\# \mu = \mu(y + \tau\nabla\phi_*(y)) \det(I_{d\times d} + \tau D^2\phi(y))$.

Hence, the eigenvalues of $I_{d\times d} + \tau D^2\phi$ roughly measure how concentrated the mass of ρ_* is compared to μ. Since one expects the difference between ρ_* and μ to be on the order of τ, it is reasonable to expect that Λ will be close to 1.

We now turn our attention to bounding the Hessian of the internal energy term $U^*(\phi)$. When U takes the form (4.28), its convex conjugate can be written as

$$U^*(\phi) = \int_\Omega u_m^*(\phi(x) - V(x))\,dx,$$

where

$$u_m^*(p) = \gamma^{-\frac{1}{m-1}}\Big(\frac{(m-1)p+\gamma}{m}\Big)_+^{\frac{m}{m-1}}$$

and $(\cdot)_+ = \max(\cdot, 0)$. Now it is clear that the Hessian of U^* is given by

$$\delta^2 U^*(\phi)(h,h) = \int_\Omega (u_m^*)''\big(\phi(x) - V(x)\big)|h(x)|^2\,dx. \tag{4.31}$$

When $1 \le m \le 2$, the bounds are straightforward as $(u_m^*)''(p)$ is increasing with respect to p. Hence, in this case, we have

$$\delta^2 U^*(\phi)(h,h) = \int_\Omega (u_m^*)''(\phi(x) - V(x))|h(x)|^2\,dx \le B\|h\|^2_{L^2(\Omega)},$$

where $B = \sup_{x\in\Omega}(u_m^*)''(\phi(x) - V(x))$. It was shown in [JKT20b] that the maximizer ϕ_* of J obeys a maximum type principle in the sense that

$$\phi_*(x) \le M := \sup_{x\in\Omega} \delta U(\mu)(x).$$

It is therefore natural to assume that ϕ will be bounded above by M throughout the algorithm (the gradient steps tend to diffuse pressure in the regions of highest concentration). Assuming $V(x) \ge 0$ everywhere, it now follows that

$$\delta^2 U^*(\phi)(h,h) \le (u_m^*)''(M)\|h\|^2_{L^2(\Omega)}.$$

The aforementioned maximum principle on the pressure, $\phi(x) \le M$, can be used again to write the upper bound in terms of density instead of pressure. Indeed note that

$$\rho(x) = (u_m^*)'(\phi(x) - V(x)) \le (u_m^*)'(\phi(x)) \le (u_m^*)'(M).$$

Therefore the quantity

$$\rho_{\max} := (u_m^*)'(M) \tag{4.32}$$

acts a natural upper bound on the densities. Furthermore writing $(u_m^*)''(M) = (u_m^*)''\big(u_m'(\rho_{\max})\big) = u_m''(\rho_{\max})^{-1}$, we obtain

$$\delta^2 U^*(\phi)(h,h) \leq u_m''(\rho_{\max})^{-1} \|h\|^2_{L^2(\Omega)}.$$

The case $m > 2$ is substantially more complicated. When $m > 2$, $(u_m^*)''$ is singular at zero. Hence, the integrand may be unbounded near points where $\phi(x) = V(x)$. In this case, it may not be possible to bound (4.31) in terms of the L^2 norm of h. To understand this better, let us focus on the most difficult model we consider in this paper: the incompressible limit $m \to \infty$. When $m \to \infty$, the energy u_m encodes a hard ceiling constraint on the density values, i.e.

$$u_\infty(\rho) = \begin{cases} 0 & \text{if} \quad 0 \leq \rho \leq 1, \\ \\ +\infty & \text{otherwise.} \end{cases}$$

Hence, the dual energy u_∞^* is given by

$$u_\infty^*(p) = \begin{cases} 0 & \text{if } p < 0 \\ \\ p & \text{if } p \geq 0. \end{cases}$$

We pause here to point out that u_∞^* has much better regularity than u_∞, for instance u_∞^* is continuous over $\mathbb{R}$ while u_∞ is discontinuous at 0 and 1. This illustrates once more the advantage of working with dual quantities. Nevertheless, u_∞^* is clearly not smooth in the convex sense, as there is a jump of derivative at 0. In fact, we have $(u_\infty^*)'' = \delta_0$, where δ_0 denotes the Dirac delta function at 0.

Luckily, even though U^* is built from u_∞^* which is not smooth, it is possible to bound the Hessian of U^* as long as the singularity only occurs on a small set. Indeed, if we make the assumption that $|\nabla\phi(x) - \nabla V(x)|$ stays away from zero on the surface $\{\phi = V\}$, i.e. there

exists a constant $\Gamma_0 > 0$ such that

$$\sup_{\{x\in\Omega:\phi(x)=V(x)\}} \frac{1}{|\nabla\phi(x) - \nabla V(x)|} \leq \Gamma_0$$

(note this is a quantitative way of saying that $\{\phi = V\}$ is a lower dimensional set), then we can use the coarea formula to rewrite equation (4.31) as

$$\begin{aligned}
\delta^2 U^*(\phi)(h,h) &= \int_{\mathbb{R}} (u_\infty^*)''(\alpha) \int_{\{x\in\Omega:\phi(x)-V(x)=\alpha\}} \frac{|h(x)|^2}{|\nabla\phi(x) - \nabla V(x)|}\, ds(x)\, d\alpha \\
&= \int_{\{\phi=V\}} \frac{|h(x)|^2}{|\nabla\phi(x) - \nabla V(x)|}\, ds(x) \\
&\leq \Gamma_0 \int_{\{\,\phi=V\,\}} |h(x)|^2\, ds(x),
\end{aligned} \tag{4.33}$$

where ds is the usual surface measure. Due to the fact that the integration occurs over a surface, we cannot bound the right hand side of (4.33) in terms of $\|h\|_{L^2}$. However, we can use *trace inequalities* from PDE theory to bound surface integrals by volume integrals involving a higher derivative [Eva10] (this can be essentially viewed as an inequality version of Stokes' Theorem). More precisely, there exist constants C_1, C_2 depending on the surface $\{\phi = V\}$, but independent of h such that

$$\int_{\{\phi=V\}} |h(x)|^2\, ds(x) \leq C_2 \|\nabla h\|^2_{L^2(\Omega)} + C_1 \|h\|^2_{L^2(\Omega)}.$$

From there we can immediately deduce that U^* is H-smooth, since

$$\Gamma_0 \int_{\{\phi=V\}} |h(x)|^2\, ds(x) \leq \|h\|_H^2$$

as long as we choose $\Theta_i \geq C_i\Gamma_0$, $i = 1, 2$.

Now that we have seen how to obtain Hessian bounds in the most singular case $m \to \infty$, we are ready to return to the case $2 < m < \infty$. Note that in this case, $(u_m^*)''(p)$ is zero if $p < 0$, singular at zero, and decreasing for $p > 0$. Hence, if we choose some value $\lambda > 0$ and let

$$A_\lambda = \{x \in \Omega : 0 \leq \phi(x) - V(x) \leq \lambda\},$$

then we immediately have the bound

$$\delta^2 U^*(\phi)(h,h) \leq (u_m^*)''(\lambda)\|h\|^2_{L^2(\Omega)} + \int_{A_\lambda} (u_m^*)''(\phi(x) - V(x))|h(x)|^2\, dx.$$

To estimate the second term, we proceed along the same lines as the case $m = \infty$. For any $\alpha \in \mathbb{R}$ let $\{\phi - V = \alpha\} = \{\, x \in \Omega : \phi(x) - V(x) = \alpha \,\}$. As long as we have a constant Γ_λ and trace inequality constants $C_1(\alpha), C_2(\alpha)$ such that

$$\sup_{x \in A_\lambda} \frac{1}{|\nabla\phi(x) - \nabla V(x)|} \leq \Gamma_\lambda \tag{4.34}$$

and

$$\int_{\{\phi - V = \alpha\}} |h(x)|^2\, ds(x) \leq C_2(\alpha)\|\nabla h\|^2_{L^2(\Omega)} + C_1(\alpha)\|h\|^2_{L^2(\Omega)}, \tag{4.35}$$

then we can replicate the argument from above. Combining the coarea formula and trace inequality, we get the following string of inequalities

$$\int_{A_\lambda} (u_m^*)''(\phi(x) - V(x))|h(x)|^2\, dx$$

$$\leq \Gamma_\lambda \int_0^\lambda (u_m^*)''(\alpha) \int_{\{\phi - V = \alpha\}} |h(x)|^2\, ds(x)\, d\alpha$$

$$\leq (u_m^*)'(\lambda)\Gamma_\lambda \Big(C_{2,\lambda}\|\nabla h\|^2_{L^2(\Omega)} + C_{1,\lambda}\|h\|^2_{L^2(\Omega)} \Big),$$

where

$$C_{i,\lambda} = \max_{0 \leq \alpha \leq \lambda} C_i(\alpha). \tag{4.36}$$

Thus, $-\delta^2 U^*(h,h)$ is bounded by $\|h\|^2_H$ as long as we choose

$$\Theta_1 \geq (u_m^*)''(\lambda) + (u_m^*)'(\lambda)\, \Gamma_\lambda\, C_{1,\lambda}$$

and

$$\Theta_2 \geq (u_m^*)'(\lambda)\, \Gamma_\lambda\, C_{2,\lambda}$$

where we have the freedom to choose the precise value of λ.

Our above computations are now summarized in the following lemma.

Lemma 4.3.2 (Hessian bound on the internal energy)**.** *Define* $\rho_{\max}$, Γ_λ *and* $C_{i,\lambda}$ *by* (4.32), (4.34) *and* (4.36).

1. *Case* $1 \le m \le 2$. *We have*

$$\delta^2 U^*(\phi)(h,h) \le \frac{1}{\gamma m}(\rho_{\max})^{2-m}\|h\|_{L^2}^2.$$

2. *Case* $2 < m < \infty$. *For any* $\lambda > 0$,

$$\delta^2 U^*(\phi)(h,h) \le (\gamma m')^{1-m'} C_{2,\lambda}\,\Gamma_\lambda\,\|\nabla h\|_{L^2}^2 + (\gamma m')^{1-m'}\Big(C_{1,\lambda}\,\Gamma_\lambda \lambda^{m'-1} + (m'-1)\lambda^{m'-2}\Big)\|h\|_{L^2}^2,$$

   where $m' = \frac{m}{m-1}$.

3. *Case* $m = \infty$. *We have*

$$\delta^2 U^*(\phi)(h,h) \le C_{2,0}\,\Gamma_0 \|\nabla h\|_{L^2}^2 + C_{1,0}\,\Gamma_0 \|h\|_{L^2}^2.$$

Combining Lemma 4.3.1 and 4.3.2 we directly obtain the main theorem of this section.

Theorem 4.3.3 (1-smoothness of J)**.** *Let* $1 \le m \le \infty$ *and* $U(\rho) = \int_\Omega u_m(\rho(x)) + V(x)\rho(x)\,dx$, *where* u_m *is defined by* (4.21). *Then* $J(\phi) := \int_\Omega \phi^c(x)\,\mu(x)dx - U^*(\phi)$ *satisfies the Hessian bound*

$$-\delta^2 J(\phi)(h,h) \le \Theta_2 \|\nabla h\|_{L^2}^2 + \Theta_1 \|h\|_{L^2}^2,$$

where Θ_1 *and* $\Theta_2 > 0$ *are given by the table below (Table 4.1).*

As in Lemma 4.3.1, Λ *is an upper bound on the eigenvalues of* $I_{d\times d} + \tau D^2\phi(y)$ *uniformly in* y. *Additionally* $\lambda > 0$ *is a parameter to choose and* $\rho_{\max}$, Γ_λ *and* $C_{i,\lambda}$ *are defined by* (4.32), (4.34) *and* (4.36).

Table 4.1: Constants Θ_1 and Θ_2 in Theorem 4.3.3

m	Θ_1	Θ_2
$m = 1$	$\frac{\rho_{\max}}{\gamma}$	$\tau\Lambda^{d-1}\|\mu\|_{L^\infty}$
$1 < m < 2$	$\frac{\rho_{\max}^{2-m}}{\gamma m}$	$\tau\Lambda^{d-1}\|\mu\|_{L^\infty}$
$m = 2$	$\frac{1}{2\gamma}$	$\tau\Lambda^{d-1}\|\mu\|_{L^\infty}$
$m > 2$	$(\gamma m')^{1-m'}\left(\lambda^{m'-1}C_{1,\lambda}\,\Gamma_\lambda + \frac{m'-1}{\lambda^{2-m'}}\right)$	$(\gamma m')^{1-m'}C_{2,\lambda}\,\Gamma_\lambda + \tau\Lambda^{d-1}\|\mu\|_{L^\infty}$
$m = \infty$	$C_{1,0}\,\Gamma_0$	$C_{2,0}\,\Gamma_0 + \tau\Lambda^{d-1}\|\mu\|_{L^\infty}$

In order to use Theorem 4.3.3 in the case $m > 2$, we need to be able to compute Γ_λ and $C_{i,\lambda}$ and we need to choose a value for λ when $m \in (2, \infty)$. On a discrete grid with n points, one can easily compute Γ_λ for all λ in $O(n)$ operations. On the other hand, it requires $O(n)$ operations to compute $C_1(\alpha)$ and $C_2(\alpha)$ for a single value of α (c.f. Section 4.1). Thus, for the case $m = \infty$, we can compute the constants explicitly in $O(n)$ operations. The case $2 < m < \infty$ is harder, since we cannot efficiently compute $C_{i,\lambda} = \max_{0\le\alpha\le\lambda} C_i(\alpha)$. To overcome this difficulty, we typically choose λ by minimizing

$$\lambda^* = \operatorname*{argmin}_{\lambda\ge 0} (\gamma m')^{1-m'}\Big(\lambda^{m'-1}\Gamma_\lambda C_1(0) + \frac{m'-1}{\lambda^{2-m'}}\Big),$$

which gives a reasonable estimate for the optimal choice of λ to make Θ_1 as small as possible. We then estimate $\max_{0\le\alpha\le\lambda^*} C_i(\alpha)$ by simply taking the max over $C_i(0)$ and $C_i(\lambda^*)$, which appears to work well in practice.

