Acta Numerica (2021), pp. 249–325 Printed in the United Kingdom
doi:10.1017/S0962492921000040

Optimal transportation, modelling and numerical simulation

Jean-David Benamou
INRIA Paris, Paris 12e, France
E-mail: Jean-David.Benamou@inria.fr

We present an overview of the basic theory, modern optimal transportation extensions and recent algorithmic advances. Selected modelling and numerical applications illustrate the impact of optimal transportation in numerical analysis.

CONTENTS

Foreword 249
Outline 252
Notation 254
1 Introduction 254
2 Kantorovich duality and the reflector problem 259
3 Wasserstein distance and entropic interpolation 268
4 Dynamic optimal transportation 280
5 Variational formulations for Euler equations 285
6 The Schrödinger problem and transport by diffusion 296
7 Transport distances as loss/fidelity 306
8 A few missing topics amongst many . . . 314
References 319

Foreword

The spectacular revival of mathematical optimal transportation can be traced back to the 1990s. A comprehensive presentation of the first modern developments and applications can be found in the monographs by Villani (2003, 2008) and – in a more probabilistic setting – Rachev and Rüschendorf (2006). They were mostly theory-oriented.

Offering a Riemannian metric on the space of measures defined on general manifolds, optimal transportation has spread far in mathematics and scientific computation. The construction of Wasserstein gradient flows, for instance, introduced by Jordan, Kinderlehrer and Otto (1998), led to a new understanding of

© The Author(s), 2021. Published by Cambridge University Press.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

nonlinear parabolic equations and is an active research field. Developed on the infinite-dimensional space of measures but also including discrete, possibly finite 'empirical' measures, optimal transportation theory offers an elegant bypass of the discretization process, a familiar but sometime burdensome routine for numerical analysts working on continuous models.

Use of the optimal transportation framework in the natural sciences appeared during the same period. The fundamental paper of Brenier (1991) was in fact preceded and inspired by an earlier paper, Brenier (1989), constructing weak solutions to Euler geodesics variational interpretation (Arnold 1966). Other applications began to appear in fields as diverse as medical imaging (Haker, Zhu, Tannenbaum and Angenent 2004), meteorology (Cullen and Purser 1984), astrophysics (Frisch, Matarrese, Mohayaee and Sobolevski 2002) and optics (Glimm and Oliker 2003, Wang 2004).

Similar developments occurred in parallel in economics, for which optimal transportation tools have been instrumental, particularly in the study of the famous 'principal–agent' problem; see Carlier (2001), Ekeland (2010), Figalli, Kim and McCann (2011) and Rochet and Chone (1998). The role of optimal transportation in economics is presented in depth in Galichon (2016); it will not be covered in this paper. Other important variants of optimal transportation will be omitted, for example transport occurring on discrete Markov chains (Maas 2011) and branched transport and optimal transportation networks; see Pegon (2017) and Bernot, Caselles and Morel (2008) for a review of these topics. It continues to pop up in fields as diverse as number theory (Steinerberger 2020), quantum mechanics (Golse and Paul 2021), relativity (Cavalletti and Mondino 2020), genomics (Lavenant 2021) . . .

The generic optimal transportation problem is a nonlinear optimization problem. Kantorovich relaxation (which earned Kantorovich the Nobel Prize for Economics) as a linear program was a giant leap forward for the theory but at the heavy computational price of replacing the space of transport maps with the space of couplings, hence squaring the number of unknowns. Combined with the cubic complexity of linear programming solvers, it was never used as such in the above-mentioned applications.

The tighter *convex* relaxation proposed by Benamou and Brenier (2000) offered the first convergent and tractable algorithm. It is difficult to assess its computational complexity precisely because it adds a time dimension, but one can safely estimate that it is somewhere between quadratic and cubic. The proposed *dynamic* vision of optimal transportation has also been inspirational, and has strongly influenced subsequent research in the field.

Significant progress was achieved in the early 2010s. Based on the convexity of the 'semi-dual' formulation of optimal transportation, a damped Newton method was proposed to solve the Monge–Ampère (Benamou, Froese and Oberman 2014) and semi-discrete (Mérigot 2011) settings, respectively. This approach resulted in the first linear complexity solvers. The 'damping' ensures that the optimization is kept to the domain of strict convexity of the optimal transportation cost. So far

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

this approach has essentially been limited to the most commonly used Euclidean quadratic cost (1.1), but a general analysis of the method is available in Kitagawa, Mérigot and Thibert (2019).

Another strategy, based on a convexification of Kantorovich relaxation, was initiated by Cuturi (2013) and in the economics context by Salanié and Galichon (2012) independently. It has been popular ever since for several different reasons. It does not solve the optimal transportation problem but yields its 'Schrödinger entropic approximation', a connection well known in probability and statistical physics (Léonard 2014). It is simple to implement and applies to virtually any optimal transportation generalization. There is a trade-off between the computational cost and the 'entropic bias', or the precision error, with classical optimal transportation. It is very competitive for estimating the transportation cost but not the transport plan itself. Finally it is easily parallelizable and, while naively quadratic, linear-cost off-the-shelf software can be found in Feydy (2019), for instance.

These numerical breakthroughs triggered an acceleration of the use of optimal transportation for numerical modelling and solutions in an increasing number of domains, including the recent and productive machine learning topic. Following all these developments is far beyond my mental abilities and energy. Here is a first disclaimer: this review paper will necessarily be incomplete, and biased towards my personal scientific environment. I apologize for the important contributions I have missed.

The first lecture notes oriented towards PDEs I came across – those of Evans (2001) – are still available online and useful! As optimal transportation is now taught at graduate level across the globe, a wealth of such lecture notes is available on the internet. Several books are also out: the book by Santambrogio (2015) is a modern self-contained review of the theory with a wide range of applications. As suggested by the title, *Optimal Transport for Applied Mathematicians*, this is definitely the reference book for applied mathematicians working in this field. In § 8.4.4, for instance, it provides an easy introduction to semi-geostrophic equations (Cullen 2006) and early universe reconstruction (Frisch *et al.* 2002), two important applications of optimal transportation. The paper by Peyré and Cuturi (2019) is a precious resource with precise descriptions of the algorithms and their analysis, in particular entropic regularization. It is nicely complemented by the book chapter by Mérigot and Thibert (2020), which presents the semi-discrete approach in more detail along with its link to entropic regularization. Finally, Carlier (2021) has provided a didactic presentation of the mathematics and the algorithmics of convex duality and convex optimization, which are ubiquitous in optimal transportation. As we will try to emphasize, convex Fenchel–Rockafellar duality is a cornerstone of optimal transportation. A good knowledge of these techniques, and the associated methods of solution, is compulsory for numerical analysts interested in optimal transportation as a research topic.

This paper has been designed and organized to present, in a unified and compact framework, the basic theory, the motivations and also the insight behind recent

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

optimal transportation developments, made possible by the algorithmic advances mentioned above. It synthesizes a large corpus; each section could be expanded into a longer technical survey paper. Except for the last section, the paper should be browsed or read linearly, at least the first time. Sections present – in a non-technical way – elements of the theory and algorithms together with flagship applications, and often use earlier expositions. The presentation will remain formal with some elements of proof. The chain of logic and the derivation of formulae will often be sketchy, and may be considered as exercises for the reader. I try to indicate precisely references with detailed proofs and rigorous mathematical formulations. In an attempt to keep the presentation self-contained, I have also gathered some required basic tools and results in footnotes. A list of notation is available after this introduction.

By definition, this paper should be useless for optimal transportation experts. It may, however, offer a broader vision of the topic and unseen connections between the many concepts used. For beginners, it cannot replace careful study of the monographs cited above but may instead be used as a study guide; each section roughly corresponds to one or two lectures.

A second disclaimer: my knowledge and understanding is of course the result of many years, even decades, of listening to lectures, discussing with colleagues and reading the literature. The scientific fruits will appear in some form or another in this paper.

Many thanks to Irène Wadspurger, Francis Collino, Guillaume Carlier and Gabriel Peyré for their pre-reading and correction of this manuscript.

I am deeply grateful to the Editorial Board of *Acta Numerica* for their invitation. Many thanks to Glennis Starling for copy-editing the document and improving my English.

This foreword would not be complete without acknowledging the fantastic group of colleagues and students at MOKAPLAN and our close collaborators. The list is too long to be reproduced here, but if they read these lines they will know they are part of the family. I thank them warmly for their friendship, patience, contributions, advice and vision.

Outline

Section 1 contains in a condensed form the classical presentation of the original Monge optimal transportation, the Kantorovich linear program and basic linear programming and linear assignment methods. This can be found in the general references above. Notation is introduced, and in particular the important idea that we can use the same notation for measures with continuous densities and discrete measures.

See Santambrogio (2015, §1), Carlier (2021, §6) and Peyré and Cuturi (2019, §2) for comprehensive presentations.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Section 2 introduces the Kantorovich dual problem and its role in the study of the well-posedness of optimal transportation. The key notion is the c-transform optimality of the Kantorovich potentials. This can also be used to tighten the linear Kantorovich problem into a concave 'semi-dual' problem. It leads to two modern numerical approaches: the Monge–Ampère and semi-discrete approaches. We present an application to the reflector cost.

As for Section 1, see Santambrogio (2015, §1), Carlier (2021, §6) and Peyré and Cuturi (2019, §2), and also Mérigot and Thibert (2020, §2).

Section 3 gives the basics of the distance on the space of probability measures. It also introduces entropic optimal transportation and the Sinkhorn algorithm and discusses the smoothing bias on the distance introduced by the approximation. This bias may be useful when computing interpolations or barycentres of discrete measures for the optimal transportation metric. These notions naturally lead to the introduction of time and dynamic optimal transportation in the next section.

Rigorous proofs (there are different techniques) are detailed in Santambrogio (2015, §5). See Peyré and Cuturi (2019, §4) for entropic optimal transportation.

Section 4 brings the notion of time or dynamics to optimal transportation. This led to the computational fluid dynamics formulation (also known as 'Benamou–Brenier'). This approach seems artificial when considering the classical source-to-target optimal transportation problem but was beneficial from the computational point of view. It also helps us to understand the richer models in Sections 5 and 6.1, and the Riemannian-like distance over the set of probability densities used, in particular, in the Wasserstein gradient flow theory (Section 7.2).

Classical first-order optimization 'proximal splitting' methods are applicable. See Carlier (2021, §7) and the review by Chambolle and Pock (2016).

Section 5 covers the optimal transportation treatment of Euler geodesics, its connection with dynamic optimal transportation and the use of optimal transportation solvers in this context.

I recommend Daneri and Figalli (2016) as a complete and precise review. See also Brenier (2020, §2).

Section 6 makes the connection between entropic optimal transportation and the Schrödinger problem in statistical physics. The entropic bias can also be interpreted as adding a diffusion to the transport model. The application of the Sinkhorn algorithm to variational mean field games is a good illustration. It also changes the nature of the transport from deterministic to stochastic and hyperbolic to parabolic. Quite remarkably, using the volatility as a control allows us to extend most of the classical optimal transportation results. A good example is martingale optimal transportation used in finance.

Léonard (2014) is the reference. It is a probabilistic presentation but the (long) introduction is accessible to all.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Section 7 introduces arguably the most active area of applications. We move one level up and use the optimal transportation distance as a metric (or 'loss' in the machine learning terminology) for variational problems set on the space of probability and non-negative Radon measures. It was a crucial motivation in the development of 'unbalanced' optimal transportation distances and the introduction of a simple but very efficient variation of entropic optimal transportation called *Sinkhorn divergence*. This is illustrated via gradient flow problems.

See Santambrogio (2015, §8) for Wasserstein gradient flows and Peyré and Cuturi (2019, §7-8) for losses and applications. The reference monograph on the theory of gradient flows is the book by Ambrosio, Gigli and Savare (2005).

Section 8 covers a few subjects and directions, which I have been interested in and are open or partly open.

Notation

$\mathcal{M}^+(X_0)$: non-negative Radon measures on the set X_0.
$\{x \in X \mapsto f(x) \in Y\}$: a function with domain X and valued in Y.
$D_x f(x)$: the gradient of f.
$\mathbf{1}_X$: the Lebesgue measure on X.
Id_{X_0} : identity map $x_0 \in X_0 \mapsto x_0$.
$f^\star$: the Legendre–Fenchel transform of the function f; see footnote 13.
$\iota_{\{v \le 0\}}$: the characteristic function := 0 if $v \le 0$ and $+\infty$ else.
$f \circ T$: the composition of function f with the map T.
$\dot{\mathcal{X}}_\tau$: the time derivative along the curve $\tau \mapsto \mathcal{X}_\tau$.
$T\#\mu_0 = \mu_1$: T 'pushes forward' $\mu_0 \in \mathcal{P}(X_0)$ to $\mu_1 \in \mathcal{P}(X_1)$; see footnote 2.
$X_0 \times X_1$: the tensor product of the two support sets of $\mu_{0,1}$.
$\mathcal{C}(X)$: continuous functions on X a compact metric space.
$\langle f, \alpha \rangle_\Omega$: duality product between $\alpha \in \mathcal{M}(\Omega)$ and $f \in \mathcal{C}(\Omega)$; see footnote 5.
$\mathrm{supp}\, \pi^*$: the support of π^*.
$\mathcal{P}(X)$: probability measures on X.
$c(x_0, x_1)$: displacement cost between x_0 and x_1 in X.
$\Pi(\mu_0, \mu_1)$: admissible transport plans; see (1.3).
$(\cdot)^*$: optimizer of considered variational problem.
f^c : the c-transform of the function f; see footnote 9.

1. Introduction

1.1. What is optimal transportation?

A general presentation on optimal transportation traditionally starts with a reference to Gaspard Monge's problem 'des déblais et des remblais' (1781): minimize the amount of work to shovel some mass from a source to a target 'configuration'

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

(we will define this last term more precisely). The energy is defined as the total distance the shovel, which contains exactly one unit of mass, has been carried over. Gaspard Monge was a military engineer and his motivation concerned fortifications: shovelling a ditch and using the soil to elevate a protective bank. In real life, the actual digging/building process is affected by accessibility/structural stability constraints: shovel transfers are done in sequence, and some sequences are infeasible (imagine you want to start placing the dirt on top of the earth bank with no foundations). One way to keep the Monge fortification analogy in line with modern mathematical formulations of optimal transportation is to use a 'magic' shovel. The magic shovel would be able to dig and carry all units of mass (one shovel) instantly, according to a predefined mapping, from the source to the target configuration.

We therefore look for a 'transport' map T, in the mathematical sense, between two source/target sets, denoted X_0 and X_1, *i.e.* $T\colon X_0 \to X_1$. The simplest situation, considered in most of this review, corresponds to $X_0 = X_1 = X$, a compact metric space. The source and target 'configurations' are real-valued non-negative mass measures on X_0 and X_1; we will first assume they are non-negative Radon measures,[1] denoted μ_0 and μ_1. In mathematical notation, $\mu_0 \in \mathcal{M}^+(X_0)$ and $\mu_1 \in \mathcal{M}^+(X_1)$. Because the shovel always carries the same amount of mass, the transport map is constrained to 'conserve the mass'. This translates into the following mathematical *measure-preserving* constraint.

For any μ_1-measurable set $A_1 \subset X_1$, $T^{-1}(A_1)$ is a μ_0-measurable subset of X_0 and $\mu_0(T^{-1}(A_1)) = \mu_1(A_1)$ (in the modern literature this is generally denoted $T\#\mu_0 = \mu_1$: we say T 'pushes forward' μ_0 to μ_1).[2] As a direct consequence, the total mass also needs to be preserved: $\mu_0(X_0) = \mu_1(X_1)$. Setting the total mass to 1, we may restrict the space of μ_0 and μ_1 to probability measures. We will use the notation $\mu_0 \in \mathcal{P}(X_0)$ and $\mu_1 \in \mathcal{P}(X_1)$. In fact, $\mu_{0,1}$[3] will often be discrete empirical measures (sums of Dirac masses) but might also be absolutely continuous measures with respect to Lebesgue measure. In the continuous case and *by abuse of notation* the probability densities will still be denoted $\mu_{0,1}$.

The Monge shovel 'carries a unit of mass' instead of volume because the relevant measure of work is mass times distance. In the simplest Monge configuration, densities[4] are constant and can be scaled to 1, *i.e.* mass and volume are equivalent. In this case the mass conservation then implies that X_0 and X_1 have the same

[1] Continuous linear forms over the set of continuous functions on the compact space X.

[2] It can also be expressed as $\int_{X_0} (f \circ T)\, \mu_0 \,\mathrm{d}x_0 = \int_{X_1} f\, \mu_1 \,\mathrm{d}x_1$ for any continuous test function $f \in \mathcal{C}(X)$. If T is smooth and one-to-one, the change of variable $x_1 = T(x_0)$ gives

$$\int_{X_0} (f \circ T)\, \mathrm{d}\mu_0 \,\mathrm{d}x_0 = \int_{X_0} (f \circ T)\, \det(DT)\, (\mu_1 \circ T)\, \mathrm{d}x_0,$$

a variational formulation for the Jacobian equation $\det(DT)\,(\mu_1 \circ T) = \mu_0$.

[3] The $(\cdot)_{0,1}$ will systematically denote the ordered pair $((\cdot)_0, (\cdot)_1)$.

[4] That is, mass per volume or how much mass is located at each element of the sets.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

cardinality (a finite number of elements) or volume (infinite number of elements) for the set of admissible maps T to be non-empty. If we want to tackle non-constant densities, the shovel has to pile several shovel/mass units on the same target location in order to reach the prescribed mass density and dig several units of mass from the same place, then send them to different locations. The second requirement rules out the map representation as each element in X_0 may only possess one image $T(x_0) \in X_1$.

1.2. Kantorovich relaxation

The mathematical framework to deal with this limitation was invented by Leonid Kantorovich in 1954. Instead of a map he proposed looking for *couplings*. These represent the amount of mass transported between all pairs $(x_0, x_1) \in X_0 \times X_1$. These couplings are called *Kantorovich transport plans*. They are probability measures, here denoted π, on the product space $X_0 \times X_1$: $\pi \in \mathcal{P}(X_0 \times X_1)$. For all measurable subsets $(A_0, A_1) \subset X_0 \times X_1$, $\pi(A_0, A_1)$ is the amount of mass transported from A_0 to A_1. The mass conservation is now automatically enforced: you receive exactly what you send. The original formulation of the problem requires the transported mass to reach the exact target configuration. Therefore you cannot send more or less than what is available, and likewise for the reception. This translates into marginal constraints on the plans:[5] $\langle \mathbf{1}_{X_1}, \pi \rangle_{X_1} = \mu_0$ and $\langle \mathbf{1}_{X_0}, \pi \rangle_{X_0} = \mu_1$ (see the footnote for the definition of the brackets). This is often written using the canonical projection $P_{X_0} : (x_0, x_1) \to x_0$ (not to be mistaken for $\mathcal{P}(X)$, the probability space over X) and the push-forward notation $P_{X_0}\#\pi = \mu_0$ and $P_{X_1}\#\pi = \mu_1$. The last ingredient we need to describe our first formal optimal transportation problem is the *displacement cost* $c : (x_0, x_1) \in X \times X \to c(x_0, x_1)$, *i.e.* the cost of moving the shovel from x_0 to x_1. The mathematical and modelling properties of c are of course of tantamount importance and we will return later to this point. In line with Monge modelling, this is often linked to the metric associated with X. The standard case is the square of the Euclidean distance and we will mostly restrict to this case in our presentation:

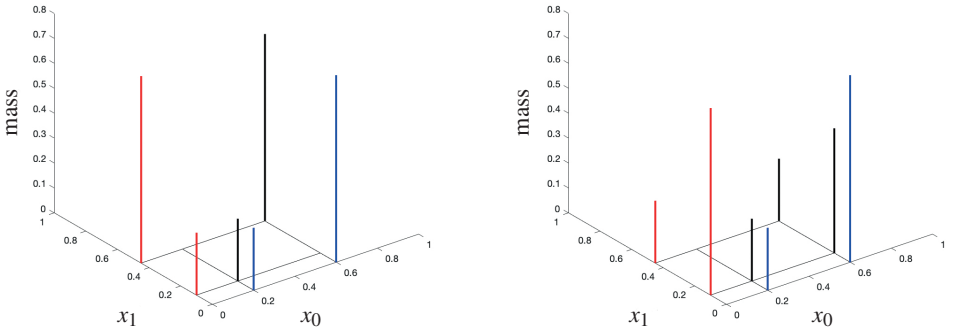
$$X_{0,1} = \text{compact subsets of } \mathbb{R}^d, \quad c(x_0, x_1) = \frac{1}{2}\|x_1 - x_0\|^2. \tag{1.1}$$

Kantorovich's problem will be the first formal optimal transportation problem discussed in this paper. Optimizers will always carry a $(\cdot)^*$ throughout the paper, that is,

$$\pi^* \in \underset{\pi \in \Pi(\mu_0, \mu_1)}{\arg\inf} \langle c, \pi \rangle_{X_0 \times X_1}, \tag{1.2}$$

[5] The notation $\langle f, \alpha \rangle_\Omega$ will stand for the duality product $\int_\Omega f \, \mathrm{d}\alpha$ between continuous functions $f \in \mathcal{C}(\Omega)$ and probability measures $\alpha \in \mathcal{P}(\Omega)$; the Ω subscript will often be omitted if we just consider a unit constant function.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) $\alpha_0 = \frac{1}{4},\ \alpha_1 = \frac{1}{4}$ (b) $\alpha_0 = \frac{1}{4},\ \alpha_1 = \frac{3}{4}$

Figure 1.1. Blue, $\mu_0 = \alpha_0\,\delta_{a_0} + (1-\alpha_0)\,\delta_{b_0}$; red, $\mu_1 = \alpha_1\,\delta_{a_1} + (1-\alpha_1)\,\delta_{b_1}$; black, π^*. The optimal plan cannot be reduced to a map when $\alpha_0 \neq (\alpha_1, 1-\alpha_1)$.

where $\Pi(\mu_0, \mu_1)$ is the following set of admissible transport plans:

$$\Pi(\mu_0, \mu_1) := \{\pi \in \mathcal{P}(X_0 \times X_1)\colon P_{X_1}\#\pi = \mu_0 \text{ and } P_{X_0}\#\pi = \mu_1\}. \tag{1.3}$$

We stress that (1.2)–(1.3) is a linear optimization problem with linear constraints in π. Minimizers exist but may not be unique simply on the condition that $c \in \mathcal{C}(X_0 \times X_1)$ and $\Pi(\mu_0, \mu_1)$ is compact for weak-$\star$ topology of measures.[6]

Let us give a trivial example. Choose $\mu_0 = \delta_{a_0}$ and $\mu_1 = \delta_{a_1}$, two Dirac masses at a_0 and a_1. We can set the unit of mass to 1 and Monge transport is done in just one shovel displacement between a_0 and a_1. The transport cost is the displacement cost, $c(a_0, a_1)$. Generalizing to $\mu_0 = (1-\alpha_0)\,\delta_{a_0} + \alpha_0\,\delta_{b_0}$ ($\alpha \in]0,1[$) and the same μ_1, the transport map is again trivial. There will be $(1-\alpha_0)$ shovel displacement between a_0 and a_1 and α_0 between b_0 and a_1. The cost $(1-\alpha_0)\,c(a_0, a_1) + \alpha_0\,c(b_0, a_1)$ is bilinear in the weight α_0 and the distance between points. Let us complicate this slightly to $\mu_1 = (1-\alpha_1)\,\delta_{a_1} + \alpha_1\,\delta_{b_1}$. If $\alpha_0 \notin \{\alpha_1, 1-\alpha_1\}$, there is no measure-preserving Monge map. We can, however, find Kantorovich transport plans. They are 2×2 matrices assigning mass between couples in $(a_0, b_0) \times (a_1, b_1)$. The Kantorovich formulation is more general than the Monge formulation and subsumes Monge. See also Figure 1.1.

1.3. Linear programming

The first and most obvious numerical method to attack the Kantorovich problem numerically is the linear programming approach. It assumes that the data are given in discrete form, for $k = 0, 1$, $\mu_k = \sum_{i=1}^{N_k} w_{k,i}\delta_{x_{k,i}}$. The measures μ_k may be empirical measures set on N_k points in X: $X_k = \{x_{k,i}\}_{i=1,N_k}$ or a discretization of

[6] A sequence of measures μ_n converges to μ for the weak-$\star$ topology if $\lim_{n\to+\infty}\langle f, \mu_n\rangle_X = \langle f, \mu\rangle_X$ for all $f \in \mathcal{C}(X)$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

probability densities on this discrete set representing a Cartesian grid, for instance. The weights $\{w_{k,i}\}_{i=1,N_k}$ are strictly positive and sum to 1.

Plugging the discrete data and support into (1.2)–(1.3), brackets become sums and we obtain the finite-dimensional *linear program*

$$\pi^* \in \underset{\pi \in \Pi(\mu_0,\mu_1)}{\arg\inf} \sum_{(i,j)\in N_0\times N_1} c(x_{0,i}, x_{1,j})\, \pi_{i,j}. \tag{1.4}$$

Note that $\Pi(\mu_0, \mu_1)$ can be rewritten as a set of $N_0 \times N_1$ matrices with constraints on sums of rows and

$$\Pi(\mu_0, \mu_1) := \left\{ (\pi_{i,j}) \in \mathcal{M}_{N_0\times N_1}(\mathbb{R}^+) : \sum_{j\in[1,N_1]} \pi_{i,j} = w_{0,i} \text{ and } \sum_{i\in[1,N_0]} \pi_{i,j} = w_{1,j} \right\}. \tag{1.5}$$

Problem (1.4-1.5) is a discrete linear programming problem in $N_0\times N_1$ unknowns and $N_0 + N_1$ constraints. Numerical resolution with linear solvers which have cubic complexity in the number of unknowns is therefore out of reach for reasonable discretizations (typically $N_k < 100$ or even less depending on machine memory).

As discussed in the example at the end of the previous subsection, Monge solutions are not guaranteed to exist: they will appear when $N_0 = N_1 = N$ and a permutation matrix $\pi_{i,j=\sigma(i)}$ (σ is a permutation of $[\![1, N]\!]$) is feasible. There is only one non-zero element per line and column and the permutation can be interpreted as a map. This implicitly constrains the weights to be able to match exactly $\{w_{0,i} = w_{1,\sigma(i)}\}_{i=1,N}$.

When the weights are all equal (to $1/N$), the Birkhoff–von Neumann theorem (see Carlier 2021, §6.4, or Peyré and Cuturi 2019, §2.3) states that the permutation matrices are the extreme points of $\Pi(\mu_0, \mu_1)$. Problem (1.4)–(1.5) is then known as the *linear assignment* problem.

Assuming all weights $\{w_{k,i}\}_{i=1,N}$ are integer multiples of the elemental mass $E := \min_k\{1/N_k\}$, it is always possible to rewrite (1.4) as a linear assignment problem by piling Dirac masses at the same point, $w_{0,i}\delta_{x_{0,i}} = E \sum_{l=1}^{w_{0,i}/E} \delta_{x_{0,i,l}}$ (and performing the same decomposition on μ_1). The new Dirac masses are at the same location but counted $w_{0,i}$ times. The weights are units but the size of the problem and the cost matrix may be larger. Optimal permutations σ^* then characterize the support of π^*, $\operatorname{supp}\pi^* = \{(x_{0,i}, x_{1,\sigma^*(i)})\}_{i=1,N'}$, where $N' \geq N$ is the new bigger number of unit Dirac masses.

The optimality over permutations translates into an important property called cyclical monotonicity (Carlier 2021, §6.4): for every subset $\{x_{0,i}, x_{1,i}\}_{i=1,I} \in \operatorname{supp}\pi^*$, we have

$$\sum_{i\in[1,I]} c(x_{0,i}, x_{1,i}) \leq \sum_{i\in[1,I]} c(x_{0,i+1}, x_{1,i}), \tag{1.6}$$

where by convention $x_{0,I+1} = x_{0,1}$. This property is in fact also a sufficient condition for the optimality.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

If we apply (1.6) to a simple pair (i, i') and use the quadratic displacement cost (1.1), we find (after simple computations)[7]

$$(x_{0,i} - x_{0,i'}) \cdot (x_{1,\sigma^*(i)} - x_{1,\sigma^*(i')}) \geq 0. \tag{1.7}$$

This is a simple monotonicity condition on the transport with the important consequence that the affine trajectories of the planned optimal transport paths $x_{0,i} \mapsto x_{1,\sigma^*(i)}$ and $x_{0,i'} \mapsto x_{1,\sigma^*(i')}$ cannot intersect.

It allows us to reconcile Monge's ordinary (not magic) shovel with the mathematical formulation. The optimality ensures that the map corresponds to a structurally feasible sequence of shovel transport: there will be no obstruction.

2. Kantorovich duality and the reflector problem

2.1. Kantorovich duality and semi-duality

Kantorovich duality is the application of linear programming duality theory to (1.2). For simplicity we assume X_0 and X_1 to be compact; for a general formulation, see Santambrogio (2015, Theorem 1.39). The dual Kantorovich problem is[8]

$$(u_0^*, u_1^*) \in \underset{(u_0,u_1)\in C_D}{\arg\sup} \langle u_0, \mu_0\rangle_{X_0} + \langle u_1, \mu_1\rangle_{X_1}, \tag{2.1}$$

where $u_0 \oplus u_1(x_0, x_1) = u_0(x_0) + u_1(x_1)$ and

$$C_D := \{(u_0, u_1) \in \mathcal{C}(X_0) \times \mathcal{C}(X_1) \colon u_0 \oplus u_1 \leq c\}. \tag{2.2}$$

The derivation of (2.1) is classical and can be found for instance in Santambrogio (2015, §1.2) or Carlier (2021, §6.4.1). The dual variables (u_0, u_1) are the Lagrange multipliers of the marginal constraints in (1.3). There is a nice economics interpretation in terms of optimal prices: μ_0 and μ_1 now represent two different types of goods, their mass and storage location. For some reason it is decided that the goods must be swapped, either via ground transport or by selling locally produced goods instead of moving the already available but misplaced ones. Before deciding on the method, the government needs to evaluate the feasibility of producing the goods locally. This depends on the maximal revenue a (single) producer may expect by selling the goods at price u_0 and u_1 respectively; this is given by (2.1). Of course selling will only happen if (2.2) is satisfied, that is, the sale price of a unit of mass on each side $u_0 \oplus u_1$ is less than the fixed price c of swapping a unit of mass through transport. Assuming maximizers (u_0^*, u_1^*) exist, the maximal revenue is the transport cost,

$$\langle c, \pi^*\rangle_{X_0\times X_1} = \langle u_0^*, \mu_0\rangle_{X_0} + \langle u_1^*, \mu_1\rangle_{X_1}, \tag{2.3}$$

[7] $(\cdot)\cdot(\cdot)$ denotes the standard scalar product in $\mathbb{R}^d$.

[8] $\oplus$ is the direct sum.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

and the optimal prices saturate the constraint when $\pi^* > 0$:

$$\begin{aligned} &u_0^*(x_0) + u_1^*(x_1) = c(x_0, x_1) \quad \text{for all } (x_0, x_1) \in \operatorname{supp} \pi^*, \\ &u_0^*(x_0) + u_1^*(x_1) < c(x_0, x_1) \quad \text{else.} \end{aligned} \tag{2.4}$$

In the literature (u_0^*, u_1^*) are known as the *Kantorovich potentials*. They are always defined up to a constant as $(u_0^* + C, u_1^* - C)$ is also a solution for any $C \in \mathbb{R}$. The optimality condition (2.4) imposes a strong structure on the maximizers, in that they are *c-transforms*[9] of one another:

$$u_0^* = (u_1^*)^c \quad \text{and} \quad u_1^* = (u_0^*)^c. \tag{2.5}$$

As (2.5) is a necessary condition for optimality, the *semi-dual* formulation idea is to leverage $u_1 = (u_0)^c$ to optimize only on u_0. This is a key part of the theory as the set of potentials $\{(u_0, u_0^c)\}$ is compact while C_D is not (see Santambrogio 2015, §1.6). The other condition $u_0 = (u_1)^c$ requires u_0 to be c-concave, hence $u_0 = u_0^{cc} = ((u_0)^c)^c = (u_1)^c$. Finally the domain of $(u_0)^c$ must match X_1, the support of μ_1, where the mass is supported.

