Optimal Transport on Discrete Domains

Justin Solomon

Abstract. Inspired by the matching of supply to demand in logistical problems, the optimal transport (or Monge–Kantorovich) problem involves the matching of probability distributions defined over a geometric domain such as a surface or manifold. In its most obvious discretization, optimal transport becomes a large-scale linear program, which typically is infeasible to solve efficiently on triangle meshes, graphs, point clouds, and other domains encountered in graphics and machine learning. Recent breakthroughs in numerical optimal transport, however, enable scalability to orders-of-magnitude larger problems, solvable in a fraction of a second. Here, we discuss advances in numerical optimal transport that leverage understanding of both discrete and smooth aspects of the problem. State-of-the-art techniques in discrete optimal transport combine insight from partial differential equations (PDE) with convex analysis to reformulate, discretize, and optimize transportation problems. The end result is a set of theoretically-justified models suitable for domains with thousands or millions of vertices. Since numerical optimal transport is a relatively new discipline, special emphasis is placed on identifying and explaining open problems in need of mathematical insight and additional research.

1. Introduction

Many tools from discrete differential geometry (DDG) were inspired by practical considerations in areas like computer graphics and vision. Disciplines like these require fine-grained understanding of geometric structure and the relationships between different shapes—problems for which the toolbox from smooth geometry can provide substantial insight. Indeed, a triumph of discrete differential geometry is its incorporation into a wide array of computational pipelines, affecting the way artists, engineers, and scientists approach problem-solving across geometry-adjacent disciplines.

A key but neglected consideration hampering adoption of ideas in DDG in fields like computer vision and machine learning, however, is *resilience* to noise and uncertainty. The view of the world provided by video cameras, depth sensors, and other equipment is extremely unreliable. Shapes do not necessarily come to a computer as complete, manifold meshes but rather may be scattered clouds of points that represent e.g. only those features visible from a single position. Similarly, it may be impossible to pinpoint a feature on a shape exactly; rather, we may receive only a fuzzy signal indicating where a point or feature of interest *may* be located.

2010 *Mathematics Subject Classification.* Primary: 65K10; Secondary: 52C99, 49Q20, 49M29.

Such uncertainty only increases in high-dimensional statistical contexts, where the presence of geometric structure in a given dataset is itself not a given. Rather than regarding this messiness as an "implementation issue" to be coped with by engineers adapting DDG to imperfect data, however, the challenge of developing principled yet noise-resilient discrete theories of shape motivates new frontiers in mathematical research.

Probabilistic language provides a natural means of formalizing notions of uncertainty in the geometry processing pipeline. In place of representing a feature or shape directly, we might instead use a probability distribution to encode a rougher notion of shape. Unfortunately, this proposal throws both smooth and discrete constructions off their foundations: We must return to the basics and redefine notions like distance, distortion, and curvature in a fashion that does not rely on knowing shape with infinite precision and confidence. At the same time, we must prove that the classical case is recovered as uncertainty diminishes to zero.

The mathematical discipline of *optimal transport* (OT) shows promise for making geometry work in the probabilistic regime. In its most basic form, optimal transport provides a means of lifting distances between points on a domain to distances between probability distributions *over* the domain. The basic construction of OT is to interpret probability distributions as piles of sand; the distance between two such piles of sand is defined as the amount of work it takes to transform one pile into the other. This intuitive construction gave rise to an alternative name for OT in the computational world: The "earth mover's distance" (EMD) [**94**]. Indeed, the basic approach in OT is so natural that it has been proposed and re-proposed in many forms and with many names, from OT to EMD, the Mallows distance [**55**], the Monge–Kantorovich problem [**115**], the Hitchcock–Koopmans transportation problem [**45, 50**], the Wasserstein/Vaserštein distance [**114, 36**], and undoubtedly many others.

Many credit Gaspard Monge with first formalizing the optimal transport problem in 1781 [**76**]. Beyond its early history, modern understanding of optimal transport dates back only to the World War II era, through the Nobel Prize-winning work of Leonid Kantorovich [**47**]. Jumping forward several decades, while many branches of DDG are dedicated to making centuries-old constructions on smooth manifolds work in the discrete case, optimal transport has the distinction of continuing to be an active area of research in the mathematical community whose basic properties are still being discovered. Indeed, the computational and theoretical literature in this area move in lock-step: New theoretical constructions often are adapted by the computational community in a matter of months, and some key theoretical ideas in transport were inspired by computational considerations and constructions.