To conclude this discussion we turn our attention to the other functional I for which a similar analysis can be made. First we define

$$p(x) = (\psi^{\bar{c}} - V)(T_{\psi^{\bar{c}}}(x)).$$

Next, for $\lambda > 0$, we define

$$\tilde{\Gamma}_\lambda = \sup_{x:0\le p(x)\le\lambda} \frac{1}{|\nabla p(x)|}. \tag{4.37}$$

Finally, we define trace constants $\tilde{C}_i(\alpha)$ such that

$$\int_{\{p=\alpha\}} |h(x)|^2 \, ds(x) \le \tilde{C}_2(\alpha)\|\nabla h\|_{L^2}^2 + \tilde{C}_1(\alpha)\|h\|_{L^2}^2,$$

and then set

$$\tilde{C}_{i,\lambda} = \sup_{0\le\alpha\le\lambda} \tilde{C}_i(\alpha). \tag{4.38}$$

Now we can state our result bounding the Hessian of I.

Theorem 4.3.4. *Let $I(\psi) = \int_\Omega \psi(x)\,\mu(x)dx - U^*(\psi^{\bar{c}})$, with $U(\rho) = \int_\Omega u_m(\rho(x)) + V(x)\rho(x)\,dx$, u_m is defined by (4.21) and $1 \le m \le \infty$. The Hessian of I can be written*

$$-\delta^2 I(\psi)(h,h) = \delta^2 U^*(\psi^c)(h\circ S_\psi, h\circ S_\psi) + \tau\int_\Omega \nabla h(x)\cdot \mathrm{cof}(I_{d\times d} - \tau D^2\psi(x))\nabla h(x)\,\delta U^*(\psi^c)(x-\tau\nabla\psi(x))\,dx.$$

It satisfies the bound

$$-\delta^2 I(\psi)(h,h) \le \Theta_2\|\nabla h\|_{L^2}^2 + \Theta_1\|h\|_{L^2}^2,$$

where Θ_1 and $\Theta_2 > 0$ are given by the table below (Table 4.2). Here Λ is an upper bound on the eigenvalues of $I_{d\times d} - \tau D^2\psi(x)$ uniformly in x. Additionally $\lambda > 0$ is a parameter to choose and $\rho_{\max}$ is defined by (4.32), $\tilde{\Gamma}_\lambda$ by (4.37) and $\tilde{C}_{i,\lambda}$ by (4.38).

4.3.3 Back-and-forth for non-convex U

In this section, we will discuss how to extend our method when U is not convex with respect to ρ. The trick is to appeal to convexity splitting [Eyr98], a well-known technique for simulating gradient flows with non-convex energies. The idea behind convexity splitting is to write U as a sum of a convex function and a concave function, i.e.

$$U(\rho) = U_1(\rho) + U_0(\rho),$$

Table 4.2: Constants Θ_1 and Θ_2 in Theorem 4.3.4

m	Θ_1	Θ_2
$m = 1$	$\dfrac{\Lambda^d \rho_{\max}}{\gamma}$	$\tau \Lambda^{d-1} \rho_{\max}$
$1 < m < 2$	$\dfrac{\Lambda^d (\rho_{\max})^{2-m}}{\gamma m}$	$\tau \Lambda^{d-1} \rho_{\max}$
$m = 2$	$\dfrac{\Lambda^d}{2\gamma}$	$\tau \Lambda^{d-1} \rho_{\max}$
$m > 2$	$\Lambda^d (\gamma m')^{1-m'} \left(\tilde{C}_{1,\lambda} \tilde{\Gamma}_\lambda \lambda^{m'-1} + \dfrac{m'-1}{\lambda^{2-m'}} \right)$	$\Lambda^d (\gamma m')^{1-m'} \tilde{C}_{2,\lambda} \tilde{\Gamma}_\lambda + \tau \Lambda^{d-1} \rho_{\max}$
$m = \infty$	$\Lambda^d \tilde{C}_{1,0} \tilde{\Gamma}_0$	$\Lambda^d \tilde{C}_{2,0} \tilde{\Gamma}_0 + \tau \Lambda^{d-1} \rho_{\max}$

where U_1 is convex and U_0 is concave. Thanks to the concavity of U_0, given any fixed density $\bar{\rho}$, we have the inequality

$$U(\rho) \leq U_1(\rho) + U_0(\bar{\rho}) + (\delta U_0(\bar{\rho}), \rho - \bar{\rho}). \tag{4.39}$$

Crucially, the right-hand-side of equation (4.39) is a convex function. As such, if we replace the JKO scheme with the relaxed scheme

$$\rho^{(n+1)} = \operatorname*{argmin}_{\rho} \; U_1(\rho) + U_0(\rho^{(n)}) + (\delta U_0(\rho^{(n)}), \rho - \rho^{(n)}) + \frac{1}{2\tau} W_2^2(\rho, \rho^{(n)}), \tag{4.40}$$

then we obtain a convex variational problem. The beauty of convexity splitting is that the relaxed scheme is still unconditionally energy stable. Combining (4.39) and (4.40) we have the string of inequalities

$$U(\rho^{(n+1)}) + \frac{1}{2\tau} W_2^2(\rho^{(n+1)}, \rho^{(n)}) \leq$$

$$U_1(\rho^{(n+1)}) + U_0(\rho^{(n)}) + (\delta U_0(\rho^{(n)}), \rho^{(n+1)} - \rho^{(n)}) + \frac{1}{2\tau} W_2^2(\rho^{(n+1)}, \rho^{(n)}) \leq$$

$$\inf_{\rho} \; U_1(\rho) + U_0(\rho^{(n)}) + (\delta U_0(\rho^{(n)}), \rho - \rho^{(n)}) + \frac{1}{2\tau} W_2^2(\rho, \rho^{(n)}).$$

By choosing $\rho = \rho^{(n)}$ in the last line, we can conclude that

$$U(\rho^{(n+1)}) + \frac{1}{2\tau} W_2^2(\rho^{(n+1)}, \rho^{(n)}) \leq U(\rho^{(n)}).$$

Thus, we see that the energy is still decreasing along the iterates of the relaxed scheme.

Now let us turn to solving the relaxed problem (4.40). Since the energy term in (4.40) is convex, we can solve the problem using the dual approach outlined above. The twin dual problems associated to (4.40), which we shall denote as $\tilde{J}$ and $\tilde{I}$, are given by

$$\tilde{J}(\phi) := \int_\Omega \phi^c(x)\, \rho^{(n)}(x) dx - \tilde{U}^*(\phi), \tag{4.41}$$

$$\tilde{I}(\psi) := \int_\Omega \psi(x)\, \rho^{(n)}(x) dx - \tilde{U}^*(\psi^{\bar{c}}), \tag{4.42}$$

where

$$\tilde{U}^*(\phi) := U_1^*\big(\phi - \delta U_0(\rho^{(n)})\big) + (\delta U_0(\rho^{(n)}), \rho^{(n)}) - U_0(\rho^{(n)})$$

is the convex conjugate of $U_1(\rho) + U_0(\rho^{(n)}) + (\delta U_0(\rho^{(n)}), \rho - \rho^{(n)})$. We can then find the dual maximizers $(\phi^{(n+1)}, \psi^{(n+1)})$ of (4.41) and (4.42) using Algorithm 6 along with the Hessian bounds developed in the previous subsection. As before, one can recover the solution $\rho^{(n+1)}$ of (4.40) through the duality relation $\rho^{(n+1)} = \delta\tilde{U}^*(\phi^{(n+1)})$.

4.4 Numerical implementation and experiments

4.4.1 Implementation details

In this section, we use the back-and-forth method to numerically simulate equation (4.1) for a wide variety of internal energies U. Throughout this section we will assume that the domain $\Omega = [-1/2, 1/2]^2$ is the unit square in $\mathbb{R}^2$, discretized using a regular rectangular grid. The numerical simulations in this section were coded in C++ and were run on 2019 MacBook Pro with 2.6 GHz 6-core and 16 GB RAM.

Following the approach in [JL20], we will compute the forward and backward c-transforms using the fast Legendre transform (FLT) algorithm [Luc97]. On a regular

rectangular grid with n points, the FLT algorithm can be used to compute either the forward or backward c-transform in $O(n)$ operations. See [JL20] for more detail on the equivalence of the c-transform and the Legendre transform.

When computing gradients with respect to the weighted norm (4.20), we will need to solve a Poisson equation with zero Neumann boundary condition. We will solve this equation numerically via the fast Fourier transform (FFT). All FFTs were calculated using the free FFTW C++ library.

To compute the gradients of I and J, we will also need to compute pushforwards. Given a density μ and an invertible map $Z\colon \Omega \to \Omega$ we can compute the pushforward $Z_{\#}\mu$ via the Jacobian formula

$$Z_{\#}\mu(x) = \frac{\mu\big(Z^{-1}(x)\big)}{|\det\big(DZ(Z^{-1}(x))\big)|} = \mu\big(Z^{-1}(x)\big)|\det\big(D(Z^{-1})(x)\big)|.$$

In our case, we will only need to compute pushforwards with respect to the maps T_ϕ and S_ψ that are induced by the forward and backward c-transforms respectively. Thanks to the structure of BFM, we only need to compute $T_{\phi\#}\rho^{(n)}$ and $S_{\psi\#}\delta U^*(\psi^{\bar{c}})$ when ϕ and ψ are c-convex and c-concave respectively. As a result, we have the simple formulas $T_\phi^{-1}(y) = y + \tau\nabla\phi(y)$ and $S_\psi^{-1}(x) = x - \tau\nabla\psi(x)$. Therefore,

$$T_{\phi\#}\rho^{(n)}(y) = \rho^{(n)}\big(y + \tau\nabla\phi(y)\big)\det\big(I_{d\times d} + \tau D^2\phi(y)\big),$$

and

$$S_{\psi\#}\delta U^*(\psi^{\bar{c}})(x) = \Big(\delta U^*\big(\psi^{\bar{c}}\big)\circ\big(x - \tau\nabla\psi(x)\big)\Big)\det\big(I_{d\times d} - \tau D^2\psi(x)\big).$$

When implementing our algorithm, we compute these quantities using a simple centered difference scheme.

Finally, let us briefly explain how to compute the trace inequality constants $C_i(\alpha)$ defined in equation (4.35). From Lemma A.2.1 and Corollary A.2.2 in the Appendix, we see that

$C_i(\alpha)$ can be computed from the solution u to the Eikonal equation

$$\begin{cases} |\nabla u(x)| = 1 & \text{if} \quad \phi(x) - V(x) \neq \alpha, \\ u(x) > 0 & \text{if} \quad \phi(x) - V(x) < \alpha, \\ u(x) < 0 & \text{if} \quad \phi(x) - V(x) > \alpha. \end{cases}$$

Note that

$$|u(x)|^2 = \min_{\{y:\phi(y)-V(y)=\alpha\}} |x-y|^2,$$

which is nothing but a c-transform of the indicator function

$$\chi_\alpha(y) = \begin{cases} 0 & \text{if} \quad \phi(y) - V(y) = \alpha, \\ +\infty & \text{else.} \end{cases}$$

Therefore, $|u|^2$ can be computed in $O(n)$ operations using the Fast Legendre transform, and from there one can recover u. Once one has u, it is straightforward to compute the constants in Corollary A.2.2 in $O(n)$ operations.

4.4.2 Experiments

We present four sets of numerical experiments. In the first set of experiments, we demonstrate the speed and accuracy of our method by comparing to the so-called Barrenblat solutions, a special case of equation (4.1) where closed-form solutions are available. In the next set of experiments, we simulate the porous media equation $\partial_t \rho = \Delta(\rho^m) + \nabla \cdot (\rho \nabla V)$ for various interesting functions $V \colon \Omega \to \mathbb{R} \cup \{+\infty\}$ and values of m. Note that if V takes the value $+\infty$ on some closed set $E \subset \Omega$, then ρ can never enter E. Hence, this is equivalent to solving (4.1) on the more complicated domain $\Omega \setminus E$. In the third set of experiments, we use the splitting scheme from Section 4.3.3 to simulate (4.1) when U is nonconvex. In this case, the non-convexity will come from an interaction energy of the form $\mathcal{W}(\rho) = \int_\Omega \int_\Omega W(x-y)\rho(x)\rho(y)\,dy\,dx$. Finally, in the last set of experiments, we study incompressible flows where U encodes the hard constraint $\rho \leq 1$ everywhere. In this case,

the dual energy U^* will have a very singular Hessian at the boundary of the support of ρ. Nonetheless, we are still able to simulate the evolution even on very fine grids.