The semi-dual optimal transportation problem reads

$$\sup_{u_0 \in C_{SD}} SD(u_0) := \langle u_0, \mu_0 \rangle_{X_0} + \langle u_0^c, \mu_1 \rangle_{X_1}, \tag{2.6}$$

where

$$C_{SD} := \{u_0 \text{ is } c\text{-concave and } \operatorname{Dom}(u_0^c) = X_1\}. \tag{2.7}$$

We will comment on the set of constraints in the simpler quadratic cost formalism in Section 2.3. Expanding the c-transform in (2.6),[10] we get

$$SD(u_0) = \langle u_0, \mu_0 \rangle_{X_0} + \left\langle \left\{ x_1 \mapsto \inf_{x_0 \in X_0} c(x_0, x_1) + u_0(x_0) \right\}, \mu_1 \right\rangle_{X_1}. \tag{2.8}$$

The concavity of the cost function $u_0 \to SD(u_0)$ follows directly from the concavity of $u_0(\cdot) \mapsto u_0^c(\cdot) := \inf_{x_0} \{c(x_0, \cdot) + u_0(x_0)\}$ and the linearity of (2.6).

We move from a linear to a concave program at the price of constraint (2.7). The computation of the critical point of $u_0 \mapsto SD(u_0)$ is the basis of the Monge–Ampère and semi-discrete formulations detailed below.

[9] A quick summary: f^c, the c-transform of a function $f : X_0 \to \mathbb{R} \cup +\infty$ is defined as $f^c(x_1) := \inf_{x_0 \in X_0} c(x_0, x_1) - f(x_0)$ (note that the definition depends on the domain of f). The domain of f^c is $\partial_c f(X_0) = \cup_{x_0 \in X_0} \{\partial_c f(x_0)\}$, where $\partial_c f$ is the c-subdifferential of f, *i.e.* the multivalued map

$$\partial_c f(x_0) = \{x_1 \in X : c(x_0, x_1) - f(x_0) \le c(x_0', x_1) - f(x_0') \text{ for all } x_0' \in X_0\},$$

that is, where the c-transform is defined; this set may be empty. A function g is c-concave if there exists f such that $g = f^c$ and then the domain is given by $\operatorname{Dom}(g^c) = \partial_c f^{cc}(X_0)$. If f is also c-concave then $f^{cc} = f$.

[10] $\{x \in X \mapsto f(x) \in Y\}$, a function with domain X and valued in Y.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

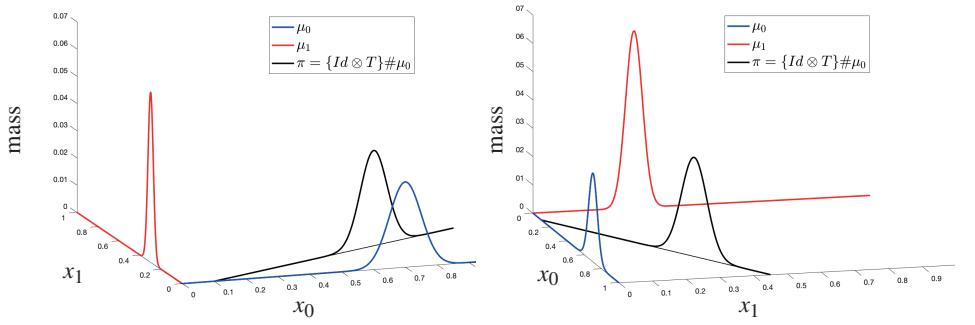


Figure 2.1. The simplest non-trivial optimal transport maps are affine and correspond to data in the form of a translation and a dilation $a\mu_1(a\,x_0 + b) = \mu_0(x_0)$ (here $a = 0.5$ and $b = -0.05$); π is null outside the graph $\{(x_0, x_1 = T(x_0) := a\,x_0 + b)),\ x_0 \in X_0\}$.

2.2. *Existence of Monge map uniqueness of Kantorovich plans*

The optimality condition (2.4) is also central to establishing the existence and uniqueness of Monge maps. This is one of the fundamental contributions of Brenier (1991) in the Euclidean quadratic cost case (1.1) and it was generalized by Gangbo and McCann (1996). The idea is straightforward (once you have it!) and relies on the *twist condition* on the cost:

$$\text{for all } x_0 \in X_0, \text{ the map } x_1 \in X_1 \mapsto D_{x_0} c(x_0, x_1) \in \mathcal{T}_{x_0} X_0 \tag{2.9}$$

is one-to-one.[11] Using (2.9), (2.4) can be solved as an implicit equation in x_1; the solution for all $x_0 \in X_0$ defines the unique optimal Monge map

$$T\colon x_0 \to x_1^* := \{\{x_1 \mapsto D_{x_0} c(x_0, x_1)\}\}^{-1}(D_{x_0} u_0^*(x_0)). \tag{2.10}$$

Note that the gradient[12] gets rid of the additive constant underdetermination. Precise theorems can be found in the literature (*e.g.* Theorem 2.9 of (Dafni, McCann and Stancu 2013, §6)). We give a few important consequences of this result. First, condition (2.9) implicitly assumes that c is continuous and at least differentiable with respect to $x \in X$. Defining this differential for all $x_0 \in X_0$ rules out discrete measures: μ_0 has to be absolutely continuous with respect to $\mathbf{1}_X$, the Lebesgue measure on X (remember the Dirac counterexample in Section 1.2). The twist condition (2.9) is obviously satisfied by the Euclidean quadratic cost (1.1), but also by the more complex reflector cost presented at the end of this section; it only needs to hold between X_0 and X_1.

The second important consequence of (2.4) is that the unique transport map is sufficient to characterize the unique Kantorovich solution. For all Borel sets

[11] $D_{x_0} c$ is the gradient with respect to the first variable x_0, and $\mathcal{T}_{x_0} X_0$ is a tangent vector to X_0 at x_0.
[12] $D_x f(x)$ is the gradient of f at x.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

$A_{0,1} \subset X_{0,1}$ and because π^*, while satisfying (1.3), vanishes outside the graph $\{x_0, x_1 = T(x_0)\}$,

$$\pi^*(A_0, A_1) = \pi^*(A_0 \cap T^{-1}(A_1), X_1) = \mu_0(A_0 \cap T^{-1}(A_1)). \tag{2.11}$$

In compact form, (2.11) is usually written $\pi^* = (\mathrm{Id}_{X_0}, T)\#\mu_0$, where Id_{X_0} is the identity map on X_0 and $(\mathrm{Id}_{X_0}, T) := \{x_0 \mapsto (x_0, T(x_0))\}$ is a map from X_0 to $X_0 \times X_1$. See Figure 2.1 for an example.

2.3. *Monge–Ampère solutions*

For the quadratic cost (1.1), c-concavity can be expressed, after a transformation of the potentials, using the classical Legendre–Fenchel transform[13] (denoted $(\cdot)^\star$ and not to be mistaken for the optimizer notation $(\cdot)^*$):

$$\begin{aligned}\left\{x_0 \mapsto \frac{1}{2}\|x_0 - .\|^2 - \phi(x_0)\right\}^\star(x_1) &:= \sup_{x_0 \in X} x_1 \cdot x_1 - \left(\frac{1}{2}\|x_0\|^2 - \phi(x_0)\right) \\ &= \frac{1}{2}\|x_1\|^2 - \phi^c(x_1).\end{aligned} \tag{2.12}$$

The functions

$$\phi_i = \left\{x_i \mapsto \frac{1}{2}\|x_i\|^2 - u_i(x)\right\}, \quad i = 0, 1, \tag{2.13}$$

correspond to the seminal setting of Brenier (1991). To optimize (2.6) over the $\{\phi_{0,1}\}$ instead of the $\{u_{0,1}\}$, change the maximization into a minimization[14] and add the constant second moments of μ_0 and μ_1 (omitted below):

$$\phi_0^* := \inf_{\phi_0 \in C_{MA}} SD(\phi_0), \quad SD(\phi_0) := \langle \phi_0, \mu_0 \rangle_{X_0} + \langle \phi_0^\star, \mu_1 \rangle_{X_1}. \tag{2.14}$$

The connection between convexity and the Legendre–Fenchel transform are well known (see footnote 13). In this setting (2.7) becomes

$$C_{MA} = \{\phi_0 \text{ convex and } \partial\phi_0(X_0) = X_1\}, \tag{2.15}$$

where

$$\partial\phi_0(x_0) := \{x_1,\ x_0 \cdot x_1 - \phi_0(x_0) \geq x_0' \cdot x_1 - \phi_0(x_0') \text{ for all } x_0' \in X_0\} \tag{2.16}$$

is the usual subgradient of ϕ_0 at x_0 (and $\partial\phi_0(X_0) = \{\partial\phi_0(x_0)\}_{x_0 \in X_0}$). As ϕ_0^* is convex, subgradients are gradients almost everywhere. This defines the single-valued Monge map (2.10)

$$x_0 \mapsto x_1 = T(x_0) = D\phi_0^*(x_0) = \mathrm{Id}_{X_0} - Du_0^*(x_0). \tag{2.17}$$

[13] The Legendre–Fenchel transform $f^\star$ of f for a function $f : X_0 \to \mathbb{R} \cup +\infty$ is defined by $f^\star(x_1) := \sup_{x_0 \in X_0} (x_0 \cdot x_1) - f(x_0)$ (note that the definition depends on the domain of f); $f^\star$ is always convex. If f is convex , $(f^\star)^\star = f$, f and $f^\star$ are a.e. differentiable and $Df \circ Df^\star = Df^\star \circ Df = \mathrm{Id}$.

[14] Use $\inf\{\cdot\} = -\sup -\{\cdot\}$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Likewise $(\phi_0^*)^\star$ is convex and its domain is constrained to be X_1. For almost every $x_1 \in X_1$ there is a unique $x_0 = D(\phi_0^*)^\star(x_1)$ achieving the maximum in the Legendre–Fenchel transform:

$$(\phi_0^*)^\star(x_1) = x_1 \cdot D(\phi_0^*)^\star(x_1) - \phi_0^* \circ D(\phi_0^*)^\star(x_1). \tag{2.18}$$

Plugging (2.18) into (2.14), a formal application of Danskin's theorem[15] to the function $\Phi(x_0, x_1) = x_0 \cdot x_1 - \phi_0(x_0)$ yields, for any functional variation $\delta\phi_0$ such that $\phi_0^* + \delta\phi_0$ remains in (2.15),

$$SD(\phi_0^* + \delta\phi_0) - SD(\phi_0^*) = \langle \delta\phi_0, \mu_0\rangle_{X_0} - \langle \delta\phi_0 \circ D(\phi_0^*)^\star, \mu_1\rangle_{X_1} + O(\|\delta\phi_0\|^2). \tag{2.19}$$

Because of the change of variables (2.13), the maximization is now a minimization and SD is convex instead of concave. We conclude that the optimal potential satisfies for all $\delta\phi_0$ the equation

$$\langle \delta\phi_0, \mu_0\rangle_{X_0} - \langle \delta\phi_0 \circ D(\phi_0^*)^\star, \mu_1\rangle_{X_1} = 0. \tag{2.20}$$

The change of variable $x_1 = D\phi_0^*(x_0)$ gives (again formally)

$$\langle \delta\phi_0, \mu_0\rangle_{X_0} - \langle \delta\phi_0, (\mu_1 \circ D\phi_0^*)\ \det(D^2\phi_0^*)\rangle_{X_0} = 0. \tag{2.21}$$

Therefore the optimal transportation potential, together with the constraint (2.15), may be understood in a weak sense as a solution of the (so-called) *second boundary value problem for the Monge–Ampère equation*:

$$\det(D^2\phi_0) = \frac{\mu_0}{\mu_1 \circ D\phi_0} \quad \text{on } X_0. \tag{2.22}$$

The link with the Monge–Ampère operator has triggered an important theoretical activity on the regularity theory of such equations; see Caffarelli (1992) and Figalli (2017). A short summary is as follows. Assuming X_1 is convex (this is imposed by $\partial\phi_0(X_0) = X_1$) and $\mu_{0,1}$ bounded above and below away from 0, we gain two orders of regularity in Hölder spaces for ϕ_0 with respect to the data $\mu_{0,1}$, as is classically expected of second-order elliptic equations. The bounds on the densities force the right-hand side of (2.22) to be strictly positive and ϕ_0 to be strictly convex. This is necessary for the variational formulation (2.21) to allow $\phi_0^* + \delta\phi_0$ to remain in C for all 'small' $\delta\phi_0$(s).

The PDE (2.22)–(2.15) may be discretized and the resulting nonlinear system solved with a damped Newton's method yielding, experimentally at least, a linear

[15] We give a simplified formulation. Let $\Phi\colon (x_0, x_1) \in X_0 \times X_1 \to \mathbb{R}$ be continuous and convex in x_1 for all x_0 and X_0 be compact. Set $\Psi(x_1) = \sup_{x_0} \Phi(x_0, x_1)$. Then, if $\{x_0^* := \arg\sup_{x_0} \Phi(x_0, x_1)\}$ is a singleton, Ψ is differentiable at x_1 and

$$D_{x_1}\Psi(x_1) = \partial_{x_1}\Phi(x_0^*, x_1).$$

The assumptions on Φ may be considerably relaxed: see Carlier (2021, §5.3).

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

cost (see Benamou *et al.* 2014, Benamou and Duval 2019, and the references therein[16]). The discretization is delicate: it has to preserve a notion of strict convexity to keep a well-defined Hessian for the discretization of the convex $\phi_0 \mapsto SD(\phi_0)$, *i.e.* the Jacobian of (2.22). The Newton's method damping ensures the right-hand side of (2.22) remains strictly positive; more on this can be found in Benamou, Collino and Mirebeau (2016*d*).

Finally let us mention that the Monge–Ampère formulation has been generalized to other costs satisfying (2.9); see Section 2.5 below for an example.

2.4. *Semi-discrete solutions*

When $\mu_1 \in \mathcal{L}^1(X_1)$ is simply integrable and $\mu_0 = \sum_{n\in[1,N]} \alpha_n \delta_{x_{0,n}}$ is a finite sum of weighted Dirac masses, (2.6) becomes an unconstrained finite-dimensional concave optimization problem called *semi-discrete optimal transportation* (not to be confused with *semi-dual*).

We discuss the two components in (2.6) separately. First we have

$$\langle u_0, \mu_0\rangle_{X_0} = \sum_{n\in[1,N]} \alpha_n\, u_0(x_{0,n}),$$

so we only need to set the N prices $\{u_{0,n}\} := \{u_0(x_{0,n})\}$ where the mass is located. Now, by construction and for all $x_1 \in X_1$,

$$u_0^c(x_1) = \inf_{n\in[1,N]} \{c(x_{0,n}, x_1) - u_{0,n}\} \tag{2.23}$$

(remember that $X_0 = \mathrm{supp}(\mu_0) = \{x_{0,n}\}$). We could plug this expression directly into (2.6), but it is helpful to introduce the sets of *Laguerre cells*. They depend on the vector $u_0 = \{u_{0,n}\} \in \mathbb{R}^N$, for all $m \in [1, N]$:

$$\mathrm{Lag}_m(u_0) := \{x_1 \in X_1 : u_{0,m} - c(x_{0,m}, x_1) \le u_{0,n} - c(x_{0,n}, x_1) \text{ for all } n \in [1, N]\}. \tag{2.24}$$

In our economics illustration, Lag_m is the subset of X_1 where the price is set to $u_0^c(x_1) = c(x_{0,m}, x_1) - u_{0,m}$ (the infimum is reached for m in (2.23)). We now plug this into (2.6) by using the decomposition $X_1 = \cup_{n\in[1,N]} \mathrm{Lag}_n(u_0)$:

$$\begin{aligned}
&\langle u_0, \mu_0\rangle_{X_0} + \langle u_0^c, \mu_1\rangle_{X_1} \\
&\quad = \sum_{n\in[1,N]} \alpha_n\, u_{0,n} + \langle u_0^c, \mu_1\rangle_{\cup_{n\in[1,N]}\mathrm{Lag}_n} \\
&\quad = \sum_{n\in[1,N]} \alpha_n\, u_{0,n} + \sum_{n\in[1,N]} \langle c(x_{0,n}, \cdot) - u_{0_n}, \mu_1\rangle_{\mathrm{Lag}_n} \\
&\quad = \sum_{n\in[1,N]} (\alpha_n - \mu_1(\mathrm{Lag}_n))\, u_{0,n} + \sum_{n\in[1,N]} \langle c(x_{0,n}, \cdot), \mu_1\rangle_{\mathrm{Lag}_n}.
\end{aligned} \tag{2.25}$$

[16] A two-dimensional code is available at https://gforge.inria.fr/scm/browser.php?group_id=9995.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The last term is constant and can be removed from the maximization. The second term expresses the mass balance of what is swapped between Lag_n and $x_{1,n}$ and the actual mass to be transported α_n. Optimality of this concave program is therefore obtained by solving the set of N nonlinear equations

$$\alpha_n = \mu_1(\mathrm{Lag}_n), \quad n \in [1, N]. \tag{2.26}$$

The constraints (2.7) are automatically taken into account by this formulation. First, the X_1 support constraint is implicitly enforced by the computation of the Laguerre cells (2.24). The c-concavity condition is more subtle and can be transformed into a local constraint. If

$$\mathrm{Lag}_n(u_0) \neq \emptyset, \quad n \in [1, N], \tag{2.27}$$

then $u_0^{cc}(x_{0,n}) = u_{0,n}$ for all $n \in [1, N]$ and we can use the function u_0^{cc}, which is by construction c-concave.

The formulation, analysis and solution method is explained in great detail in Mérigot and Thibert (2020). A more computationally oriented survey is that of Lévy and Schwindt (2018). In the case of the Euclidean quadratic cost (1.1), Laguerre cells can be computed[17] in $O(N \log(N))$ operations. The integration of the measure μ_0 on each cell is the only numerical approximation in the method. Finally, Mérigot (2011) has shown that (2.26) was amenable to a damped Newton's method forcing the volume condition (2.27) and yielding a linear cost optimal transportation solver for the Euclidean quadratic cost.

2.5. *The far field reflector cost*

Analysts and differential geometers have gone a long way to extend optimal transportation to Riemannian manifolds with a significant impact on pure mathematics. The place to delve further is Villani (2008). It is also important for numerical analysts. We illustrate this generalization with an optimal transportation application set on the $(d-1)$-dimensional unit sphere $\mathbb{S}^{d-1}$. Here $X_0 \subset (\mathbb{S}^{d-1})^+$ and $X_1 \subset (\mathbb{S}^{d-1})^-$ are connected domains in the northern and southern hemispheres respectively; $\mu_0 \in X_0$ represents a given illuminance from a point source (the $\mathbb{R}^d$ origin); rays carry $\mu_0(x_0)$ light intensity in the direction x_0. The other prescribed marginal is the illumination: $\mu_1(x_1)$ is the light intensity in the direction x_1. The reflector 'freeforming' problem is to find a surface in $\mathbb{R}^d$ such that the specular law of reflection maps μ_0 to μ_1.

This problem has an elegant optimal transportation formulation introduced independently by Glimm and Oliker (2003) and Wang (2004), as follows.

[17] Available software: https://www.cgal.org/, http://alice.loria.fr/index.php/software/4-library/75-geogram.html, https://github.com/mrgt/MongeAmpere.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Let us apply the Monge theory to the displacement cost:[18]

$$c(x_0, x_1) = -\log(1 - x_0 \cdot x_1), \quad (x_0, x_1) \in (X_0 \times X_1). \tag{2.28}$$

The TWIST condition (2.9) is satisfied; note that the restriction to $X_{0,1}$ is essential to preserve the regularity of the cost. A transport map exists, given by (2.10). Taking the exponential of the equation/inequation over the support of the optimal transportation plan in (2.4), we get

$$\frac{e^{-u_1^*(T(x_0))}}{1 - x_0 \cdot T(x_0)} = e^{u_0^*(x_0)} \leq \frac{e^{-u_1^*(x_1))}}{1 - x_0 \cdot x_1} \quad \text{for all } x_1 \in X_1. \tag{2.29}$$

We now notice that

$$x_0 \in X_0 \subset (\mathbb{S}^{d-1})^+ \to \mathrm{Par}_{x_1}(x_0) := x_0 \frac{e^{-u_1^*(x_1))}}{1 - x_0 \cdot x_1} \tag{2.30}$$

is a family of parabolic reflectors in $\mathbb{R}^d$ with axis $x_1 \in X_1$, focal at the origin and focal length $\frac{1}{2} e^{-u_1^*(x_1)}$.[19] A parabolic mirror reflects all incoming light rays in its axis direction. Equation (2.29) therefore shows that a ray shot in direction x_0 will first touch the parabola $\mathrm{Par}_{T(x_0)}$ after travelling the distance

$$e^{u_0^*(x_0)} = \inf_{x_1 \in X_1} \left\{ \frac{e^{-u_1^*(x_1))}}{1 - x_0 \cdot x_1} \right\} \tag{2.31}$$

and then be reflected in the $x_1^* = T(x_0)$ direction. The transport map enforces the law of specular reflection of the surface

$$\mathrm{Refl} = \{x_0 \, e^{u_0^*(x_0)}, \ x_0 \in X_0\}.$$

By construction $\mu_1 = T\#\mu_0$, so the illuminance and illumination constraints are satisfied. The simplest example corresponds to a single Dirac target $\mu_1 = \delta_{a_1}$, the solution being the parabolic mirror Par_{a_1}: all the light from μ_0 is reflected in the a_1 direction. The case of two Dirac masses $\mu_1 = \alpha_1 \, \delta_{a_1} + (1 - \alpha_1) \, \delta_{b_1}$ is sufficient to get the general idea. According to (2.31) rays will hit the closest parabola, and the mirror is the inf-envelope of Par_{a_1} and Par_{b_1}. The focal lengths, depending on the 'prices' $(u_1(a_1), u_1(b_1))$, determine their unique intersection point. It partitions the support of μ_0 into the rays reflecting in the a_1 and b_1 direction respectively. It is adjusted so that the energy carried splits according to the Dirac weights $(\alpha_1, 1-\alpha_1)$. This is easily generalized to M Dirac masses and, as the reflection depends on the local tangent to the reflector, to densities.

The corresponding c-Monge–Ampère equation has been solved numerically using a B-spline collocation method and a multi-scale approach in Brix *et al.* (2015).

[18] Written for vectors in the ambient space $\mathbb{R}^d$.

[19] A parabola reflects all incoming rays in the direction of its axis.
See https://en.wikipedia.org/wiki/Parabola#Proof_of_the_reflective_property.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a)



(b)

Figure 2.2. (a) Different μ_1 distributions for a fixed source μ_0. (b) Ray-tracing resimulation via Monge–Ampère computation of the reflector. Figure reproduced from Brix, Hafizogullari and Platen (2015) with permission. Copyright © 2015 World Scientific.

The method is tested by providing an image on a projection plane, computing the corresponding reflector, resimulating via ray-tracing, and finally comparing the two pictures: see Figure 2.2.

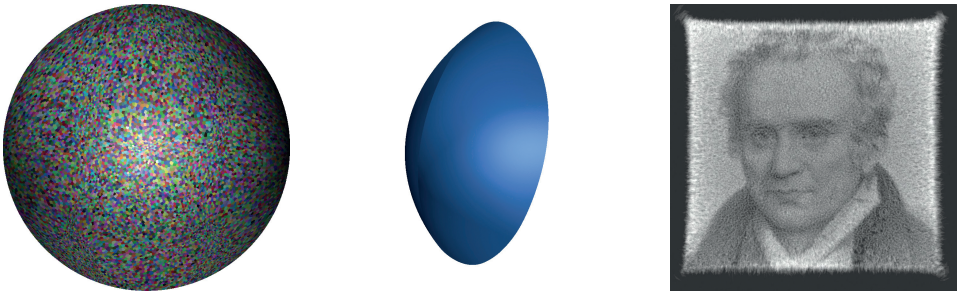
Semi-discrete optimal transportation has been applied to the reflector problem in de Castro *et al.* (2016). Equation (2.31) is simply the u_1 c-transform. When μ_1 is discrete we have a finite collection of parabolae, each sending rays to one fixed direction $x_{1,n}$. The Laguerre cells contain the ray directions of the source touching the corresponding parabola. Adjusting the focal length $\mathrm{e}^{-u_1(x_1)}$ amounts to moving the parabola closer to or further away from the origin such that the amount of light energy carried by the rays in the Laguerre cell is exactly the weight of the corresponding target direction; see Figure 2.3.

Use of the Sinkhorn algorithm and the entropic regularization method has been investigated in Benamou, Ijzerman and Rukhaia (2020).

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) initial $(\lambda_i)_{1\le i\le N}$ (b) final $(\lambda_i)_{1\le i\le N}$ (c) rendering



(d) final $(\lambda_i)_{1\le i\le N}$ (e) reflector (f) rendering

Figure 2.3. The target distribution μ_1 is a picture of Monge. Calculations were done with $N = 1000$ paraboloids for the first row and $N = 15\,000$ paraboloids for the second row. (a) Paraboloid intersection diagram (the Laguerre cells) for an initial $(u_{0,n})_{1\le n\le N}$. (b,d) Final intersection diagram after optimization. (e) Reflector surface defined by the intersection of paraboloids. (c,f) Simulation of the illumination at infinity from a point light source illuminating the northern hemisphere $(\mathbb{S}^2)^+$ uniformly, using LuxRender, a physically accurate ray-tracing engine. Figure reproduced from de Castro, Mérigot and Thibert (2016) with permission. Copyright © 2016 Springer.

3. Wasserstein distance and entropic interpolation

3.1. Displacement interpolation and $\mathcal{W}_2$ metric

In this section we will restrict ourselves to the classical Euclidean quadratic cost (1.1) and show that the optimal value function, or Monge optimal transport 'work'

$$\mathcal{W}_2(\mu_0,\mu_1) := \sqrt{\langle c,\pi^*\rangle_{X_0\times X_1}} = \sqrt{\int \|T^*(x_0)-x_0\|^2\, d\mu_0(x_0)}, \tag{3.1}$$

defines a distance, in the mathematical sense on $\mathcal{P}(X)$, known as the Wasserstein-2 distance. Identity (nothing moves) and symmetry (transport back) are immediate.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The triangle inequality is a consequence of the same property for c, the displacement cost. Let $T_{\mu_0\to\rho}$ and $T_{\rho\to\mu_1}$, respectively, be the *optimal Monge maps* from μ_0 to ρ and ρ to μ_1, and assume they are all absolutely continuous densities of probabilities. Then $T_{\rho\to\mu_1}\circ T_{\mu_0\to\rho}$ is an admissible transport from μ_0 to μ_1. We have

$$\begin{aligned}\mathcal{W}_2(\mu_0,\mu_1) &= \sqrt{\int \|T_{\rho\to\mu_1}(T_{\mu_0\to\rho}(x_0)) - x_0\|^2\, d\mu_0(x_0)}\\ &\le \sqrt{\int \|T_{\rho\to\mu_1}(T_{\mu_0\to\rho}(x_0)) - T_{\mu_0\to\rho}(x_0)\|^2\, d\mu_0(x_0)}\\ &\quad + \sqrt{\int \|T_{\mu_0\to\rho}(x_0) - x_0\|^2\, d\mu_0(x_0)}.\end{aligned}$$

The last term is $\mathcal{W}_2(\mu_0,\rho)$ and the change of variable $x_1 = T_{\mu_0\to\rho}(x_0)$ in the first one gives $\mathcal{W}_2(\rho,\mu_1)$ (see Santambrogio 2015, §5.1).

We are now going to make a detour through *displacement interpolation* and an incursion into *multi-marginal optimal transportation*.

We assume that X_0 and X_1 are subsets of a larger X. The interpolation in $\mathcal{P}(X)$ with respect to the transport cost is

$$\rho_\tau := \operatorname*{arg\,inf}_{\rho\in\mathcal{P}(X)}\{(1-\tau)\,\mathcal{W}_2^2(\mu_0,\rho) + \tau\,\mathcal{W}_2^2(\rho,\mu_1)\},\quad \tau\in(0,1). \tag{3.2}$$

The mapping $\rho\mapsto\mathcal{W}_2^2(\mu_0,\rho)$ is continuous, differentiable and convex on $\mathcal{P}(X)$. These properties are further discussed in Section 7.2. The interpolation ρ_τ is well-defined. It is the distribution of mass closest to μ_0 and μ_1 in the sense of the weighted average transport cost. We have two nested optimization problems in (3.2), that is,

$$\inf_{\rho\in\mathcal{P}(X)}\left\{(1-\tau)\inf_{\pi_{0\to\tau}\in\Pi(\mu_0,\rho)}\langle c,\pi_{0\to\tau}\rangle_{X_0\times X} + \tau\inf_{\pi_{\tau\to1}\in\Pi(\rho,\mu_1)}\langle c,\pi_{\tau\to1}\rangle_{X\times X_1}\right\}. \tag{3.3}$$

The plans $\pi_{0\to\tau}$ and $\pi_{\tau\to1}$ specify the quantity of mass flowing from $x_0\in X_0$ to some $x_\tau\in X$ and then from $x_\tau\in X$ to $x_1\in X_1$. We can rewrite this problem in the Kantorovich spirit. Instead of looking for coupling, let us search for 'triplings': the quantity of mass, denoted π_3, flowing across triplets $(x_0,x_\tau,x_1)\mapsto\pi_3(x_0,x_\tau,x_1)$ in the bigger space $\mathcal{P}(X_0\times X\times X_1)$:

$$\inf_{\rho\in\mathcal{P}(X)}\left\{\inf_{\pi_3\in\Pi_3(\mu_0,\rho,\mu_1)}\langle c_3,\pi_3\rangle_{X_0\times X\times X_1}\right\}, \tag{3.4}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

where[20]

$$\Pi_3(\mu_0, \rho, \mu_1) \tag{3.5}$$
$$:= \{\pi_3 \in \mathcal{P}(X_0 \times X \times X_1) \colon P_{X_0}\#\pi_3 = \mu_0, P_{X_1}\#\pi_3 = \mu_1 \text{ and } P_X\#\pi_3 = \rho\}$$

is the set of 3-marginal plans and c_3 is given by (3.6) below. For all couples of 2-plans $\pi_{0\to\tau} \in \Pi(\mu_0, \rho)$ and $\pi_{\tau\to 1} \in \Pi(\rho, \mu_1)$ there exists a 3-plan $\pi_3 \in \Pi_3(\mu_0, \rho, \mu_1)$ such that $\pi_{0\to\tau} = P_{X_1}\#\pi_3$ and $\pi_{\tau\to 1} = P_{X_0}\#\pi_3$ (see Lemma 5.5 in Santambrogio 2015) and conversely. Problems (3.3) and (3.4) are equivalent.

The intermediate marginal ρ support can be determined with the help of the displacement cost:

$$c_3(x_0, x_\tau, x_1) = (1-\tau)\frac{1}{2}\|x_\tau - x_0\|^2 + \tau\frac{1}{2}\|x_1 - x_\tau\|^2. \tag{3.6}$$

If $\pi_3(x_0, x_\tau, x_1) > 0$ and because c_3 is a strictly convex function of x_τ, mass travelling from x_0 to x_1 necessarily goes through its minimum $x_\tau = x_0 + \tau(x_1 - x_0)$. Straight trajectories are optimal. The 3-marginal cost c_3 therefore simplifies to the 2-marginal cost c (1.1) for the mass carried by minimizers π_3^* of (3.5)–(3.4). The integration $P_{X_0\times X_1}\#\pi_3^* = \langle 1, \pi_3\rangle_X$ coincides with the Kantorovich 2-plan $\pi_{0\to 1}^*$ solution of (1.2).[21] The 3-plan π_3^* is fully determined by $\pi_3^*(x_0, x_\tau, x_1) = \pi_{0\to 1}^*(x_0, x_1)$ if $x_\tau = x_0 + \tau(x_1 - x_0)$ and 0 else; remember that optimal transport paths do not cross. The outer minimization in (3.4) can be eliminated and the problem relaxed to the simpler

$$\inf_{\pi_3 \in \Pi_2(\mu_0,\mu_1)} \langle c_3, \pi_3\rangle_{X_0\times X\times X_1}, \tag{3.7}$$

where

$$\Pi_2(\mu_0, \mu_1) := \{\pi_3 \in \mathcal{P}(X_0 \times X \times X_1) \colon P_{X_0}\#\pi_3 = \mu_0, P_{X_1}\#\pi_3 = \mu_1\}. \tag{3.8}$$

We are back to the 2-marginal constraints, but optimize on the 3-marginal plans. The solutions of (3.3) are recovered via marginal integrations: $\rho_\tau = P_X\#\pi_3^*$, $\pi_{0\to\tau}^* = P_{X_0\times X}\#\pi_3^*$ and $\pi_{\tau\to 1}^* = P_{X\times X_1}\#\pi_3^*$.