Here, we aim to provide some intuition about transport and its relevance to the discrete differential geometry world. While a complete survey of work on OT or even just its computational aspects is worthy of a full textbook, here we focus on the narrower problem of how to "make transport work" on a discretized piece of geometry amenable to representation on a computer. The primary aim is to highlight the challenges in transitioning from smooth to discrete, to illustrate some basic constructions that have been proposed recently for this task, and—most importantly—to expose the plethora of open questions remaining in the relatively

young discipline of computational OT. No-doubt incomplete references are provided to selected intriguing ideas in computational OT, each of which is worthy of far more detailed discussion.

Additional reference. Those readers with limited experience in related disciplines may wish to begin by reading [**103**], a shorter survey by the author on the same topic, intended for a generalist audience.

Disclaimer. These notes are intended as a short, intuitive, and *extremely* informal introduction. Optimal transport is a popular topic in mathematical research, and interested readers should refer to surveys such as [**115, 116**] for more comprehensive discussion. The recent text [**96**] provides discussion targeted to the applied world. A few recent surveys also are targeted to computational issues in optimal transport [**57, 90**].

The author of this tutorial offers his sincere apology to those colleagues whose foundational work is undoubtedly yet accidentally omitted from this document. A "venti"-sized caffeinated beverage is humbly offered in exchange for those readers' forgiveness and understanding.

2. Motivation: From Probability to Discrete Geometry

To motivate the construction of optimal transport in the context of geometry processing, we begin by considering the case of smooth probability distributions over the real numbers $\mathbb{R}$. Here, the geometry is extremely simple, described by values $x \in \mathbb{R}$ equipped with the distance metric $d(x, y) := |x - y|$. Then we expand to define the transport problem in more generality and state a few useful properties.

2.1. The Transport Problem. Define the space of probability measures over $\mathbb{R}$ as $\text{Prob}(\mathbb{R})$. Without delving into the formalities of measure theory, these are roughly the functions $\mu \in \text{Prob}(\mathbb{R})$ assigning probabilities to sets $S \subseteq \mathbb{R}$ such that $\mu(S) \geq 0$ for all measurable S, $\mu(\mathbb{R}) = 1$, and $\mu(\cup_{i=1}^k S_i) = \sum_{i=1}^k \mu(S_i)$ for disjoint sets $\{S_i \subseteq \mathbb{R}\}_{i=1}^k$. If μ is absolutely continuous, then it admits a *distribution function* $\rho(x) : \mathbb{R} \to \mathbb{R}$ assigning a probability density to every point:

$$\mu(S) = \int_S \rho(x)\, dx.$$

Measure theory, probability, and statistics each are constructed from slightly different interpretations of the set of probability distributions $\text{Prob}(\mathbb{R})$. Adding to the mix, we can think of optimal transport as a *geometric* theory of probability. In particular, as illustrated in Figure 1, roughly a probability distribution over $\mathbb{R}$ can be thought of as a superposition of points in $\mathbb{R}$, whose weights are determined by $\rho(x)$. We can recover a (complicated) representation for a single point $x \in \mathbb{R}$ as a Dirac δ-measure centered at x.

From a physical perspective, we can think of distributions geometrically using a physical analogy. Suppose we are given a bucket of sand whose total mass is one pound. We could distribute this pound of sand across the real numbers by stacking it all at a single point, concentrating it at a few points, or spreading it out smoothly. The height of the pile of sand expresses a geometric feature: Lots of sand at a point $x \in \mathbb{R}$ indicates we think a feature is located at x.