4.4.2.1 Accuracy: Barenblatt solutions

In this experiment, we use our back-and-forth algorithm to solve the PME,

$$\partial_t \rho = \gamma \Delta(\rho^m), \tag{4.43}$$

with the initial data

$$\rho(0, x) = M\delta_0(x).$$

Here, $\gamma > 0$ is a constant that controls the speed of the diffusion, $M > 0$ is the total initial mass and δ_0 is the standard Dirac distribution centered at zero. When $m > 1$, this equation is the Wasserstein gradient flow of the energy $U(\rho) = \int_\Omega \frac{\gamma}{m-1}\rho(x)^m \, dx$. Thanks to the simplicity of the initial data, on the domain $\mathbb{R}^2$ the equation has a closed form solution, known as the Barenblatt solution [Bar96, Bar03],

$$\rho(t, x) = \left(\left(\frac{M}{4\pi m t \gamma} \right)^{\frac{m-1}{m}} - \frac{(m-1)}{4m^2 t \gamma} |x|^2 \right)_+^{\frac{1}{m-1}}, \tag{4.44}$$

where $(\cdot)_+ = \max(\cdot, 0)$. The Barenblatt solution is compactly supported, therefore, it agrees with the solution on the square $[-1/2, 1/2]^2$ up until the time $t_c = \frac{m-1}{16m^2\gamma}(\frac{\pi(m-1)}{4mM})^{m-1}$ when the mass hits the boundary of the square.

Using the Barenblatt solution as a benchmark, we can test the accuracy and efficiency of our scheme. We will simulate the equation for the exponents $m = 2, 4, 6$. Since the Dirac delta function is challenging to work with numerically, we shall instead fix a height $h_0 > 0$ and start the flow at a time $t_0 > 0$, where t_0 is chosen so that $\|\rho(t_0, \cdot)\|_{L^\infty} = h_0$. Note that the value of t_0 will depend on the exponent m, and can be found explicitly from equation (4.44). In addition, we will only consider the flow within in the time interval $[t_0, t_c]$, since the Barenblatt solution is only valid on the unit square up to time t_c.

In all of our benchmark experiments, we shall set $M = 0.5$, $h_0 = 15$ and $\gamma = 10^{-3}$. Note that the small value of γ is just a time rescaling to ensure that the flow occurs on a macroscopic time interval. We will compute the evolution between the times $t_0 \leq t \leq 2 + t_0$ with different step sizes $\tau = 0.4, 0.2, 0.1, 0.05, 0.025$ (one can check that with our parameter choices $t_0 + 2 < t_c$ for $m = 2, 4, 6$). Running the experiments with various time step sizes allows us to verify that the scheme becomes more accurate as the time step is decreased. We shall measure the accuracy of the solution using the L^1 norm, which is very natural in the context of Wasserstein gradient flows (see for instance [JKT20b]). The precise formula for our error estimate is

$$\text{error} = \frac{1}{N_\tau} \sum_{n=0}^{N_\tau} \int_\Omega |\rho(n\tau + t_0, x) - \rho^{(n)}(x)| \, dx, \tag{4.45}$$

where $N_\tau = \lfloor \frac{2}{\tau} \rfloor$, $\rho(n\tau + t_0, x)$ is the Barenblatt solution and $\rho^{(n)}$ is the n^{th} JKO iterate starting from the initial data $\rho^{(0)}(x) = \rho(t_0, x)$. When solving for $\phi^{(n+1)}$, we will run Algorithm 6 until the residual $\|T_\phi - \delta U^*(\phi)\|_{L^1(\Omega)}$ is less than $\epsilon = 10^{-3}$.

The results of these experiments are displayed in Table 4.3 and Figure 4.1. Table 4.3 displays the error (4.45) and the total computation time for all of the aforementioned experiments. In Figure 4.1, we plot a cross section of our solutions and the exact solution at various time snapshots. The cross section is taken along the horizontal line $\{(x_1, 0) : x_1 \in [-1/2, 1/2]\}$. One can see that as the time step is decreased, our solution is in excellent agreement with the exact solution for all exponents $m = 2, 4, 6$. Figure 4.1 also shows that our method correctly captures the discontinuity of $\nabla\rho$ at the boundary of the support of ρ. This is notable as most other numerical methods smooth out the discontinuity. The reason that we are able to correctly capture the discontinuity is due to the fact that we recover the density through the duality relation $\rho^{(n+1)} = \delta U^*(\phi^{(n+1)}) = \left(\frac{m-1}{m\gamma} \max(\phi, 0)\right)^{\frac{1}{m-1}}$. The function $s(x) = \max(x, 0)^{\frac{1}{m-1}}$ has discontinuous derivatives at zero, therefore even when $\phi^{(n+1)}$ is smooth, $\nabla\rho$ will still have a discontinuity at the boundary of its support.

Table 4.3: Barenblatt solution test case (grid size 512×512)

τ	N_τ	$m = 2$		$m = 4$		$m = 6$	
		Error	Time (s)	Error	Time (s)	Error	Time (s)
0.4	5	6.35×10^{-2}	14.54	1.19×10^{-1}	23.11	1.13×10^{-1}	22.02
0.2	10	3.72×10^{-2}	22.16	7.95×10^{-2}	30.34	7.48×10^{-1}	30.41
0.1	20	2.08×10^{-2}	36.57	5.03×10^{-2}	48.41	4.74×10^{-2}	43.95
0.05	40	1.18×10^{-2}	55.64	3.06×10^{-2}	77.03	2.90×10^{-2}	80.10
0.025	80	8.26×10^{-3}	77.67	1.89×10^{-2}	140.38	1.79×10^{-2}	164.89

4.4.2.2 Slow diffusion with drifts and obstacles.

In our next set of experiments, we add spatially varying potentials to the energy functional. The resulting equations are a type of drift-diffusion equations. The energy takes the specific form

$$U(\rho) = \int_\Omega \frac{\gamma}{m-1}\rho^m(x) + V(x)\rho(x)dx,$$

where V is a given function.

In the first set of experiments, we consider an example where the initial density is the characteristic function of a star shaped region normalized to have mass 1, and we use the fixed potential function

$$V_1(x) = 1 - \sin(5\pi x_1)\sin(3\pi x_2). \tag{4.46}$$

The initial data and the potential V_1 are shown in Figure 4.2.

Using this setup, we run two different experiments, one where $m = 2$ and another where $m = 4$. In both cases, we set $\gamma = 0.1$ and use the time step $\tau = 0.001$. We run the equations until we reach a state that is essentially stationary. The flow for $m = 2$ is run from time $t = 0$ to time $t = 5$, and the flow for $m = 4$ is run from time $t = 0$ to time $t = 2$. The flow for the $m = 2$ case is shown in Figure 4.3 and the $m = 4$ case is shown in Figure 4.4. The

solutions show the density is drawn to regions where the potential is small, while avoiding concentration due to the ρ^m term. Notice that the steady state for $m = 4$ is much more diffuse than the steady state for $m = 2$, this is because ρ^4 penalizes concentration much more than ρ^2.

Next, we consider a different potential function:

$$V_2(x) = 10\left((x_1 - 0.4)^2 + (x_2 - 0.4)^2\right) + \iota_{\Omega\backslash E}(x) \tag{4.47}$$

where E is a given subset of Ω and $\iota_{\Omega\backslash E} : \Omega \to \mathbb{R} \cup \{+\infty\}$ is the indicator function

$$\iota_{\Omega\backslash E}(x) = \begin{cases} 0 & \text{if } x \in \Omega\backslash E \\ +\infty & \text{if } x \in E. \end{cases}$$

With this setup, the set E represents an obstacle that the density is not allowed to penetrate. During the flow, the density diffuses and drifts towards the lower level sets of V_2, all while avoiding the set E.

In Figure 4.5 and Figure 4.6, we display two different experiments with different obstacles E, but the same diffusion exponent $m = 4$. In both experiments, the starting density is the characteristic function of a square centered at $(-0.3, -0.3)$ with side length 0.2 renormalized to have unit mass. In Figure 4.5, the obstacle is a disc with radius 0.2 centered at the origin, and in Figure 4.6, the obstacle is a star shaped region centered at the origin. In both experiments, we set $\tau = 0.001$, $\gamma = 0.0075$ and we run the flow until time $t = 2$. An interesting difference between the two flows is that the non-convexity of the star shaped obstacle results in some mass being trapped between the arms of the star. It is not entirely clear if the mass eventually escapes as time goes to infinity. This is because the PME allows for compactly supported solutions (in contrast to say the behavior of the heat equation).

4.4.2.3 Non-convex U (aggregation-diffusion)

In this experiment, we simulate (4.1) with an energy functional U that is not a convex with respect to ρ. Specifically, we consider the energy

$$U(\rho) = \mathcal{W}(\rho) + \int_\Omega \frac{1}{60}\rho^3(x)\,dx, \tag{4.48}$$

where

$$\mathcal{W}(\rho) := \frac{1}{2}\int_\Omega \int_\Omega |x-y|^2 \rho(x)\rho(y)\,dy\,dx.$$

By separating out the square, one can check $\mathcal{W}$ is concave with respect to ρ.

While convex energies U encourage mass diffusion, non-convex energies allow for both aggregation and diffusion phenomena. Indeed, one can see that $\mathcal{W}(\rho)$ encourages the density to concentrate while the ρ^3 term encourages the density to diffuse. Due to the convolution, $\mathcal{W}$ can be viewed as a "lower order" term as compared to ρ^3. However, since the coefficients in front of the convolution is much larger than the coefficient in front of the ρ^3 term, the aggregation effect will dominate until the density reaches a certain saturation level.

Here we run a single experiment starting with an initial density that is the sum of the characteristic function of four squares with side lengths 0.2 centered at each combination of $(\pm 0.3, \pm 0.3)$ and renormalized to have total mass equal to one. We set $\tau = .005$ and run the flow from time $t = 0$ to $t = 10$, at which time the evolution appears to have reached a steady state.

The results of the experiment are displayed in Figures 4.7 and 4.8. Figure 4.7 displays a heat map of the density evolution, while Figure 4.8 gives a 3 dimensional plot showing the height of the density. Throughout the evolution, one can see the competing effects of aggregation and diffusion. The heights of the four densities decrease due to diffusion, however aggregation pulls the four separate components together towards the center of the domain.

4.4.2.4 Incompressible projections and flows

In our last set of experiments, we consider incompressible flows, which have applications to crowd motion models and fluid mechanics. Here the energy takes the form

$$U(\rho) = s_\infty(\rho) + \int_\Omega V(x)\rho(x)\,dx, \tag{4.49}$$

where

$$s_\infty(\rho) = \begin{cases} 0 & \text{if } 0 \le \rho(x) \le 1 \text{ for a.e. } x \in \Omega, \\ \infty & \text{otherwise,} \end{cases}$$

and V is a fixed potential function. Note that $s_\infty(\rho)$ can be seen as the limit of the energy

$$s_m(\rho) = \frac{1}{m-1}\int_\Omega \rho^m(x)\,dx$$

as $m \to \infty$.

We will run our experiments, using the potential energy

$$V(x) = \frac{1}{2}\big((x_1 - \frac{3}{10})^2 + (x_2 - \frac{3}{10})^2\big) + \iota_{\Omega\setminus E}(x) \tag{4.50}$$

where E is a closed set that represents an impenetrable obstacle. We run two simulations using two different obstacles

$$E_1 = B_{\frac{1}{4}}(\frac{1}{5}, -\frac{1}{5}) \cup B_{\frac{1}{4}}(-\frac{1}{5}, \frac{1}{5})$$

and

$$E_2 = B_{\frac{1}{10}}(0, \frac{1}{5}) \cup B_{\frac{1}{10}}(0, -\frac{1}{5}) \cup B_{\frac{1}{10}}(\frac{1}{5}, 0) \cup B_{\frac{1}{10}}(-\frac{1}{5}, 0),$$

where $B_r(x_1, x_2)$ denotes the closed ball of radius r centered at (x_1, x_2). In both experiments, we choose an initial density $\rho^{(0)}$, which equals 1 on a ball of a radius 0.15 centered at $(-0.3, -0.3)$ and is equal to 0 elsewhere.

The results of our experiments are displayed in Figures 4.9 and 4.10. Figure 4.9 uses the obstacle E_1, while Figure 4.10 uses the obstacle E_2. In the figures, the yellow pixels represent

the density $\rho^{(n)}$ and white pixels represents the obstacle. In both experiments we use a time step $\tau = 0.05$ and run the evolution from time $t = 0$ to time $t = 20$. Both experiments are conducted on 1024×1024 pixel grids.

Notably, in both of the simulations depicted in Figures 4.9 and 4.10, there is a sharp interface separating the regions $\rho = 1$ and $\rho = 0$. This matches the expected behavior of the flow with our chosen potentials. In general, it is difficult for numerical methods to correctly capture sharp interfaces. Again, the reason that our method is able to do so is because of our dual approach. By recovering the density through the duality relation $\rho^{(n+1)} \in \delta U^*(\phi^{(n+1)})$ we automatically produce a discontinuity at the level set $\{y \in \Omega : \phi^{(n+1)}(y) - V(y) = 0\}$.