At the Monge level, McCann (1997) has shown that, given the map $x_0 \mapsto T_{0\to 1}(x_0)$ between μ_0 and μ_1, the map

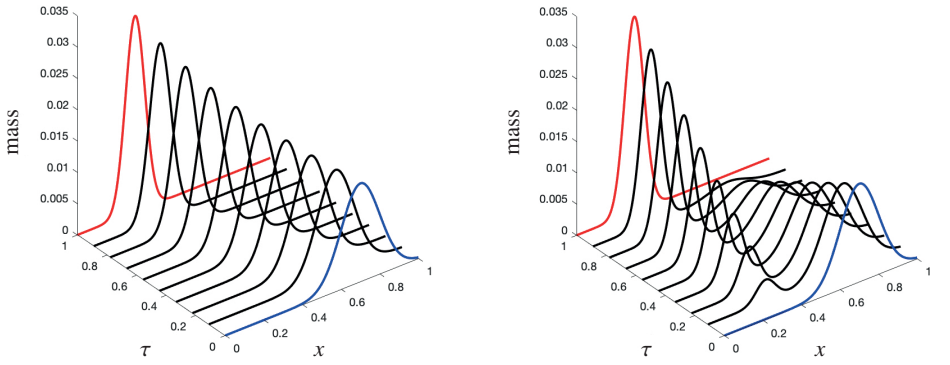
$$x_0 \mapsto T_{0\to\tau}(x_0) := x_0 + \tau\,(T_{0\to 1}(x_0) - x_0)\ (= x_\tau \text{ above}) \tag{3.9}$$

is the Monge map between μ_0 and ρ_τ, the solution of the interpolation problem (3.2). He called this process *displacement interpolation*. It can be used to generate a smooth dynamic $\tau \in (0, 1) \to \rho_\tau$ sequence of probability measures morphing μ_0 to μ_1 continuously. This process will, in particular, displace the support of ρ_τ

[20] The marginal distributions are defined using the canonical projections $P_{X_0} : (x_0, x_\tau, x_1) \to x_0$ and $P_{X_0}\#\pi_3 = \langle 1, \pi_3\rangle_{M\times X_1}$ and likewise $P_{X_1}\#\pi_3 = \langle 1, \pi_3\rangle_{X_0\times X}$, $P_X\#\pi_3 = \langle 1, \pi_3\rangle_{X_0\times X_1}$.

[21] P is again a canonical projection: $P_{X_0,X_1} : (x_0, x_\tau, x_1) \to (x_0, x_1)$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) (b)

Figure 3.1. Blue, μ_0; red, μ_1; black, ρ_τ. The interpolation for (a) the $\mathcal{W}_2$ distance and (b) the $\mathcal{L}^2$ distance.

smoothly and 'horizontally'. Assuming μ_0 and μ_1 are integrable, it is interesting to compare with the $\mathcal{L}^2$ interpolation of functions, that is,

$$\rho_\tau := \arg\inf_{\rho} \{(1-\tau)\,\|\mu_0 - \rho\|^2_{\mathcal{L}^2} + \tau\,\|\rho - \mu_1\|^2_{\mathcal{L}^2}\},$$

which yields a pointwise interpolated sequence $\rho_\tau = (1-\tau)\,\mu_0 + \tau\,\mu_1$. The mass will only move 'vertically'. Figure 3.1 illustrates these fundamentally different behaviours.

An important remark is that the curve $\tau \in (0,1) \mapsto \rho_\tau$ is a *geodesic curve* for the $\mathcal{W}_2$ distance. For any two times $(\tau_0, \tau_1) \in [0,1]^2$, $\tau \in]\tau_0, \tau_1[\to \rho_\tau$ is the displacement interpolation between ρ_{τ_0} and ρ_{τ_1}. The piece of the map (3.9) $x_{\tau_0} \to x_{\tau_1}$ is also the Monge map between ρ_{τ_0} and ρ_{τ_1}.

Displacement interpolation is harder to compute than $\mathcal{L}^2$ interpolation, but its ability to model smooth displacements has caught the attention of many researchers in signal processing. All is far from perfect, though. Signals are not necessarily probability distributions. The measure-preserving/mass-conservation property also rules out any 'rigidity' or 'geometric' constraint on $\mathrm{supp}\,(\rho_\tau)$, but it may be a no-go for many morphing applications. We discuss these restrictions further in Sections 4 and 7.

Finally, let us mention attempts at formulating higher-order optimal transportation interpolations, *e.g.* Benamou, Gallouët and Vialard (2019*c*) and Chen, Conforti and Georgiou (2018).

3.2. Sinkhorn algorithm

A method based on an 'entropic regularization' (Cuturi 2013) was introduced to speed up optimal transportation computations at the price of an approximation.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The method comes with a parallelization-friendly[22] computational method usually known as the *Sinkhorn algorithm*. It can be traced back to much older works. We will present a statistical physics interpretation in Section 6 but it is also linked to exponential barrier functions in discrete optimization (Cominetti and Martín 1994).

A quick and somewhat dirty derivation of the algorithm (the literature usually starts with the primal problem) starts from the dual problem (2.1) in its equivalent formulation

$$\sup_{\{u_0,u_1\}} \langle u_0,\mu_0\rangle_{X_0} + \langle u_1,\mu_1\rangle_{X_1} + \langle \iota_{\{u_0\oplus u_1-c\le 0\}},\mu_0\otimes\mu_1\rangle_{X_0\times X_1}, \tag{3.10}$$

where ι is the characteristic function.[23] The integration of the constraint against the product measure $\mu_0\otimes\mu_1$ will be discussed for the primal regularized problem in Section 3.3 below. Entropic regularization replaces the hard constraint (2.2) with a smooth differentiable barrier function depending on a small parameter[24] ϵ:

$$\begin{aligned} OT_\epsilon(\mu_0,\mu_1) &:= \sup_{\{u_{\epsilon,0},u_{\epsilon,1}\}} J_\epsilon(u_{\epsilon,0},u_{\epsilon,1}), \\ J_\epsilon(u_{\epsilon,0},u_{\epsilon,1}) &:= \langle u_{\epsilon,0},\mu_0\rangle_{X_0} + \langle u_{\epsilon,1},\mu_1\rangle_{X_1} - \epsilon\,\langle e^{(u_{\epsilon,0}\oplus u_{\epsilon,1}-c)/\epsilon},\mu_0\otimes\mu_1\rangle_{X_0\times X_1}. \end{aligned} \tag{3.11}$$

The penalization is strictly concave. As in the semi-dual approach, this is a large gain as this formulation is amenable to the solver arsenal of smooth convex/concave optimization. The Sinkhorn algorithm is the simplest iterative (in k) coordinate ascent method (we drop the $.\epsilon$ notation):

$$\begin{aligned} u_0^k &= \arg\sup_{u_0} \langle u_0,\mu_0\rangle_{X_0} - \epsilon\,\langle e^{(u_0\oplus u_1^{k-1}-c)/\epsilon},\mu_0\otimes\mu_1\rangle_{X_0\times X_1}, \\ u_1^k &= \arg\sup_{u_1} \langle u_1,\mu_1\rangle_{X_1} - \epsilon\,\langle e^{(u_0^k\oplus u_1-c)/\epsilon},\mu_0\otimes\mu_1\rangle_{X_0\times X_1}. \end{aligned} \tag{3.12}$$

The maximization can be expressed in closed form as[25]

$$u_0^k = LSE^\epsilon_{\mu_1,X_1}(u_1^{k-1}), \quad u_1^k = LSE^\epsilon_{\mu_0,X_0}(u_0^k), \tag{3.13}$$

where we introduced log/sum/exp (*LSE*), a classical and optimized function in modern scientific software (*e.g.* PyTorch):

$$LSE^\epsilon_{\rho,X}(u) := \{x_1 \mapsto -\epsilon\,\log(\langle e^{(u-c)/\epsilon},\rho\rangle_X)\}, \tag{3.14}$$

which sends, alternately, a function or vector defined on X_0 to one defined on X_1, and reciprocally (remember c is defined on $X_0\times X_1$). At convergence, (3.13) is an

[22] An efficient off-the-shelf GPU implementation is available at https://www.kernel-operations.io/geomloss/.

[23] $\iota_{\{v\le 0\}} := 0$ if $v\le 0$ and $+\infty$.

[24] $\lim_{\epsilon\to 0}\epsilon\,e^{v/\epsilon} = \iota_{\{v<0\}}$.

[25] This is a good exercise.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

ϵ-smooth version of (2.5)[26] up to a constant shift of the potentials. The formulation (3.12) holds for both continuous and discrete measures. Using the discretization of Section 1.3, one iteration of (3.13) is quadratic in N. A nice feature of the algorithm is its formal independence on c, unlike the Monge–Ampère and semi-discrete approaches. Details and references can be found in the survey by Vialard (2019) and the recent papers by Di Marino and Gerolin (2020) and Peyré and Cuturi (2019, §4.2). It gives a linear convergence rate in[27]

$$O(1 - \mathrm{e}^{L_c(\operatorname{Diam}\mu_0)/\epsilon})\, O(1 - \mathrm{e}^{L_c(\operatorname{Diam}\mu_1)/\epsilon}),$$

where L_c is the Lipschitz constant of c. For the Euclidean quadratic cost (1.1) this is the friendly constant 1, but, as we can see, the dependence on ϵ is not so nice.

We finish this section with several important remarks for the numerical analysts interested in using (3.13).

The kernel appearing in the iterations scales like $\mathrm{e}^{L_c(\operatorname{Diam}\mu_{0,1})/\epsilon}$. In finite precision this leads to computer underflow and overflow when ϵ is too small or L_c Diam $\mu_{0,1}$, the scale of the ground displacement, is too large (or both). Several stabilization techniques are available: see Schmitzer (2019).

The operation cost of Sinkhorn iterations can be drastically reduced for the Euclidean quadratic cost (1.1). Rewrite $\langle \mathrm{e}^{(u-c)/\epsilon}, \rho\rangle_X$ in the *LSE* function as

$$\{y \mapsto \langle \mathrm{e}^{u(x)/\epsilon}\gamma_\epsilon(y-x)), \rho\rangle_X\},$$

where

$$\gamma_\epsilon(z) = \frac{1}{(2\pi\epsilon)^{d/2}}\mathrm{e}^{-\|z\|^2/(2\epsilon)} \tag{3.15}$$

is a Gaussian with variance ϵ. The normalization constant is not important at this stage (and we will assume in this section it is absorbed by the potentials (u_0, u_1)) but will be useful in Section 6.1. If $X_0 = X_1$ are d-dimensional regular Cartesian grids,

$$\|z\|^2 = \|z_1\|^2 + \|z_1\|^2 + \cdots + \|z_d\|^2$$

and the number of z_i points is $N^{1/d}$. The exponential matrix tensorizes and the operation cost of one iteration reduces to $O(N^{1+1/d})$.

When the problem admits a Monge solution, it is possible to take advantage of the sparsity of the entropic plan to obtain an $O(N\log(N))$ algorithm via a strategy that is multiscale in ϵ; see Schmitzer (2019).

Much more is to be found in Peyré and Cuturi (2019, §4).

[26] *LSE* is also known as 'soft-min': $\lim_{\epsilon\to 0} LSE^\epsilon_{\mathbf{1},X}(g)(\cdot) = \inf_{x\in X} c(\cdot, x) - g(x)$.

[27] Diam ρ is the diameter of supp ρ.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

3.3. Entropic bias

The OT_ϵ solution is, however, biased by the penalization term. The entropic bias is better understood for the Fenchel–Rockafellar primal corresponding to (3.11):

$$OT_\epsilon(\mu_0, \mu_1) := \inf_{\pi_\epsilon \in \Pi(\mu_0,\mu_1)} \langle c, \pi_\epsilon \rangle_{X_0 \times X_1} + \epsilon\ KL(\pi_\epsilon | \mu_0 \otimes \mu_1), \tag{3.16}$$

where $KL(\nu|\nu_0) := \langle \log(\nu/\nu_0) - 1, \nu \rangle + \langle 1, \nu_0 \rangle$ is the Kullback–Leibler divergence.[28] The compact notation may be a bit confusing: ν/ν_0 implies that μ is absolutely continuous with respect to the measure ν_0 (continuous or discrete), otherwise KL takes, by convention, an infinite value. So $\operatorname{supp} \nu \subset \operatorname{supp} \nu_0$ and ν/ν_0 is a density.

The penalization $\nu \mapsto KL(\nu|\nu_0)$ is strictly convex, positive and vanishes for $\nu = \nu_0$. The penalization $\epsilon\ KL(\pi_\epsilon|\mu_0 \otimes \mu_1)$ biases the entropic plan π_ϵ^* towards the measure $\mu_0 \otimes \mu_1$. Figure 3.2 illustrates this effect. The choice of this particular reference measure is guided by the following consideration: it is our best direct guess for π_ϵ^* in $\Pi(\mu_0, \mu_1)$ in terms of both support and value. We therefore expect to reduce the bias with the classical optimal transportation solution introduced by the regularization. Note that it is possible to choose a reference measure such that $\Pi(\mu_0, \mu_1)$ will be empty of plans with the corresponding support. The problem then becomes infeasible.

A few lines of computations rewrite (3.16) as

$$\inf_{\pi_\epsilon \in \mathcal{M}(X_0 \times X_1)} \iota_{\{P_{X_0}\#\pi_\epsilon = \mu_0\}} + \iota_{\{P_{X_1}\#\pi_\epsilon = \mu_1\}} + \epsilon\ KL(\pi_\epsilon | \pi_\epsilon^0), \tag{3.17}$$

where $\mathcal{M}$ is the set of Radon measure and

$$(x_0, x_1) \mapsto \pi_\epsilon^0(x_0, x_1) = \gamma_\epsilon(x_1 - x_0)\mu_0(x_0)\mu_1(x_1). \tag{3.18}$$

It is important to note that this formulation is unconstrained. The (strict) positivity of π_ϵ is enforced by the KL penalization and its mass sums to 1 because of the marginal constraints. Optimizers are necessarily in $\mathcal{P}(X_0 \times X_1)$.

The equivalence between (3.10) and (3.17) is a direct application of *Fenchel–Rockafellar convex duality*; see Theorem 6.3 in Carlier (2021). Specifically, let $(\mathcal{E}, \mathcal{E}')$ and $(\mathcal{F}, \mathcal{F}')$ be a pair of dual normed spaces. Let $A\colon \mathcal{E} \to \mathcal{F}$ be a continuous linear operator and A' its adjoint. Let F, G be proper, convex and lower semi-continuous functions defined on $\mathcal{E}$ and $\mathcal{F}$ respectively. If there exists $\sigma \in \operatorname{Dom}(F)$ such that G is continuous at $A\sigma$, then

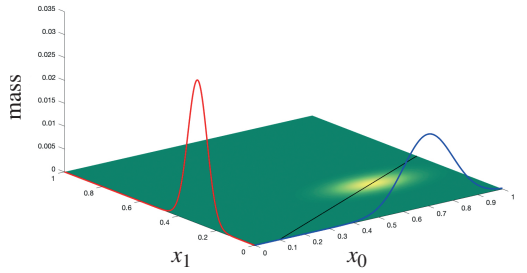
$$\sup_{\sigma \in E} -F(-\sigma) - G(A\sigma) = \inf_{q \in F'} F^\star(A'q) + G^\star(q) \tag{3.19}$$

and the infimum is attained. Moreover, if there exists a maximizer $\sigma^* \in E$, then there exists $q^* \in F'$ such that

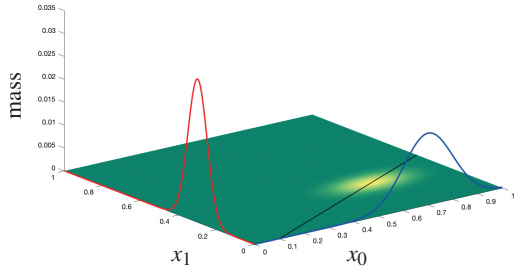
$$A\sigma^* \in \partial G(q^*) \quad \text{and} \quad A'q^* \in \partial F(-\sigma^*). \tag{3.20}$$

[28] Also known as the *relative entropy* between ν and ν_0. It is strictly convex, vanishes and takes its minimum at ν_0, and has an infinite slope at 0. Its Fréchet derivative is formally given by $\langle \delta\, KL(\nu|\nu_0), \delta\nu \rangle = \langle \log(\nu/\nu_0), \delta\nu \rangle$.

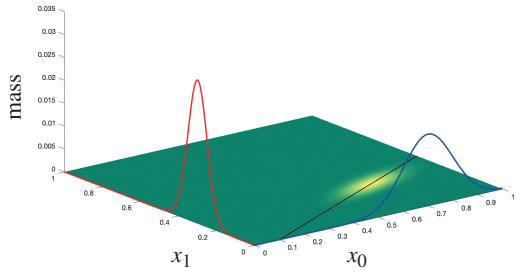
https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



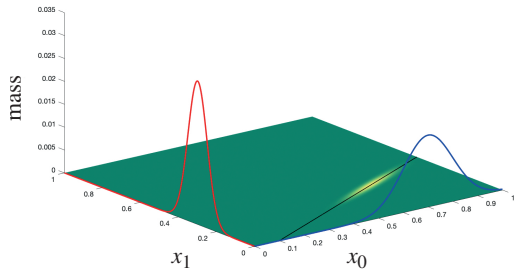
(a) $\epsilon = 1$



(b) $\epsilon = 0.1$



(c) $\epsilon = 0.01$



(d) $\epsilon = 0.001$

Figure 3.2. Blue, μ_0; red, μ_1. The colour map corresponds to π^*_ϵ for different values of ϵ. The black line is the optimal transport map T ($\epsilon = 0$). Observe how the entropic plan 'turns', 'aligns' and 'shrinks' from the product measure $\mu_0 \otimes \mu_1$ to the graph of the transport map as ϵ decreases to 0.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

We now apply (3.19) with

$$
\begin{aligned}
&A\colon \sigma := (u_0, u_1) \in \mathcal{C}(X_0) \times \mathcal{C}(X_1) \mapsto u_0 + u_1 \in \mathcal{C}(X_0 \times X_1),\\
&A'\colon q := \pi_\epsilon \in \mathcal{M}(X_0 \times X_1) \mapsto (P_{X_0}\#\pi_\epsilon, P_{X_1}\#\pi_\epsilon) \in \mathcal{M}(X_0) \times \mathcal{M}(X_1),\\
&F\colon (u_0, u_1) \in \mathcal{C}(X_0) \times \mathcal{C}(X_1) \mapsto \langle u_0, \mu_0\rangle_{X_0} + \langle u_1, \mu_1\rangle_{X_1},\\
&F^\star\colon (\rho_0, \rho_1) \in \mathcal{M}(X_0) \times \mathcal{M}(X_1) \ \mapsto \iota_{\{\rho_0=\mu_0\}} + \iota_{\{\rho_1=\mu_1\}},\\
&G\colon u \in \mathcal{C}(X_0 \times X_1) \mapsto \{q \to KL(q|\pi_\epsilon^0)\}^\star(u) = -\epsilon\, \langle \mathrm{e}^{(u-c)/\epsilon}, \mu_0 \otimes \mu_1\rangle_{X_0\times X_1},\\
&G^\star\colon q \in \mathcal{M}(X_0 \times X_1) \mapsto G^\star(q) = \epsilon\ KL(q|\pi_\epsilon^0).
\end{aligned}
\tag{3.21}
$$

Remember that $F = F^{\star\star}$ and $G = G^{\star\star}$ are convex though not necessarily continuous. In particular, $F^\star$ above takes infinite values. Using this particular choice the primal–dual optimality conditions (3.20) take the form

$$
\pi_\epsilon^* = \mathrm{e}^{(u^*_{\epsilon,0} \oplus u^*_{\epsilon,1} - c)/\epsilon} \mu_0 \otimes \mu_1 \quad \text{and} \quad (P_{X_0}\#\pi_\epsilon, P_{X_1}\#\pi_\epsilon) = (\mu_0, \mu_1). \tag{3.22}
$$

Under the Fenchel–Rockafellar formalism, Sinkhorn iterations are the coordinate ascents

$$
\begin{aligned}
u_0^k &= \arg\sup_{u_0} -F(-(u_0, u_1^{k-1})) - G(A\,(u_0, u_1^{k-1}),\\
u_1^k &= \arg\sup_{u_1} -F(-(u_0^k, u_1)) - G(A\,(u_0^k, u_1)).
\end{aligned}
\tag{3.23}
$$

Details and generalizations of this formalism are given in Chizat, Peyré, Schmitzer and Vialard (2018*a*).

Let us restrict again to the Euclidean quadratic cost (1.1) and the discrete case (Section 1.3) to discuss (3.22). The entropic optimal transportation matrix is

$$
\pi^*_{\epsilon,(i,j)} = (w_{0,i}\, \mathrm{e}^{u^*_{0,i}/\epsilon})\, \gamma_\epsilon(x_{1,j} - x_{0,i})\, (w_{1,i}\, \mathrm{e}^{u^*_{1,i}/\epsilon}), \tag{3.24}
$$

where $\{u_{0,i}\} := \{u_{\epsilon,0}(x_{0,i})\}$ and $\{u_{1,j}\} := \{u_{\epsilon,1}(x_{1,j})\}$ are converged solutions of the (3.12) Sinkhorn iterations. The Gaussian matrix $\Gamma_\epsilon := \{\gamma_\epsilon(x_{1,j} - x_{0,i})\}$ (see (3.15)) is strictly positive and decreases exponentially away from the diagonal (if the marginal discretizations are identical). On each side there is a diagonal matrix scaling Γ_ϵ such that $\pi^*_\epsilon \in \Pi(\mu_0, \mu_1)$. Denoting the discretization scale by h, the effective bandwidth of Γ_ϵ in finite precision admits a maximum admissible displacement. It will decrease as h^2/ϵ increases. It may again cause the set of $\pi^*_\epsilon \in \Pi(\mu_0, \mu_1)$ in the form (3.24) to be empty.

Decreasing the entropic bias therefore involves h and ϵ; Berman (2020) gives quantitative results. Set $h = 1/N^{1/d} = \epsilon$ and assume the ground space discretization of the *smooth densities* $\mu_{0,1}$ is on the scale of ϵ. After running (3.13) for

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

$k_\epsilon = 1/\epsilon \log(1/\epsilon)$ iterations, we get[29]

$$\|u_{0,\epsilon}^{k_\epsilon} - u_0^*\|_\infty \leq C\,\epsilon\,\log\left(\frac{1}{\epsilon}\right). \tag{3.25}$$

As for the optimal plan, Berman (2020) gives the estimate

$$\pi_\epsilon^{k_\epsilon} \leq B_\epsilon\,\mu_0 \otimes \mu_1, \tag{3.26}$$

where

$$(x_0, x_1) \longrightarrow B_\epsilon(x_0, x_1) := \frac{p}{\epsilon^p}\,\exp\left(-\frac{c(x_1, T(x_0))}{\epsilon\,p}\right). \tag{3.27}$$

See (2.10) for the formulation of T as a function of c and u_0^*. The approximation $\pi_\epsilon^{k_\epsilon}$ is to be understood as (3.22) for the running (3.13) iterates. The parameter p is positive and depends on the marginals but is unfortunately not explicit. Estimate (3.27) is nevertheless important for two reasons. First it shows that the entropic bias decreases exponentially fast around the continuous transport map (see Figure 3.2). So even though $\pi_\epsilon^{k_\epsilon}$, in theory, has full support on the grid, it will be negligible, and even null in finite precision, for an increasing number of points as ϵ decreases. A detailed numerical description of these phenomena and a proposed heuristic multi-scale method in ϵ can be found in Schmitzer (2019). A GPU implementation can also be found in Feydy (2020). They both observe numerically the desired $O(N \log(N))$ complexity arising from the sparsity of the transport plan. A convergence proof of the multiscale method using (3.26) is given in Benamou and Martinet (2020). We are back to the performance of the Monge–Ampère and semi-discrete formulations but only when smooth Monge maps exist.

3.4. *Soft displacement interpolation*

Computing the Wasserstein interpolant should in principle be as simple as solving the original optimal transportation problem and using the Monge map (3.9) to push forward μ_0 to ρ_τ. As discussed in the introduction, Monge maps are concepts that unfortunately may be lost after discretization and will definitely be lost after the entropic regularization. So we instead resort to the simplified Kantorovich formulation (3.7)–(3.8) and will recover ρ_τ as the intermediate marginal. The discretization of (1.2) presented in Section 1.3 was straightforward. The discretization of (3.7)–(3.8) is more subtle. We have seen that the optimal plan π_3^* charges triplets $(x_0, x_0 + \tau(x_1 - x_0), x_1)$ when the original optimal transportation solution $\pi^*(x_0, x_1)$ also does. If $X_k = \{x_{k,i}\}_{i=1,N}$ for $k = 0, 1$, we need to cover all possible displacement interpolants for the discretization of the support ρ_τ. This translates into $\{x_{0,i} + \tau(x_{1,j} - x_{0,i})\}_{(i,j)\in[1,N]^2}$ with possibly N^2 points. The 3-marginal transport plan will be a $N \times N^2 \times N$ tensor. Solving the discretized (3.7)–(3.8) linear

[29] The precise result involves interpolation of the discrete potential with the c-transform formula (2.10).

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

program is not possible for reasonable discretization sizes. One may, of course, use an arbitrary intermediate grid; the discretization of all three ground spaces may be, for instance, identical regular Cartesian grids of size N. In so doing, we add a constraint on the support of ρ_τ. This often generates non-smooth oscillating results even for smooth μ_0 and μ_1.

Entropic penalization offers a win–win solution to this problem, in terms of both computational efficiency and removing the grid pollution explained above. The suggested primal entropic version of (3.7) is

$$\inf_{\pi_{3,\epsilon}\in\Pi_2(\mu_0,\mu_1)} \langle c_3, \pi_{3,\epsilon}\rangle_{X_0\times X\times X_1} + \epsilon\ KL(\pi_{3,\epsilon}|\mu_0\otimes \mathbf{1}_X\otimes\mu_1). \tag{3.28}$$

As π_3 and π share the same marginal constraints (3.8), the dual problem is shown to be

$$\sup_{(u_{\epsilon,0},u_{\epsilon,1})} \langle u_{\epsilon,0},\mu_0\rangle_{X_0} + \langle u_{\epsilon,1},\mu_1\rangle_{X_1} - \epsilon\,\langle \mathrm{e}^{(u_{\epsilon,0}\oplus u_{\epsilon,1}-c_3)/\epsilon}, \mu_0\otimes\mu_1\rangle_{X_0\times X\times X_1}. \tag{3.29}$$

It differs from (3.11) only in the final term. Using c_3 (see 3.6), the primal–dual optimality conditions (3.22) are

$$\pi^*_{3,\epsilon}(x_0,x_\tau,x_1) = \mathrm{e}^{u_0^*(x_0)/\epsilon}\,\gamma_{\epsilon/(1-\tau)}(x_\tau-x_{0,i})\,\gamma_{\epsilon/\tau}(x_1-x_\tau)\,\mathrm{e}^{u_1^*(x_1)/\epsilon}. \tag{3.30}$$

Remember that the non-entropic $P_{X_0\times X_1}\#\pi^*_3 = \langle 1,\pi_3\rangle_X$ is the Kantorovich plan $\pi^*_{0\to 1}$ solution of (1.2). Applying the same marginal integration inside (3.30) amounts to the convolution

$$\langle\{x_\tau\mapsto \gamma_{\epsilon/(1-\tau)}(x_\tau-x_0)\,\gamma_{\epsilon/\tau}(x_1-x_\tau)\},\mathbf{1}_X\rangle_X = \gamma_{\epsilon/(\tau(1-\tau))}(x_1-x_0), \tag{3.31}$$

where we are taking some liberties with the constants and used the convolution of Gaussian properties.[30] We get

$$P_{X_0\times X_1}\#\pi^*_{3,\epsilon}(x_0,x_\tau,x_1) = \mathrm{e}^{u_0^*(x_0)/\epsilon}\,\gamma_{\epsilon/(\tau(1-\tau))}(x_1-x_0)\,\mathrm{e}^{u_1^*(x_1)/\epsilon}, \tag{3.32}$$

and recognize $\pi^*_{\epsilon/(\tau(1-\tau))}$ the solution of $OT_{\epsilon'}(\mu_0,\mu_1)$, $\epsilon'=\epsilon/(\tau(1-\tau))$ because it satisfies (3.24) and also belongs to $\Pi(\mu_0,\mu_1)$ by construction. After discretization, we use (3.13) with ϵ' to compute an approximation to (3.24) and of the potentials $u^*_{0,1}$ shared by (3.32) and (3.30). Then we discretize the formula (3.30) to smoothly interpolate

$$\begin{aligned}\rho_\tau = P_X\#\pi^*_{3,\epsilon} &= \langle 1,\pi^*_{3,\epsilon}\rangle_{X_0\times X_1}\\ &= \Big\{x_\tau\mapsto \\ &\quad\Big(\sum_i (w_{0,i}\,\mathrm{e}^{u^*_{0,i}/\epsilon})\gamma_{\epsilon/(1-\tau)}(x_\tau-x_{0,i})\Big)\Big(\sum_j \gamma_{\epsilon/\tau}(x_{1,j}-x_\tau)(w_{1,i}\,\mathrm{e}^{u^*_{1,j}/\epsilon})\Big)\Big\}.\end{aligned} \tag{3.33}$$

[30] The convolution of two Gaussians is a Gaussian. The expectation is the sum of the expectations and the variance is the product of the variances.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

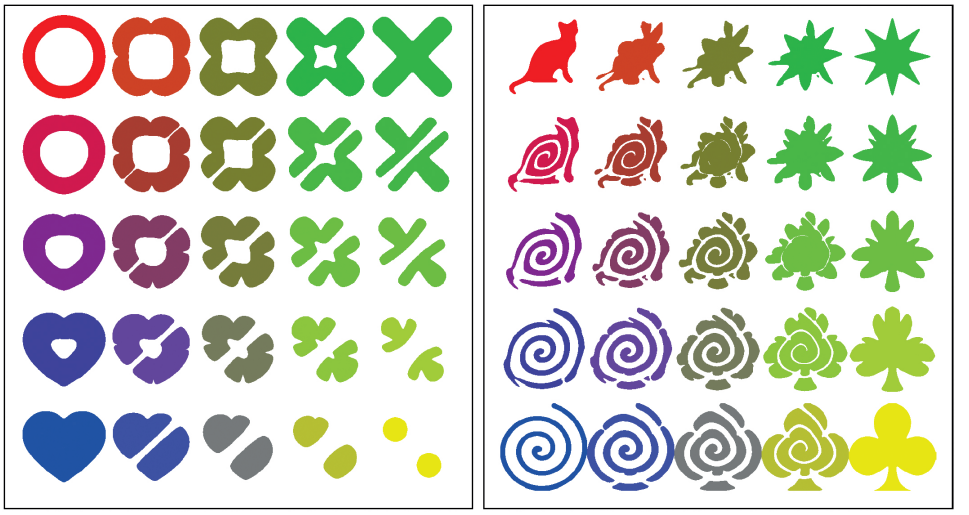


Figure 3.3. OT_ϵ barycentres (the interior) and interpolations (the edges) of four densities (the corners) for different sets of $(\lambda_p)_{p=1..4}$. Figure reproduced from Peyré and Cuturi (2019) with permission.

This is the product of two discrete convolutions with Gaussians. If the discretization h is small compared to the standard deviations $\epsilon/(1-\tau)$ and $\epsilon/(1-\tau)$, then the interpolation will be smooth. Note that the ϵ' parameter needs to be small to get a reasonable approximation of the non-entropic interpolant. This bounds the τ parameter away from 0 and 1.

Generalizing the interpolation, the computation of means or barycentres of images is a crucial step in the statistical analysis of medical images, for example, or in texture synthesis (Bonneel, Peyré and Cuturi 2016). Formally, at least, Wasserstein barycentres are a direct generalization of (3.2). Given a set of P probability measures $\{\mu_p\}_{p\in[1,P]}$,

$$\rho_\lambda := \arg\inf_{\rho\in\mathcal{P}(X)} \sum_{p=1}^{P} \lambda_p \, \mathcal{W}_2^2(\rho,\mu_p), \tag{3.34}$$

where $\{\lambda_p\}_{p\in[1,P]} \in (\mathbb{R}^+)^P$ are given and $\sum_{p=1}^P \lambda_p = 1$. The Kantorovich formulation of this problem was worked out by Agueh and Carlier (2011), who showed the equivalence of different formulation. Replacing $\mathcal{W}_2$ with OT_ϵ in (3.34) leads to a Sinkhorn-like algorithm (Benamou *et al.* 2015); see Figure 3.3 for an illustration.