If we wish to deepen this analogy and lift notions from geometry to the space $\text{Prob}(\mathbb{R})$, perhaps the most basic object we must define is a notion of *distance* between two distributions $\mu_0, \mu_1 \in \text{Prob}(\mathbb{R})$ that resembles the distance $d(x, y) =$

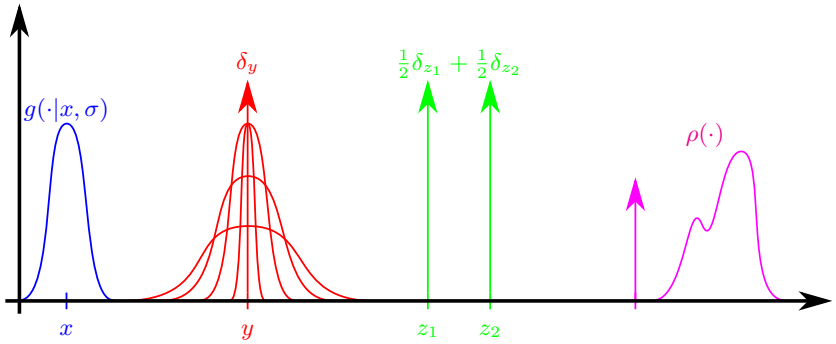


FIGURE 1. One-dimensional examples of probability distributions used to encode geometric features with uncertainty. A probability distribution like a Gaussian g with standard deviation σ can be thought of as a "fuzzy" location of a point in $x \in \mathbb{R}$. As a distribution sharpens about its mean to a δ-function δ_y, it encodes a classical piece of geometry: a point $y \in \mathbb{R}$. This language, however, is fundamentally broader, including constructions like the superposition of two points z_1 and z_2 or combining a point and a fuzzy feature into one distribution ρ.

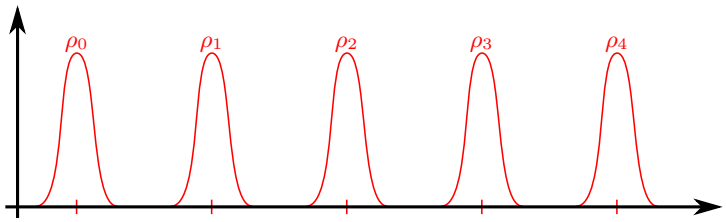


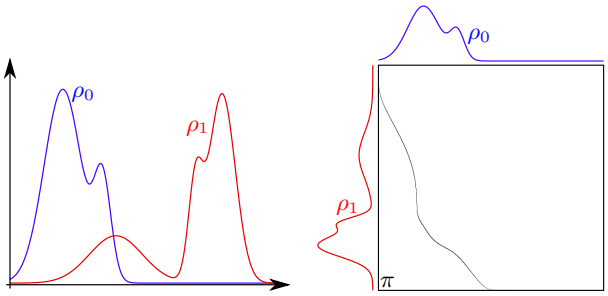
FIGURE 2. The distributions $\rho_0, \ldots, \rho_4$ are equidistant with respect to the L_1 and KL divergences, while the Wasserstein distance from optimal transport increases linearly with distance over $\mathbb{R}$.

$|x - y|$ between points on the underlying space. Supposing for now that μ_0 and μ_1 admit distribution functions ρ_0 and ρ_1, respectively, a few candidate notions of distance or divergence come to mind:

$$\begin{aligned} L_1 \text{ distance:} \qquad & d_{L_1}(\rho_0, \rho_1) := \int_{-\infty}^{\infty} |\rho_0(x) - \rho_1(x)|\, dx \\ \text{KL divergence:} \qquad & d_{\text{KL}}(\rho_0 \| \rho_1) := \int_{-\infty}^{\infty} \rho_0(x) \log \frac{\rho_0(x)}{\rho_1(x)}\, dx. \end{aligned}$$

These divergences are used widely in analysis and information theory, but they are insufficient for geometric computation. In particular, consider the distributions in Figure 2. The two divergences above give the same value for any pair of different ρ_i's! This is because they measure only the overlap; the ground distance $d(x, y) = |x - y|$ is never used in their computation.

Optimal transport resolves this issue by leveraging the physical analogy proposed above. In particular, suppose our sand is currently in arrangement ρ_0 and



(a) Source and target (b) Transport map

FIGURE 3. Two distributions over the real line (a) and the resulting transport map (b). In (b), the box is the product space $[0,1] \times [0,1]$, and dark values indicate a matching between ρ_0 and ρ_1.

we wish to *reshape* it to a new distribution ρ_1. We take a steam shovel and begin scooping up the sand at points x in ρ_0 where $\rho_0(x) > \rho_1(x)$ and dropping it places where $\rho_1(x) > \rho_0(x)$; eventually one distribution is transformed into the other.