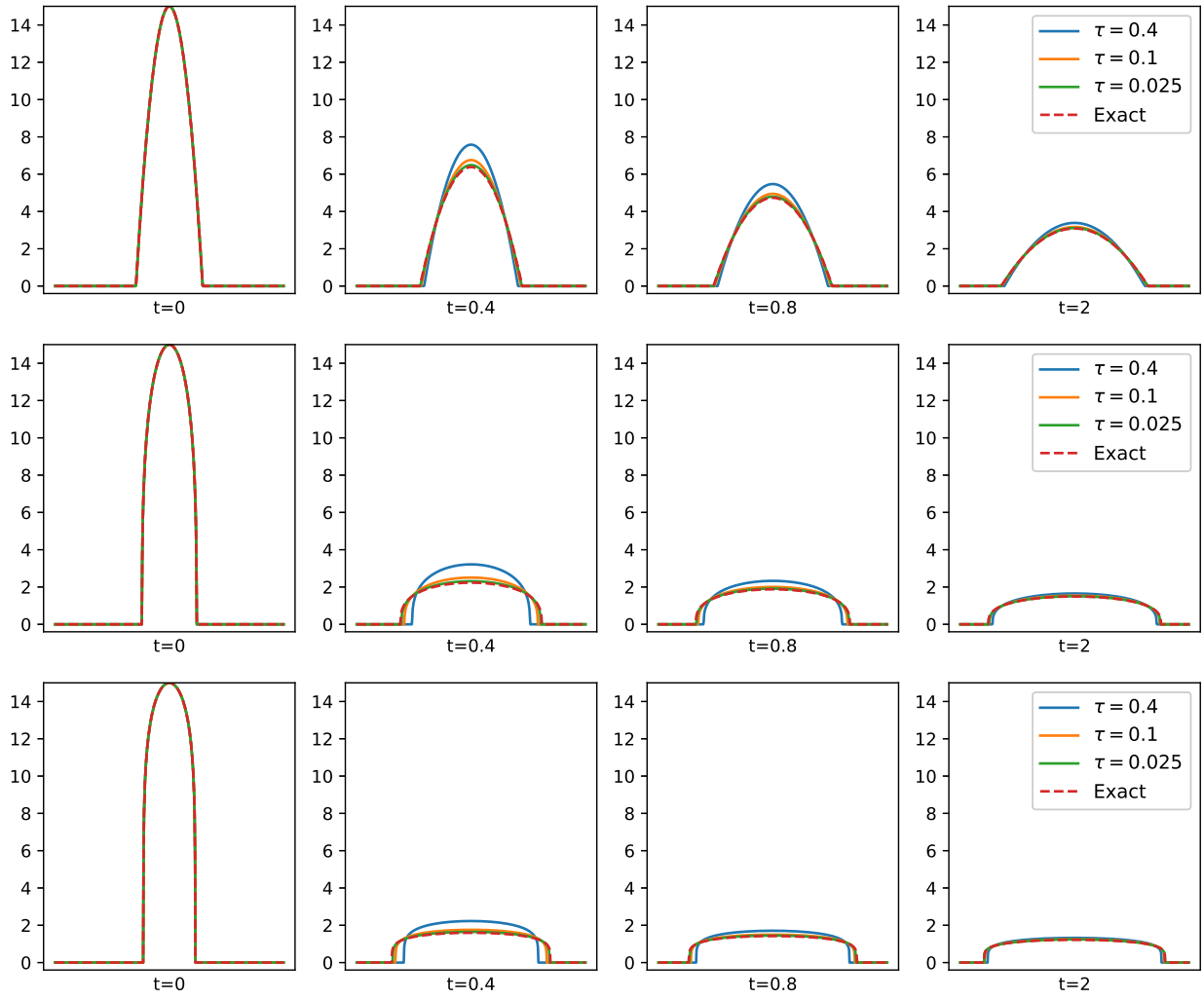
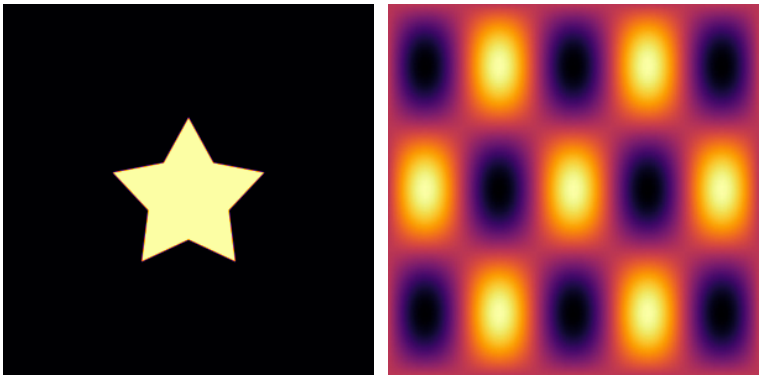


Figure 4.1: Cross sections of our computed solutions and the exact Barenblatt solution at times $t = t_0, t_0 + 0.4, t_0 + 0.8, t_0 + 2$ along the horizontal line $\{(x_1, 0) : x_1 \in [-1/2, 1/2]\}$. Row 1: $m = 2$, Row 2: $m = 4$, Row 3: $m = 6$.



Initial density The potential V_1

Figure 4.2: Higher values are depicted with brighter pixels.

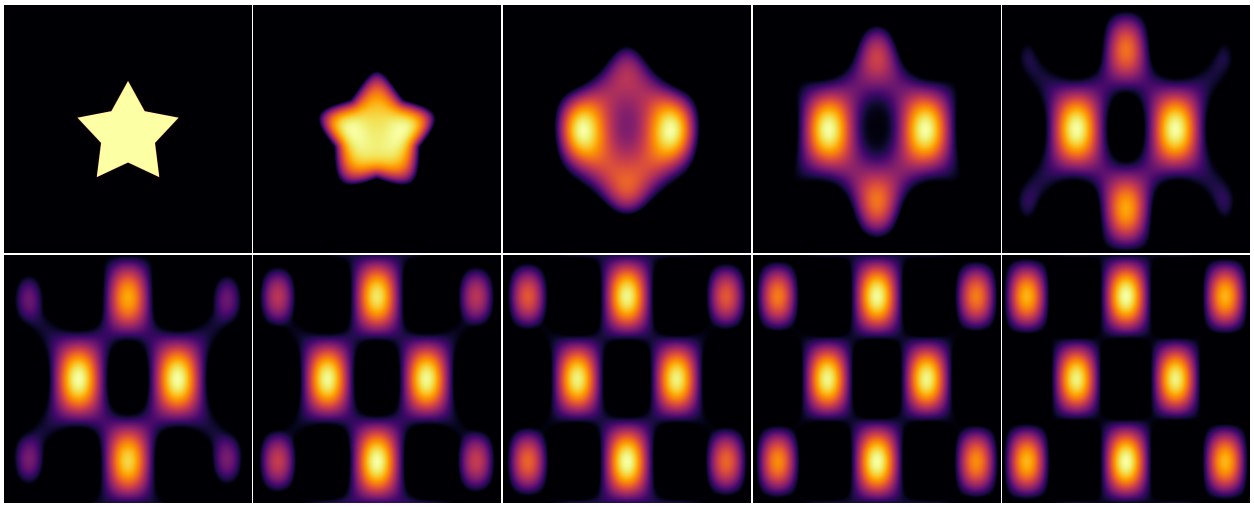


Figure 4.3: PME with exponent $m = 2$ and potential given by (4.46). The images show the evolution from time $t = 0$ to $t = 5$ (top left to bottom right). The final image is the approximate steady state. Images are 512×512 pixels. Brighter pixels indicate larger density values.

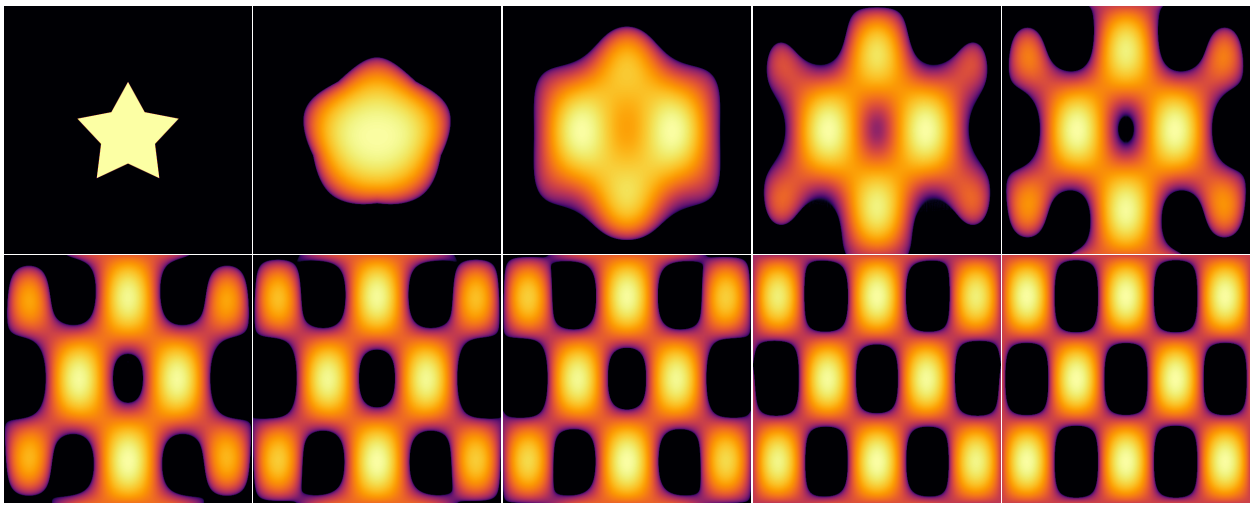


Figure 4.4: PME with exponent $m = 4$ and potential given by (4.46). The images show the evolution from time $t = 0$ to $t = 2$ (top left to bottom right). The final image is the approximate steady state. Images are 512×512 pixels. Brighter pixels indicate larger density values.

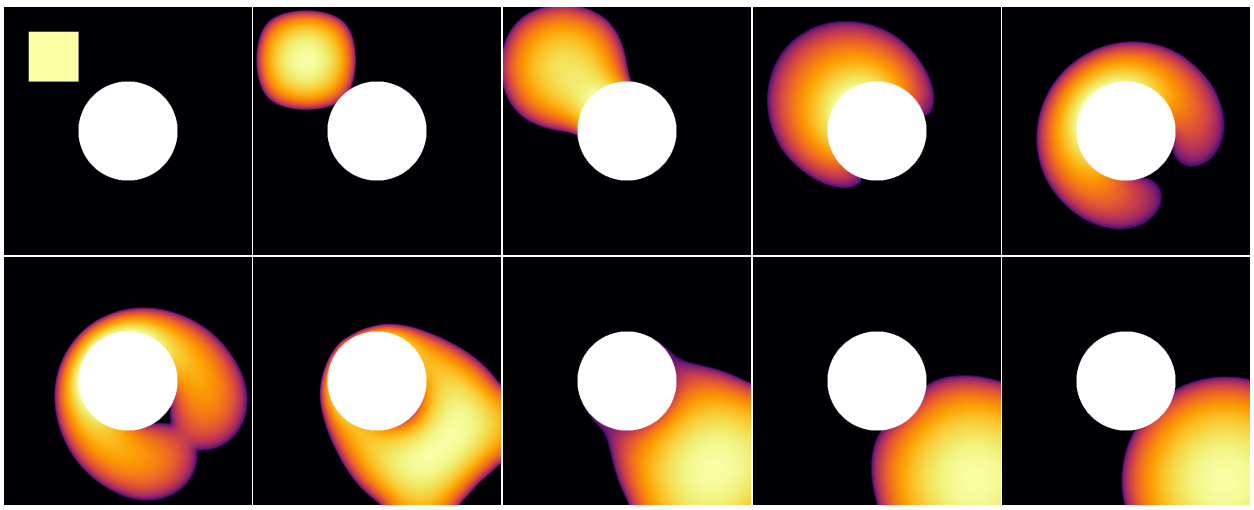


Figure 4.5: PME with exponent $m = 4$, $\gamma = .0075$ and potential given by (4.47). The obstacle E is represented by the white region. The images show the evolution from time $t = 0$ to $t = 2$ (top left to bottom right). Images are 512×512 pixels. With the exception of the obstacle, brighter pixels indicate larger density values.

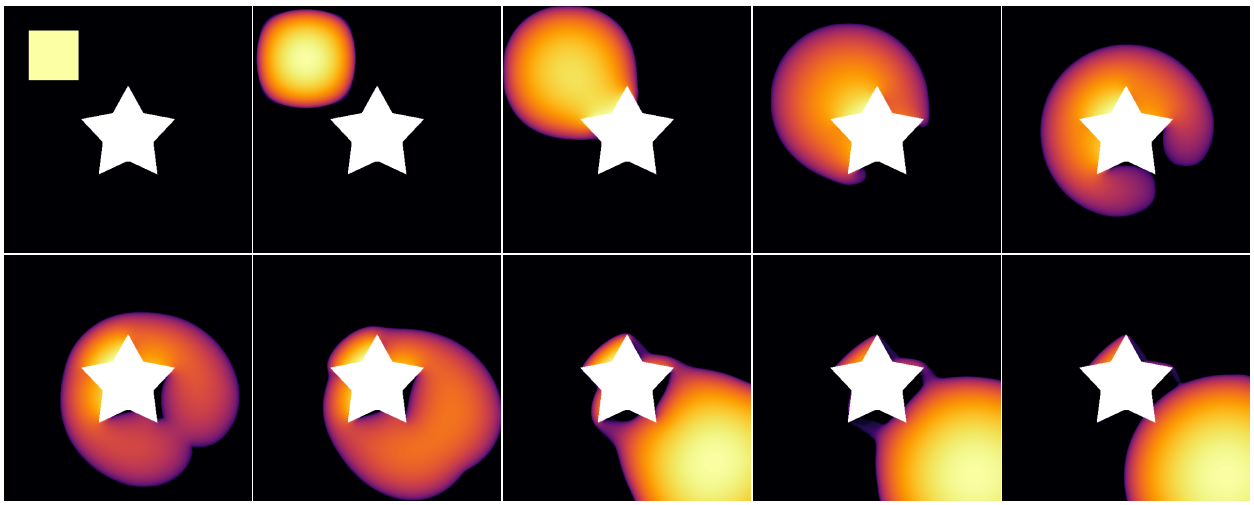


Figure 4.6: PME with exponent $m = 4$, $\gamma = .0075$ and potential given by (4.47). The obstacle E is represented by the white region. The images show the evolution from time $t = 0$ to $t = 2$ (top left to bottom right). Images are 512×512 pixels. With the exception of the obstacle, brighter pixels indicate larger density values.