For a meaningful central limit theorem see Carlier, Eichinger and Kroshnin (2020), which also uses an entropic regularization but this time of ρ.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

4. Dynamic optimal transportation

4.1. Velocity discretization and Lagrangian computational fluid dynamics formulation

The 3-marginal formulation developed in Section 3.1 can be generalized to M marginals. Let the sets $\{X_m\}$ be M identical copies of a compact set $X \in X$, and consider the problem

$$\pi_M^* \in \underset{\pi_M \in \Pi_M(\rho_0,\rho_1)}{\arg\min} \langle c_M, \pi_M \rangle_{\bigotimes_{m=0}^M X_m}, \tag{4.1}$$

where

$$c_M(x_0, x_1, \ldots, x_M) = \frac{1}{2\,M} \sum_{m=0}^{M-1} \|x_{m+1} - x_m\|^2 \tag{4.2}$$

and

$$\Pi_M(\rho_0, \rho_1) := \left\{ \pi_M \in \mathcal{P}\left(\bigotimes_{m=0}^{M} X_m \right) : P_{X_0} \# \pi_M = \rho_0, P_{X_M} \# \pi_M = \rho_M \right\}. \tag{4.3}$$

What we have discussed in the 3-marginal case remains true for M marginals, that is, the mass optimally transported from the optimal x_0^* to x_M^* travels in straight lines:

$$x_m^* = x_0^* + \frac{m}{M}(x_M^* - x_0^*), \quad m = [\![0, M]\!].$$

On such an optimal chain $\{x_m^*\}_{m=0\cdots M}$ the mass transported is constant and equal to $\rho_0(x_0^*)$. This is because $\pi_{0\to M}^* = P_{X_0 \times X_M} \# \pi_M^*$ is also the solution of (1.2).

This may all seem pointless as the underlying transport can be obtained via a static $\pi_{0\to M}$ computation, but the following Lagrangian interpretation will lead us to important variations of optimal transportation, in particular the *computational fluid dynamics* (CFD) formulation (Benamou and Brenier 2000).

The optimal chain $\{x_m^*\}_{m=1\cdots M}$ can be interpreted as a straight continuous curve $\mathcal{X}^*\colon (\tau, x_0^*) \in [0,1] \times X_0 \mapsto \mathcal{X}_\tau^*(x_0^*)$ initially at x_0^* and passing through all points $x_m^* = \mathcal{X}_{\tau_m}^*(x_0^*)$ at times $\tau_m = (1/M)[\![0, M]\!]$. We note that[31] $\dot{\mathcal{X}}_\tau^*(x_0^*) = M(x_{m+1}^* - x_m^*)$ is a constant for all $\tau \in]m/M, (m+1)/M[$ and for all m. The map $x_0^* \mapsto \mathcal{X}_\tau^*(x_0^*)$ is exactly the optimal transportation map between ρ_0 and ρ_τ obtained via displacement interpolation (3.9). Hence

$$\{x_0^* \mapsto \mathcal{X}_\tau^*(x_0^*)\} \# \rho_0 = \rho_\tau. \tag{4.4}$$

We have already explained in the discrete case (Section 1.3) that optimal transport trajectories cannot cross except perhaps at final time 1. This is also the case in the continuous case. Therefore there exists a vector field (a 'velocity' as $\mathrm{d}\tau := 1/M$ is

[31] $\dot{\mathcal{X}}_\tau$ is the time τ derivative along the curve $\tau \to \mathcal{X}_\tau(x_0)$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

a time step) $(\tau, x) \in [0,1] \times X \mapsto V_\tau^*(x) \in \mathbb{R}^d$ such that $\dot{\mathcal{X}}_\tau^*(x_0^*) = \mathcal{V}_\tau^*(\mathcal{X}_\tau^*(x_0^*)) = (1/M)(x_{m+1}^* - x_m^*)$ for all $\tau \in (m/M, (m+1)/M)$ and for all m.

Wrapping all the information above in the cost function (4.1), we have

$$\begin{aligned}\langle c_M, \pi_M^* \rangle_{\bigotimes_{m=0}^M X_m} &= \frac{M}{2} \sum_{m=0}^{M-1} \langle \|\mathrm{Id}_{X_{m+1}} - \mathrm{Id}_{X_m}\|^2, \pi_{m\to m+1}^* \rangle_{X_m \times X_{m+1}} \\ &= \frac{1}{2} \left\langle \int_0^1 \|\mathcal{V}_\tau^*(\mathcal{X}^*)\|^2 \, \mathrm{d}\tau, \rho_0 \right\rangle_{X_0}, \end{aligned} \tag{4.5}$$

where $\pi_{m\to m+1}^* := P_{X_m \times X_{m+1}} \# \pi_M^*$ is the (2-marginal) $m \to m+1$ integration of π_M^*.

It is legitimate to tighten the optimal transportation optimization to the class of Lagrangian curves and minimize their *kinetic energy*

$$\mathcal{W}_2^2(\rho_0, \rho_1) = \inf_{(\mathcal{X}, \mathcal{V}) \in CL(\rho_0, \rho_1)} \frac{1}{2} \left\langle \int_0^1 \|\mathcal{V}_\tau(\mathcal{X}_\tau)\|^2 \, \mathrm{d}\tau, \rho_0 \right\rangle_{X_0}, \tag{4.6}$$

where

$$\begin{aligned} CL(\rho_0, \rho_1) := \big\{ &(\tau, x) \in [0,1] \times X \mapsto \mathcal{V}_\tau(x) \in \mathbb{R}^d \text{ and} \\ &(\tau, x_0) \in [0,1] \times X_0 \mapsto \mathcal{X}_\tau(x_0) \in X \\ &\text{such that} \\ &\dot{\mathcal{X}}_\tau = \mathcal{V}_\tau(\mathcal{X}_\tau),\ \mathcal{X}_0 = \mathrm{Id}_{x_0},\ \mathcal{X}_1 \# \rho_0 = \rho_1 \big\}. \end{aligned} \tag{4.7}$$

This is the Lagrangian CFD formulation. What are the computational benefits or pitfalls of this dynamic optimal transportation? Let us go back to (4.1) and consider an N-point Cartesian grid discretization, denoted X_N, of the identical sets $\{X_m\}$. The marginals ρ_0 and ρ_1 are empirical measures supported on the grid and $\bigotimes_{m=0}^M X_N$ is the fully discretized space–time cylinder. Solving (4.1) then involves π_M, a probability measure over a discrete set of size N^M. This is the set of all possible chains $x_0 \to x_1 \to \cdots \to x_M$ lying on the grid and π_M mass charges these chains under the initial and final constraint (4.3). This problem is usually too large to be tackled numerically and can in any case be simplified to the simpler interpolation problem described in Section 3.1 corresponding to $M = 2$.[32] Using the same grid of points, the Lagrangian paths can be represented as N piecewise affine chains $\{x_{m_i}\}_{m=[\![1,M]\!]}$, $i = [\![1, N]\!]$. The size of the problem is now $O(N\,M)$, and (4.6) becomes

$$\frac{1}{2} \sum_{i=0}^N \left(\mathrm{d}\tau \sum_{m=0}^{M-1} \|x_{m+1,i} - x_{m,i}\|^2 \right) \rho_0(x_{0,i}). \tag{4.8}$$

[32] However, we will see in Section 6.1 that entropic regularization again offers a tractable numerical method to compute the geodesic curve $\tau \to \rho_\tau$. We will also show in Sections 5 and 6.2 that the introduction of time allows us to enrich the problem with additional constraints on the M marginals.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The mass-preserving constraint becomes

$$\rho_1(x) = \sum_{i\in[0,N],\ \text{s.t. } x=x_{M,i}} \rho_0(x_{0,i}) \quad \text{for all } x \in X_N. \tag{4.9}$$

The Lagrangian CFD formulation is an intermediate approach between the Monge–Ampère and semi-discrete methods and the Kantorovich problem (4.1) – when a Monge solution exists, of course. Instead of considering any set of paths between time 0 and 1, an Eulerian vector field governs the trajectories on the space–time cylinder $[0,1]\times X$, and no crossings are allowed. The constraint (4.9), however, makes the problem non-convex. The next section shows how the Eulerian formulation fixes the convexity.

4.2. *Eulerian computational fluid dynamics and hidden convexity*

To do so, again use the discrete time representation (4.5):

$$\begin{aligned}\langle c_M, \pi^*_M\rangle_{\bigotimes_{m=0}^M X_m} &= \frac{M}{2}\sum_{m=0}^{M-1}\langle \|\mathrm{Id}_{X_{m+1}} - \mathrm{Id}_{X_m}\|^2, \pi^*_{m\to m+1}\rangle_{X_m\times X_{m+1}}\\ &= \frac{1}{2}\sum_{m=0}^{M-1}\left\langle \mathrm{d}\tau \|\frac{1}{\mathrm{d}\tau}(\mathcal{X}^*_{\tau_{m+1}} - \mathcal{X}^*_{\tau_{m+1}})\|^2, \rho_0\right\rangle_{X_0}\\ &= \frac{1}{2}\sum_{m=0}^{M-1}\left\langle \mathrm{d}\tau \|\frac{1}{\mathrm{d}\tau}(\mathcal{X}^*_{\tau_{m+1}}\circ(\mathcal{X}^*_{\tau_m})^{-1} - \mathrm{Id}_{X_m})\|^2, \rho_m\right\rangle_{X_m}.\end{aligned} \tag{4.10}$$

The last line is obtained via the change of variable $x_m = \mathcal{X}^*_m(x_0)$ and using (4.4). We have played with the time step $\mathrm{d}\tau = 1/M$ to produce a velocity at $(\tau_m, x_m = \mathcal{X}^*_{\tau_m})$, and for small $\mathrm{d}\tau$ (we omit $\cdot^*$ for clarity),

$$\mathcal{V}_{\tau_m}(x_m) \simeq \frac{1}{\mathrm{d}\tau}(\mathcal{X}_{\tau_{m+1}}(\mathcal{X}^{-1}_{\tau_m}(x_m)) - x_m). \tag{4.11}$$

Neglecting the discretization error (which is nil for straight optimal curves), we complete the Lagrangian-to-Eulerian-coordinates transformation in (4.10) and obtain

$$\langle c_M, \pi^*_M\rangle_{\bigotimes_{m=0}^M X_m} \simeq \frac{1}{2}\sum_{m=0}^{M-1}\langle \mathrm{d}\tau\|\mathcal{V}_{\tau_m}\|^2, \rho_m\rangle_{X_m}. \tag{4.12}$$

We now turn to the measure-preserving constraint (4.4). On all time intervals $[\tau_m, \tau_{m+1}]$ we can use the geodesic property of the curve $\tau \to \rho^*_\tau$ (Section 3.1). For all $m = 0, M-1$, $T_m = \mathcal{X}^*_{\tau_{m+1}}\circ(\mathcal{X}^*_{\tau_m})^{-1}$ is the optimal map from $\rho^*_{\tau_m}$ to $\rho^*_{\tau_{m+1}}$ and in particular $T_m\#\rho^*_{\tau_m} = \rho^*_{\tau_{m+1}}$. Using the approximation (4.11) and the Jacobian

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

equation formulation (2), the constraint becomes[33]

$$\begin{aligned}\rho_{\tau_m} &= \det(D_x T_m)\,(\rho_{m+1}\circ T_m)\\ &\simeq \det(\mathbf{1}_{X_m} + \mathrm{d}\tau\, D_x \mathcal{V}_{\tau_m})\,(\rho_{m+1}\circ(\mathrm{Id}_{X_m} + \mathrm{d}\tau\,\mathcal{V}_{\tau_m}))\\ &\simeq (\mathbf{1}_{X_m} + \mathrm{d}\tau\ \mathrm{Tr}(D_x\mathcal{V}_{\tau_m}))\,(\rho_{m+1} + \mathrm{d}\tau\, D_x\rho_{m+1}\cdot\mathcal{V}_{\tau_m}).\end{aligned} \tag{4.13}$$

To first order in $\mathrm{d}\tau$, we end up with

$$\frac{1}{\mathrm{d}\tau}(\rho_{m+1}-\rho_{\tau_m}) \simeq -\rho_{m+1}\cdot D\mathcal{V}_{\tau_m} - D_x\rho_{m+1}\cdot\mathcal{V}_{\tau_m} = -\operatorname{div}_x(\rho_{m+1}\mathcal{V}_{\tau_m}). \tag{4.14}$$

The Eulerian CFD formulation is obtained by letting $\mathrm{d}\tau\to 0$ in (4.12) and (4.14):

$$\mathcal{W}_2^2(\rho_0,\rho_1) = \inf_{(\rho,\mathcal{V})\in CE(\rho_0,\rho_1)} \int_0^1 \frac{1}{2}\langle\|\mathcal{V}_\tau(x)\|^2,\rho_\tau\rangle_X\,\mathrm{d}\tau, \tag{4.15}$$

where

$$\begin{aligned}CE(\rho_0,\rho_1) := \big\{&(\tau,x)\in[0,1]\times X\mapsto(\rho_\tau(x),\mathcal{V}_\tau(x))\in\mathbb{R}^+\times\mathbb{R}^d\\ &\text{such that}\\ &\partial_\tau\rho_\tau + \operatorname{div}_x(\rho_\tau\,\mathcal{V}_\tau) = 0 \text{ and } \rho_{\tau=0,1}=\mu_{0,1}\big\}.\end{aligned} \tag{4.16}$$

The partial differential equation $\partial_\tau\rho_\tau + \operatorname{div}_x(\rho_\tau\,\mathcal{V}_\tau) = 0$ is called the *continuity equation* in computational fluid dynamics. It must be understood as the measure-preserving constraint (4.4). Rigorous proofs of the Lagrangian/Eulerian formulation require careful mathematical analysis and can be found in Brenier (2020) or Santambrogio (2015, §6.1).

The Eulerian formulation 'breaks' the Lagrangian path dynamics and replaces it with the continuity equation. The nonlinearity in the initial/final marginals (4.9) disappears. The continuity equation becomes linear under the simple change of variable, from velocity to momentum: $(\rho,\mathcal{V})\to(\rho,m:=\rho\,\mathcal{V})$. 'Hidden convexity' is revealed when plugging it into the integrand of the cost function. Indeed $(\rho,\mathcal{V})\to\frac{1}{2}\rho\|\mathcal{V}\|^2$ becomes

$$(\rho,m)\to J(\rho,m) := \frac{1}{2\,\rho}\|m\|^2. \tag{4.17}$$

Convex analysis tells us that, as a homogeneous function of degree one, this is a convex function; more precisely it is the Legendre–Fenchel dual of the characteristic function of a convex set. The change of variable holds at first glance only for $\rho>0$ but can be made legitimate by replacing J with $J^{\star\star}$, its bi-dual Legendre–Fenchel transform (see footnote 13) extended to $\mathbb{R}\times\mathbb{R}^d$. Keeping the notation J, we find that

$$\text{for all } (a,B)\in\mathbb{R}\times\mathbb{R}^d,\quad J^\star(a,B) := \chi_{a+\frac{1}{2}\|B\|^2\le 0}, \tag{4.18}$$

[33] Tr is the trace of a square matrix, and we have used the classical identity $\det(\mathrm{Id}+\mathrm{d}\tau A)\simeq 1+\mathrm{d}\tau\,\mathrm{Tr}(A)$ for small $\mathrm{d}\tau$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

and redefining J,

$$(J^{\star\star}) = J(\rho, m) := \begin{cases} \frac{1}{2\rho}\|m\|^2 & \text{if } \rho > 0 \\ 0 & \text{if } \rho = 0 \text{ and } m = 0 \\ +\infty & \text{else} \end{cases} \tag{4.19}$$

for all $(\rho, m) \in \mathbb{R} \times \mathbb{R}^d$. Remarkably, (4.15)–(4.16) also fits the abstract Fenchel–Rockafellar formalism (3.19) with

$$\begin{aligned} &A\colon \sigma := (\phi, \phi_0, \phi_1) \in \mathcal{C}([0,1], X) \times \mathcal{C}(X_0) \times \mathcal{C}(X_1) \mapsto (\partial_\tau \phi, D_x \phi), \\ &A'\colon q := (\rho, m) \in (\mathcal{M}([0,1] \times X))^{d+1} \mapsto -(\partial_\tau \rho_\tau + \operatorname{div}_x\, m, \rho_0, -\rho_1), \\ &F\colon (\phi, \phi_0, \phi_1) \mapsto \langle \phi_0, \mu_0 \rangle_{X_0} + \langle \phi_1, \mu_1 \rangle_{X_1}, \\ &F^\star\colon (a, \rho_0, \rho_1) \mapsto \chi_{a=0, \rho_0=\mu_0, \rho_1=\mu_1}, \\ &G\colon (a, b) \mapsto \mathbf{1}_{a+\frac{1}{2}\|b\|^2 \le 0}, \\ &G^\star\colon q \mapsto \int_X \int_0^1 J(q)\, \mathrm{d}x\, \mathrm{d}\tau \end{aligned} \tag{4.20}$$

(we assume above that $\mu_{0,1}$ has compact support, to avoid discussion of boundary conditions (4.16)). With this particular choice, the primal–dual optimality conditions (3.20) take the form

$$\begin{aligned} \partial_\tau \rho_\tau^* + \operatorname{div}_x(\rho_\tau^* D_x \phi^*) = 0, \quad \rho^*_{\tau=0,1} = \mu_{0,1}, \\ \partial_\tau \phi_\tau^* + \frac{1}{2}\|D_x \phi_\tau^*\|^2 = 0, \quad \rho_\tau^* \text{ a.e. on } [0,1] \times X. \end{aligned} \tag{4.21}$$

We have taken some liberties with the definition of the spaces above, but the general idea that this is a duality result between continuous functions and measures is valid. The system (4.21) must be understood in a weak sense. For a rigorous treatment see Santambrogio (2015, §6.1).

The concave dual problem $\sup_{\sigma \in E} -F(-\sigma) - G(A\sigma)$ is[34]

$$(\phi_0^*, \phi_1^*) := \operatorname*{arg\,inf}_{\{\phi\colon \partial_\tau \phi_\tau + \|D_x \phi_\tau\|^2/2 \le 0\}} \langle \phi_0, \mu_0 \rangle_{X_0} + \langle \phi_1, \mu_1 \rangle_{X_1}. \tag{4.22}$$

It is worth mentioning that the semi-dual formulation (*e.g.* Section 2.1 and equation (2.14)) can be recovered by substituting the explicit Lax–Oleinik formula solution of the Hamilton–Jacobi equation[35] (4.21) $\phi_1(x_1) = \inf_x \phi_0(x) + \frac{1}{2}\|x_1 - x\|^2$ in (4.22). The Lax–Oleinik formula also holds backwards, and setting

$$(u_0, u_1) = \left(\frac{1}{2}\|\mathrm{Id}_{X_0}\|^2 - \phi_0, \frac{1}{2}\|\mathrm{Id}_{X_1}\|^2 - \phi_1 \right) \tag{4.23}$$

[34] Use $\inf(\cdot) = -\sup -(\cdot)$.

[35] A classical result in the calculus of variations: $\phi_t(x) = \inf_{x_0} \phi_0(x_0) + t\, H^\star(t^{-1}(x - x_0))$ is the solution of the Hamilton–Jacobi equation $\partial_\tau \phi_\tau + H(D_x \phi_\tau) = 0$ if H is convex and superlinear.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

we recover the maximization problem (2.1) and c-concave duality (2.5).

The non-smooth convex optimization problem (4.15) is amenable to first-order optimization methods and in particular a technique called *proximal splitting*. A notable advantage of these methods is to preserve the positivity of the density ρ during the optimization and therefore its stability; this is detailed in Benamou and Carlier (2015). The discretization on regular grids is discussed in Papadakis, Peyré and Oudet (2014). For rigorous Galerkin discretizations in time and space of the CFD problem, see the recent papers by Lavenant (2021), Natale and Todeschi (2020) and their references. See also Hug, Maitre and Papadakis (2020), Guittet (2003) and Andreev (2017).

The recent book by Carlier (2021, §7.4) gives a comprehensive review of these methods and their convergence. The method advocated by Benamou and Brenier (2000) is a Douglas–Rachford solver called ALG2 (in Fortin and Glowinski 1985) and belongs to this family of proximal splitting methods. Even though the convergence of such an optimization algorithm is slow, the CFD approach remained state-of-the-art for a decade, the hidden convexity and the reduced size of the problem $M\,N$ instead of N^2 giving an advantage over linear programming. At this time, I am not aware of any convincing way to apply second-order optimization methods such as the damped Newton's algorithm used for the Monge–Ampère and semi-discrete methods. The difficulty seems to be connected with the ability of the CFD formulation to handle locally vanishing Wasserstein geodesics.

Other instances of hidden convexity are given in Brenier (2020) and in particular the theory of generalized incompressible flows (Brenier 1989), a precursor to modern optimal transportation and multi-marginal optimal transportation. See the next section.

5. Variational formulations for Euler equations

5.1. Euler geodesics

The general setting in this section, in particular the space–time cylinder, is identical to Section 4 above. A flow is a (possibly infinite) collection of Lagrangian particles $\{\mathcal{X}_\tau(x_0)\}$ for $(\tau, x_0) \in [0,1] \times X_0$. It is incompressible if the 'material density is constant within a fluid parcel'. This is mathematically translated into the Eulerian property $\rho_\tau = \mathbf{1}_X$; for all $\tau \in [0,1]$, ρ_τ is the mean field density of particles at time τ when their number tends to ∞. The mass per volume is uniformly constant. As in Section 4.1, a curve of densities evolves in time via push-forwards by the Lagrangian map $x_0 \mapsto \mathcal{X}_\tau(x_0)$. The Lagrangian flow interpretation of incompressibility can be written as

$$\mathcal{X}_\tau \# \mathbf{1}_{X_0} = \mathbf{1}_{X_\tau} \quad \text{for all } \tau \in [0,1]. \tag{5.1}$$

The density is constant, but an incompressible fluid does not necessarily remain still. 'Lagrangian stillness' would be $\mathcal{X}_\tau(x_0) = x_0$ for all $(\tau, x) \in [0,1] \times X_0$ or $\mathcal{V}_\tau(x) = 0$ for all $\tau \in [0,1] \times X$. For example, let us consider a fluid with non-zero

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

initial velocity $\mathcal{V}_0(x)$. If we let the particles evolve in straight lines $x_0 + \tau\,\mathcal{V}_0(x_0)$ then they may move closer to or further away from each other, hence violating the incompressibility, or exit the domain, or both. To maintain incompressibility and remain in the domain, there needs to be a change in velocity: an acceleration or deceleration. Under this condition a moving incompressible flow may still end up with a different final particle configuration at time 1, $\mathcal{X}_1 : x_0 \to \mathcal{X}^f(x_0) \neq x_0$, but the particle trajectories are not straight. The acceleration is provided by a mean pressure field generated by the particles themselves. Under the additional assumption that the flow is a diffeomorphism for all τ, the Jacobian equation interpretation (see footnote 2) of incompressibility is $\det(D_{x_0}\mathcal{X}_\tau) = 1$. The set of such Lebesgue-preserving diffeomorphisms in X is denoted $\mathbb{S}\mathrm{Diff}(X)$. As in the optimal transportation case (Section 4.1), the Eulerian vector field $\mathcal{V}_\tau$ can be defined and the familiar $\mathrm{div}_x\,\mathcal{V}_\tau = 0$ incompressibility condition can be derived from (4.14). This tells us, at the pointwise infinitesimal level, that dilations or compressions must be balanced coordinate-wise.

Two centuries after Leonhard Euler, Arnold (1966) gave a geodesic interpretation of the Euler equations. Replacing $\mathcal{V}_\tau$ with $\dot{\mathcal{X}}_\tau$, taking into account the final configuration constraint ($\mathcal{X}^f$ is given) and also the incompressibility, modify (4.6)–(4.7) to obtain

$$EG(\mathcal{X}^f) := \inf_{(\mathcal{X}_\tau,\mathcal{V}_\tau)\in C_{EG}(\mathcal{X}^f)} \left\langle \frac{1}{2}\int_0^1 \|\dot{\mathcal{X}}_\tau\|^2\,\mathrm{d}\tau, \mathbf{1}_{X_0}\right\rangle_{X_0}, \tag{5.2}$$

where

$$\begin{aligned} C_{EG}(\mathcal{X}^f) := \big\{&(\tau,x)\in[0,1]\times X \mapsto \mathcal{V}_\tau(x)\in\mathbb{R}^d \quad \text{and}\\ &(\tau,x_0)\in[0,1]\times X_0 \mapsto \mathcal{X}_\tau(x_0)\in X\\ &\text{such that}\\ &\dot{\mathcal{X}}_\tau = \mathcal{V}_\tau(\mathcal{X}_\tau),\\ &\mathcal{X}_0 = \mathrm{Id}_{X_0},\ \mathcal{X}_1 = \mathcal{X}^f,\\ &\mathcal{X}_\tau \in \mathbb{S}\mathrm{Diff}(X) \text{ for all } \tau\in[0,1]\big\}. \end{aligned} \tag{5.3}$$

The cost $EG(X_f)$ minimizes, with respect to the $\mathcal{L}^2(X,\mathbb{R}^d)$ metric, the length of the flow between Id_{X_0} and $\mathcal{X}_f$, constrained onto the set of Lebesgue-preserving diffeomorphisms $\mathbb{S}\mathrm{Diff}(X)$. This is the definition of a geodesic for the optimal flow $\tau \to \mathcal{X}^*_\tau$. The particle acceleration must be tangent to $\mathbb{S}\mathrm{Diff}(X)$ and therefore[36] the gradient of a 'pressure' potential p^*:

$$\ddot{\mathcal{X}}^*_\tau = D_x\,p^*_\tau(\mathcal{X}^*_\tau) \quad \text{for all } \tau\in[0,1]. \tag{5.4}$$

[36] This is a classical result obtained formally by integration by parts (brackets denote the $\mathcal{L}^2$ scalar product and I have skipped discussion of the boundary conditions), that is,

$$\langle \mathcal{V}, D_x\,p\rangle = -\langle \mathrm{div}_x\,\mathcal{V}, p\rangle_{\mathcal{L}^2(X)} = 0.$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Together with constraints (5.3), this is the Lagrangian formulation of the Euler equation with initial and final time boundary conditions on the particle configuration. The equivalence with the Euler partial differential equations in $(\mathcal{V}_\tau, p_\tau)$

$$\begin{aligned} \partial_\tau \mathcal{V}_\tau + \mathcal{V}_\tau \cdot D_x(\mathcal{V}_\tau) &= D_x p_\tau, \\ \operatorname{div}_x \mathcal{V}_\tau &= 0 \end{aligned} \tag{5.5}$$

is well known and can also be recovered (as in Section 4.2) using Fenchel–Rockafellar optimality conditions. This pure Eulerian formulation, however, cannot model the Lagrangian initial–final particle boundary conditions.

The existence of global-in-time solutions to Euler equations remains a challenge. Arnold's approach offers a way to characterize solutions with the variational formulation (5.2)–(5.3). However, $\mathbb{S}$Diff is not closed in $\mathcal{L}^2(X, \mathbb{R}^d)$ and the existence of an Eulerian vector field is conditional on the flow remaining in $\mathbb{S}$Diff. As such, (5.2)–(5.3) can only be used with additional regularity assumptions on $\mathcal{X}^f$; see Ebin and Marsden (1970).

5.2. *Generalized geodesics*

The idea in Brenier (1989) is to relax the constraint $\mathcal{X}_\tau \in \mathbb{S}\text{Diff}(X)$ in (5.3). Following (5.1), replace $\mathbb{S}$Diff with $\mathbb{S} = \{\mathcal{X} \in \mathcal{L}^2(X, X),\ \mathcal{X}\#\mathbf{1}_X = \mathbf{1}_X\}$, *i.e.* still measure-preserving maps but not necessarily diffeomorphisms. Particles may cross, and the existence of an Eulerian velocity $\mathcal{V}$ is not guaranteed. The constraint $\dot{\mathcal{X}}_\tau = \mathcal{V}_\tau$ is relaxed and we optimize on $\mathcal{X}_\tau$ alone:

$$GEG(\mathcal{X}^f) := \inf_{\mathcal{X}_\tau \in C_{GEG}(\mathcal{X}^f)} \left\langle \frac{1}{2} \int_0^1 \|\dot{\mathcal{X}}_\tau\|^2 \, \mathrm{d}\tau, \mathbf{1}_{X_0} \right\rangle, \tag{5.6}$$

where

$$\begin{aligned} C_{GEG}(\mathcal{X}^f) := \big\{ &(\tau, x_0) \in [0,1] \times X_0 \mapsto \mathcal{X}_\tau(x_0) \in X \\ &\text{such that} \\ &\mathcal{X}_0 = \mathrm{Id}_{X_0},\ \mathcal{X}_1 = \mathcal{X}^f, \\ &\mathcal{X}_\tau \in \mathbb{S}(X) \text{ for all } \tau \in [0,1] \big\}. \end{aligned} \tag{5.7}$$

The set $\mathbb{S}$ is closed and is even the completion of $\mathbb{S}$Diff for $d < 3$. A numerical method based on a discretization of (5.3)–(5.7) has been proposed by Mérigot and Mirebeau (2016). It uses a finite number of particles following piecewise affine paths $\{x_{m_i}\}_{m=0,\dots,M}$, $i = [\![1, N]\!]$, as in the Lagrangian CFD optimal transportation formulation (4.8). Discrete incompressibility requires that particles, while moving, remain at every time τ_m equi-distributed in space with respect to Lebesgue measure. There are many such configurations corresponding to local minima of the non-convex semi-discrete optimal transportation functional:

$$\{x_{m_i}\}_{i=1,\dots,N} \to \mathcal{W}_2^2\left(\sum_{i=1,\dots,N} \delta_{x_{m_i}}, \mathbf{1}_X \right). \tag{5.8}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Mérigot and Mirebeau (2016) propose penalizing (5.6) with (5.8) to approximately maintain the incompressibility instead of using the hard constraint in (5.7). The final configuration is also enforced through a penalization. The now unconstrained functional to minimize is

$$\inf_{\{x_{m_i}\}_{m=0,\dots,M,i=1,\dots,N}} \frac{1}{2}\sum_{i=0}^{N}\left(\mathrm{d}\tau\sum_{m=0}^{M-1}\|x_{m+1,i}-x_{m,i}\|^2\right) + \lambda_1\sum_{i=0}^{N}\|x_{M,i}-\mathcal{X}^f(x_{0,i})\|^2 + \lambda_2\sum_{m=0}^{M-1}\mathcal{W}_2^2\left(\sum_{i=1,\dots,N}\delta_{x_{m_i}},\mathbf{1}_X\right). \tag{5.9}$$

A careful scaling of the parameters $(M, N, \lambda_1, \lambda_2)$ is needed to ensure convergence to geodesics (5.2)–(5.3) when they exist. The same technique has been adapted to the Euler–Dirichlet problem (Gallouët and Mérigot 2018).

The incompressible Beltrami flow on the unit square $\mathcal{X} = [0,1]^2$ is an analytical non-trivial global-in-time solution of Euler equations. It corresponds to the stationary velocity and pressure

$$\begin{aligned}\mathcal{V}^*(x) &= (-\cos(\pi\, a)\sin(\pi\, b), \sin(\pi\, a)\cos(\pi\, b)),\\ p^*(x) &= \frac{1}{2}(\sin(\pi\, a)^2 + \sin(\pi\, b)^2),\end{aligned} \tag{5.10}$$

where $x = (a, b)$ are the Cartesian coordinates. It can be used as a test case for Euler geodesics. Figure 5.1 is reproduced from Mérigot and Mirebeau (2016). Particles are labelled with just three different colours to provide for a general description of the flow. The first line shows the precomputed configuration $\mathcal{X}^f$ for different final times; this is the classical solution of $\dot{\mathcal{X}}^*_\tau = \mathcal{V}^*$. The following lines show the Euler geodesic solution of (5.9). For small times and as predicted by the theory, the solution remains classical. For larger times a shorter geodesic is found in the larger set $\mathbb{S}$ by allowing particles close to the centre of rotation to cross and travel directly to the final configuration. The observed 'mixing' suggests that the solution is dependent on the discretization and points to a *looser relaxation* (see the next subsection).