There are many ways the steam shovel could approach its task: We could move sand efficiently, or we could drive it miles away and then drive back, wasting fuel in the process. But assuming $\rho_0 \neq \rho_1$, there is some amount of work inherent in the fact that ρ_0 and ρ_1 are not the same. We can formalize this idea by solving for an unknown measure $\pi(x, y)$ determining how much mass gets moved from x to y by the steam shovel for each (x, y) pair. The minimum amount of work is then

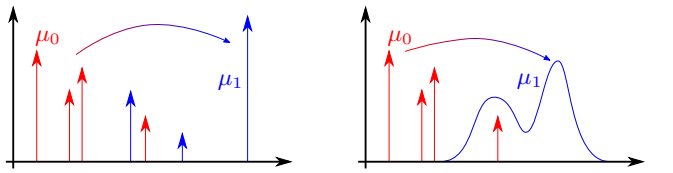
$$\mathcal{W}_1(\rho_0, \rho_1) := \begin{cases} \min_\pi \iint_{\mathbb{R}\times\mathbb{R}} \pi(x,y)|x-y|\,dx\,dy & \text{Minimize total work} \\ \text{s.t. } \pi \geq 0 \;\forall x, y \in \mathbb{R} & \text{Nonnegative mass} \\ \int_{\mathbb{R}} \pi(x,y)\,dy = \rho_0(x) \;\forall x \in \mathbb{R} & \text{Starts from } \rho_0 \\ \int_{\mathbb{R}} \pi(x,y)\,dx = \rho_1(y) \;\forall y \in \mathbb{R} & \text{Ends at } \rho_1. \end{cases} \tag{2.1}$$

This optimization problem quantifies the minimum amount of work—measured as mass $\pi(x, y)$ times distance traveled $|x - y|$—required to transform ρ_0 into ρ_1. We can think of the unknown function π as the instructions given to the laziest possible steam shovel tasked with dropping one distribution onto another. This amount of work is known as the *1-Wasserstein distance* in optimal transport; in one dimension, it equals the L_1 distance between the cumulative distribution functions of ρ_0 and ρ_1. An example of ρ_0, ρ_1, and the resulting π is shown in Figure 3.

Generalizing slightly, we can define the p-Wasserstein distance:

$$[\mathcal{W}_p(\rho_0, \rho_1)]^p := \begin{cases} \min_\pi & \iint_{\mathbb{R}\times\mathbb{R}} \pi(x,y)|x-y|^p\,dx\,dy \\ \text{s.t.} & \pi \geq 0 \;\forall x, y \in \mathbb{R} \\ & \int_{\mathbb{R}} \pi(x,y)\,dy = \rho_0(x) \;\forall x \in \mathbb{R} \\ & \int_{\mathbb{R}} \pi(x,y)\,dx = \rho_1(y) \;\forall y \in \mathbb{R}. \end{cases} \tag{2.2}$$

In analogy to Euclidean space, many properties of $\mathcal{W}_p$ are split into cases $p < 1$, $p = 1$, and $p > 1$; for instance, it satisfies the triangle inequality any time $p \geq 1$. The $p = 2$ case is of particular interest in the literature and corresponds to a "least-squares" version of transport that minimizes kinetic energy rather than work (see §2.4). Generalizing (2.2) even more, if we replace $|x-y|^p$ with a generic cost $c(x, y)$ we recover the *Kantorovich problem* [47].



(a) Fully discrete transport (b) Semidiscrete transport

FIGURE 4. Discrete (a) and semidiscrete (b) optimal transport in one dimension.