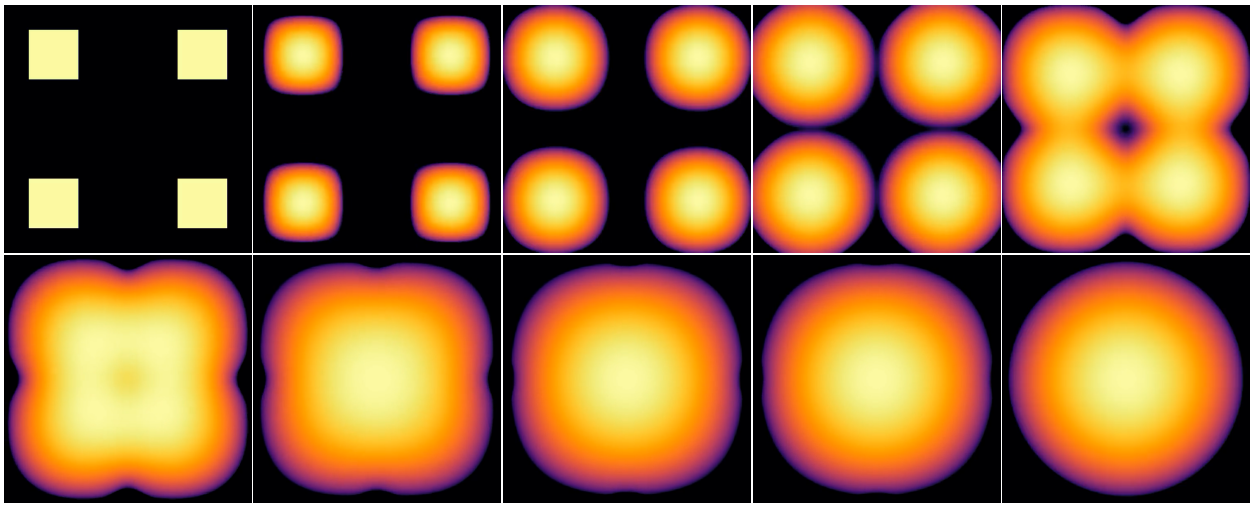


Figure 4.7: Aggregation-diffusion equation with an energy given by (4.48). The images show the evolution from time $t = 0$ to $t = 10$ (top left to bottom right). The final image is the approximate steady state. Images are 512×512 pixels. Brighter pixels indicate larger density values.

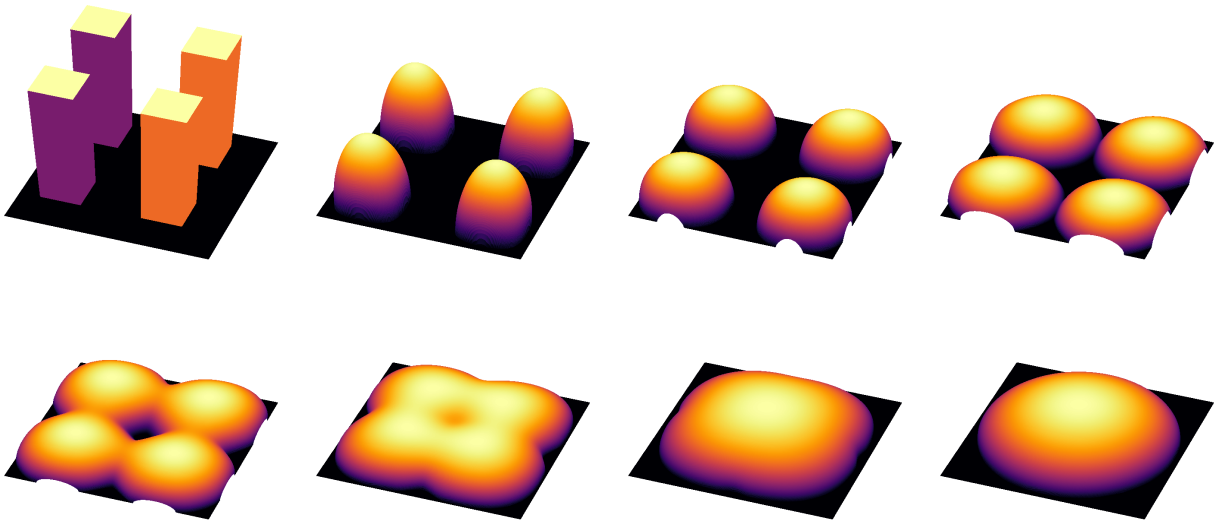


Figure 4.8: Aggregation-diffusion equation with an energy given by (4.48). The images show a 3-d surface plot of the evolution from time $t = 0$ to $t = 10$ (top left to bottom right). The final image is the approximate steady state. Images are 512×512 pixels.

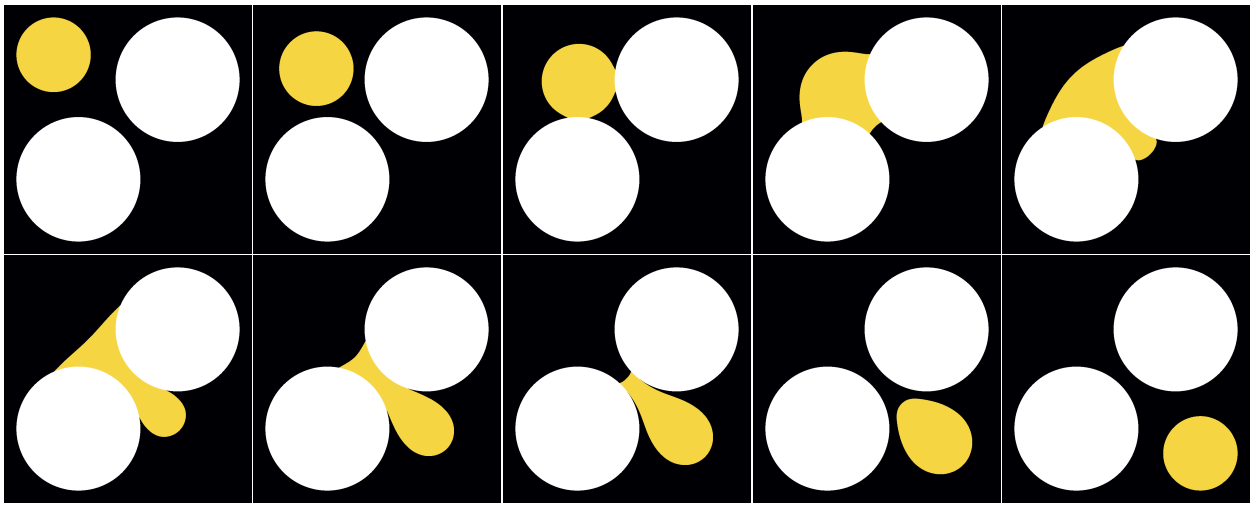


Figure 4.9: Incompressible flow with the energy (4.49), potential (4.50), and obstacle E_1. The images show the evolution from time $t = 0$ to $t = 20$ (top left to bottom right). The final image is the approximate steady state. Images are 1024×1024 pixels. Yellow pixels represents the density and white pixels represents the obstacle.

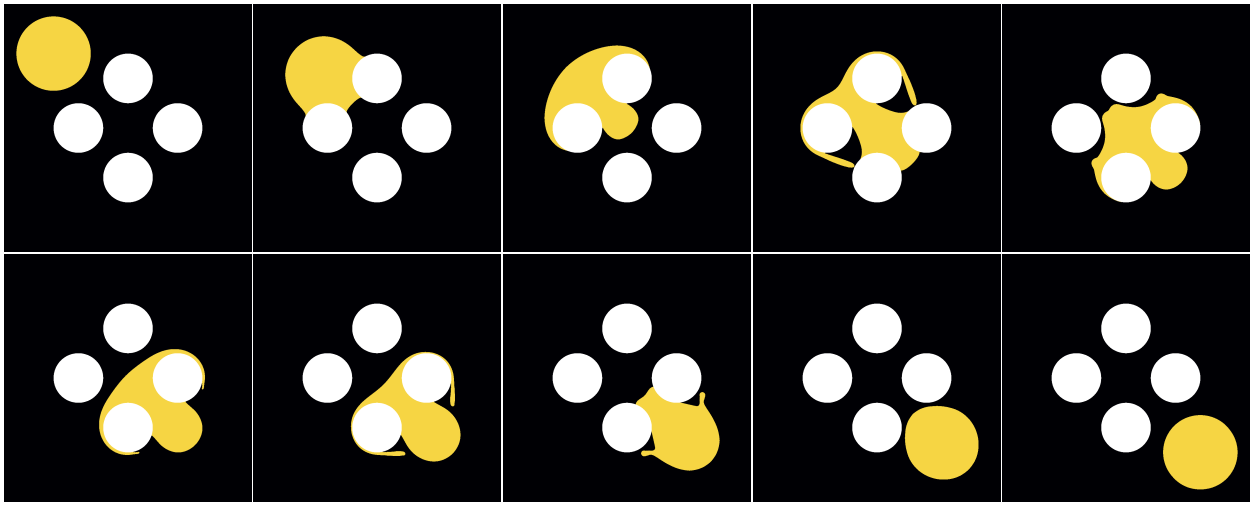


Figure 4.10: Incompressible flow with the energy (4.49), potential (4.50), and obstacle E_2. The images show the evolution from time $t = 0$ to $t = 20$ (top left to bottom right). The final image is the approximate steady state. Images are 1024×1024 pixels. Yellow pixels represents the density and white pixels represents the obstacle.

APPENDIX A

Supplementary materials

A.1 Chapter 3 supplementary materials

Proof of Proposition 3.2.1. From the saddle point problem (3.6), we can rewrite the problem as

$$\begin{aligned}
&\inf_{(\rho_i,m_i)_{i\in\{S,I,R\}}} \sup_{\phi} \mathcal{L}((\rho_i,m_i,\phi_i)_{i\in\{S,I,R\}}) \\
&= \inf_{(\rho_i,m_i)_{i\in\{S,I,R\}}} \sup_{\phi} P((\rho_i,m_i)_{i\in\{S,I,R\}}) - \int_0^T \int_\Omega \sum_{i\in\{S,I,R\}} \phi_i \left(\partial_t\rho_i + \nabla\cdot m_i - \frac{\eta_i^2}{2}\Delta\rho_i\right) dx\,dt \\
&\quad + \int_0^T \int_\Omega \mathcal{Q}((\rho_i,\phi_i)_{i\in\{S,I,R\}})\,dx\,dt
\end{aligned} \tag{A.1}$$

where

$$\mathcal{Q}((\rho_i,\phi_i)_{i\in\{S,I,R\}}) = \beta\rho_S(\phi_I-\phi_S)K*\rho_I + \gamma\rho_I(\phi_R-\phi_I).$$

If $((\rho_i,m_i,\phi_i)_{i\in\{S,I,R\}})$ is the saddle point of the problem, the differential of Lagrangian with respect to ρ_i, m_i, ϕ_i ($i\in\{S,I,R\}$), and $\rho_I(T,\cdot)$ equal to zero. Thus, from $\frac{\delta\mathcal{L}}{\delta\phi_i}=0$ we have

$$\partial_t\rho_i + \nabla\cdot m_i - \frac{\eta_i^2}{2}\Delta\rho_i - \frac{\delta\mathcal{Q}}{\delta\phi_i} = 0, \quad (t,x)\in(0,T)\times\Omega, \quad i=S,I,R.$$

Using integration by parts, we reformulate the Lagrangian function (A.1) as follows.

$$\begin{aligned}
&\mathcal{L}((\rho_i, m_i, \phi_i)_{i\in\{S,I,R\}}) \\
=&E(\rho_I(T,\cdot)) + \int_0^T \int_\Omega \frac{c}{2}(\rho_S+\rho_I+\rho_R)^2 + \int_0^T \int_\Omega \mathcal{Q}((\rho_i,\phi_i)_{i\in\{S,I,R\}})\, dx\, dt \\
&+ \sum_{i=S,I,R} \int_0^T \int_\Omega \frac{\alpha_i |m_i|^2}{2\rho_i} + m_i\cdot\nabla\phi_i + \frac{\eta_i^2}{2}\rho_i\Delta\phi_i\, dx\, dt + \sum_{i=S,I,R}\int_0^T\int_\Omega \rho_i\partial_t\phi_i\, dx\, dt \\
&+ \sum_{i=S,I,R}\int_\Omega \rho_i(0,x)\phi_i(0,x) - \rho_i(T,x)\phi_i(T,x)dx.
\end{aligned}$$