5.3. *Generalized incompressible flows*

So far we have only dropped the diffeomorphism constraint. To go further, and following the Kantorovich paradigm, we now allow the unit of mass at all x_0 to split and be carried by more than one path. Let us consider the set $\Omega = \mathcal{C}^0([0,1], X)$ of all possible continuous paths in the space–time cylinder. The mass carried by every path $\tau \in [0,1] \to \omega(\tau)$ will be modelled as $\pi(\omega)$, where $\pi \in \mathcal{P}(\Omega)$ now denotes probabilities on the space of curves Ω. The Eulerian density of paths $\rho_\tau \in \mathcal{P}(X)$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

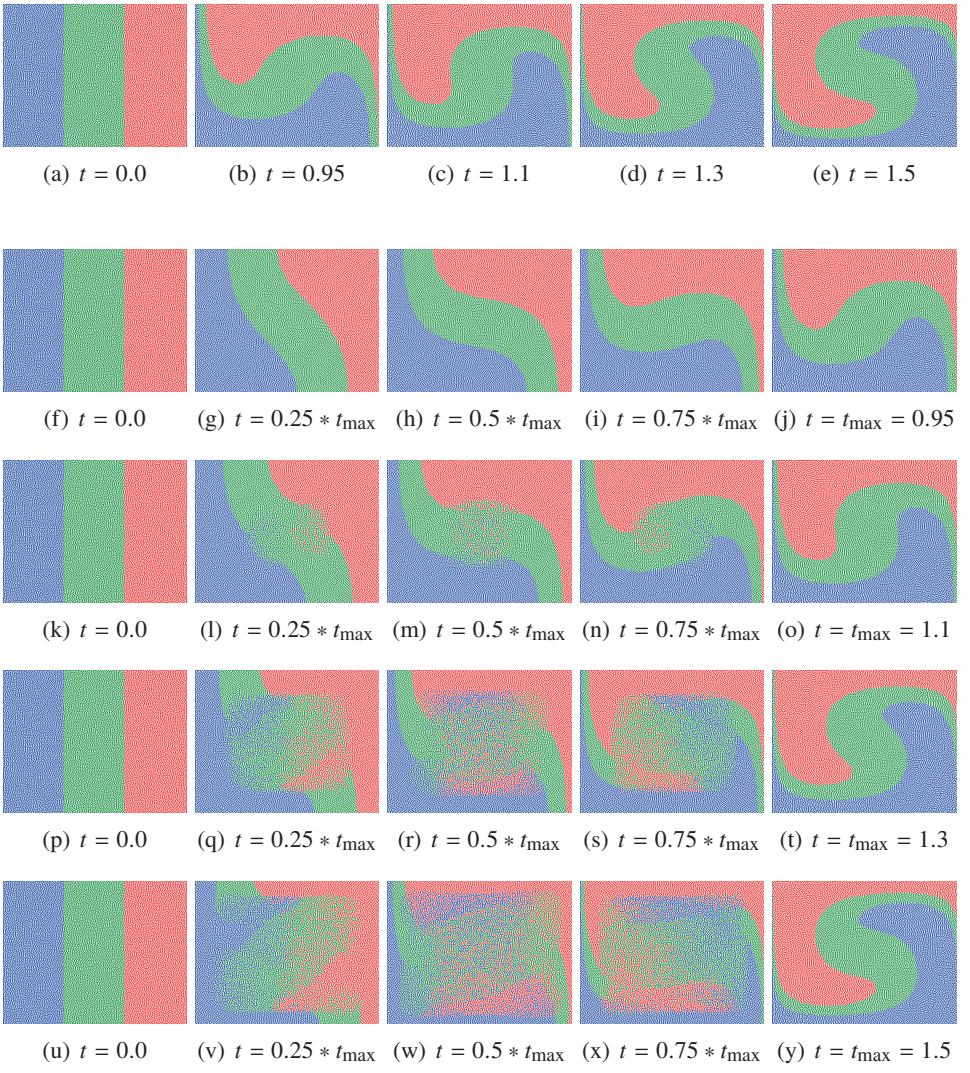


Figure 5.1. (a–e) Beltrami flow in the unit square at various time steps, a classical solution to Euler's equation. The colour of the particles depends on their initial position. (f–j, k–o, p–t, u–y) Generalized fluid flows that are reconstructed for different final times $t_{\max}$, *i.e.* using boundary conditions displayed in the first and last columns. When $\tau := t_{\max} < 1$, we recover the classical flow, while for $t_{\max} \geq 1$ the solution is no longer classical and includes mixing. Figure reproduced from Mérigot and Mirebeau (2016) with permission. Copyright © 2016 Society for Industrial and Applied Mathematics. All rights reserved.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

at each time τ can be recovered as the 'time' marginal[37]

$$\rho_\tau := e_\tau \# \pi, \quad \text{where } e_\tau : \omega \in \Omega \to \omega(\tau) \in X. \tag{5.11}$$

The transport plan between times τ_0 and τ_1 is given by[38]

$$\pi_{\tau_0 \to \tau_1} := (e_{\tau_0}, e_{\tau_1}) \# \pi. \tag{5.12}$$

If $\pi_{0\to 1}(x_0, \mathcal{X}^f(x_0)) = \mathbf{1}_{X_0}$, the unit of mass initially at x_0 ends up finally at $\mathcal{X}^f(x_0)$ and only there.[39] The final configuration Lebesgue-preserving map $\mathcal{X}_f : X_0 \mapsto X_1$ therefore corresponds to the plan

$$\pi_{0\to 1} = (\mathrm{Id}_{X_0}, \mathcal{X}^f) \# \mathbf{1}_{X_0}. \tag{5.13}$$

A generalized incompressible flow will be the solution of

$$GIF(\mathcal{X}^f) := \inf_{\pi \in C_{GIF}(\mathcal{X}^f)} \left\langle \frac{1}{2} \int_0^1 \|\dot\omega(\tau)\|^2 \, \mathrm{d}\tau, \pi \right\rangle_\Omega, \tag{5.14}$$

where

$$\begin{aligned} C_{GIF}(\mathcal{X}^f) := \big\{ & \pi \in \mathcal{P}(\Omega) : \\ & (e_0, e_1) \# \pi = (\mathrm{Id}_{X_0}, \mathcal{X}^f) \# \mathbf{1}_{X_0} \\ & e_\tau \# \pi = \mathbf{1}_{X_\tau} \text{ for all } \tau \in [0,1] \big\}. \end{aligned} \tag{5.15}$$

In line with the Kantorovich formulation of optimal transportation, we are back to a 'simple' Eulerian linear programming problem but set on a 'large' infinite-dimensional space $\mathcal{P}(\Omega)$. The last line in (5.15) is Eulerian incompressibility and the middle line the initial/final boundary conditions. Note that even though $(e_0, e_1)\#\pi$ is fully described by the map $\mathcal{X}_f$, it does not imply that the generalized flow itself concentrates on a diffeomorphism in $\mathbb{S}\mathrm{Diff}$ for all time. Mass can divide and flow from x_0 to $\mathcal{X}^f(x_0)$ along an arbitrary number of paths, possibly crossing with different velocities.

We follow the discretization in Section 4.1: let $\Omega_{N,M} = \bigotimes_{m=0}^{M} X_N$ be the discrete (in space and time) cylinder, where X_N is a fixed N-point grid and $\{\tau_m\}, m = [\![1, M]\!]$ is a time discretization. The discretization of π is a multi-marginal plan $\pi_{M,N} \in \mathcal{P}(\Omega_{N,M})$.

[37] By definition (see Section 1) and for any measurable subset $A \subseteq X$, $e_\tau \# \pi(A) = \pi(\{\omega \in \Omega,\ \omega(\tau) \in A\})$.

[38] For any pair of measurable subsets $(A_{\tau_0}, A_{\tau_1}) \in X_{\tau_0} \times X_{\tau_1}$, we have $(e_{\tau_0}, e_{\tau_1})\#\pi(A_{\tau_0}, A_{\tau_1}) = \mathrm{d}\pi(\{\omega \in \Omega,\ (\omega(\tau_0), \omega(\tau_1)) \in A_{\tau_0} \times A_{\tau_1}\})$.

[39] Remember that $\rho_1 = P_{X_1} \# \pi_{0\to 1}$ is a probability measure and therefore necessarily $\pi_{0\to 1}(x_0, x_1) = 0$ for all $x_1 \neq \mathcal{X}^f(x_0)$. We have already discussed (see (2.11)) how a transport map can be encoded into a plan.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Resorting once more to a piecewise affine approximation of paths (see (4.1)–(4.2)) and using an N-point grid X_N for the space discretization, we end up with

$$\inf_{\pi_{M,N}\in\Pi_{M,N}(\mathcal{X}^f)} \langle c_M, \pi_{M,N}\rangle_{\bigotimes_{m=0}^{M} X_N}, \tag{5.16}$$

where[40]

$$\begin{aligned}\Pi_{M,N}(\mathcal{X}^f) := \big\{&\pi_{M,N}\in\mathcal{P}(\Omega_{M,N}):\\ &P_{X_0\times X_1}\#\pi_{M,N} = (\mathrm{Id}_{X_0}, \mathcal{X}^f)\#\mathbf{1}_{X_N},\\ &P_{X_m}\#\pi_{M,N} = \mathbf{1}_{X_N} \text{ for all } m = 0\ldots,M\big\}.\end{aligned} \tag{5.17}$$

This problem minimizes the same action as the CFD optimal transportation problem (see (4.1)), but the incompressibility constraint is active on all marginals. We will see another example of *multi-marginal constraints* in Section 6.2. It is convenient to eliminate the 2-marginal 'plan' constraint (second line in (5.17)). It can be relaxed by adding a penalization to the displacement cost ($\lambda > 0$ and 'large'):

$$c_{M,\mathcal{X}^f}(x_0,, x_1, \ldots, x_M) = \frac{1}{2M}\sum_{m=0}^{M-1} \|x_{m+1} - x_m\|^2 + \lambda\,\|x_M - \mathcal{X}^f(x_0)\|^2. \tag{5.18}$$

The problem becomes

$$\inf_{\pi_{M,N}\in\Pi_{M,N}} \langle c_{M,\mathcal{X}^f}, \pi_{M,N}\rangle_{\bigotimes_{m=0}^{M} X_N}, \tag{5.19}$$

where

$$\begin{aligned}\Pi_{M,N} = \big\{&\pi_{M,N}\in\mathcal{P}(\Omega_{M,N}):\\ &P_{X_m}\#\pi_{M,N} = \mathbf{1}_{X_N} \text{ for all } m = 0,\ldots,M\big\}\end{aligned} \tag{5.20}$$

no longer depends on $\mathcal{X}^f$. From the numerical point of view, this problem suffers from the same difficulties as the discrete Kantorovich problem (1.4) on an even larger scale (Section 1.3).

5.4. *Entropic regularization for generalized incompressible flows*

Following Benamou, Carlier and Nenna (2019*a*), we propose to use the entropic regularization (as presented in Section 3.3) and adapt the Sinkhorn algorithm (3.12) to solve (5.19)–(5.20). Dropping some of the indices for clarity,

$$\inf_{\pi_\epsilon\in\Pi_{M,N}} \langle c_{M,\mathcal{X}^f}, \pi_\epsilon\rangle_{\bigotimes_{m=0}^{M} X_N} + \epsilon\, KL(\pi_\epsilon|\mathbf{1}_{\Omega_{M,N}}) \tag{5.21}$$

[40] $P_{X_m}\#\pi_{M,N}$ is the discrete analogue of $e_{\tau_m}\#\pi$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

can be rewritten as

$$\inf_{\pi_\epsilon \in \mathbb{R}^{M\times N}} \sum_{m=0}^{M} \chi_{P_{X_m}\#\pi_\epsilon = \mathbf{1}_{X_N}} + \epsilon\ KL(\pi_\epsilon | \pi_\epsilon^0 \mathbf{1}_{\Omega_{M,N}}), \tag{5.22}$$

where (γ_ϵ is a Gaussian kernel: see (3.15); note that we have normalized the constant to 1) and

$$\pi_\epsilon^0(x_0, x_1, \ldots, x_M) := \gamma_{\epsilon/\lambda}(x_M - \mathcal{X}^f(x_0)) \prod_{m=0}^{M-1} \gamma_{\epsilon/\mathrm{d}\tau}(x_{m+1} - x_m) \tag{5.23}$$

for $(x_0, x_1, \ldots, x_M) \in \Omega_{N,M}$ and $\mathrm{d}\tau = 1/(2M)$ a time step. Note that this is now a discrete optimization problem over $M \times N$ real tensors.

The generalization of Fenchel–Rockafellar duality (3.19) (see also (3.21) on the larger multi-marginal space) is relatively straightforward:

$$\begin{aligned}
&A\colon \sigma := (u_0, u_1, \ldots, u_M) \in (\mathbb{R}^N)^M \mapsto \sum_{m=0}^{M} u_m \in \mathbb{R}^{M\times N},\\
&A'\colon q := \pi_\epsilon \in \mathbb{R}^{N\times M} \mapsto (P_{X_0}\#\pi_\epsilon, P_{X_1}\#\pi_\epsilon, \ldots, P_{X_M}\#\pi_\epsilon) \in (\mathbb{R}^N)^M,\\
&F\colon (u_0, u_1, \ldots, u_M) \in (\mathbb{R}^N)^M \mapsto \sum_{m=0}^{M} \langle u_m, \mathbf{1}_{X_N}\rangle,\\
&F^\star\colon (\rho_0, \rho_1, \ldots, \rho_M) \in (\mathbb{R}^N)^M \mapsto \sum_{m=0}^{M} \chi_{\rho_m = \mathbf{1}_{X_N}},\\
&G\colon u \in \mathbb{R}^{M\times N} \mapsto G(u) = \epsilon \langle \mathrm{e}^{u/\epsilon}, \pi_\epsilon^0 \mathbf{1}_{\Omega_{M,N}}\rangle_{\bigotimes_{m=0}^{M} X_m},\\
&G^\star\colon q \in \mathbb{R}^{N\times M} \mapsto G^\star(q) = \epsilon\ KL(\pi_\epsilon | \pi_\epsilon^0 \mathbf{1}_{\Omega_{M,N}}).
\end{aligned} \tag{5.24}$$

The primal–dual optimality conditions (3.20) take the form

$$\pi_\epsilon^* = \mathrm{e}^{(\sum_{m=0}^{M} u_m^*)/\epsilon} \pi_\epsilon^0\, \mathbf{1}_{\Omega_{M,N}}, \quad P_{X_m}\#\pi_\epsilon^* = \mathbf{1}_X \quad \text{for all } m. \tag{5.25}$$

Likewise the coordinate ascent Sinkhorn algorithm (3.23) generalizes to

$$\begin{aligned}
u_m^k = \arg\sup_{u_m} &-F(-\{u_0^k, \ldots, u_{m-1}^k, u_m, u_m^{k-1}, \ldots, u_M^{k-1}\})\\
&- G(A\{u_0^k, \ldots, u_{m-1}^k, u_m, u_m^{k-1}, \ldots, u_M^{k-1}\})
\end{aligned} \tag{5.26}$$

for all $m = 0, \ldots, M$ at each k iteration. As for the 2-marginal Sinkhorn, these are strictly concave unconstrained problems, leading to the nonlinear set of equations:

$$\begin{aligned}
&\partial_{u_m} F(-\{u_0^k, \ldots, u_{m-1}^k, u_m, u_m^{k-1}, \ldots, u_M^{k-1}\})\\
&\quad = \partial_{u_m} G(A\{u_0^k, \ldots, u_{m-1}^k, u_m, u_m^{k-1}, \ldots, u_M^{k-1}\}).
\end{aligned} \tag{5.27}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The left-hand side $\partial_{u_m} F = \mathbf{1}_{X_N}$ is given. The right-hand side simplifies to

$$\begin{aligned} u_m \to \partial_{u_m} G(A\{u_0, \dots, u_{m-1}, u_m, u_m, \dots, u_M\}) \\ = \langle \mathrm{e}^{(\sum_{m'=0}^{M} u'_m)/\epsilon}, \pi_\epsilon^0 \mathbf{1}_{\Omega_{M,N}} \rangle_{X_0 \times \cdots \times X_{m-1} \times X_{m+1} \times \cdots X_M} \\ = \mathrm{e}^{u_m/\epsilon} \langle \mathrm{e}^{(\sum_{m' \neq m} u_m)/\epsilon}, \pi_\epsilon^0 \mathbf{1}_{\Omega_{M,N}} \rangle_{X_0 \times \cdots \times X_{m-1} \times X_{m+1} \times \cdots \times X_M}. \end{aligned} \tag{5.28}$$

As in the classic Sinkhorn algorithm, the equation (5.27) is explicit in u_m. The numerical limits of this algorithm ultimately rest on the computational cost of the sum in (5.28). The product structure of (5.23) involves only two successive (in a circular way) marginal kernels, and the chain of summations in the expression above is 'broken' at m: $m+1 \to \cdots \to x_M \to x_0 \to \cdots \to x_{m-1}$. Computing (5.28) consists in M independent matrix vector products

$$\sum_j \gamma_{\epsilon/\mathrm{d}\tau}(x_{p,j} - x_{p+1,i})\, \mathrm{e}^{u_{p,j}/\epsilon},$$

where p runs over the chain and the indices (p, j) and $(p+1, i)$ over the space discretization at times p and $p+1$. We finally find that computing (5.28) for all $\{m\}$ requires $O(M^2 N^2)$ operations which can be reduced to $O(M^2 N^{1+1/d})$ using the separability of Euclidean quadratic cost (1.1). See Benamou *et al.* (2019*a*) for more on the convergence of the entropic regularization as $\epsilon \to 0$. For the convergence rate of multi-marginal Sinkhorn, see Di Marino and Gerolin (2020).

We applied this method to the numerical resolution of the Beltrami flow (5.10) with $M = 16$ and $N = 64^2$. As in Section 5.2, we use three colours to label the mass initially. It is split into three non-overlapping subdomains of X_0 called R, G and B (red, green and blue). This is the first line in Figures 5.2 and 5.3. The first column shows the exact Beltrami flow for increasing times. The last three columns show where this mass has been sent by plotting[41]

$$x_m \to P_{X_0 \times X_m} \# \pi_\epsilon(R/G/B, x_m), \tag{5.29}$$

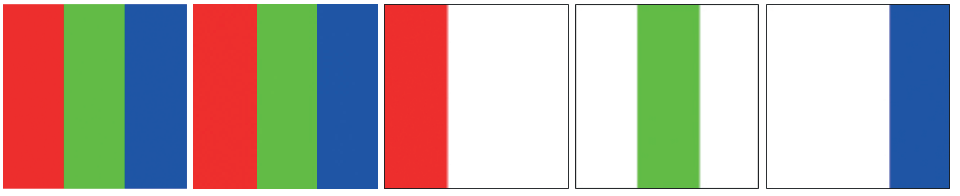
using a variable transparency depending[42] on (5.29). The second column adds the three *RGB* channels.

Figures 5.2 and 5.3 correspond to two different final configurations $\mathcal{X}^f$ given by the exact Beltrami flow at (small) time $T = 0.9$ and (large) time $T = \pi$. The diffusion induced by entropic regularization ($\epsilon = 1E-04$) smoothing can be observed on the short time geodesic (Figure 5.2). As in Figure 5.1, mass takes a shortest path in the large time simulation (Figure 5.3). There is also mass splitting and mixing, suggesting *GIF* is indeed a looser relaxation than generalized geodesics (Section 5.2).

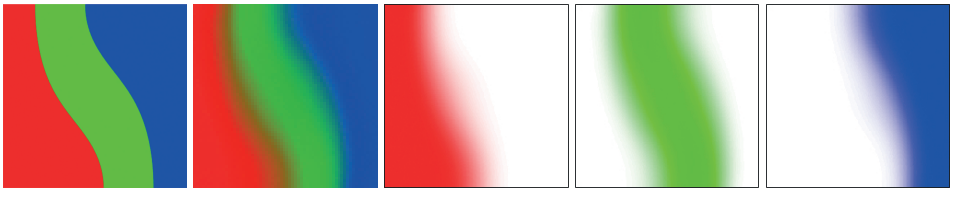
[41] π_ϵ is given by (5.25), and (5.29) is the measure of the mass transported from the grid points in regions $R/G/B$, respectively, to the grid point x_m at the discrete time τ_m.

[42] No mass has perfect transparency: one sees the white background, full mass 1 is the true opaque colour (blue, red, green), and if the mass is in between it is partially transparent.

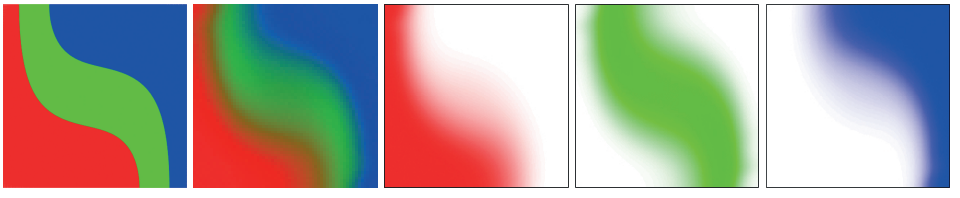
https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



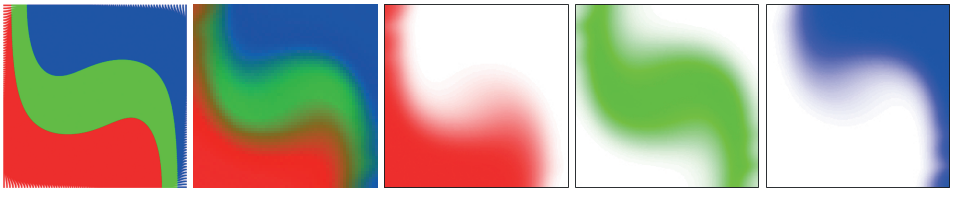
(a) $t = 0$



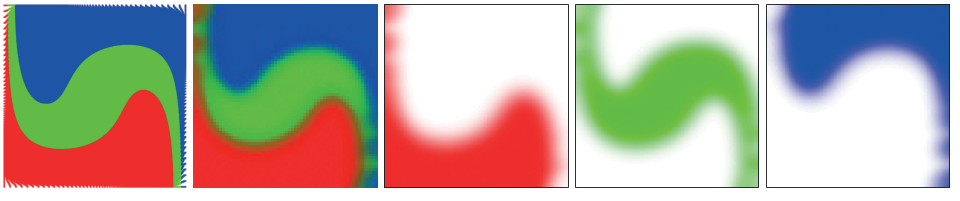
(b) $t = \frac{T}{4}$



(c) $t = \frac{T}{2}$



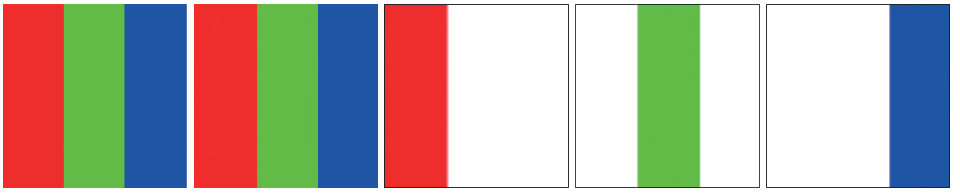
(d) $t = \frac{3T}{4}$



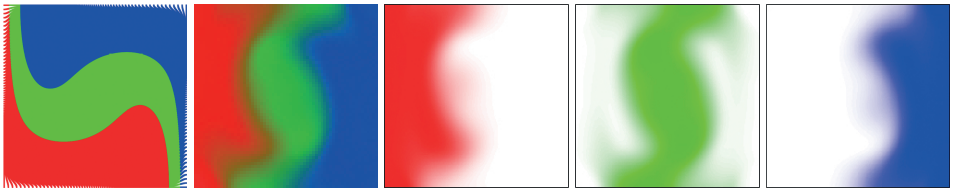
(e) $t = T$

Figure 5.2. Final time, $T = 0.9$. Columns, classical colour tracking of the Lagrangian solution with no mixing in the first column, $P_R + P_G + P_B$ in the second column and $P_R/P_G/P_B$ in the remaining three columns. Rows (a–e), time evolution. The final Lagrangian configuration at the bottom left is the final datum X_T in $\pi_{0,T} = (\mathrm{Id}, X_T)_{\#}\mathbf{1}$. Figure reproduced from Benamou *et al.* (2019*a*) with permission. Copyright © 2019 Springer.

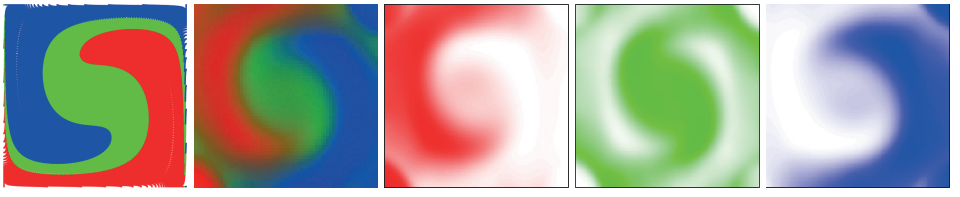
https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



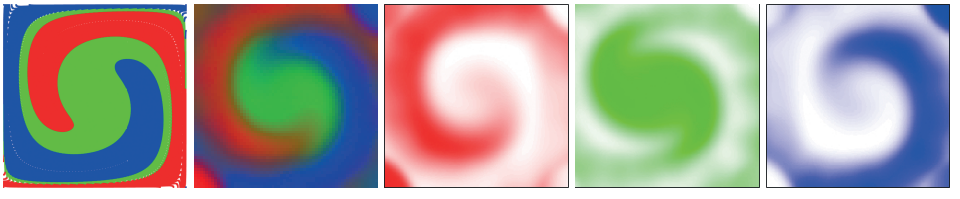
(a) $t = 0$



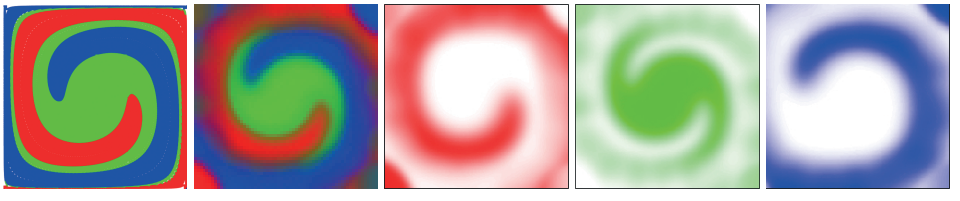
(b) $t = \frac{T}{4}$



(c) $t = \frac{T}{2}$



(d) $t = \frac{3T}{4}$



(e) $t = T$

Figure 5.3. Final time, $T = \pi$. Columns, classical colour tracking of the Lagrangian solution with no mixing in the first column, $P_R + P_G + P_B$ in the second column and $P_R/P_G/P_B$ in the remaining three columns. Rows (a–e), time evolution. The final Lagrangian configuration at the bottom left is the final datum X_T in $\pi_{0,T} = (\mathrm{Id}, X_T)_{\#}\mathbf{1}$. Figure reproduced from Benamou *et al.* (2019*a*) with permission. Copyright © 2019 Springer.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

6. The Schrödinger problem and transport by diffusion

6.1. The Schrödinger problem

At convergence, Sinkhorn equations (3.13) are

$$u_0^* = LSE^{\epsilon}_{\mu_1,X_1}(u_1^*), \quad u_1^* = LSE^{\epsilon}_{\mu_0,X_0}(u_0^*). \tag{6.1}$$

This is a direct consequence of the optimality conditions (3.22). As mentioned in Section 3.2, the entropic regularization machinery holds in the continuous setting and this will make the exposition easier here. We will also assume that $\mu_{0,1}$ are smooth densities over $\mathbb{R}^d$ with finite second moments. The change of variable

$$(f_0^*, g_1^*) = (2\pi\epsilon)^{d/2}\,(\mathrm{e}^{u_0^*/\epsilon}\,\mu_0, \mathrm{e}^{u_1^*/\epsilon}\,\mu_1) \tag{6.2}$$

in (6.1) gives (after a few lines)

$$f_0^*\, g_0 = \mu_0 \quad \text{and} \quad f_1\, g_1^* = \mu_1, \tag{6.3}$$

where (f_1, g_0) are obtained using forward and backward time integration of a heat flow, respectively. The heat flow is expressed using the Laplace operator $\Delta = \mathrm{div}_x(D_x)$ (in our previous notation) as the generator of the semi-group,[43] that is,

$$g_\tau = \mathrm{e}^{-(\epsilon\tau/2)\Delta}\, g_1^*, \quad \tau\colon 1 \to 0 \quad \text{and} \quad f_\tau = \mathrm{e}^{(\epsilon\tau/2)\Delta} f_0^*, \quad \tau\colon 0 \to 1. \tag{6.4}$$

This reformulation relies on the interpretation of (3.15) as a heat kernel. It implies that the problem is now set on $X = \mathbb{R}^d$ (the periodic torus with the corresponding periodic cost c is also possible). The heat flows (6.4) have a *Lagrangian stochastic differential equation* interpretation, (f_τ, g_τ) are the probability laws[44] of two independent random processes $\mathcal{X}_\tau^{\pm}$ following the standard Brownian motions $\mathcal{B}_\tau$:

$$\mathrm{d}\mathcal{X}_\tau^{\pm} = \pm\sqrt{\epsilon}\,\mathrm{d}\mathcal{B}_\tau, \quad \mathcal{X}_0^{+} \sim f_0^* \quad \text{and} \quad \mathcal{X}_1^{-} \sim g_1^*. \tag{6.5}$$

The Sinkhorn algorithm (3.12) may be interpreted as Picard-type iterations to determine the initial and final laws (f_0^*, g_1^*). We explain below how the density of transport in time is retrieved through the 'interference' product between the forward and backward probability laws (see also Figure 6.1). A similar approach was proposed by Guéant (2012) in the variational mean field games context (next section).

We know that (f_0^*, g_1^*) are the solutions of the dual problem (3.11). The Schrödinger problem is a probabilistic interpretation of the primal problem associated to (3.11) under the change of variable (6.2). It is a reformulation (again a few lines of calculations) of (3.16):

$$\pi_\epsilon^* := \underset{\pi_\epsilon \in \Pi(\mu_0,\mu_1)}{\arg\inf} \; \epsilon\, KL(\pi_\epsilon | \pi_\epsilon^0\, \mathbf{1}_{X_0\times X_1}), \tag{6.6}$$

[43] $h_\tau(x) = (\mathrm{e}^{(\epsilon\tau/2)\Delta} h_0)(x) := \int \gamma_{(\epsilon\tau/2)}(x - x')\, h_0(x')\,\mathrm{d}x'$.

[44] $\mathcal{X}_\tau \sim \rho_\tau$ is to be understood as 'the probability law of $\mathcal{X}_\tau$ is ρ_τ': $P(\mathcal{X}_\tau \in A) = \rho_\tau(A)$ for any measurable $A \subset X$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

where the transport part of the cost is embedded in

$$\pi^0_\epsilon(x_0, x_1) := \gamma_\epsilon(x_1 - x_0). \tag{6.7}$$

As in Section 5.3, let us use the time flow in τ. Let $R_\epsilon \in \mathcal{P}(\Omega)$ denote the Wiener measure.[45] Then π^0_ϵ is the density of $R^{0\to 1}_\epsilon := (e_0, e_1)\#R_\epsilon$, the $0 \to 1$ transition probability measure associated to R_ϵ. If, for example, $\mu_{0,1} = \mathbf{1}_X$, *i.e.* Lebesgue measure, then the solution to (6.6) is simply $\pi^*_\epsilon = R^{0\to 1}_\epsilon$. More generally, $\pi^*_\epsilon = (e_0, e_1)\#Q^*_\epsilon$, where Q^*_ϵ solves a *dynamic* version of (6.6):

$$Q^*_\epsilon := \underset{Q_\epsilon \in \mathcal{P}(\Omega),\ (e_0,e_1)\#Q_\epsilon \in \Pi(\mu_0,\mu_1)}{\arg\inf} \epsilon\ KL(Q_\epsilon | R_\epsilon). \tag{6.8}$$

This last step requires a careful analysis, which can be found in Léonard (2014). The *Lagrangian stochastic differential equation* interpretation is interesting. The minimizer of (6.8) is Markovian and the law of a diffusion process with constrained initial and final time law:

$$\mathrm{d}\mathcal{X}^*_\tau = -D_x\phi^*(\mathcal{X}^*_\tau)\,\mathrm{d}\tau + \sqrt{\epsilon}\,\mathrm{d}\mathcal{B}_\tau, \quad \mathcal{X}^*_{0,1} \sim \mu_{0,1}. \tag{6.9}$$

The law of $\mathcal{X}^*_\tau \sim e_\tau\#Q^*_\epsilon = \rho^*_\tau$ is the entropic optimal transportation interpolation. The drift $D_x\phi^*$ and volatility can be deduced from the PDE interpretation (6.12) below. As in the case of the generalized incompressible flows, the mass at x_0 can be split along C^0-path solutions of (6.9). The measure Q^*_ϵ is closest to R^ϵ in the sense of KL entropy (known in this context as the Boltzmann–Shannon entropy) and constrained to have $\mu_{0,1}$ as the initial and final densities in time, respectively. The value function can also be interpreted (using Sanov's theorem: see Léonard 2014) as (minus the log of) the event probability that, given $\mu_{0,1}$, the Wiener measure satisfies $(e_0, e_1)\#R_\epsilon \in \Pi(\mu_0, \mu_1)$.