It is important to note an alternative formulation of the transport problem (2.2), which historically was posed first but does not always admit a solution. Rather than optimizing for a function $\pi(x, y)$ with an unknown for every possible (x, y) pair, one could consider an alternative in which instead the variable is a single function $\phi(x)$ that "pushes forward" ρ_0 onto ρ_1; this corresponds to choosing a single destination $\phi(x)$ for every source point x. In this case, the objective function would look like

$$\int_{-\infty}^{\infty} |\phi(x) - x|^p \rho_0(x)\, dx, \tag{2.3}$$

and the constraints would ask that the image of ρ_0 under ϕ is ρ_1, notated $\phi_\sharp \rho_0 = \rho_1$. While this version corresponds to the original version of transport proposed by Monge, sometimes for the transport problem to be solvable it is necessary to split the mass at a single source point to multiple destinations. A triumph of theoretical optimal transport, however, shows that $\pi(x, y)$ is nonzero only on some set $\{(x, \phi(x)) : x \in \mathbb{R}\}$ whenever ρ_0 is absolutely continuous, linking the two problems.

2.2. Discrete Problems in One Dimension. So far our definitions have not been amenable to numerical computation: Our unknowns are functions $\pi(x, y)$ with *infinite* numbers of variables (one value of π for each (x, y) pair in $\mathbb{R} \times \mathbb{R}$)—certainly more than can be stored on a computer with finite capacity. Continuing to work in one dimension, we suggest some special cases where we can solve the transport problem with a finite number of variables.

Rather than working with distribution functions $\rho(x)$, we will relax to the more general case of transport between measures $\mu_0, \mu_1 \in \text{Prob}(\mathbb{R})$. Define the Dirac δ-measure centered at $x \in \mathbb{R}$ via

$$\delta_x(S) := \begin{cases} 1 & \text{if } x \in S \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to check that $\delta_x(\cdot)$ is a probability measure.

Suppose $\mu_0, \mu_1 \in \text{Prob}(\mathbb{R})$ can be written as *superpositions* of δ measures:

$$\mu_0 := \sum_{i=1}^{k_0} a_{0i} \delta_{x_{0i}} \qquad \text{and} \qquad \mu_1 := \sum_{i=1}^{k_1} a_{1i} \delta_{x_{1i}}, \tag{2.4}$$

where $1 = \sum_{i=1}^{k_0} a_{0i} = \sum_{i=1}^{k_1} a_{1i}$ and $a_{0i}, a_{1i} \geq 0$ for all i. Figure 4(a) illustrates this case; all the mass of μ_0 and μ_1 is concentrated at a few isolated points.

In the case where the source and target distributions are composed of δ's, we only can move mass between pairs of points $x_{0i} \mapsto x_{1j}$. Taking T_{ij} the total mass

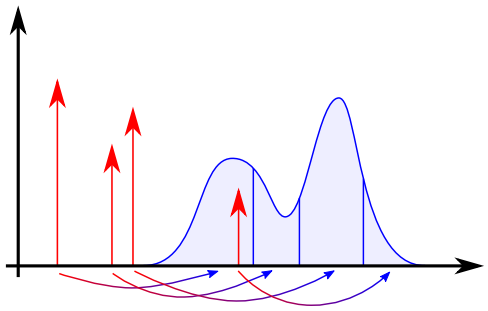


FIGURE 5. Solving 1D semidiscrete transport from Figure 4(b); every Dirac δ-function mass in the source μ_0 gets mapped to a contiguous interval worth of mass in the target μ_1.

moved from x_{0i} to x_{1j}, we can solve for $\mathcal{W}_p^p$ as

$$[\mathcal{W}_p(\mu_0,\mu_1)]^p = \left\{ \begin{array}{rl} \min_{T\in\mathbb{R}^{k_0\times k_1}} & \sum_{ij} T_{ij}|x_{0i}-x_{1j}|^p \\ \text{s.t.} & T \geq 0 \\ & \sum_j T_{ij} = a_{0i} \\ & \sum_i T_{ij} = a_{1j}. \end{array} \right. \tag{2.5}$$

This is an optimization problem in k_0k_1 variables T_{ij}: No need for an infinite number of $\pi(x,y)$'s! In fact, it is a *linear program* solvable using many classic algorithms, such as the simplex or interior point methods.