From $\frac{\delta\mathcal{L}}{\delta\rho_i} = 0$ $(i\in\{S,I,R\})$,

$$c(\rho_S+\rho_I+\rho_R) + \frac{\delta\mathcal{Q}}{\delta\rho_i}((\rho_i,\phi_i)_{i\in\{S,I,R\}}) - \frac{\alpha_i|m_i|^2}{2\rho_i^2} + \frac{\eta_i^2}{2}\Delta\phi_i + \partial_t\phi_i = 0 \quad (t,x)\in(0,T)\times\Omega$$

From $\frac{\delta\mathcal{L}}{\delta\rho_I(T,\cdot)} = 0$,

$$\frac{\delta E}{\delta\rho_I(T,\cdot)}(\rho_I(T,\cdot)) = \phi_I(T,\cdot).$$

From $\frac{\delta\mathcal{L}}{\delta m_i} = 0$ $(i\in\{S,I,R\})$,

$$\frac{\alpha_i m_i}{\rho_i} = -\nabla\phi_i \quad (t,x)\in(0,T)\times\Omega, \quad i\in\{S,I,R\}.$$

By replacing $\frac{\alpha_i m_i}{\rho_i} = -\nabla\phi_i$ in $\frac{\delta\mathcal{L}}{\delta\rho_i} = 0$ and $\frac{\delta\mathcal{L}}{\delta\phi_i} = 0$, we derive the result. □

Proof of Lemma 3.3.2. Let $q = (u,p)$. By the definition of $M^{(k)}$, we have

$$\langle q, M^{(k)}q\rangle = \frac{1}{\tau^{(k)}}\|u\|_{L^2}^2 + \frac{1}{\sigma^{(k)}}\|p\|_{\mathcal{H}^{(k)}}^2 - 2\langle u, A_{u^{(k)}}^T p\rangle_{L^2}.$$

Using Young's inequality and Lemma 3.3.1,

$$\begin{aligned}
&\le \left(\frac{1}{\tau^{(k)}}+1\right)\|u\|_{L^2}^2 + \left(\frac{1}{\sigma^{(k)}}+1\right)\|p\|_{\mathcal{H}^{(k)}}^2 \\
&\le \left(\frac{1}{\tau^{(k)}}+1\right)\|u\|_{L^2}^2 + C^2\left(\frac{1}{\sigma^{(k)}}+1\right)\|p\|_{L^2}^2 \le \Theta^2\|q\|_{L^2}^2.
\end{aligned}$$

We are left to show the lower bound. Let $\epsilon > 0$ be such that $\tau^{(k)}\sigma^{(k)} = (1-\epsilon)^2$. Then using Hölder's inequality,

$$\begin{aligned}
\langle q, M^{(k)}q\rangle &\ge \frac{1}{\tau^{(k)}}\|u\|_{L^2}^2 + \frac{1}{\sigma^{(k)}}\|p\|_{\mathcal{H}^{(k)}}^2 - 2\|u\|_{L^2}\|p\|_{\mathcal{H}^{(k)}} \\
&= \frac{1}{\tau^{(k)}}\|u\|_{L^2}^2 + \frac{1}{\sigma^{(k)}}\|p\|_{\mathcal{H}^{(k)}}^2 - \frac{2(1-\epsilon)}{\sqrt{\tau^{(k)}\sigma^{(k)}}}\|u\|_{L^2}\|p\|_{\mathcal{H}^{(k)}}.
\end{aligned}$$

Again, using Young's inequality and Lemma 3.3.1,

$$\geq \frac{\epsilon}{\tau^{(k)}}\|u\|_{L^2}^2 + \frac{\epsilon}{\sigma^{(k)}}\|p\|_{\mathcal{H}^{(k)}}^2 \geq \frac{\epsilon}{\tau^{(k)}}\|u\|_{L^2}^2 + \frac{c^2\epsilon}{\sigma^{(k)}}\|p\|_{L^2}^2 \geq \theta^2\|q\|_{L^2}^2.$$

This proves the claim. □

A.2 Chapter 4 supplementary materials

Proof of Lemma 4.3.1.

Step 1: Derivation of the Hessian. In order to obtain the Hessian of F let us start with the first derivative. We have

$$F(\phi + h) - F(\phi) = \int_\Omega \left[(\phi + h)^c(x) - \phi^c(x)\right] \mu(x)dx.$$

Assume that ϕ is c-convex. Then Proposition 4.2.2 tells us how to differentiate the c-transform, so that we may write

$$\int_\Omega \left[(\phi + h)^c(x) - \phi^c(x)\right] \mu(x)dx = \int_\Omega h(T_\phi(x))\, \mu(x)dx + o(h).$$

Therefore $\delta F(\phi)(h) = \int_\Omega h(T_\phi(x))\, \mu(x)dx$. To derive the Hessian of F we similarly compute

$$\delta F(\phi + h)(h) - \delta F(\phi)(h) = \int_\Omega \left[h(T_{\phi+h}(x)) - h(T_\phi(x))\right] \mu(x)dx.$$

We must now differentiate the maps T_ϕ with respect to ϕ. By Proposition 4.2.2 we know that $T_\phi(x) = x - \tau\nabla\phi^c(x)$. As a consequence

$$\begin{aligned} T_{\phi+h}(x) - T_\phi(x) &= -\tau\nabla[(\phi + h)^c - \phi^c](x) \\ &= -\tau\nabla(h \circ T_\phi)(x) + o(h) \\ &= -\tau DT_\phi(x)^T\nabla h(T_\phi(x)) + o(h). \end{aligned}$$

Note that $DT_\phi = I_{d\times d} - \tau D^2\phi^c$ is a symmetric matrix. We deduce from the above computations

that

$$\delta F(\phi+h)(h) - \delta F(\phi)(h) = \int_\Omega \nabla h(T_\phi(x)) \cdot (-\tau) DT_\phi(x) \nabla h(T_\phi(x))\, \mu(x) dx + o(h),$$

from which we conclude that

$$\delta^2 F(\phi)(h,h) = -\tau \int_\Omega \nabla h(T_\phi(x)) \cdot DT_\phi(x) \nabla h(T_\phi(x))\, \mu(x) dx.$$

Since our goal is to bound this Hessian by a norm of h we do the change of variable $y = T_\phi(x)$, or equivalently $x = S_{\phi^c}(y)$ since S_{ϕ^c} is the inverse of T_ϕ, see Proposition 4.2.2. We obtain

$$\delta^2 F(\phi)(h,h) = -\tau \int_\Omega \nabla h(y) \cdot DT_\phi(S_{\phi^c}(y)) \nabla h(y)\, \mu(S_{\phi^c}(y)) \det DS_{\phi^c}(y) dy.$$

Note that DS_{ϕ^c} is a positive semi-definite matrix and therefore no absolute value is needed on the determinant term. Moreover we have $DT_\phi(S_{\phi^c}(y)) = DS_{\phi^c}(y)^{-1}$ and putting this term together with the determinant we can form the cofactor matrix $\mathrm{cof}(DS) = \det(DS) DS^{-1}$. As a result we obtain the expression

$$\delta^2 F(\phi)(h,h) = -\tau \int_\Omega \nabla h(y) \cdot \mathrm{cof}(DS_{\phi^c}(y)) \nabla h(y)\, \mu(S_{\phi^c}(y)) dy.$$

Step 2: Hessian bounds. Since ϕ is c-convex, $\phi = \phi^{c\bar{c}}$ and therefore $S_{\phi^c}(y) = y + \tau \nabla \phi(y)$. The c-convexity of ϕ also implies that the symmetric matrix $DS_{\phi^c}(y) = I_{d\times d} + \tau D^2\phi(y)$ is positive semi-definite. Assume now that $I_{d\times d} + \tau D^2\phi(y) \leq \Lambda\, I_{d\times d}$ for all $y \in \Omega$. Then $I_{d\times d} + \tau D^2\phi(y)$ is a symmetric matrix with eigenvalues between 0 and Λ. By general properties of the cofactor matrix the eigenvalues of $\mathrm{cof}(DS_{\phi^c}(y))$ lie between 0 and Λ^{d-1} where d is the space dimension. We immediately deduce

$$-\delta^2 F(\phi)(h,h) \leq \tau \Lambda^{d-1} \|\mu\|_{L^\infty} \int_\Omega |\nabla h(y)|^2\, dy.$$

□

Lemma A.2.1. *Suppose that $E \subset \mathbb{R}^d$ is a bounded set with C^2 boundary and let $R := Reach(\partial E)$. Let $u_0 : \mathbb{R}^d \to \mathbb{R}$ be a solution to the Eikonal equation $|\nabla u_i| = 1$ where $u_0 < 0$ inside E and $u_0 > 0$ outside E and set $u_1 = -u_0$. Let*

$$E_r^0 = \{x \in \mathbb{R}^d : u_0(x) \in (0, r)\}.$$

and

$$E_r^1 = \{x \in \mathbb{R}^d : u_1(x) \in (0, r)\}.$$

If $g : \mathbb{R}^d \to \mathbb{R}$ is a smooth function, then for $i = 0, 1$

$$\int_{\partial E} |g(x)| ds(x) \leq \inf_{0<r<R} \Big(\int_{E_r^i} |\nabla g(x)| \, dx + C_i(E, r) \int_{E_r^i} |g(x)| \, dx \Big)$$

where

$$C_i(E, r) = \inf_{0<r'<r} \frac{1}{r'} + \sup_{x \in E_{r'}^i} (\Delta u_i(x))_+$$

Remark A.2.1. The reach of ∂E is the largest number r such that the characteristics of u_0 do not cross in $E_r^0 \cup E_r^1$. When ∂E is C^2, the reach must be strictly positive and the Laplacian Δu must be bounded on $E_r^0 \cup E_r^1$ for all r smaller than the reach of ∂E.

Remark A.2.2. If E is a convex set, then $C_1(E, r) = \frac{1}{r}$.

Proof. Note that if $x \in \partial E$ and $n(x)$ is the outward facing normal at x, then $\nabla u_0(x) = n(x)$. Therefore,

$$\int_{\partial E} |g(x)| \, ds(x) = \int_{\partial E} |g(x)| \nabla u_0(x) \cdot n(x) \, ds(x)$$

For some $r \in (0, R)$ let $\alpha_r : \mathbb{R} \to \mathbb{R}$ be a function such that

$$\alpha_r'(t) = \begin{cases} 1 & \text{if } t \geq 0, \\ 1 + \frac{t}{r} & \text{if } t \in (-r, 0), \\ 0 & \text{if } t \leq -r. \end{cases}$$

We then have

$$\int_{\partial E} |g(x)| \nabla u_0(x) \cdot n(x) \, ds(x) = \int_{\partial E} |g(x)| \nabla \Big(\alpha_r \big(u_0(x) \big) \Big) \cdot n(x) \, ds(x) =$$

$$\int_E \nabla \cdot \Big(|g(x)| \nabla \Big(\alpha_r \big(u_0(x) \big) \Big) \Big) \, dx$$

where the last equality follows from Stokes Theorem. Expanding out the derivatives and noting that $\alpha_r'(t) \in [0,1]$, $\alpha_r''(t) \in [0, \frac{1}{r}]$ and $\alpha'(u_0(x)), \alpha''(u_0(x))$ both vanish for x outside of E_r^0, we get

$$\int_{\partial E} |g(x)| ds(x) \leq \int_{E_r^0} |\nabla g(x)| \, dx + \int_{E_r^0} |g(x)| \big((\Delta u_0(x))_+ + \frac{1}{r} \big) \, dx \leq$$

$$\int_{E_r^0} |\nabla g(x)| \, dx + C_0(E,r) \int_{E_r^0} |g(x)| \, dx$$

Our choice of r was arbitrary, thus we can take an inf over $r \in (0, R)$ to conclude the result when $i = 0$.

To tackle the case $i = 1$, we will employ a nearly identical argument, except we will use Stokes Theorem to convert the boundary integral into an integral over $\mathbb{R}^d \setminus E$. Since $\nabla u_1(x) \cdot n(x) = -1$ for $x \in \partial E$, we have

$$\int_{\partial E} |g(x)| \, ds(x) = - \int_{\partial E} |g(x)| \nabla \Big(\alpha_r \big(u_1(x) \big) \Big) \cdot n(x) \, ds(x) =$$

$$\int_{\mathbb{R}^d \setminus E} \nabla \cdot \Big(|g(x)| \nabla \Big(\alpha_r \big(u_1(x) \big) \Big) \Big) \, dx$$

Now an identical argument to the one above gives the bound for the case $i = 1$. □

Corollary A.2.2. *Suppose that $E \subset \Omega$ is a set with C^2 boundary and let $R := \min(\mathit{Reach}(\partial E), \mathit{dist}(E, \partial\Omega))$. Define u_i, E_r^i, and $C_i(E,r)$ as in Lemma A.2.1, and let*

$$C(E, \Omega) = \min_{i \in \{0,1\}} \inf_{0 < r < R} C_i(E, r).$$

If $h : \Omega \to \mathbb{R}$ is an H^1 function, then

$$\int_{\partial E} |h(x)|^2 \, dx \leq \frac{1}{C} \int_\Omega |\nabla h(x)|^2 + 2C \int_\Omega |h(x)|^2 \, dx,$$

where

$$C = \max(1, C(E, \Omega)).$$

Proof. Suppose first that $h : \Omega \to \mathbb{R}$ is a smooth function. By Lemma A.2.1, we have

$$\int_{\partial E} |h(x)|^2 \, ds(x) \leq \inf_{0<r<R} \Big(\int_{E_r^i} 2|h(x)\nabla h(x)| \, dx + C_i(E,r) \int_{E_r^i} |h(x)|^2 \, dx \Big)$$

for $i = 0, 1$. Clearly this is bounded from above by

$$\int_{\Omega} 2|h(x)\nabla h(x)| \, dx + \inf_{0<r<R} C_i(E,r) \int_{\Omega} |h(x)|^2 \, dx$$

Taking a minimum over $i = 0, 1$, we can conclude that

$$\int_{\partial E} |h(x)|^2 \, ds(x) \leq \int_{\Omega} 2|h(x)\nabla h(x)| \, dx + C \int_{\Omega} |h(x)|^2 \, dx.$$

We can then use Cauchy-Schwarz to get

$$\int_{\partial E} |h(x)|^2 \, ds(x) \leq \frac{1}{C} \int_{\Omega} |\nabla h(x)| \, dx + 2C \int_{\Omega} |h(x)|^2 \, dx.$$

The result extends to H^1 functions thanks to the continuity of the trace operator over H^1. □

Proof of Theorem 4.3.4.