The Schrödinger interpretation of optimal transportation is somewhat technical for numerical analysts not well versed in probability theory. There is a more formal PDE interpretation that will, as in the non-entropic case, lead to a CFD-like model and link entropic optimal transportation with diffusion.

The optimality conditions (6.4) are rewritten as

$$(-\partial_\tau - \epsilon\,\Delta)\,g_\tau = 0,\ \ g_1 = g^*_1 \quad \text{and} \quad (\partial_\tau - \epsilon\,\Delta)\,f_\tau = 0,\ \ f_0 = f^*_0. \tag{6.10}$$

We apply the 'Hopf–Cole-type' transformation

$$\{f_\tau, g_\tau\} \to \left\{\rho_\tau = f_\tau\, g_\tau, g_\tau = \exp\left(-\frac{1}{2\,\epsilon}\,\phi_\tau\right)\right\} \tag{6.11}$$

[45] R_ϵ is a measure on $\Omega = \mathcal{C}([0,1], X)$,

$$R \sim \frac{1}{(2\pi\epsilon)^{d/2}} \int \mathrm{Law}(x + \sqrt{\epsilon}\,\mathcal{B})\,\mathrm{d}x,$$

where $\mathcal{B}$ is the standard Brownian motion starting at 0, *i.e.* the Markov process whose generator is the operator above $\epsilon\Delta$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

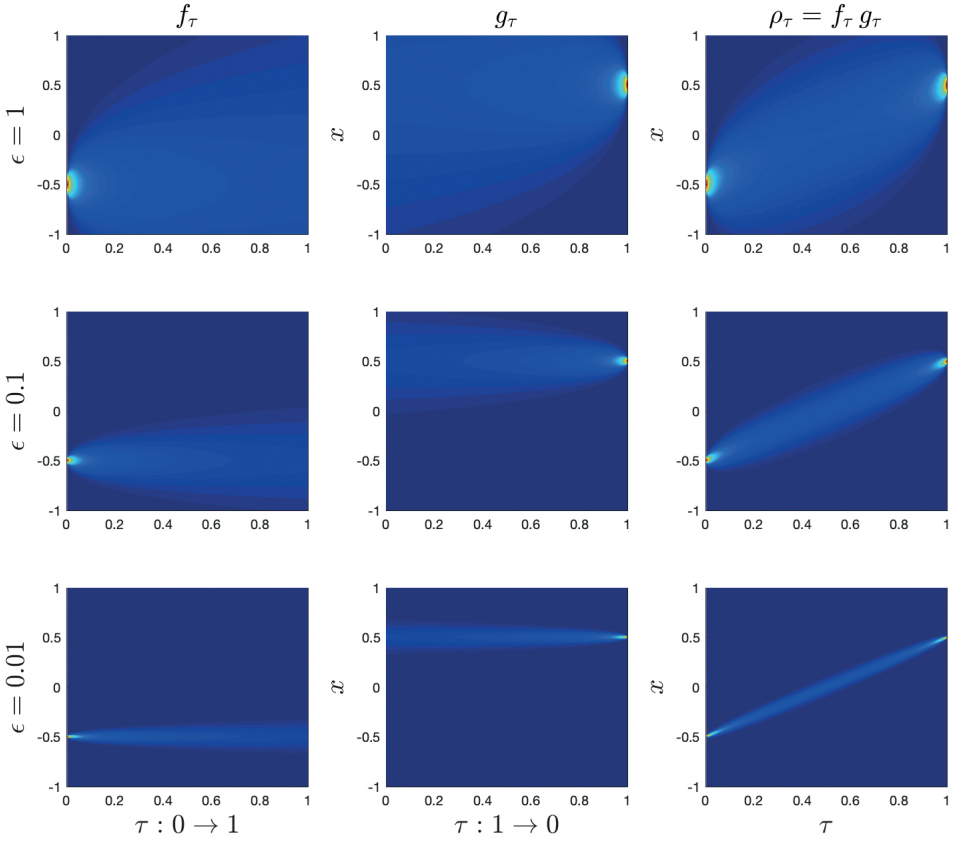


Figure 6.1. This is an illustration of the 'interference product' in (6.11) for two Dirac masses $\mu_{0,1}$ (one at the initial time and the other at the final time) and a diminishing entropy/diffusion parameter ϵ. It also explains the numerical stability limit of the entropic approach. The normalization is achieved by the potentials; when the heat kernel is numerically 0 instead of extremely small, it is no longer possible.

to (6.10) and obtain the *Eulerian version* of (6.9):

$$\begin{aligned} (\partial_\tau - \epsilon\,\Delta)\,\rho_\tau + \mathrm{div}_x(\rho_\tau\, D_x\phi_\tau) &= 0, \\ (-\partial_\tau - \epsilon\,\Delta)\,\phi_\tau + \frac{1}{2}\|D_x\phi_\tau\|^2 &= 0. \end{aligned} \tag{6.12}$$

The second equation is a Hamilton–Jacobi–Bellman equation and the first equation a Fokker–Planck equation for the density ρ_τ. The optimality condition (6.3) becomes the familiar marginal constraints

$$\rho_{0,1} = \mu_{0,1}. \tag{6.13}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The Fenchel–Rockafellar formalism (4.20) can be used simply by replacing A with $A^\epsilon : (\phi, \phi_0, \phi_1) \mapsto ((-\partial_\tau - \epsilon\,\Delta)\,\phi, D_x\phi)$, the adjoint becoming $(A^\epsilon)' : (\rho, m) \mapsto -((\partial_\tau - \epsilon\,\Delta)\,\rho_\tau + D_x\,m, \rho_0, -\rho_1)$. The solution (ρ^*, ϕ^*) of (6.12)–(6.13) is the (unique) minimizer of

$$\inf_{(\rho,\mathcal{V})\in FP(\rho_0,\rho_1)} \int_0^1 \frac{1}{2}\langle \|\mathcal{V}_\tau(x)\|^2, \rho_\tau\rangle_X \,\mathrm{d}\tau, \tag{6.14}$$

where

$$\begin{aligned} FP(\rho_0,\rho_1) := \big\{&(\tau, x) \in [0,1]\times X \mapsto (\rho_\tau(x), \mathcal{V}_\tau(x)) \in \mathbb{R}^+\times\mathbb{R}^d : \\ &(\partial_\tau - \epsilon\,\Delta)\,\rho_\tau + \mathrm{div}_x(\rho_\tau\,\mathcal{V}_\tau) = 0 \text{ and } \rho_{\tau=0,1} = \mu_{0,1}\big\}. \end{aligned} \tag{6.15}$$

It is similar to the CFD formulation (4.15)–(6.15) but, as expected, with the entropic regularization the interpretation in terms of transport maps is completely lost, and the mass is transported along the stochastic paths (6.9).

6.2. *Variational mean field games*

Variational mean field games are a specific subclass of mean field games, introduced in Lasry and Lions (2007), which can be tackled with the numerical and theoretical tools used in dynamic optimal transportation. For a general introduction to mean field games and their link with Nash equilibria of multi-agent systems, see Achdou *et al.* (2020). We will restrict ourselves to the simplest *variational* mean field games generalization of (6.14)–(6.15):

$$\inf_{(\rho,\mathcal{V})\in FP(\rho_0)} \int_0^1 \frac{1}{2}\langle \|\mathcal{V}_\tau(x)\|^2, \rho_\tau\rangle_X \,\mathrm{d}\tau + \int_0^1 H(\rho_\tau)\,\mathrm{d}\tau + H_1(\rho_1), \tag{6.16}$$

where H and H_1 are convex function in $\mathcal{C}(P(X), \mathbb{R})$ and

$$\begin{aligned} FP(\rho_0) := \big\{&(\tau, x) \in [0,1]\times X \mapsto (\rho_\tau(x), \mathcal{V}_\tau(x)) \in \mathbb{R}^+\times\mathbb{R}^d : \\ &(\partial_\tau - \epsilon\,\Delta)\,\rho_\tau + \mathrm{div}_x(\rho_\tau\,\mathcal{V}_\tau) = 0 \text{ and } \rho_{\tau=0} = \mu_0\big\}. \end{aligned} \tag{6.17}$$

Note that ρ is now completely determined by the velocity $\mathcal{V}_\tau$ though the resolution of the initial value problem in (6.17). The final density ρ_1 appears in the cost $H_1(\rho_1)$ in (6.16). This formulation belongs to the well-known class of *optimal control of a system governed by partial differential equations* Lions (1971), here a Fokker–Planck equation representing (in the mean field games paradigm) the density of players whose trajectories are subject to white noise and optimized to achieve global minimum cost. Individual trajectories are not observable, only the mean field density. This formulation also holds when $\epsilon = 0$, under the name of 'deterministic mean field games'. The Fokker–Planck equation becomes the continuity equation of the CFD formulation.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

With a slight variation of (4.20),

$$\delta_\tau \to (\partial_\tau - \epsilon\,\Delta) \text{ in } A',$$
$$F^\star \to \{(a,\rho_0) \mapsto \iota_{\{a=0,\rho_0=\mu_0\}}\},$$
$$G^\star \to \left\{ q := (\rho, P) \mapsto \int_X \int_0^1 J(q)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^1 H(\rho_\tau)\,\mathrm{d}\tau + H_1(\rho_1) \right\},$$

we can once more apply Fenchel–Rockafellar duality. The optimality system (4.21) becomes

$$\begin{aligned} &(\partial_\tau - \epsilon\,\Delta)\rho^*_\tau + \mathrm{div}_x(\rho^*_\tau\, D_x\phi^*) = 0, \quad \rho^*_0 = \mu_0, \\ &(-\partial_\tau - \epsilon\,\Delta)\,\phi^*_\tau + \frac{1}{2}\|D_x\phi^*_\tau\|^2 = \frac{\partial}{\partial_\rho} H(\rho_\tau), \\ &\phi^*_1 = \frac{\partial}{\partial_\rho} H_1(\rho_1), \quad \rho^*_\tau \text{ a.e. on } [0,1]\times X. \end{aligned} \tag{6.18}$$

This is the general form of our mean field games. Picking $H := 0$ and the characteristic function $H_1 := \iota_{\{\rho_1=\mu_1\}}$, for example, brings us back to the Schrödinger problem (Section 6.1).

As in the entropic treatment of the generalized incompressible flows (Section 5.4), we are going to relax (6.16)–(6.17) to measures on curves. By analogy with the PDE interpretation (4.15)–(6.15) of the Schrödinger problem in its dynamic form (6.8), (6.16)–(6.17) has its own dynamic Schrödinger version:

$$\inf_{Q_\epsilon \in \mathcal{P}(\Omega),\ e_0\#Q_\epsilon = \mu_0} \epsilon\ KL(Q_\epsilon | R_\epsilon) + \int_0^1 H(\rho_\tau)\,\mathrm{d}\tau + H_1(\rho_1). \tag{6.19}$$

Going back to the discretization-in-space-and-time technique developed in Section 5.4 and re-using all the notation of Section 5, we get from (5.22)

$$\inf_{\pi_\epsilon \in \mathbb{R}^{M\times N}} \epsilon\ KL(\pi_\epsilon | \pi^0_\epsilon \mathbf{1}_{\Omega_{M,N}}) + \sum_{m=0}^{M} H_m(P_{X_m}\#\pi_\epsilon). \tag{6.20}$$

This is a slightly generalized version as H may now depend on time and H_M is the new notation for H_1. The $\{H_m\}$ are assumed to be convex. The initial condition can be enforced with $H_0(\rho) := \iota_{\{\rho=\mu_0\}}$. The gamma-convergence to variational mean field games (6.16)–(6.17) is established in Benamou, Carlier, Di Marino and Nenna (2019*b*). We again apply the Fenchel–Rockafellar duality (3.19):

$$A\colon \sigma := (u_0, u_1, \ldots, u_M) \in (\mathbb{R}^N)^M \mapsto \sum_{m=0}^{M} u_m \in \mathbb{R}^{M\times N},$$
$$A'\colon q := \pi_\epsilon \in \mathbb{R}^{N\times M} \mapsto (P_{X_0}\#\pi_\epsilon, P_{X_1}\#\pi_\epsilon, \ldots, P_{X_M}\#\pi_\epsilon) \in (\mathbb{R}^N)^M,$$
$$F\colon (u_0, u_1, \ldots, u_M) \in (\mathbb{R}^N)^M \mapsto \sum_{m=0}^{M} H^\star_m(u_m),$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

$$F^\star\colon (\rho_0,\rho_1,\dots,\rho_M)\in(\mathbb{R}^N)^M \mapsto \sum_{m=0}^{M} H_m(\rho_m),$$
$$G\colon u\in\mathbb{R}^{M\times N}\mapsto G(u)=\epsilon\langle \mathrm{e}^{u/\epsilon},\pi_\epsilon^0\mathbf{1}_{\Omega_{M,N}}\rangle_{\bigotimes_{m=0}^M X_m},$$
$$G^\star\colon q\in\mathbb{R}^{N\times M}\mapsto G^\star(q)=\epsilon\ KL(\pi_\epsilon|\pi_\epsilon^0\mathbf{1}_{\Omega_{M,N}}). \tag{6.21}$$

We recognize (5.24) with an abstract version of F. The primal–dual optimality condition (5.25) and Sinkhorn algorithm (5.27) are unchanged but for an important point. The chain of dependence of marginals in the cost $x_0\to x_1\to\cdots\to x_M$ is no longer circular. If one is willing to pay the memory cost, the operation cost of one sweep of the Sinkhorn iteration can be taken down to $O(M\,N^2)$, and to $O(M\,N^{1+1/d})$ when using the separability of the Euclidean quadratic cost (1.1). Of course one still needs to iterate to reach convergence.

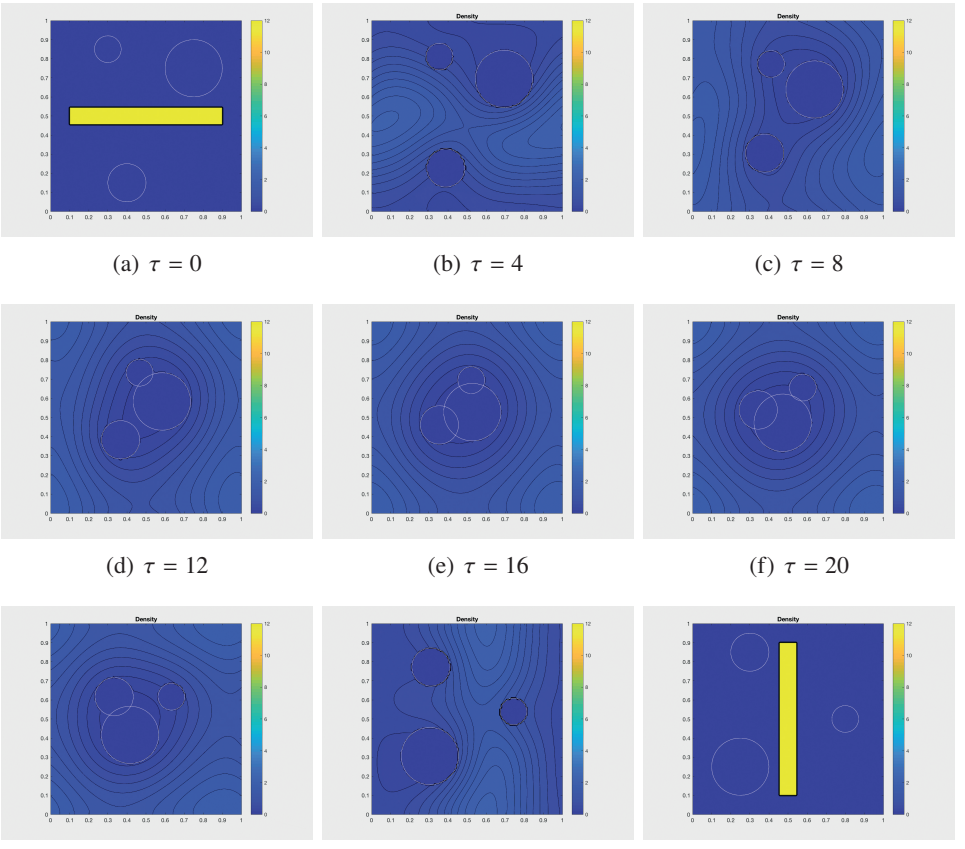
We illustrate this numerical approach with a simulation from Benamou *et al.* (2019*b*). Initial and final densities are prescribed with $H_{0,M}(\rho)=\iota_{\{\rho=\mu_{0,1}\}}$. The 'agents' must avoid multiple obstacles moving in time and this is modelled letting $H_m(\rho)=\langle\iota_{X\setminus O_m},\rho\rangle_X$, O_m be the moving sets,[46] and we are paying an infinite price if some mass (and therefore agents) is present. In this case F is linear, but entropic regularization still yields a strictly convex minimization problem. The boundaries of the obstacles are the white circles in the snapshots displayed in Figures 6.2–6.4. They correspond to different levels of diffusions: $\epsilon=1,10^{-1},10^{-2}$. The discretization is $M=32$ and $N=128^2$.

6.3. *Martingale optimal transportation and transport by diffusion*

Motivated by applications in finance (see Beiglböck and Juillet 2016, Beiglböck, Henry-Labordère and Touzi 2017, Ghoussoub, Kim and Lim 2019 and the references therein), martingale optimal transportation is a recent branch of optimal transportation where the transport plan is constrained to satisfy a *martingale constraint*. As with 'standard' optimal transportation, it can be formulated as a dynamic problem, but we will start with the static version to simplify the exposition. The vector space $X\subset\mathbb{R}^d$ describes all possible prices for d underlying assets in a portfolio. The marginals $\mu_{0,1}\in\mathcal{P}(X_{0,1})$ represent the state of the market at times 0 and 1, *i.e.* the distribution over the set of asset prices of a portfolio. The set of transport plans $\pi\in\mathcal{P}(X_0\times X_1)$ describes all market changes for the distribution of the portfolio between times 0 and 1. The displacement cost c is now interpreted as the pay-off (a fixed gain or loss depending on the prices $x_0\in X_0\subset X$ and $x_1\in X_1\subset X$). Assuming the market changes are given by a fixed transport plan π, also called a *model*, the Monge–Kantorovich cost $\langle c,\pi\rangle_{X_0\times X_1}$ is the yield of the option (*i.e.* the right to buy the portfolio μ_1 at time 1 knowing μ_0). The buying

[46] The characteristic function $\iota_B(x)=0$ if $x\in B$ and $+\infty$ else. $X\setminus B$ is the complement of the set B in X.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) $\tau = 0$ (b) $\tau = 4$ (c) $\tau = 8$

(d) $\tau = 12$ (e) $\tau = 16$ (f) $\tau = 20$

(g) $\tau = 24$ (h) $\tau = 28$ (i) $\tau = 32$

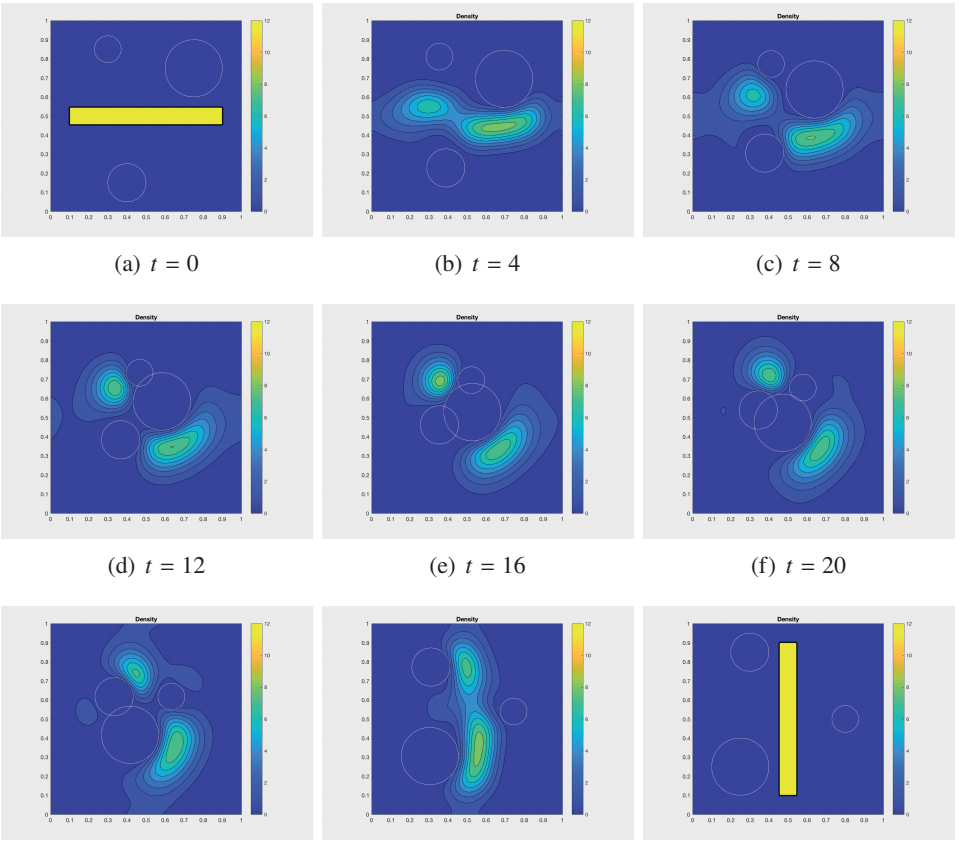
Figure 6.2. Planning mean field games on the torus with moving obstacles and densities at different time steps; $\epsilon = 1$ and 32 time steps. Figure reproduced from Benamou *et al.* (2019*b*) with permission. Copyright © 2019 World Scientific.

price of this option should not exceed the yield in order to prevent loss. If nothing is known about the market mechanisms, this price can be bounded below by

$$\inf_{\pi \in \Pi(\mu_0, \mu_1)} \langle c, \pi \rangle_{X_0 \times X_1}. \tag{6.22}$$

We recognize the familiar Kantorovich problem (1.2). This approach is called 'model-free hedging' because the sets of admissible changes on the market are really 'free': there are no constraints other than the marginals. The real world is not that simple, and the optimal price $\langle c, \pi^* \rangle_{X_0 \times X_1}$ may be greatly underestimated. It seems generally agreed that, at least, a 'no-arbitrage' constraint must be added. Using the notation in this paper and assuming we are dealing with densities, it takes

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) $t = 0$ (b) $t = 4$ (c) $t = 8$

(d) $t = 12$ (e) $t = 16$ (f) $t = 20$

(g) $t = 24$ (h) $t = 28$ (i) $t = 32$

Figure 6.3. Planning mean field games on the torus with moving obstacles, fixed initial and final densities similar to Figure 6.2 and densities at different time steps; $\epsilon = 0.1$ and 32 time steps. Figure reproduced from Benamou *et al.* (2019*b*) with permission. Copyright © 2019 World Scientific.

the form

$$\langle \mathrm{Id}_{X_1}, \pi \rangle_{X_1} = \mathrm{Id}_{X_0}\, \mu_0. \tag{6.23}$$

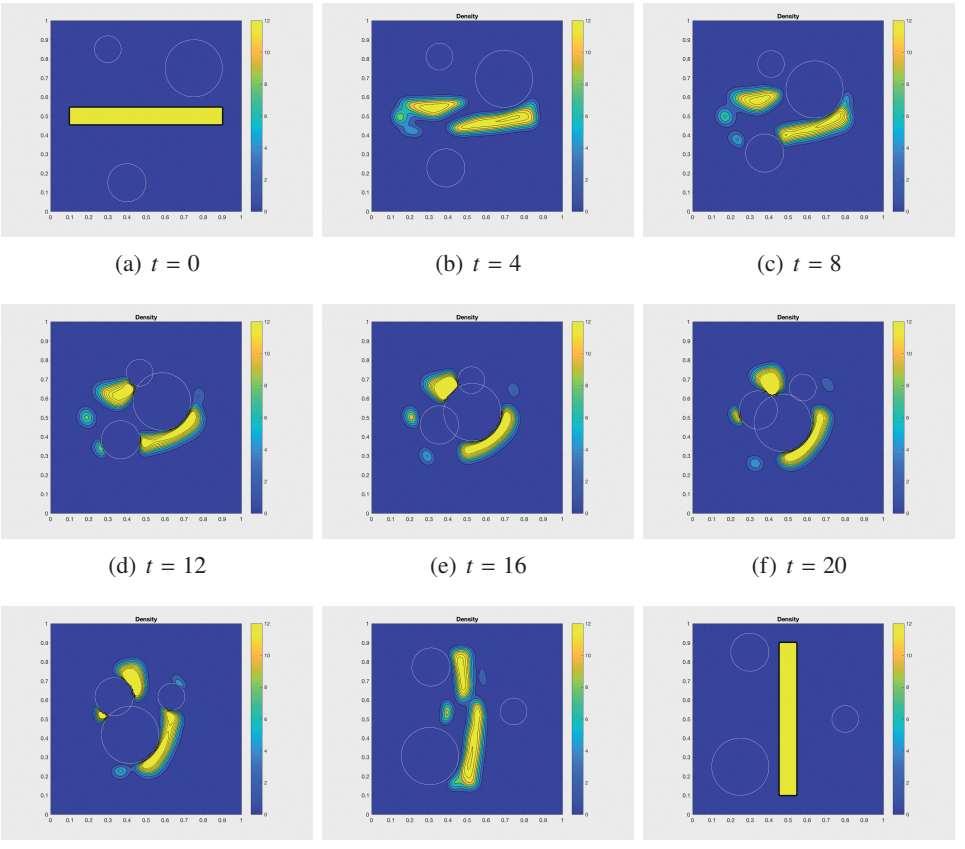
This is the so-called *martingale constraint*. It states that considering the initial wealth $x_0\, \mu_0(x_0)$ at x_0, markets cannot 'arbitrage' a strategy to generate more (or less) wealth at time 1 from the portfolio components previously at x_0. Wrapping up, martingale optimal transportation is

$$\pi^* := \underset{\pi \in \Pi_{Mrt}(\mu_0, \mu_1)}{\arg\inf} \langle c, \pi \rangle_{X_0 \times X_1}, \tag{6.24}$$

where

$$\Pi_{Mrt}(\mu_0, \mu_1) := \{\pi \in \Pi(\mu_0, \mu_1),\ \langle \mathrm{Id}_{X_1}, \pi \rangle_{X_1} = \mathrm{Id}_{X_0}\, \mu_0\}. \tag{6.25}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) $t = 0$ (b) $t = 4$ (c) $t = 8$

(d) $t = 12$ (e) $t = 16$ (f) $t = 20$

(g) $t = 24$ (h) $t = 28$ (i) $t = 32$

Figure 6.4. Planning mean field games on the torus with moving obstacles and densities at different time steps; $\epsilon = 0.01$ and 32 time steps. Figure reproduced from Benamou *et al.* (2019*b*) with permission. Copyright © 2019 World Scientific.

Because of the additional martingale constraints, $\Pi_{Mrt}(\mu_0, \mu_1)$ may be empty. Characterizing the class of $\mu_{0,1}$ such that there exists a martingale $\pi \in \Pi_{Mrt}(\mu_0, \mu_1)$ is the subject of a classical result called Strassen's theorem.[47]

In the mathematical finance literature, martingale optimal transportation is written in probabilistic notation. Let $\mathcal{X}_0 \sim \mu_0$ and $\mathcal{X}_1 \sim \mu_1$ be two random variables on X. Transport plans π are the set of joint laws of $(\mathcal{X}_0, \mathcal{X}_1)$. Then (6.22) is reformulated as the mathematical expectation

$$\inf_{\mathcal{X}_{0,1} \sim \mu_{0,1}} \mathcal{E}(c(\mathcal{X}_0, \mathcal{X}_1)). \tag{6.26}$$

[47] Formally, $\Pi_{Mrt}(\mu_0, \mu_1) \neq \emptyset$ if μ_1 dominates μ_0 in the convex order, meaning $\langle \phi, \rho_1 \rangle_{X_1} \geq \langle \phi, \rho_0 \rangle_{X_0}$ for all convex functions ϕ.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

The martingale constraint (6.23) is given using the conditional expectation

$$\mathcal{E}(\mathcal{X}_1|\mathcal{X}_0) = \mathcal{X}_0, \quad \mu_0 \text{ a.e.} \tag{6.27}$$

Departing from finance, let us look at the martingale optimal transportation problem when the displacement cost is the Euclidean quadratic cost (1.1). In the probabilistic formalism (6.26) and using (6.27), we get

$$\inf_{\mathcal{X}_{0,1} \simeq \mu_{0,1}} \mathcal{E}(\|\mathcal{X}_1 - \mathcal{E}(\mathcal{X}_1|\mathcal{X}_0)\|^2|\mathcal{X}_0), \tag{6.28}$$

the minimization of the conditional variance of $\mathcal{X}_1$ knowing $\mathcal{X}_0$. This is in line with Monge optimization of mass transport: if mass follows a stochastic path $\mathcal{X}_0 \to \mathcal{X}_1$, and $(\mathcal{X}_0, \mathcal{X}_1)$ is a martingale, minimizing the transport work may only be achieved by acting on the variance or equivalently the volatility of the stochastic process.

In a striking parallel with Sections 4.1 and 6.1, Huesmann and Trevisan (2019) have provided a dynamic interpretation that sheds new light on the martingale constraint. All martingales $(\mathcal{X}_0, \mathcal{X}_1) \sim \pi \in \Pi_{Mrt}(\mu_0, \mu_1)$ can be represented as a joint probability $\pi = (e_0, e_1)\#Q$, where Q is the law of a diffusion process

$$\mathrm{d}\mathcal{X}_\tau = \sqrt{\alpha_\tau}\, \mathrm{d}\mathcal{B}_\tau, \quad \mathcal{X}_{0,1} \sim \mu_{0,1}. \tag{6.29}$$

Martingale optimal transportation is controlled by the volatility parameter α_τ. The law of $\mathcal{X}_\tau$, again called ρ_τ, therefore satisfies the diffusion equation and initial/final boundary conditions:

$$(\partial_\tau - \alpha_\tau \Delta)\, \rho_\tau = 0, \quad \rho_{0,1} = \mu_{0,1}. \tag{6.30}$$

Assuming a displacement cost in the general form $(x, y) \to c(y - x)$ (see Huesmann and Trevisan 2019 for the precise mathematical statement), a CFD generalization of martingale optimal transportation holds:

$$\inf_{\pi \in \Pi_M(\mu_0, \mu_1)} \langle c, \pi \rangle_{X_0 \times X_1} = \inf_{(\alpha_\tau, \rho_\tau) \text{ satisfies (6.30)}} \int_0^1 \langle c(\alpha_\tau), \rho_\tau \rangle_X \, \mathrm{d}\tau. \tag{6.31}$$

Needless to say, the mathematical tool is again convex duality, and the full optimal transportation machinery described in Section 6 applies. At this time, the interplay between the entropic-regularization-induced diffusion, if one uses the Sinkhorn algorithm, and the controlling diffusion is still unclear.

Classical dynamic optimal transport optimizes the velocities, while transport by diffusion optimizes the volatilities. Both can be combined into *semi-martingale* optimal transportation, where stochastic paths are also controlled using a deterministic drift $\{\tau \mapsto \beta_\tau\}$:

$$\mathrm{d}\mathcal{X}_\tau = \beta_\tau \, \mathrm{d}\tau + \sqrt{\alpha_\tau}\, \mathrm{d}\mathcal{B}_\tau, \quad \mathcal{X}_{0,1} \sim \mu_{0,1}. \tag{6.32}$$

If $\alpha_\tau := 0$, we are back to standard optimal transportation. If α_τ is fixed and equal to ϵ, this is entropic optimal transportation, and if $\beta_\tau := 0$, this is the martingale optimal transportation above. For these generalizations, see Léonard (2014) and Guo and Loeper (2018) and the references therein.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

7. Transport distances as loss/fidelity

7.1. Variational Russian dolls

So far we have described static and dynamic optimal transportation problems as *linear or convex optimization problems* set on the configuration space X and the rich links existing between various deterministic or stochastic models, in particular in the dynamic case (Sections 4, 5 and 6). The optimizers of these variational problems were characterized as measures on X for the *primals* and continuous functions on X for the *duals*, respectively. This is our smallest Russian doll.

The intermediate Russian doll is at the $\mathcal{P}(X)$ level. In Section 3 we saw that freezing one marginal (say the first ρ_0) in the $\mathcal{W}_2$ cost gives a mathematical metrization on the space of probability measures $\mathcal{P}(X)$. The distance-to-ρ_0 ($\rho_0 \in \mathcal{P}(X)$) information given by the *Wasserstein loss* functional

$$\rho \in \mathcal{P}(X) \mapsto \mathcal{W}_2^2(\rho_0, \rho) \tag{7.1}$$

was sufficient to define meaningful (and useful) notions such as interpolation and barycentres in $\mathcal{P}(X)$. A comprehensive review of transport-based statistical applications on this space can be found in Kolouri *et al.* (2017). Statistical (or machine) learning can be seen as a powerful inference tool in $\mathcal{M}^+(X)$: it naturally explains the explosion of work connecting optimal transportation to this domain. Finally, the theory of *Wasserstein gradient flows*, presented below in Section 7.2, may now be seen as an autonomous research topic and is the subject of several surveys; see Ambrosio *et al.* (2005) or Santambrogio (2015, §8).

The last (and largest) fascinating Russian doll is the Gromov–Wasserstein distance on the space of metric-measured spaces introduced by Mémoli (2011) and Sturm (2020). Using our notation, a metric-measured space is a triplet (X, μ, c), where μ is a reference measure on X and the displacement cost c a distance. Gromov–Wasserstein is a generalization of the optimal transportation distance between two such metric-measured spaces (X_0, μ_0, c_0) and (X_1, μ_1, c_1). It is important for applications where the spaces supporting the distributions $\mu_{0,1}$ cannot be embedded into a common X. The problem is known to be NP-hard; see Peyré, Cuturi and Solomon (2016), who propose an approximate entropic regularization method.

Below we discuss the middle Russian doll.

7.2. $\mathcal{W}_2$ gradient flows

We are concerned with functionals defined on $\mathcal{P}(X)$. Let us first recap the properties of (7.1). Using the dual formulation (2.1), this is equivalent to

$$\rho \in \mathcal{P}(X) \mapsto \sup_{(u_0,u_1)\in C_D} \langle u_0, \rho_0\rangle_{X_0} + \langle u_1, \rho\rangle_X. \tag{7.2}$$

The Wasserstein loss functional is the upper envelope of linear functionals in ρ,

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

$\rho \mapsto \langle u_0, \rho_0\rangle_{X_0} + \langle u_1, \rho\rangle_X$, and it is therefore convex. Formally, again, Danskin's theorem (see footnote 15) implies that (7.2) is differentiable and its functional gradient[48] (we are in the intermediate Russian doll environment) is simply given by the Kantorovich optimal potential u_1^*:

$$\frac{\partial\{\rho \mapsto \mathcal{W}_2^2(\rho_0, \rho)\}}{\partial \rho} = u_1^*. \tag{7.3}$$

Looking back at the economics interpretation of the dual Kantorovich problem (Section 2.1), it makes sense. If the producer wants to increase revenue by lobbying the government into modifying the repartition of one of the goods, they just need to look at the local prices.

A gradient flows in X aims to find a minimizer of some convex differentiable 'energy' $x \in X \mapsto F(x) \in \mathbb{R}$. This is achieved by solving, from some initial $\mathcal{X}_0 = x_0$, the ordinary differential equation $\dot{\mathcal{X}}_\tau = -D_x F(\mathcal{X}_\tau)$, until it becomes stationary and therefore reaches a minimum. Using an implicit discretization in time, we get

$$\mathcal{X}_{\tau+\mathrm{d}\tau} - \mathcal{X}_\tau = -\,\mathrm{d}\tau\, D_x F(\mathcal{X}_{\tau+\mathrm{d}\tau}), \tag{7.4}$$

which, if F is strictly convex, admits the well-known variational formulation

$$\mathcal{X}_{\tau+\mathrm{d}\tau} := \operatorname*{arg\,inf}_{\mathcal{X}\in X} \frac{1}{2}\|\mathcal{X}_\tau - \mathcal{X}\|^2 + \mathrm{d}\tau\, F(x). \tag{7.5}$$

Starting with Jordan *et al.* (1998), this concept has been lifted to the space of probability measures $\mathcal{P}(X)$ using the Wasserstein distance instead of the Euclidean quadratic cost (1.1):

$$\mu_{\tau+\mathrm{d}\tau} := \operatorname*{arg\,inf}_{\mu\in\mathcal{P}(X)} \frac{1}{2}\mathcal{W}_2^2(\mu_\tau, \mu) + \mathrm{d}\tau F(\mu). \tag{7.6}$$

According to the dynamic optimal transportation formulation, the density μ minimizes its kinetic energy plus some penalization depending on the additional energy F. A sequence of probability measures $\{\rho_{\tau_m}\}_{m=0,\dots}$ ($\tau_m = m\,\mathrm{d}\tau$) minimizing the energy F is built iterating (7.6). The minimization (7.6) is a particular instance of the mean field game (6.16)–(6.17) with $H := 0$ and $H_1 := \mathrm{d}\tau\, F$, where we have interpreted the $\mathcal{W}_2$ distance with its CFD formulation. Using the mean field game optimality condition (6.18), there is, on all intervals (τ_m, τ_{m+1}), a curve in time of

[48] The Fréchet derivative of $\rho \in \mathcal{P}(\mathcal{X}) \mapsto F(\rho)$, if it exists, is defined as, for any variation ξ such that $\rho + \xi \in \mathcal{P}(X)$,

$$\lim_{\|\xi\|\to 0} \frac{1}{\|\xi\|} F(\rho+\xi) - F(\rho) = \left\langle \frac{\partial F}{\partial \rho}(\rho), \xi \right\rangle_X.$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

probability measures $\{\mu_\tau^*\}$. Reparametrizing time, it satisfies

$$\begin{aligned} &\partial_\tau \rho_\tau^* + \mathrm{div}_x\left(\rho_\tau^* \frac{D_x \phi^*}{\mathrm{d}\tau}\right) = 0, \quad \rho^*_{\tau=\tau_m,\tau_{m+1}} = (\mu_m, \mu_{m+1}), \\ &\frac{1}{\mathrm{d}\tau}\phi^*_{\tau_{m+1}} = \frac{\partial}{\partial_\rho} F(\mu_{\tau_{m+1}}). \end{aligned} \tag{7.7}$$

In the limit $\mathrm{d}\tau \to 0$, it is possible to show rigorously (see Santambrogio 2015, §8) that the curve $\tau \to \rho_\tau$ satisfies, in a weak sense, the initial value problem

$$\partial_\tau \rho_\tau - \mathrm{div}_x\left(\rho_\tau D_x\left(\frac{\partial F}{\partial \rho}(\rho_\tau)\right)\right) = 0, \quad \rho_0 \text{ given.} \tag{7.8}$$

This is a gradient descent of the energy F with respect to the geometry of the support space $\mathcal{P}(X)$ described by the Wasserstein distance. The simplest and famous example in Jordan *et al.* (1998), mathematically consistent with the second law of thermodynamics,[49] is the negative Gibbs entropy $F(\rho) = KL(\rho|\mathbf{1}_X)$, from which one recovers the dissipative heat equation; see also Gentil (2020) for a review of entropy and gradient flows.

This approach to functional gradient flows is well documented and has been applied to many nonlinear dissipation/diffusion models, in particular to derive theoretical rates of convergence to equilibria. From the numerical point of view, all optimal transportation numerical techniques, with or without entropic regularization, can be applied and provide, naturally, mass conservation and non-negativity of the density; see Benamou, Carlier and Laborde (2016*a*), Cancès, Gallouët and Todeschi (2020), Matthes and Osberger (2014), Benamou, Carlier, Mérigot and Oudet (2016*c*) and Peyré (2015), among others.

The power of the Wasserstein variational approach is nicely illustrated in a model of crowd motion under congestion proposed in Maury, Roudneff-Chupin, Santambrogio and Venel (2011). Individuals try to exit a room X. Let the door be a part of the boundary, and denote it by $\mathcal{D}$. They follow the 'closest exit' direction signs posted everywhere in X. This is given as the gradient of the eikonal $D_x E$, where

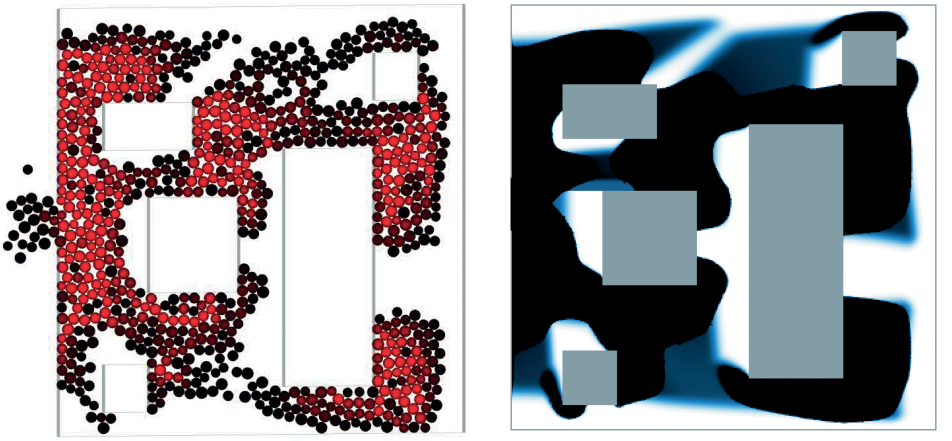
$$E := \left\{ x \mapsto \inf_{x_0 \in \mathcal{D}} \frac{1}{2}\|x - x_0\| \right\}. \tag{7.9}$$

At the 'microscopic' level, people are modelled by hard spheres of positive radius R, which cannot overlap. At the macroscopic level, one asks that the *mean field density* does not exceed a fixed threshold, say 1. The density of the crowd of people may vary but it has a compressibility hard limit. The proposed energy for the Wasserstein gradient flow is

$$F := \rho \in \mathcal{P}(X)) \mapsto \begin{cases} \langle E, \rho \rangle_X & \text{if } \rho \le 1 \text{ a.e. in } X, \\ +\infty & \text{else.} \end{cases} \tag{7.10}$$

[49] A system particle evolves towards thermodynamic equilibrium by maximizing Gibbs entropy.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) (b)

Figure 7.1. (a) 'Microscopic' agent simulation at a fixed time. The colour map indicates the pressure. (b) The 'macroscopic' gradient flow at the same time. The density saturates to 1 in the darkest regions. Figure reproduced from Maury *et al.* (2011) with permission. Copyright © 2011 American Institute of Mathematical Sciences.

While moving, the density strives to minimize the distance to the exit but cannot exceed the compressibility limit. The energy is convex but obviously not differentiable. It is nevertheless shown in Maury *et al.* (2011) that the curve-of-densities solution of (7.6) converges to a (weak) solution of

$$\begin{aligned} \partial_\tau \rho_\tau - \operatorname{div}_x(\rho_\tau\,(D_x\,E - D_x\,p)) = 0, \quad & \rho_0 \text{ given}, \\ 0 \le \rho \le 1,\ p \ge 0,\ p\,(1-\rho) = 0, \quad & \text{a.e. in } X. \end{aligned} \tag{7.11}$$

The velocity $D_x\,E$ points to the exit but, because of the congestion constraint and reminiscent of the Euler problem (Section 5), a pressure p kicks in locally when the density reaches 1, to correct the trajectories. Figure 7.1 shows a convincing comparison with a microscopic simulation.

7.3. Inverse problems and unbalanced optimal transportation

The classical least-squares approach to the solution of ill-posed linear system $A\,x = b^*$ is the variational problem $x^* := \arg\inf_{x\in X} \frac{1}{2}\,\|Ax - b^*\|^2$. Likewise, defining an *abstract model*

$$\text{Mod} := (\theta, \rho_0) \in \Theta \times \mathcal{P}(\mathcal{X}) \mapsto \rho_1 = T_\theta \# \rho_0 \in \mathcal{P}(X) \tag{7.12}$$

as a family of transport map $T_\theta : \mathcal{P}(X) \to \mathcal{P}(X)$ parametrized by a set of parameters $\theta \in \Theta$, the Wasserstein loss can be used in at least two ways. Assuming a computational definition of the model in terms of θ and ρ_0 (the forward map)

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

is known, then, given some observed input and output of the model (ρ_0, ρ_1), the *inverse problem*

$$\theta^* := \arg\inf_{\theta\in\Theta} \frac{1}{2}\,\mathcal{W}_2^2(T_\theta\#\rho_0, \rho_1) \tag{7.13}$$

stands for the set of model parameters best approximating the observation density. If, on the contrary, the model's dependence on θ is unknown or too complicated, another method, now called 'supervised learning', is to replace it with a surrogate *convolutional neural network* (CNN) model $\mathcal{N}_\theta(\rho_0) \simeq \mathrm{Mod}(\rho_0)$ for all ρ_0. The parameters, still denoted by θ, are now independent of the model and characterize the CNN. Given a collection of observations $(\rho_{0,i}, \rho_{1,i})$,

$$\theta^* := \arg\inf_{\theta\in\Theta} \frac{1}{2}\sum_i \mathcal{W}_2^2(\mathcal{N}_\theta(\rho_{0,i}), \rho_{1,i}) \tag{7.14}$$

is the optimal set of parameters for the CNN $\mathcal{N}_{\theta^*}$ to approximate the model Mod.

The advantages of the transport distance have already been discussed in Section 3. First, as indicated by the name, it is a distance on the space of probability measures allowing a generalization to $\mathcal{P}(X)$ of interpolation, barycentres, geodesics and more. It also 'metrizes weak convergence'.[50] This property is important in supervised learning, for example, as the learning samples $(\rho_{0,i}, \rho_{1,i})$ are usually discrete observations: histograms or empirical measures. One expects the loss used in (7.14) to be at least continuous with respect to this sampling process.

There are limits to the use of standard transport distances for such inverse problems. On the curse of dimensionality suffered by the accuracy of the sampling process for $\mathcal{W}_2$, see Chizat *et al.* (2020) Vacher, Muzellec, Rudi, Bach and Vialard (2021) and the references therein.

Another serious issue is linked to mass conservation, or rather the frequent lack of it for realistic models. Noise or discrete approximations may be a reason but the model itself may not be conservative: $\langle 1_X, \mathrm{Mod}(\rho_0)\rangle \neq \langle 1_X, \rho_0\rangle$. The simplest fix is to introduce an additional normalization, *i.e.* replace (7.12) with

$$\mathrm{Mod}_{\text{normalized}} := (\theta, \rho_0) \mapsto \rho_1 = \frac{\mathrm{Mod}(\rho_0)}{\langle 1_X, \mathrm{Mod}(\rho_0)\rangle}, \tag{7.15}$$

but it may seriously modify the mass distribution and changes the modelling.

A different approach was proposed in Benamou (2003), where the term 'unbalanced transport' was introduced. It corresponds to the mean field game (6.16) (Section 7.2),

$$\inf_{(\rho,\mathcal{V})\in FP(\rho_0)} \int_0^1 \frac{1}{2}\langle \|\mathcal{V}_\tau(x)\|^2, \rho_\tau\rangle_X \,\mathrm{d}\tau + H_1(\rho_1), \tag{7.16}$$

[50] For a sequence $\rho_n \overset{*}{\rightharpoonup} \rho$ in $\mathcal{P}(X)$, $\lim_{n\to+\infty} \mathcal{W}_2^2(\rho_n, \rho) = 0$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

where $H := 0$ and

$$H_1 : \rho_1 \to \frac{1}{2}\|\rho_1 - \mu_1\|^2_{\mathcal{L}^2} \tag{7.17}$$

is a relaxation of the target marginal condition. The Fokker–Planck equation constraint (6.17) corresponding to a noisy transport may be used and solved using the entropic optimal transportation machinery presented in Section 6. It is also possible to replace $(\partial_\tau - \epsilon\,\Delta)$ with ∂_τ, *i.e.* $\epsilon = 0$, giving the *deterministic* continuity equation constraints

$$\begin{aligned} CE(\rho_0) := \big\{(\tau, x) \in [0,1] \times X \mapsto (\rho_\tau(x), \mathcal{V}_\tau(x)) \in \mathbb{R}^+ \times \mathbb{R}^d : \\ \partial_\tau \rho_\tau + \mathrm{div}_x(\rho_\tau\, \mathcal{V}_\tau) = 0 \text{ and } \rho_{\tau=0} = \mu_0\big\}, \end{aligned} \tag{7.18}$$

and use the proximal splitting methods mentioned in Section 4.2. This is the method followed in Benamou (2003). In (7.16)–(7.18) all the mass initially distributed as μ_0 is transported. The relaxation (7.17) is the price to pay for the mass default. This part of the cost is static and allows for mass to be created or destroyed anywhere independently of the support ρ_0. It is, however, known to fail to define a distance between non-negative Radon measures $(\mu_0, \mu_1) \in \mathcal{M}^+ \times \mathcal{M}^+$.

The fix, independently proposed by Chizat, Peyré, Schmitzer and Vialard (2018*b*), Liero, Mielke and Savaré (2016) and Kondratyev, Monsaingeon and Vorotnikov (2016), is based firstly on introducing a reaction term in the constraint allowing for mass creation/destruction:

$$\begin{aligned} T_R(\rho_0) := \big\{(\tau, x) \in [0,1] \times X \mapsto (\rho_\tau(x), \mathcal{V}_\tau(x), r_\tau(x)) \in \mathbb{R}^+ \times \mathbb{R}^d \times \mathbb{R} : \\ \partial_\tau \rho_\tau + \mathrm{div}_x(\rho_\tau\, \mathcal{V}_\tau) = r_\tau\, \rho_\tau \text{ and } \rho_{\tau=0,1} = \mu_{0,1}\big\}. \end{aligned} \tag{7.19}$$

Note that, thanks to the reaction term, the equation is not conservative and we can impose the usual initial/final boundary conditions in time.

Secondly, competition is established between transport and reaction in the cost function:

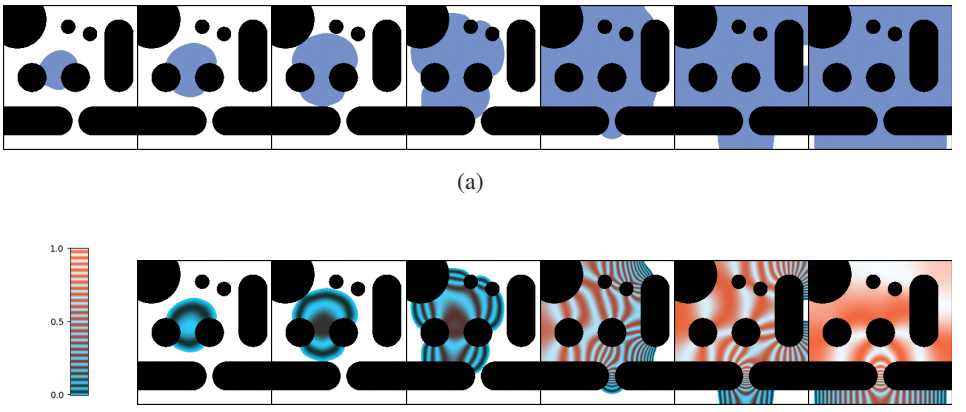
$$\mathcal{W}^2_{FR}(\mu_0, \mu_1) := \inf_{(\rho, \mathcal{V}, r) \in T_R(\rho_0)} \int_0^1 \frac{1}{2} \left\langle \|\mathcal{V}_\tau(x)\|^2 + \frac{1}{4}\, \|r_\tau(x)\|^2, \rho_\tau \right\rangle_X \mathrm{d}\tau. \tag{7.20}$$

The problem (7.20)–(7.19) remains convex and the *Wasserstein–Fisher–Rao* distance $\mathcal{W}_{FR}(\mu_0, \mu_1)$ is a well-defined distance on $\mathcal{M}^+(X)$. It shares a lot of the theory: a static Kantorovich formulation, entropic version, geodesics and barycentres. A gradient flow illustration follows.

Di Marino and Chizat (2020) derived the free boundary 'Hele-Shaw' tumour growth model proposed in Mellet, Perthame and Quirós (2017) as the $\mathcal{W}_{FR}$ gradient flow (replace $\mathcal{W}_2$ with $\mathcal{W}_{FR}$ in (7.6)) for the energy

$$F := \rho \in \mathcal{M}^+(X) \mapsto \begin{cases} -\lambda\, \rho(X) & \text{if } \rho \ll \mathbf{1}_X, \\ +\infty & \text{else.} \end{cases} \tag{7.21}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a)

(b)

Figure 7.2. Evolution on a non-convex two-dimensional domain ('bone' obstacles in black). (a) Evolution of the density. The colour map is linear from white to blue as the density goes from 0 to 1. (b) Pressure represented by a striped colour map to make the level sets apparent. The white area corresponds to $p = \rho = 0$. Figure reproduced from Di Marino and Chizat (2020) with permission. Copyright © 2020 EDP Sciences.

In the limit it yields[51]

$$\begin{aligned} &\partial_\tau \rho_\tau - \operatorname{div}_x(\rho_\tau D_x p) = (\lambda - p)_+ \rho_\tau, && \rho_0 \text{ given}, \\ &0 \le \rho \le 1,\ p \ge 0,\ p(1-\rho) = 0, && \text{a.e. in } X. \end{aligned} \tag{7.22}$$

Minimizing F increases the proportion of dead cells modelled by ρ. It expands according to the 'geometry' of the reaction diffusion equation. As in (7.10), there is a compressibility limit as the proportion of dead cells cannot exceed 1, resulting in the apparition of a pressure governing the flow of cells. The parameter λ, called the 'homeostatic pressure', corresponds to the equilibrium between natural cell division and cell death. The reaction rate is positive only when the pressure is below λ and new cells appear; see Figure 7.2.

7.4. Sinkhorn divergence

Use of the entropic $OT_\epsilon(\mu_0, \mu_1)$ (defined in (3.11)) as a proxy for $OT(\mu_0, \mu_1)$ is widespread, in particular for the applications presented in the above sections. Unfortunately OT_ϵ is not a distance on $\mathcal{P}(X)$, and in particular $OT_\epsilon(\mu_0, \mu_0) > 0$. The identity Monge map cannot be represented by a diffuse entropic plan in the form (3.22): when $\epsilon \to +\infty$, π^*_ϵ will tend to the most diffuse admissible transport plan, $\mu_0 \otimes \mu_0$ and $OT_\epsilon(\mu_0, \mu_0) \to \langle c, \mu_0 \otimes \mu_0 \rangle$. Assuming, without loss of generality,

[51] $(\cdot)_+ = \max(0, .)$ is the positive part.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

that the mean of μ_0 is the origin and c is the Euclidean quadratic cost (1.1), let us define the sequence of dilated distributions

$$\mu_\lambda := x \mapsto \frac{1}{\lambda^d}\mu_0\left(\frac{x}{\lambda}\right) \quad \text{for } \lambda > 1.$$

Then $\langle c, \mu_0 \otimes \mu_\lambda\rangle = 1/2\,(1 + 1/\lambda^2)\, m_2(\mu)$[52] is strictly decreasing as $\lambda \to +\infty$. In particular,

$$\langle c, \mu_0 \otimes \mu_0\rangle > \langle c, \mu_0 \otimes \mu_\lambda\rangle. \tag{7.23}$$

So OT_ϵ obviously does not metrize weak convergence (see footnote 50), but also μ_0 is not even guaranteed to be a minimizer of $\mu \mapsto OT_\epsilon(\mu_0, \mu)$ (depending of course on ϵ, λ).

The tempting way to fix the problem is to decrease ϵ to reduce the entropic bias, but the stability and rate of convergence seriously deteriorate when $\epsilon \to 0$ (see Section 3.2). One remedy (see Feydy *et al.* 2019 and the references therein) is to replace OT_ϵ with

$$S_\epsilon(\mu_0, \mu_1) = OT_\epsilon(\mu_0, \mu_1) - \frac{1}{2}(OT_\epsilon(\mu_0, \mu_0) + OT_\epsilon(\mu_1, \mu_1)). \tag{7.24}$$

We immediately see that, at least, the identity bias discussed previously is removed: $S_\epsilon(\mu, \mu) = 0$. This loss is called *Sinkhorn divergence*, and Feydy *et al.* (2019) proved that it is symmetric in μ_0 and μ_1, and remains positive and convex with respect to ρ_0 and ρ_1. It also metrizes the weak convergence of measures. It yields a better optimal transportation cost proxy than OT_ϵ at the same computational cost. Details, references and simulations can be found in Ramdas, Garcia and Cuturi (2017). Figure 7.3 shows the comparison of a gradient descent

$$\mu_{\tau+\mathrm{d}\tau} := \mu_{\tau+\mathrm{d}\tau} - \mathrm{d}\tau\, \frac{\partial}{\partial \mu}\,\mathrm{Loss}(\mu, \mu_1) \tag{7.25}$$

for different losses and $\mu_0 \neq \mu_1$. It tests the ability of the Loss to 'drive' the distribution μ_0 to the 'minimizer' μ_1 and whether it metrizes the weak convergence. As expected, for the same regularization $\epsilon = 0.1$, S_ϵ performs much better than OT_ϵ.

Asymptotic results on OT_ϵ as $\epsilon \to 0$ (see Conforti and Tamanini 2021, Pal 2019 and the references therein) offer a complementary view of the construction of the Sinkhorn divergence loss S_ϵ. For small ϵ,

$$\begin{aligned} &OT_\epsilon(\mu_0, \mu_1) - OT(\mu_0, \mu_1) \\ &\quad = d\,\epsilon\sqrt{2\pi\,\epsilon} + \epsilon\,(KL(\mu_0|\mathbf{1}_X) + KL(\mu_1|\mathbf{1}_X)) + \epsilon^2\, I(\mu_0, \mu_1) + o(\epsilon^2). \end{aligned} \tag{7.26}$$

The densities are decoupled in the first-order term, which is therefore removed together with the constant by the Sinkhorn divergence. The coefficient in the second-order term is the Fisher information along the displacement interpolation between μ_0 and μ_1.

[52] $m_2(\mu) = \langle \mathrm{Id}^2, \mu\rangle$ is the second moment of μ.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

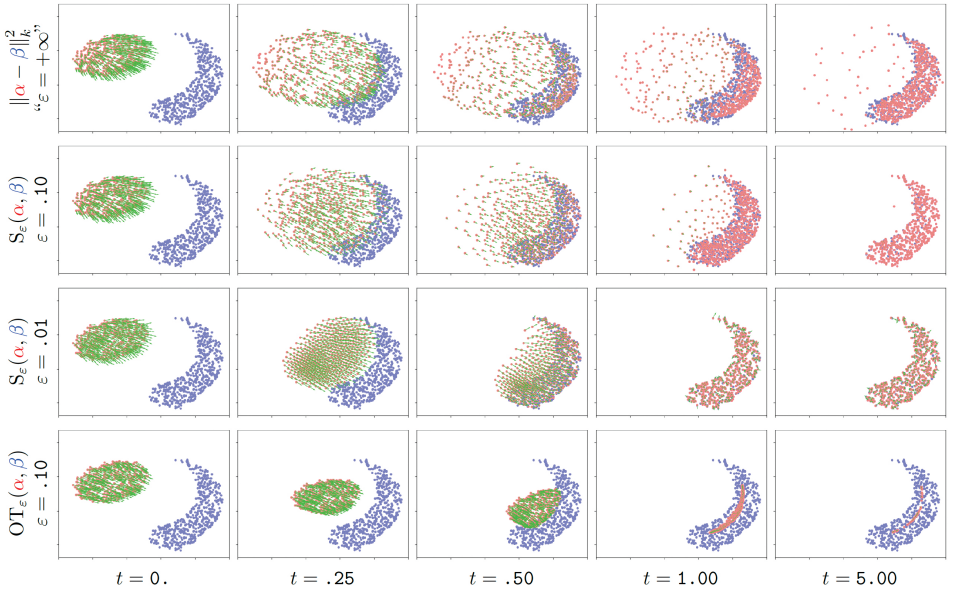


Figure 7.3. Gradient descent (7.25) for different losses; μ_0 is in green, μ_1 in blue and μ_τ in red. Figure reproduced from Feydy *et al.* (2019) with permission.

8. A few missing topics amongst many ...

8.1. Multi-marginal optimal transportation and DFT

The prototype M-marginal problem is

$$\inf_{\pi_M \in \Pi_M} \langle c_M, \pi_M \rangle_{\bigotimes_{m=0}^M X_m}, \tag{8.1}$$

where

$$\Pi_M := \left\{ \pi_M \in \mathcal{P}\left(\bigotimes_{m=1}^M X_m \right) : P_{X_m} \# \pi_M = \mu_m,\ m = 1, \dots, M \right\}, \tag{8.2}$$

and the $\{\mu_m\}$ are given probability measures. Multi-marginal optimal transportation appeared in Section 3 and underlies all dynamic optimal transportation models (Sections 4, 5 and 6). The multi-marginal cost (4.2) there arises from the discretization of the kinetic energy ((4.15) reverting to (4.12)) and is closely linked to the Euclidean quadratic cost (1.1).

The Coulomb 'repulsive cost'

$$c_M(x_1, \dots, x_M) = \sum_{i=1}^M \sum_{j>i}^M \frac{1}{|x_i - x_j|} \tag{8.3}$$

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

appears in *density functional theory*, a branch of quantum chemistry; see Friesecke *et al.* (2013) and Buttazzo, De Pascale and Gori-Giorgi (2012). All marginals μ_m are a single ρ, the identical density probability distribution for the M electrons in a given molecule. The optimal transportation value function is the Coulomb repulsive energy whose minimizers characterize the state of *strictly correlated electrons*. The dimensionality of this problem is daunting: it is naturally set in dimension $d = 3$ but the plan itself is a probability measure over $(\mathbb{R}^d)^{\times M}$. The sparsity of multi-marginal DFT plans is studied in Friesecke and Vögler (2017) For a review of results on the sparsity – or lack – of multi-marginal optimal plans for general costs, see Pass (2015) and Di Marino, Gerolin and Nenna (2017). The global interactions between marginals prevents the tensorization gains used for the kernel (5.23); Altschuler and Boix-Adsera (2020) studied the complexity of available algorithms with respect to the structure of the multi-marginal cost. An entropic numerical approach was carried out by Benamou, Carlier and Nenna (2016*b*). Relaxations/simplifications are still under investigation; see Alfonsi, Coyaud, Ehrlacher and Lombardi (2021), Friesecke and Vögler (2017, 2018) and Cotar *et al.* (2015), among others.

8.2. $\mathcal{L}^1$ optimal transportation, the Beckman problem and optic flows

Contrary to the Euclidean quadratic cost (1.1), the $\mathcal{L}^1$ optimal transportation cost (originally considered by Monge)

$$c(x_0, x_1) = \|x_1 - x_0\| \tag{8.4}$$

does not lead to a well-posed primal Kantorovich problem (1.2) and Monge map solutions (2.10). This is well documented in Santambrogio (2015, §3.1) and Peyré and Cuturi (2019, §6).

The uniqueness issue is easily explained using the one-dimensional 'bookshelf' example. Books (all with the same weight) are arranged on a shelf and there is just one free spot at the furthest right. The librarian is a maniac and only admits free spots at the furthest left. You are in charge of fixing this problem but you are lazy. Will you pick the book on the furthest left and place it at the furthest right? Or will you shift all the books on the shelf one by one? The Euclidean quadratic cost (1.1) will tell you the second solution is optimal while the cost (8.4) says both strategies require the same amount of work. The only important choice is the direction of transport, and for $d = 1$ this is an easy binary choice.