Recall that $I(\psi) = \int_\Omega \psi(x) \, \mu(x) dx - U^*(\psi^{\bar{c}})$.

Step 1: formula for the Hessian of I. The derivation of the Hessian of I is similar to the one of J (see for instance the proof of Lemma 4.3.1). Using the formulas for the first variation of the c-transform in Proposition 4.2.2 we can check that

$$\delta I(\psi)h = -\delta U^*(\psi^{\bar{c}})(h \circ S_\psi),$$

for any test function h. To obtain the Hessian of I, we need to differentiate S_ψ. As in the proof of Lemma 4.3.1 we can show that $S_{\psi+h}(y) - S_\psi(y) = \tau DS_\psi(y)^T \nabla h(S_\psi(y)) + o(h)$. This implies

$$\delta^2 I(\psi)(h,h) = -\delta^2 U^*(\psi^{\bar{c}})(h \circ S_\psi, h \circ S_\psi) - \tau \int_\Omega \eta(y) \, \nabla h(S_\psi(y)) \cdot DS_\psi(y) \nabla h(S_\psi(y)) \, dy,$$

where we set $\eta = \delta U^*(\psi^{\bar{c}})$. Thus as for J, the Hessian of I contains two terms which we can bound separately, $\delta^2 I(\psi)(h,h) = -(A) - (B)$.

Step 2: Bound on (B). Do the change of variables $x = S_\psi(y)$, i.e. $y = T_{\psi^{\bar{c}}}(x)$ in (B). We obtain

$$(B) = \tau \int_\Omega \eta(T_{\psi^{\bar{c}}}(x)) \nabla h(x) \cdot \operatorname{cof} DT_{\psi^{\bar{c}}}(x) \nabla h(x)\, dx,$$

which can be bounded above by $\tau \|\eta\|_{L^\infty} \Lambda^{d-1} \|\nabla h\|_{L^2}$ in the same spirit as in the proof of Lemma 4.3.1. Moreover $\|\eta\|_{L^\infty} \leq \rho_{\max}$. Indeed, assuming $V(x) \geq 0$ we have for all $x \in \Omega$

$$\eta(x) = \delta U^*(\psi^{\bar{c}})(x) = (u_m^*)'(\psi^{\bar{c}}(x) - V(x)) \leq (u_m^*)'(\psi^{\bar{c}}(x)) \leq \rho_{\max},$$

by monotonicity of $(u_m^*)'$ and by definition of $\rho_{\max}$. As a consequence

$$(B) \leq \tau \rho_{\max} \Lambda^{d-1} \|\nabla h\|_{L^2}^2.$$

Step 3: Bound on (A). We have

$$(A) = \delta U^*(\psi^{\bar{c}})(h \circ S_\psi, h \circ S_\psi) = \int_\Omega (u_m^*)''(\psi^{\bar{c}}(y) - V(y)) |h(S_\psi)|^2\, dy.$$

Do again the change of variables $y = T_{\psi^{\bar{c}}}(x)$ to obtain

$$(A) = \int_\Omega (u_m^*)''(p(x)) |h(x)|^2 \det(DT_{\psi^{\bar{c}}}(x))\, dx,$$

where we recall that $p(x) = \psi^{\bar{c}}(T_{\psi^{\bar{c}}}(x)) - V(T_{\psi^{\bar{c}}}(x))$. We bound the determinant term by Λ^d. Then, to go further we must distinguish between the three cases $1 \leq m \leq 2$, $2 < m < \infty$ and $m = \infty$.

When $1 \leq m \leq 2$, the function $(u_m^*)''$ is increasing and therefore

$$(u_m^*)''(p(x)) \leq (u_m^*)''(M) = u_m''(\rho_{\max})^{-1},$$

where $M = \sup_x \delta U(\mu)(x)$ (see the maximum principle and the related discussion when $\rho_{\max}$ is defined in equation (4.32)). To sum up,

$$(A) \leq u''(\rho_{\max})^{-1} \Lambda^d \|h\|_{L^2}^2.$$

When $2 < m \leq \infty$, one can follow the same line of proof as in the case of J, using now the function $p(x)$ instead of $\phi(x) - V(x)$ which modifies the related constants accordingly.

□

REFERENCES

[ACB17a] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein GAN." *arXiv:1701.07875 [cs, stat]*, 2017a.

[ACB17b] M. Arjovsky, S. Chintala, and L. Bottou. "Wasserstein generative adversarial networks." In *International conference on machine learning*, pp. 214–223. PMLR, 2017b.

[AKS18] L. M. Briceño Arias, D. Kalise, and F. J. Silva. "Proximal methods for stationary mean field games with local couplings." *SIAM J. Control Optim.*, **56**(2):801–836, 2018.

[AKY14] D. Alexander, I. Kim, and Y. Yao. "Quasi-static evolution and congested crowd transport." *Nonlinearity*, **27**(4):823–858, 2014.

[AL19] Y. Achdou and J.-M. Lasry. "Mean field games for modeling crowd motion." In *Contributions to partial differential equations and applications*, volume 47 of *Comput. Methods Appl. Sci.*, pp. 17–42. Springer, Cham, 2019.

[All17] L. J. Allen. "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis." *Infectious Disease Modelling*, **2**(2):128–142, 2017.

[Aro77] D. Aronson. "The asymptotic speed of propagation of a simple epidemic." In *Nonlinear diffusion*, volume 14, pp. 1–23. Pitman London, 1977.

[Bar96] G. I. Barenblatt. *Scaling, self-similarity, and intermediate asymptotics*, volume 14 of *Cambridge Texts in Applied Mathematics*. Cambridge University Press, Cambridge, 1996. With a foreword by Ya. B. Zeldovich.

[Bar03] G. I. Barenblatt. *Scaling*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 2003. With a foreword by Alexandre Chorin.

[BB00] 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.

[BC15] J.-D. Benamou and G. Carlier. "Augmented Lagrangian methods for transport optimization, mean field games and degenerate elliptic equations." *J. Optim. Theory Appl.*, **167**(1):1–26, 2015.

[BCL16] 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*, pp. 1–17. EDP Sci., Les Ulis, 2016.

[BCM16] J.-D. Benamou, G. Carlier, Q. Mérigot, and E. Oudet. "Discretization of functionals involving the Monge–Ampère operator." *Numer. Math.*, **134**(3):611–636, 2016.

[BCW10] M. Burger, J. A. Carrillo, and M.-T. Wolfram. "A mixed finite element method for nonlinear diffusion equations." *Kinet. Relat. Models*, **3**(1):59–83, 2010.

[BDM13] M. Burger, M. Di Francesco, P. Markowich, and M.-T. Wolfram. "Mean field games with nonlinear mobilities in pedestrian dynamics." *arXiv preprint arXiv:1304.5201*, 2013.

[BFM20] A. L. Bertozzi, E. Franco, G. Mohler, M. B. Short, and D. Sledge. "The challenges of modeling and forecasting the spread of covid-19." *arXiv preprint arXiv:2004.04741*, 2020.

[Bre91] Y. Brenier. "Polar factorization and monotone rearrangement of vector-valued functions." *Comm. Pure Appl. Math.*, **44**(4):375–417, 1991.

[BRR20] H. Berestycki, J.-M. Roquejoffre, and L. Rossi. "Propagation of epidemics along lines with fast diffusion." *arXiv preprint arXiv:2005.01859*, 2020.

[BT09] A. Beck and M. Teboulle. "A fast iterative shrinkage-thresholding algorithm for linear inverse problems." *SIAM journal on imaging sciences*, **2**(1):183–202, 2009.

[CC10] S. Chinviriyasit and W. Chinviriyasit. "Numerical modelling of an sir epidemic model with diffusion." *Applied Mathematics and Computation*, **216**(2):395–409, 2010.

[CCW19] J. A. Carrillo, K. Craig, L. Wang, and C. Wei. "Primal dual methods for Wasserstein gradient flows." *arXiv preprint arXiv:1901.08081*, 2019.

[CCY19] J. A. Carrillo, K. Craig, and Y. Yao. "Aggregation-diffusion equations: dynamics, asymptotics, and singular limits." In *Active Particles, Volume 2*, pp. 65–108. Springer, 2019.

[CDP17] G. Carlier, V. Duval, G. Peyré, and B. Schmitzer. "Convergence of entropic schemes for optimal transport and gradient flows." *SIAM J. Math. Anal.*, **49**(2):1385–1418, 2017.

[CGC02] T. Caraco, S. Glavanakov, G. Chen, J. E. Flaherty, T. K. Ohsumi, and B. K. Szymanski. "Stage-structured infection transmission and a spatial epidemic: a model for lyme disease." *The American Naturalist*, **160**(3):348–359, 2002.

[CGN17] Y. Chen, T. Georgiou, L. Ning, and A. Tannenbaum. "Matricial Wasserstein-1 Distance." *IEEE Control Systems Letters*, **PP**:1–1, 2017.

[CGT19] Y. Chen, T. T. Georgiou, and A. Tannenbaum. "Interpolation of matrices and matrix-valued densities: The unbalanced case." *European Journal of Applied Mathematics*, **30**(3):458–480, 2019.

[CL18] L. Chayes and H. K. Lei. "Transport and equilibrium in non-conservative systems." *Advances in Differential Equations*, **23**(1/2):1–64, 2018.

[CLO18] Y. T. Chow, W. Li, S. Osher, and W. Yin. "Algorithm for Hamilton-Jacobi equations in density space via a generalized Hopf formula." *arXiv:1805.01636 [math]*, 2018.

[CM10] J. A. Carrillo and J. S. Moll. "Numerical simulation of diffusive and aggregation phenomena in nonlinear continuity equations by evolving diffeomorphisms." *SIAM J. Sci. Comput.*, **31**(6):4305–4329, 2009/10.

[CP11a] A. Chambolle and T. Pock. "A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging." *Journal of Mathematical Imaging and Vision*, **40**(1):120–145, 2011a.

[CP11b] A. Chambolle and T. Pock. "A first-order primal-dual algorithm for convex problems with applications to imaging." *J. Math. Imaging Vision*, **40**(1):120–145, 2011b.

[CP16] A. Chambolle and T. Pock. "On the ergodic convergence rates of a first-order primal-dual algorithm." *Math. Program.*, **159**(1-2, Ser. A):253–287, 2016.

[CPS15] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. "Unbalanced Optimal Transport: Geometry and Kantorovich Formulation." *arXiv:1508.05216 [math]*, 2015.

[CPS18] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. "An Interpolating Distance Between Optimal Transport and Fisher–Rao Metrics." *Foundations of Computational Mathematics*, **18**(1):1–44, 2018.

[CV17a] C. Clason and T. Valkonen. "Primal-dual extragradient methods for nonlinear nonsmooth pde-constrained optimization." *SIAM Journal on Optimization*, **27**(3):1314–1339, 2017a.

[CV17b] C. Clason and T. Valkonen. "Stability of saddle points via explicit coderivatives of pointwise subdifferentials." *Set-Valued and Variational Analysis*, **25**(1):69–112, 2017b.

[CWX20] J. A. Carrillo, L. Wang, W. Xu, and M. Yan. "Variational asymptotic preserving scheme for the Vlasov–Poisson–Fokker–Planck system." *arXiv preprint arXiv:2007.01969*, 2020.

[Die79] O. Diekmann. "Run for your life. a note on the asymptotic speed of propagation of an epidemic." *Journal of Differential Equations*, **33**(1):58–73, 1979.

[DMS16] G. De Philippis, A. R. Mészáros, F. Santambrogio, and B. Velichkov. "BV estimates in optimal transportation and applications." *Arch. Ration. Mech. Anal.*, **219**(2):829–860, 2016.

[DWH15] F. De Goes, C. Wallez, J. Huang, D. Pavlov, and M. Desbrun. "Power particles: an incompressible fluid solver based on power diagrams." *ACM Trans. Graph.*, **34**(4):50–1, 2015.

[EHL18] W. E, J. Han, and Q. Li. "A Mean-Field Optimal Control Formulation of Deep Learning." *arXiv:1807.01083 [cs, math]*, 2018.

[Eva10] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[EY18] B. Engquist and Y. Yang. "Seismic Inversion and the Data Normalization for Optimal Transport." *arXiv:1810.08686 [math]*, 2018.

[Eyr98] D. J. Eyre. "Unconditionally gradient stable time marching the Cahn–Hilliard equation." *MRS Proceedings*, **529**:39, 1998.