The optimality conditions (2.4) still hold. Using the distance property of c and assuming $X_0 = X_1 = X$, the Kantorovich potentials can be shown to be 1-Lipschitz and complementary: $u_0^* = -u_1^*$. The dual Kantorovich problem (2.1) may be tightened to

$$\mathcal{W}_1(\mu_0, \mu_1) := \sup_{\{u_0,\, \|D_x u_0\| \le 1\}} \langle u_0, \mu_1 - \mu_0 \rangle. \tag{8.5}$$

This new formulation has a well-posed primal called the Beckman problem. We

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

can again use the Fenchel–Rockafellar formalism (3.19) and adapt the dynamic CFD version (4.20) to the 'static' version:

$$\begin{aligned} &A\colon \sigma := \phi \in \mathcal{C}(X) \mapsto -D_x\phi,\\ &A'\colon q \in X(X)^d \mapsto \operatorname{div}_x q,\\ &F\colon \phi \mapsto \langle \phi, \mu_1 - \mu_0\rangle_X,\\ &F^\star\colon a, \mapsto \chi_{a=\mu_1-\mu_0},\\ &G\colon b \mapsto \mathbf{1}_{\|b\|\le 1},\\ &G^\star\colon q \mapsto \int_X \|q\|\,\mathrm{d}x\,\mathrm{d}\tau. \end{aligned} \tag{8.6}$$

With this particular choice the primal–dual optimality conditions (3.20) take the form

$$\begin{aligned} \operatorname{div}_x(q^*) &= \mu_1 - \mu_0, \qquad \text{on } X,\\ q^* &= \|q^*\| D_x\phi^*, \quad \|q\| \text{ a.e. on } X. \end{aligned} \tag{8.7}$$

For a rigorous derivation of these equations and the space settings, see Santambrogio (2015, §4.3); boundary issues in particular are delicate. The system (8.7) can be interpreted[53] as the eikonal and transport equations of geometric optics

$$\|D_x\phi^*\| = 1, \quad \operatorname{div}_x(\|q^*\| D_x\phi^*) = \mu_1 - \mu_0 \tag{8.8}$$

arising from the high-frequency wave equation asymptotic ansatz

$$f(\tau, x) \simeq A(\tau, x) S(\tau - \phi(x))$$

in homogeneous space.[54] Here ϕ^* is the static phase and $\|q^*\| = \int_0^1 A^2(\tau, x)\,\mathrm{d}\tau$ is the energy travelling through x. Problem (1.2) finds the optimal transportation kinematics ϕ^* with prescribed initial/final amplitudes $A^2(\{0,1\},\cdot) = \mu_{0,1}$.

A different high-frequency wave asymptotic analysis linking the CFD formulation to a paraxial approximation of the Helmholtz equation has been carried out by Rubinstein and Wolansky (2004).

The link with optics is natural as the cost (8.4) is simply the length of rays $(x_0 \to x_1)$, *i.e.* the travel time with index of refraction $g := 1$. It can be replaced by the general Riemannian distance

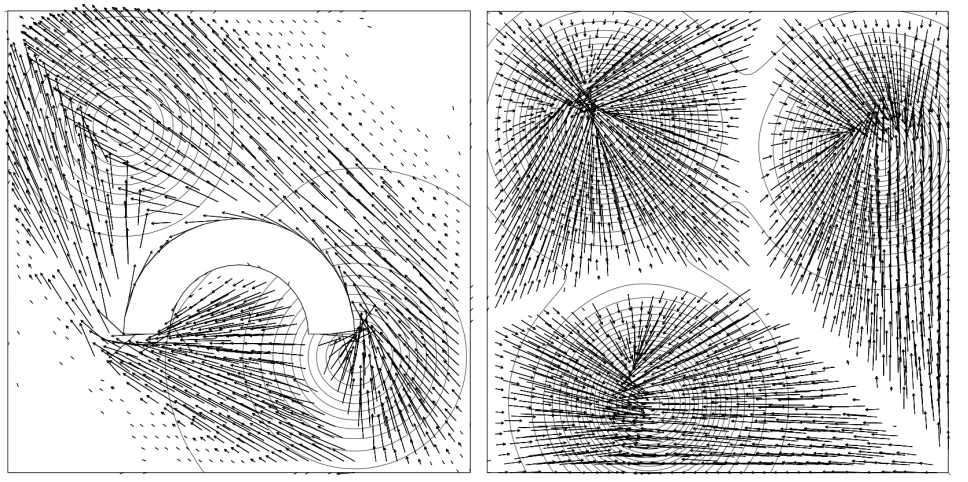
$$c := d_g(x_0, x_1) = \inf_{\{\mathcal{X}_\tau \in W^{1,1}([0,1],X),\ \mathcal{X}_0 = x_0,\ \mathcal{X}_1 = x_1\}} \int_0^1 g(\mathcal{X}_\tau)\|\dot{\mathcal{X}}_\tau\|\,\mathrm{d}\tau, \tag{8.9}$$

yielding minimum travel times of rays from x_0 to x_1. The geometry of the rays depends on g: the eikonal equation in (8.8) generalizes to $\|D_x\phi^*\| = g$. See Figure 8.1 for an illustration in different configurations.

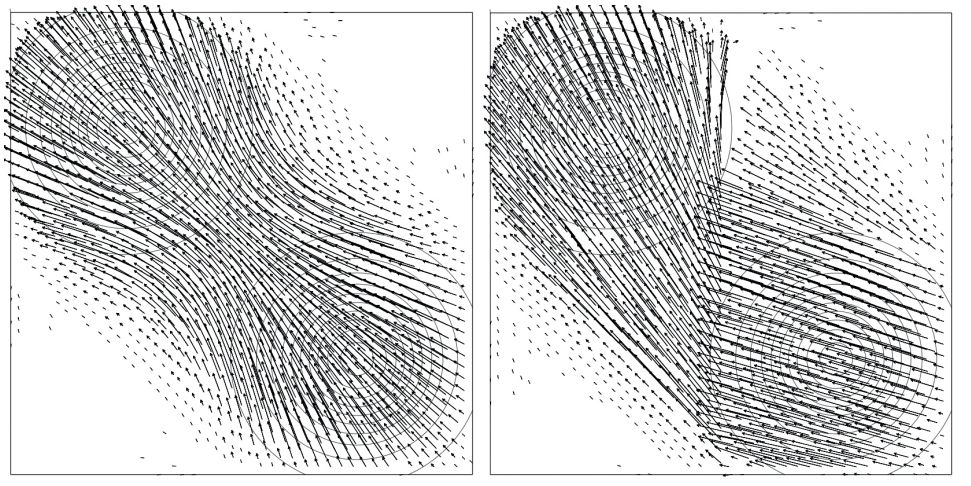
[53] The calculation can be done using high-frequency asymptotics (see Symes 1998, §5, for example).

[54] The wave speed is constant equal to 1, and $\tau \mapsto S(\tau)$ is a wavelet in time triggered at a point source x_S. The acoustic wave equation is $(\partial_\tau^2 - \Delta) f = S(\tau)\,\delta_{x_S}$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press



(a) (b)



(c) (d)

Figure 8.1. Arrow directions correspond to $(1/g)\, D_x \phi^*$ and their length is proportional to $\|q^*\|$. Level curves correspond to the right-hand side density term of the divergence $\mu_{0,1}$ source target data to be transported (Gaussian densities except in (b), where the target is a collection of three Dirac masses). (a,b) $g := 1$; (c,d) g is a lens (c) and a two-layer (d) medium. Figure reproduced from Benamou and Carlier (2015) with permission. Copyright © 2015 Springer Nature.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

More general metrics are possible, for example Finsler metrics (Benamou, Carlier and Hatchi 2018) and nonlinear congestion problems (Benmansour, Carlier, Peyré and Santambrogio 2009), where g depends on the energy $\|q^*\|$.[55]

8.3. *Signed measures and full seismic waveform inversion*

Transport distances have been used recently to attack the problem of *full seismic waveform inversion*. This is a well-known nonlinear inverse problem:

$$\theta^* := \arg\inf_{\theta} \frac{1}{2} \|\mu_0 - \mu_1\|^2_{\mathcal{L}^2}, \quad \mu_0 := \mathrm{Mod}_\theta(S_t), \tag{8.10}$$

where one tries to recover the underground parameters θ^* of a 'forward' wave propagation model Mod_{θ^*} mapping a wave source S_t (a surface explosion at a fixed source) to a 'seismogram' μ_1 (time $\times$ surface signal recordings at fixed receivers). The classical least-squares loss in (8.10) is known to lead to an ill-posed minimization problem with many local minima. A pathology is seen, called 'cycle skipping', linked to phase shifts in the θ^*-observed (or θ-guessed) oscillatory signals. A simple computation shows that a shift in the distribution support $\tau \mapsto \mathcal{W}_2^2(\mu_0(\cdot), \mu_0(. + \tau)) = (\tau^2)/2$ is convex and even quadratic for the $\mathcal{W}_2$ distance. Yang and Engquist (2018) therefore proposed replacing $\mathcal{L}^2$ with $\mathcal{W}_2$ (8.10) to convexify the problem and eliminate cycle skipping. Time signals are, however, not probability measures and some data transformation is needed. There is little understanding of the interpretation of time oscillatory signals in terms of mass or probability measures. In particular, they are not positive. Just considering the energy of the signal to fix the problem (but losing the phase information), the source/target may also not have the same total mass because of acquisition noise or dissipation in the forward model Mod_θ.

The problem of defining an optimal transport distance for general signed measures is discussed in Mainini (2012). The simplest idea is to split the positive and the negative part[56] of the signals and consider the sum of the Wasserstein distances

$$(\mu_0, \mu_1) \mapsto \mathcal{W}_2^2(\mu_0^+, \mu_1^+) + W_2^2(\mu_0^-, \mu_1^-). \tag{8.11}$$

If this expression makes sense, *i.e.* the positive and negative parts are 'balanced' (the same total mass), then this cost remains a distance on $\mathcal{P}(X)$. If not, then one may try replacing $\mathcal{W}_2$ with the unbalanced $\mathcal{W}_{FR}$ distance (7.20) as in Li, Lamoureux and Liao (2020), but it does not define a proper distance. Several signal transformations and normalizations are investigated in Yang and Engquist (2018).

The second idea is to use the $\mathcal{W}_1$ distance (8.5). It remains well-defined and a distance for signed measures. There are no positivity constraints but the (signed)

[55] I am still wondering if it may be linked to high-frequency asymptotics of a nonlinear auto-(de)focusing optic model.

[56] $\mu^+ = \max\{\mu, 0\}$ and $\mu^- = \max\{-\mu, 0\}$.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

total mass balance

$$\langle \mathbf{1}, \mu_1 - \mu_0 \rangle = 0 \tag{8.12}$$

needs to be satisfied.

Using positive and negative parts, a simple computation shows that

$$\mathcal{W}_1(\mu_0, \mu_1) = \mathcal{W}_1(\mu_0^+ + \mu_1^-, \mu_0^- + \mu_1^+). \tag{8.13}$$

It is not clear how to interpret this formula, as the distance optimizes transport between composite distributions aggregating information from the source and the target data. It was used with convincing results in Métivier *et al.* (2016), where the total mass balanced constraint is relaxed.

The geophysical community represents seismograms as collections of independent time 'lines' (the time recordings at each location). Instead of considering the total acquisition as a time × surface image, summing line-wise the optimal transportation distance between observed and simulated lines at the same receivers seems to perform well, and to decrease the dimension of the optimal transportation problems to be solved. Lines may also be interpreted as shapes in the time × signal amplitude 'graph space' (Métivier, Brossier, Mérigot and Oudet 2019). After a discretization in time, they give empirical measures living in $\mathbb{R}_{\text{time}} \times \mathbb{R}_{\text{amplitude}}$ space.

References

Y. Achdou, P. Cardaliaguet, F. Delarue, A. Porretta and F. Santambrogio (2020), *Mean Field Games*, Springer.

M. Agueh and G. Carlier (2011), Barycenters in the Wasserstein space, *SIAM J. Math. Anal.* **43**, 904–924.

A. Alfonsi, R. Coyaud, V. Ehrlacher and D. Lombardi (2021), Approximation of optimal transport problems with marginal moments constraints, *Math. Comp.* **90**, 689–737.

J. M. Altschuler and E. Boix-Adsera (2020), Polynomial-time algorithms for multimarginal optimal transport problems with decomposable structure.
Available at arXiv:2008.03006.

L. Ambrosio, N. Gigli and G. Savare (2005), *Gradient Flows in Metric Spaces and in the Space of Probability Measures*, Birkhäuser.

R. Andreev (2017), Preconditioning the augmented Lagrangian method for instationary mean field games with diffusion, *SIAM J. Sci. Comput.* **39**, A2763–A2783.

V. Arnold (1966), Sur la géométrie différentielle des groupes de Lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits, *Ann. Inst. Fourier* **16**, 319–361.

M. Beiglböck and N. Juillet (2016), On a problem of optimal transport under marginal martingale constraints, *Ann. Probab.* **44**, 42–106.

M. Beiglböck, P. Henry-Labordère and N. Touzi (2017), Monotone martingale transport plans and Skorokhod embedding, *Stochastic Process. Appl.* **127**, 3005–3013.

J.-D. Benamou (2003), Numerical resolution of an 'unbalanced' mass transport problem, *ESAIM Math. Model. Numer. Anal.* **37**, 851–868.

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

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

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

J.-D. Benamou and V. Duval (2019), Minimal convex extensions and finite difference discretisation of the quadratic Monge–Kantorovich problem, *Europ. J. Appl. Math.* **30**, 1041–1078.

J.-D. Benamou and M. Martinet (2020), Capacity constrained entropic optimal transport, Sinkhorn saturated domain out-summation and vanishing temperature. Available at hal-02563022.

J.-D. Benamou, G. Carlier and R. Hatchi (2018), A numerical solution to Monge's problem with a Finsler distance as cost, *ESAIM Math. Model. Numer. Anal.* **52**, 2133–2148.

J.-D. Benamou, G. Carlier and M. Laborde (2016*a*), An augmented Lagrangian approach to Wasserstein gradient flows and applications, *ESAIM Proc. Surveys* **54**, 1–17.

J.-D. Benamou, G. Carlier and L. Nenna (2016*b*), A numerical method to solve multi-marginal optimal transport problems with Coulomb cost, in *Splitting Methods in Communication, Imaging, Science, and Engineering* (R. Glowinski, S. J. Osher and W. Yin, eds), Springer, pp. 577–601.

J.-D. Benamou, G. Carlier and L. Nenna (2019*a*), Generalized incompressible flows, multi-marginal transport and Sinkhorn algorithm, *Numer. Math.* **142**, 33–54.

J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna and G. Peyré (2015), Iterative Bregman projections for regularized transportation problems, *SIAM J. Sci. Comput.* **2**, A1111–A1138.

J.-D. Benamou, G. Carlier, S. Di Marino and L. Nenna (2019*b*), An entropy minimization approach to second-order variational mean-field games, *Math. Models Methods Appl. Sci.* **29**, 1553–1583.

J.-D. Benamou, G. Carlier, Q. Mérigot and E. Oudet (2016*c*), Discretization of functionals involving the Monge–Ampère operator, *Numer. Math.* **134**, 611–636.

J.-D. Benamou, F. Collino and J.-M. Mirebeau (2016*d*), Monotone and consistent discretization of the Monge–Ampère operator, *Math. Comp.* **85**, 2743–2775.

J.-D. Benamou, B. D. Froese and A. M. Oberman (2014), Numerical solution of the optimal transportation problem using the Monge–Ampère equation, *J. Comput. Phys.* **260**, 107–126.

J.-D. Benamou, T. O. Gallouët and F.-X. Vialard (2019*c*), Second-order models for optimal transport and cubic splines on the Wasserstein space, *Found. Comput. Math.* **19**, 1113–1143.

J.-D. Benamou, W. L. Ijzerman and G. Rukhaia (2020), An entropic optimal transport numerical approach to the reflector problem. Available at hal-02539799.

F. Benmansour, G. Carlier, G. Peyré and F. Santambrogio (2009), Numerical approximation of continuous traffic congestion equilibria, *Netw. Heterog. Media* **4**, 605–623.

R. J. Berman (2020), The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations, *Numer. Math.* **145**, 771–836.

M. Bernot, V. Caselles and J. M. Morel (2008), *Optimal Transportation Networks: Models and Theory*, Vol. 1955 of Lecture Notes in Mathematics, Springer.

N. Bonneel, G. Peyré and M. Cuturi (2016), Wasserstein barycentric coordinates: Histogram regression using optimal transport, *ACM Trans. Graphics* **35**, 71.

Y. Brenier (1989), The least action principle and the related concept of generalized flows for incompressible perfect fluids, *J. Amer. Math. Soc.* **2**, 225–255.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

Y. Brenier (1991), Polar factorization and monotone rearrangement of vector-valued functions, *Comm. Pure Appl. Math.* **44**, 375–417.
Y. Brenier (2020), Examples of hidden convexity in nonlinear PDEs. Available at hal-02928398.
K. Brix, Y. Hafizogullari and A. Platen (2015), Solving the Monge–Ampère equations for the inverse reflector problem, *Math. Models Methods Appl. Sci.* **25**, 803–837.
G. Buttazzo, L. De Pascale and P. Gori-Giorgi (2012), Optimal-transport formulation of electronic density-functional theory, *Phys. Rev. A* **85**, 062502.
L. Caffarelli (1992), The regularity of mappings with a convex potential, *J. Amer. Math. Soc.* **5**, 99–104.
C. Cancès, T. Gallouët and G. Todeschi (2020), A variational finite volume scheme for Wasserstein gradient flows, *Numer. Math.* **146**, 437–480.
G. Carlier (2001), A general existence result for the principal-agent problem with adverse selection, *J. Math. Econ.* **35**, 129–150.
G. Carlier (2021), *Classical and Modern Optimization*, Imperial College Press. To appear.
G. Carlier, K. Eichinger and A. Kroshnin (2020), Entropic-Wasserstein barycenters: PDE characterization, regularity and CLT. Available at hal-03084049.
F. Cavalletti and A. Mondino (2020), Optimal transport in Lorentzian synthetic spaces, synthetic timelike Ricci curvature lower bounds and applications. Available at arXiv:2004.08934.
A. Chambolle and T. Pock (2016), An introduction to continuous optimization for imaging, in *Acta Numerica*, Vol. 25, Cambridge University Press, pp. 161–319.
Y. Chen, G. Conforti and T. T. Georgiou (2018), Measure-valued spline curves: An optimal transport viewpoint, *SIAM J. Math. Anal.* **50**, 5947–5968.
L. Chizat, G. Peyré, B. Schmitzer and F.-X. Vialard (2018*a*), Scaling algorithms for unbalanced optimal transport problems, *Math. Comput.* **87**, 2563–2609.
L. Chizat, G. Peyré, B. Schmitzer and F.-X. Vialard (2018*b*), Unbalanced optimal transport: Dynamic and Kantorovich formulations, *J. Funct. Anal.* **274**, 3090–3123.
L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard and G. Peyré (2020), Faster Wasserstein distance estimation with the Sinkhorn divergence. Available at arXiv:2006.08172.
R. Cominetti and J. S. Martín (1994), Asymptotic analysis of the exponential penalty trajectory in linear programming, *Math. Program.* **67**, 169–187.
G. Conforti and L. Tamanini (2021), A formula for the time derivative of the entropic cost and applications, *J. Funct. Anal.* **280**, 108964.
C. Cotar, C. Cotar, G. Friesecke, G. Friesecke, B. Pass and B. Pass (2015), Infinite-body optimal transport with Coulomb cost, *Calc. Var. Partial Differential Equations* **54**, 717–742.
M. J. P. Cullen (2006), *A Mathematical Theory of Large-Scale Atmosphere/Ocean Flow*, Imperial College Press.
M. J. P. Cullen and R. J. Purser (1984), An extended Lagrangian theory of semigeostrophic frontogenesis, *J. Atmos. Sci.* **41**, 1477–1497.
M. Cuturi (2013), Sinkhorn distances: Lightspeed computation of optimal transport, in *Advances in Neural Information Processing Systems 26 (NIPS 2013)* (C. J. C. Burges *et al.*, eds), Curran Associates, pp. 2292–2300.
G. D. Dafni, R. J. McCann and A. Stancu (2013), *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the 50th Séminaire de Mathématiques Supérieures (SMS)*, American Mathematical Society.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

S. Daneri and A. Figalli (2016), *Variational Models for the Incompressible Euler Equations*, Vol. 7 of AIMS on Applied Mathematics, American Institute of Material Sciences, pp. 1–48.
P. M. M. de Castro, Q. Mérigot and B. Thibert (2016), Far-field reflector problem and intersection of paraboloids, *Numer. Math.* **134**, 389–411.
S. Di Marino and L. Chizat (2020), A tumor growth model of Hele-Shaw type as a gradient flow, *ESAIM Control Optim. Calc. Var.* **26**, 103.
S. Di Marino and A. Gerolin (2020), An optimal transport approach for the Schrödinger bridge problem and convergence of Sinkhorn algorithm, *J. Sci. Comput.* **85**, 27.
S. Di Marino, A. Gerolin and L. Nenna (2017), Optimal transportation theory with repulsive costs, in *Topological Optimization and Optimal Transport: In the Applied Sciences*, Vol. 17 of Radon Series on Computational and Applied Mathematics, De Gruyter, pp. 204–256.
D. G. Ebin and J. Marsden (1970), Groups of diffeomorphisms and the motion of an incompressible fluid, *Ann. of Math.* **92**, 102–163.
I. Ekeland (2010), Notes on optimal transportation, *Economic Theory* **42**, 437–459.
L. C. Evans (2001), Partial differential equations and Monge–Kantorovich mass transfer. Available at https://math.berkeley.edu/~evans/Monge-Kantorovich.survey.pdf.
J. Feydy (2019), Geometric loss functions between sampled measures, images and volumes. Available at https://www.kernel-operations.io/geomloss/.
J. Feydy (2020), Analyse de données géométriques, au delà des convolutions. PhD thesis, Mathématiques appliquées, Université Paris–Saclay.
J. Feydy, T. Séjourné, F.-X. Vialard, S.-i. Amari, A. Trouvé and G. Peyré (2019), Interpolating between optimal transport and MMD using Sinkhorn divergences, in *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS 2019)* (K. Chaudhuri and M. Sugiyama, eds), Vol. 89 of Proceedings of Machine Learning Research, PMLR, pp. 2681–2690.
A. Figalli (2017), *The Monge–Ampère Equation and its Applications*, Zurich Lectures in Advanced Mathematics, European Mathematical Society.
A. Figalli, Y.-H. Kim and R. J. McCann (2011), When is multidimensional screening a convex program?, **146**, 454–478.
M. Fortin and R. Glowinski (1985), *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*, Vol. 15 of Studies in Mathematics and its Applications, North-Holland.
G. Friesecke and D. Vögler (2017), Breaking the curse of dimension in multi-marginal Kantorovich optimal transport on finite state spaces. Available at arXiv:1801.00341.
G. Friesecke and D. Vögler (2018), Breaking the curse of dimension in multi-marginal Kantorovich optimal transport on finite state spaces, *SIAM J. Math. Anal.* **50**, 3996–4019.
G. Friesecke, C. B. Mendl, B. Pass, C. Cotar and C. Klüppelberg (2013), N-density representability and the optimal transport limit of the Hohenberg–Kohn functional, *J. Chem. Phys.* **139**, 164109.
U. Frisch, S. Matarrese, R. Mohayaee and A. Sobolevski (2002), A reconstruction of the initial conditions of the universe by optimal mass transportation, *Nature* **417**, 260–262.
A. Galichon (2016), *Optimal Transport Methods in Economics*, first edition, Princeton University Press.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

T. O. Gallouët and Q. Mérigot (2018), A Lagrangian scheme à la Brenier for the incompressible Euler equations, *Found. Comput. Math.* **18**, 835–865.
W. Gangbo and R. J. McCann (1996), The geometry of optimal transportation, *Acta Math.* **177**, 113–161.
I. Gentil (2020), The entropy, from Clausius to functional inequalities. Available at arXiv:2011.05206.
N. Ghoussoub, Y.-H. Kim and T. Lim (2019), Structure of optimal martingale transport plans in general dimensions, *Ann. Probab.* **47**, 109–164.
T. Glimm and V. Oliker (2003), Optical design of single reflector systems and the Monge–Kantorovich mass transfer problem, *J. Math. Sci.* **117**, 4096–4108.
F. Golse and T. Paul (2021), Quantum and semiquantum pseudometrics and applications. Available at arXiv:2102.05184.
O. Guéant (2012), Mean field games equations with quadratic Hamiltonian: A specific approach, *Math. Models Methods Appl. Sci.* **22**, 1250022.
K. Guittet (2003), On the time-continuous mass transport problem and its approximation by augmented Lagrangian techniques, *SIAM J. Numer. Anal.* **41**, 382–399.
I. Guo and G. Loeper (2018), Path dependent optimal transport and model calibration on exotic derivatives, *SSRN Electron. J.* doi:10.2139/ssrn.3302384.
S. Haker, L. Zhu, A. Tannenbaum and S. Angenent (2004), Optimal mass transport for registration and warping, *Internat. J. Comput. Vision* **60**, 225–240.
M. Huesmann and D. Trevisan (2019), A Benamou–Brenier formulation of martingale optimal transport, *Bernoulli* **25**, 2729–2757.
R. Hug, E. Maitre and N. Papadakis (2020), On the convergence of augmented Lagrangian method for optimal transport between nonnegative densities, *J. Math. Anal. Appl.* **485**, 123811.
R. Jordan, D. Kinderlehrer and F. Otto (1998), The variational formulation of the Fokker–Planck equation, *SIAM J. Math. Anal.* **29**, 1–17.
J. Kitagawa, Q. Mérigot and B. Thibert (2019), Convergence of a Newton algorithm for semi-discrete optimal transport, *J. Eur. Math. Soc.* **21**, 2603–2651.
S. Kolouri, S. Park, M. Thorpe, D. Slepcev and G. K. Rohde (2017), Transport-based analysis, modeling, and learning from signal and data distributions, *IEEE Signal Process. Magazine* **34**, 43–59.
S. Kondratyev, L. Monsaingeon and D. Vorotnikov (2016), A new optimal transport distance on the space of finite Radon measures, *Adv. Diff. Equations* **21**, 1117–1164.
J.-M. Lasry and P.-L. Lions (2007), Mean field games, *Japan. J. Math.* **2**, 229–260.
H. Lavenant (2021), Unconditional convergence for discretizations of dynamical optimal transport, *Math. Comp.* **90**, 739–786.
C. Léonard (2014), A survey of the Schrödinger problem and some of its connections with optimal transport, *Discrete Contin. Dyn. Syst.* **34**, 1533–1574.
B. Lévy and E. L. Schwindt (2018), Notions of optimal transport theory and how to implement them on a computer, *Comput. Graph.* **72**, 135–148.
D. Li, M. P. Lamoureux and W. Liao (2020), Application of an unbalanced optimal transport distance and a mixed L1/Wasserstein distance to full waveform inversion. Available at arXiv:2004.05237.
M. Liero, A. Mielke and G. Savaré (2016), Optimal transport in competition with reaction: The Hellinger–Kantorovich distance and geodesic curves, *SIAM J. Math. Anal.* **48**, 2869–2911.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

J. L. Lions (1971), *Optimal Control of Systems Governed by Partial Differential Equations*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Springer.
J. Maas (2011), Gradient flows of the entropy for finite Markov chains, *J. Functional Analysis* **261**, 2250–2292.
E. Mainini (2012), A description of transport cost for signed measures, *J. Math. Sci.* **97**, 837–855.
D. Matthes and H. Osberger (2014), Convergence of a variational Lagrangian scheme for a nonlinear drift diffusion equation, *ESAIM Math. Model. Numer. Anal.* **48**, 697–726.
B. Maury, A. Roudneff-Chupin, F. Santambrogio and J. Venel (2011), Handling congestion in crowd motion modeling, *Netw. Heterog. Media* **6**, 485.
R. J. McCann (1997), A convexity principle for interacting gases, *Adv. Math.* **128**, 153–179.
A. Mellet, B. Perthame and F. Quirós (2017), A Hele-Shaw problem for tumor growth, *J. Funct. Anal.* **273**, 3061–3093.
F. Mémoli (2011), Gromov–Wasserstein distances and the metric approach to object matching, *Found. Comput. Math.* **11**, 417–487.
Q. Mérigot (2011), A multiscale approach to optimal transport, *Computer Graphics Forum* **30**, 1583–1592.
Q. Mérigot and J.-M. Mirebeau (2016), Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport, *SIAM J. Numer. Anal.* **54**, 3465–3492.
Q. Mérigot and B. Thibert (2020), Optimal transport: Discretization and algorithms. Available at hal-02494446.
L. Métivier, R. Brossier, Q. Mérigot and E. Oudet (2019), A graph space optimal transport distance as a generalization of L^p distances: Application to a seismic imaging inverse problem, *Inverse Problems* **35**, 085001.
L. Métivier, R. Brossier, Q. Mérigot, E. Oudet and J. Virieux (2016), An optimal transport approach for seismic tomography: Application to 3D full waveform inversion, *Inverse Problems* **32**, 115008.
A. Natale and G. Todeschi (2020), A mixed finite element discretization of dynamical optimal transport. Available at hal-02501634.
S. Pal (2019), On the difference between entropic cost and the optimal transport cost. Available at arXiv:1905.12206.
N. Papadakis, G. Peyré and E. Oudet (2014), Optimal transport with proximal splitting, *SIAM J. Imaging Sci.* **7**, 212–238.
B. Pass (2015), Multi-marginal optimal transport: Theory and applications, *ESAIM Math. Model. Numer. Anal.* **49**, 1771–1790.
P. Pegon (2017), Transport branché et structures fractales. PhD thesis, Mathématiques appliquées, Université Paris–Saclay (ComUE).
G. Peyré (2015), Entropic approximation of Wasserstein gradient flows, *SIAM J. Imaging Sci.* **8**, 2323–2351.
G. Peyré and M. Cuturi (2019), Computational optimal transport, *Found. Trends Mach. Learning* **11**, 355–607.
G. Peyré, M. Cuturi and J. Solomon (2016), Gromov–Wasserstein averaging of kernel and distance matrices, in *Proceedings of the 33rd International Conference on Machine Learning (ICML 2016)* (M. F. Balcan and K. Q. Weinberger, eds), Vol. 48 of Proceedings of Machine Learning Research, PMLR, pp. 2664–2672.
S. T. Rachev and L. Rüschendorf (2006), *Mass Transportation Problems: Applications*, Probability and its Applications, Springer.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

A. Ramdas, N. Garcia and M. Cuturi (2017), On Wasserstein two sample testing and related families of nonparametric tests, *Entropy* **19**, 47.
J.-C. Rochet and P. Chone (1998), Ironing, sweeping, and multidimensional screening, *Econometrica* **66**, 783–826.
J. Rubinstein and G. Wolansky (2004), A variational principle in optics, *J. Opt. Soc. Amer. A* **21**, 2164–2172.
B. Salanié and A. Galichon (2012), Cupid's invisible hand: Social surplus and identification in matching models. Available at hal-01053710.
F. Santambrogio (2015), *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, Vol. 87 of Progress in Nonlinear Differential Equations and Their Applications, Springer.
B. Schmitzer (2019), Stabilized sparse scaling algorithms for entropy regularized transport problems, *SIAM J. Sci. Comput.* **41**, A1443–A1481.
S. Steinerberger (2020), On a Kantorovich–Rubinstein inequality. Available at arXiv:2010.12946.
K.-T. Sturm (2020), The space of spaces: Curvature bounds and gradient flows on the space of metric measure spaces. Available at arXiv:1208.0434.
W. Symes (1998), Mathematics of reflection seismology. Available at http://wwsorcas.com/book0/book0.pdf.
A. Vacher, B. Muzellec, A. Rudi, F. Bach and F.-X. Vialard (2021), A dimension-free computational upper-bound for smooth optimal transport estimation. Available at arXiv:2101.05380.
F.-X. Vialard (2019), An elementary introduction to entropic regularization and proximal methods for numerical optimal transport. Available at hal-02303456.
C. Villani (2003), *Topics in Optimal Transportation*, Graduate Studies in Mathematics, American Mathematical Society.
C. Villani (2008), *Optimal Transport: Old and New*, Vol. 338 of Grundlehren der mathematischen Wissenschaften, Springer.
X.-J. Wang (2004), On the design of a reflector antenna II, *Calc. Var. Partial Differential Equations* **20**, 329–341.
Y. Yang and B. Engquist (2018), Analysis of optimal transport and related misfit functions in full-waveform inversion, *Geophys.* **83**, A7–A12.

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press

https://doi.org/10.1017/S0962492921000040 Published online by Cambridge University Press