[FH16] I. Faragó and R. Horváth. "Qualitatively adequate numerical modelling of spatial sirs-type disease propagation." *Electronic Journal of Qualitative Theory of Differential Equations*, **2016**(12):1–14, 2016.

[Gal16] A. Galichon. *Optimal transport methods in economics*. Princeton University Press, 2016.

[Gan94] W. Gangbo. "An elementary proof of the polar factorization of vector-valued functions." *Archive for rational mechanics and analysis*, **128**(4):381–399, 1994.

[Gan95a] W. Gangbo. *Quelques problèmes d'analyse non convexe. Habilitation à diriger des recherches en mathématiques.* Habilitation, Université de Metz, 1995a.

[Gan95b] W. Gangbo. "Quelques problemes d'analyse non convexe." *Habilitation à diriger des recherches en mathématiques. Université de Metz (Janvier 1995)*, 1995b.

[GBK01] B. T. Grenfell, O. N. Bjórnstad, and J. Kappey. "Travelling waves and spatial hierarchies in measles epidemics." *Nature*, **414**(6865):716–723, 2001.

[GHL19] A. Garbuno-Inigo, F. Hoffmann, W. Li, and A. M. Stuart. "Interacting Langevin Diffusions: Gradient Structure And Ensemble Kalman Sampler." *arXiv:1903.08866 [math]*, 2019.

[GLO19] W. Gangbo, W. Li, S. Osher, and M. Puthawala. "Unnormalized optimal transport." *Journal of Computational Physics*, **399**:108940, 2019.

[GM96] W. Gangbo and R. J. McCann. "The geometry of optimal transportation." *Acta Math.*, **177**(2):113–161, 1996.

[GM18] T. O. Gallouët and Q. Mérigot. "A lagrangian scheme à la brenier for the incompressible euler equations." *Foundations of Computational Mathematics*, **18**(4):835–865, 2018.

[GNP15] D. A. Gomes, L. Nurbekyan, and E. A. Pimentel. *Economic models and mean-field games theory*. IMPA Mathematical Publications. Instituto Nacional de Matemática Pura e Aplicada (IMPA), Rio de Janeiro, 2015.

[Gom14] D. A. Gomes et al. "Mean field games models—a brief survey." *Dynamic Games and Applications*, **4**(2):110–154, 2014.

[HI95] Y. Hosono and B. Ilyas. "Traveling waves for a simple diffusive epidemic model." *Mathematical Models and Methods in Applied Sciences*, **5**(07):935–966, 1995.

[HMC06] M. Huang, R. P. Malhamé, and P. E. Caines. "Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle." *Commun. Inf. Syst.*, **6**(3):221–251, 2006.

[JC14] A. Jaichuang and W. Chinviriyasit. "Numerical modelling of influenza model with diffusion." *International Journal of Applied Physics and Mathematics*, **4**(1):15, 2014.

[JKL20] J. Jang, H.-D. Kwon, and J. Lee. "Optimal control problem of an sir reaction–diffusion model with inequality constraints." *Mathematics and Computers in Simulation*, **171**:136–151, 2020.

[JKO98] R. Jordan, D. Kinderlehrer, and F. Otto. "The variational formulation of the Fokker–Planck equation." *SIAM journal on mathematical analysis*, **29**(1):1–17, 1998.

[JKT20a] M. Jacobs, I. Kim, and J. Tong. "Darcy's law with a source term." *Archive for Rational Mechanics and Analysis*, pp. 1–45, 2020a.

[JKT20b] M. Jacobs, I. Kim, and J. Tong. "The L^1-contraction principle in optimal transport." *arXiv preprint arXiv:2006.09557*, 2020b.

[JL20] M. Jacobs and F. Léger. "A fast approach to optimal transport: the back-and-forth method." *Numerische Mathematik*, pp. 1–32, 2020.

[JLL19] M. Jacobs, F. Léger, W. Li, and S. Osher. "Solving large-scale optimization problems with a convergence rate independent of grid size." *SIAM Journal on Numerical Analysis*, **57**(3):1100–1123, 2019.

[JLL21] Jacobs, Matt, Lee, Wonjun, and Léger, Flavien. "The back-and-forth method for wasserstein gradient flows." *ESAIM: COCV*, **27**:28, 2021.

[JR88] B. Jovanovic and R. W. Rosenthal. "Anonymous sequential games." *Journal of Mathematical Economics*, **17**(1):77 – 87, 1988.

[Kal84] A. Källén. "Thresholds and travelling waves in an epidemic model for rabies." *Nonlinear Analysis: Theory, Methods & Applications*, **8**(8):851–856, 1984.

[Kan06] L. V. Kantorovich. "On the translocation of masses." *Journal of mathematical sciences*, **133**(4):1381–1382, 2006.

[Ken65] D. G. Kendall. "Mathematical models of the spread of infection." *Mathematics and computer science in biology and medicine*, pp. 213–225, 1965.

[KM27] W. O. Kermack and A. G. McKendrick. "A contribution to the mathematical theory of epidemics." *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, **115**(772):700–721, 1927.

[LES18] A. Lahrouz, H. El Mahjour, A. Settati, and A. Bernoussi. "Dynamics and optimal control of a non-linear epidemic model with relapse and cure." *Physica A: Statistical Mechanics and its Applications*, **496**:299–317, 2018.

[LFL20] A. T. Lin, S. W. Fung, W. Li, L. Nurbekyan, and S. J. Osher. "Apac-net: Alternating the population and agent control via two neural networks to solve high-dimensional stochastic mean field games.", 2020.

[Li18] W. Li. "Geometry of probability simplex via optimal transport." *arXiv:1803.06360 [math]*, 2018.

[LJL20] S. Liu, M. Jacobs, W. Li, L. Nurbekyan, and S. J. Osher. "Computational methods for nonlocal mean field games with applications." *arXiv preprint arXiv:2004.12210*, 2020.

[LL06a] J.-M. Lasry and P.-L. Lions. "Jeux à champ moyen. I. Le cas stationnaire." *C. R. Math. Acad. Sci. Paris*, **343**(9):619–625, 2006a.

[LL06b] J.-M. Lasry and P.-L. Lions. "Jeux à champ moyen. II. Horizon fini et contrôle optimal." *C. R. Math. Acad. Sci. Paris*, **343**(10):679–684, 2006b.

[LL07] J.-M. Lasry and P.-L. Lions. "Mean field games." *Japanese journal of mathematics*, **2**(1):229–260, 2007.

[LL19] F. Léger and W. Li. "Hopf-Cole transformation via generalized Schr\"odinger bridge problem." *arXiv:1901.09051 [math]*, 2019.

[LLL16] A. Lachapelle, J.-M. Lasry, C.-A. Lehalle, and P.-L. Lions. "Efficiency of the price formation process in presence of high frequency participants: a mean field game analysis." *Mathematics and Financial Economics*, **10**(3):223–262, 2016.

[LLL21a] W. Lee, R. Lai, W. Li, and S. Osher. "Generalized unnormalized optimal transport and its fast algorithms." *Journal of Computational Physics*, **436**:110041, 2021a.

[LLL21b] W. Lee, S. Liu, W. Li, and S. Osher. "Mean field control problems for vaccine distribution." *arXiv preprint arXiv:2104.11887*, 2021b.

[LLO18] A. T. Lin, W. Li, S. Osher, and G. Montufar. "Wasserstein proximal of GANs." 2018.

[LLT21] W. Lee, S. Liu, H. Tembine, W. Li, and S. Osher. "Controlling propagation of epidemics via mean-field control." *SIAM Journal on Applied Mathematics*, **81**(1):190–207, 2021.

[LMS18] M. Liero, A. Mielke, and G. Savaré. "Optimal Entropy-Transport problems and a new Hellinger–Kantorovich distance between positive measures." *Inventiones mathematicae*, **211**(3):969–1117, 2018.

[LMS20] H. Leclerc, Q. Mérigot, F. Santambrogio, and F. Stra. "Lagrangian discretization of crowd motion and linear diffusion." *SIAM Journal on Numerical Analysis*, **58**(4):2093–2118, 2020.

[LRO18] W. Li, E. K. Ryu, S. Osher, W. Yin, and W. Gangbo. "A Parallel Method for Earth Mover's Distance." *Journal of Scientific Computing*, **75**(1):182–197, 2018.

[Luc97] Y. Lucet. "Faster than the fast Legendre transform, the linear-time Legendre transform." *Numer. Algorithms*, **16**(2):171–185, 1997.

[LYO18] W. Li, P. Yin, and S. Osher. "Computations of Optimal Transport Distance with Fisher Information Regularization." *Journal of Scientific Computing*, **75**(3):1581–1595, 2018.

[LZM19] K. Li, G. Zhu, Z. Ma, and L. Chen. "Dynamic stability of an siqs epidemic network and its optimal control." *Communications in Nonlinear Science and Numerical Simulation*, **66**:84–95, 2019.

[Mon81] G. Monge. "Mémoire sur la théorie des déblais et des remblais." *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[MRS15] J. Maas, M. Rumpf, C. Schönlieb, and S. Simon. "A generalized model for optimal transport of images including dissipation and density modulation." *arXiv:1504.01988 [math]*, 2015.

[Mur01] J. Murray. *Mathematical biology II: spatial models and biomedical applications*. Springer New York, 2001.

[Nes83] Y. E. Nesterov. "A method for solving the convex programming problem with convergence rate o (1/kˆ 2)." In *Dokl. akad. nauk Sssr*, volume 269, pp. 543–547, 1983.

[Nes13] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[Ott01] F. Otto. "The geometry of dissipative evolution equations: the porous medium equation." *Comm. Partial Differential Equations*, **26**(1-2):101–174, 2001.

[PB14] N. Parikh and S. Boyd. "Proximal algorithms." *Foundations and Trends® in Optimization*, **1**(3):127–239, 2014.

[PC18] G. Peyré and M. Cuturi. "Computational Optimal Transport." *arXiv:1803.00567 [stat]*, 2018.

[Pey15] G. Peyré. "Entropic approximation of Wasserstein gradient flows." *SIAM J. Imaging Sci.*, **8**(4):2323–2351, 2015.

[PR14] B. Piccoli and F. Rossi. "Generalized Wasserstein Distance and its Application to Transport Equations with Source." *Archive for Rational Mechanics and Analysis*, **211**(1):335–358, 2014.

[PR16] B. Piccoli and F. Rossi. "On Properties of the Generalized Wasserstein Distance." *Archive for Rational Mechanics and Analysis*, **222**(3):1339–1365, 2016.

[PW08] O. Pele and M. Werman. "A linear time histogram metric for improved sift matching." In *European conference on computer vision*, pp. 495–508. Springer, 2008.

[RLY17] E. K. Ryu, W. Li, P. Yin, and S. Osher. "Unbalanced and Partial L1 Monge–Kantorovich Problem: A Scalable Parallel First-Order Method." *Journal of Scientific Computing*, pp. 1–18, 2017.

[ROL19] L. Ruthotto, S. Osher, W. Li, L. Nurbekyan, and S. W. Fung. "A machine learning framework for solving high-dimensional mean field game and mean field control problems.", 2019.

[Rua07] S. Ruan. "Spatial-temporal dynamics in nonlocal epidemiological models." In *Mathematics for life science and medicine*, pp. 97–122. Springer, 2007.

[RW09] R. T. Rockafellar and R. J.-B. Wets. *Variational analysis*, volume 317. Springer Science & Business Media, 2009.

[San15] F. Santambrogio. *Optimal transport for applied mathematicians*, volume 87 of *Progress in Nonlinear Differential Equations and their Applications*. Birkhäuser/Springer, Cham, 2015. Calculus of variations, PDEs, and modeling.

[SS78] S. P. Sethi and P. W. Staats. "Optimal control of some simple deterministic epidemic models." *Journal of the Operational Research Society*, **29**(2):129–136, 1978.

[Thi77] H. Thieme. "A model for the spatial spread of an epidemic." *Journal of Mathematical Biology*, **4**(4):337–351, 1977.

[TPK17a] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev. "A transportation l^p distance for signal analysis." *Journal of mathematical imaging and vision*, **59**(2):187–210, 2017a.

[TPK17b] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepuaźev. "A Transportation L^P Distance for Signal Analysis." *J. Math. Imaging Vis.*, **59**(2):187–210, 2017b.

[Vaz07] J. L. Vázquez. *The porous medium equation: mathematical theory*. Oxford University Press, 2007.

[Vil09] C. Villani. *Optimal Transport: Old and New*. Number 338 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin, 2009.

[WOS10] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, and G. K. Rohde. "An optimal transportation approach for nuclear structure-based pathology." *IEEE transactions on medical imaging*, **30**(3):621–631, 2010.

[WSB13] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde. "A linear optimal transportation framework for quantifying and visualizing variations in sets of images." *International journal of computer vision*, **101**(2):254–269, 2013.

[WW10] Z.-C. Wang and J. Wu. "Travelling waves of a diffusive kermack–mckendrick epidemic model with non-local delayed transmission." *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **466**(2113):237–261, 2010.

[YES18] Y. Yang, B. Engquist, J. Sun, and B. F. Hamfeldt. "Application of optimal transport and the quadratic Wasserstein metric to full-waveform inversion." *GEOPHYSICS*, **83**(1):R43–R62, 2018.

[ZYH07] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum. "An image morphing technique based on optimal mass preserving mapping." *IEEE Transactions on Image Processing*, **16**(6):1481–1495, 2007.