There is a more subtle case where we can still represent the unknown in optimal transport using a finite number of variables. Suppose $\mu_0 \in \text{Prob}(\mathbb{R})$ is a superposition of δ measures and $\mu_1 \in \text{Prob}(\mathbb{R})$ is absolutely continuous, implying μ_1 admits a distribution function $\rho_1(x)$:

$$\mu_0 := \sum_{i=1}^{k} a_i \delta_{x_i} \qquad \text{and} \qquad \mu_1(S) := \int_S \rho_1(x)\,dx. \tag{2.6}$$

This situation is illustrated in Figure 4(b); it corresponds to transporting from a distribution whose mass is concentrated at a few points to a distribution whose distribution is more smooth. In the technical literature, this setup is known as *semidiscrete* transport.

Returning to the transport problem in (2.2), in this semidiscrete case we can think of the coupling π as decomposing into a set of measures $\pi_1, \pi_2, \ldots, \pi_k \in \text{Prob}(\mathbb{R})$ where each term in the sum (2.6) has its own target distribution: $\delta_{x_i} \mapsto \pi_i$. As a sanity check, note that $\mu_1 = \sum_i a_i \pi_i(x)$.

Without loss of generality, we can assume the x_i's are sorted, that is, $x_1 < x_2 < \cdots < x_k$. Suppose $1 \leq i < j \leq k$, and hence $x_i < x_j$. In one dimension, it is easy to see that the optimal transport map π should never "leapfrog" mass, that is, the delivery target of the mass at x_i when transported to ρ_1 should be *to the left* of the target of mass at x_j, as illustrated in Figure 5. This monotonicity property implies the existence of intervals $[b_1,c_1],[b_2,c_2],\ldots,[b_k,c_k]$ such that π_i is supported in $[b_i,c_i]$ and $c_i < b_j$ whenever $i < j$; the mass $a_i\delta_{x_i}$ is distributed according to $\rho_1(x)$ in the interval $[b_i,c_i]$.

The semidiscrete transport problem corresponds to another case where we can solve a transport problem with a finite number of variables, the b_i's and c_i's. Of

course, in one dimension these can be read off from the cumulative distribution function (CDF) of ρ_1, but in higher dimensions this will not be the case. Instead, the intervals $[b_i, c_i]$ will be replaced with *power cells*, a generalization of a Voronoi diagram (§4.3).

While our discussion above gives two cases in which a computer could plausibly solve the transport problem, they do not correspond to the usual situation for DDG in which the geometry itself—in this case the real line $\mathbb{R}$—is discretized. As we will see in the discussion in future sections, there currently does not exist consensus about what to do in this case but several possible adaptations to this case have been proposed.

2.3. Moving to Higher Dimensions. We are now ready to state the optimal transport problem in full generality. Following [**115**, §1.1.1], take (X, μ) and (Y, ν) to be probability spaces, paired with a nonnegative measurable cost function $c(x, y)$. Define a *measure coupling* $\pi \in \Pi(\mu, \nu)$ as follows:

Definition 2.1 (Measure coupling). A *measure coupling* $\pi \in \text{Prob}(X \times Y)$ is a probability measure on $X \times Y$ satisfying

$$\pi(A \times Y) = \mu(A)$$
$$\pi(X \times B) = \nu(B)$$

for all measurable $A \subseteq X$ and $B \subseteq Y$. The set of measure couplings between μ and ν is denoted $\Pi(\mu, \nu)$.

With this piece of notation, we can write the Kantorovich optimal transport problem as follows:

$$\boxed{\text{OT}(\mu, \nu; c) := \min_{\pi \in \Pi(\mu,\nu)} \iint_{X \times Y} c(x, y)\, d\pi(x, y)} \tag{2.7}$$

Here, we use some notation from measure theory: $d\pi(x, y)$ denotes integration against probability measure π. Note if π admits a distribution function $p(x, y)$ then we can write $d\pi(x, y) = p(x, y)\, dx\, dy$; the more general notation allows for δ measures and other objects that cannot be written as functions.

We note a few interesting special cases below:

Discrete transportation. Suppose $X = \{1, 2, \ldots, k_1\}$ and $Y = \{1, 2, \ldots, k_2\}$. Then, $\mu \in \text{Prob}(X)$ can be written as a vector $v \in \mathbb{S}_{k_1}$ and $\nu \in \text{Prob}(Y)$ can be written as a vector $w \in \mathbb{S}_{k_2}$, where $\mathbb{S}_k$ denotes the k-dimensional probability simplex:

$$\mathbb{S}_k := \left\{ v \in \mathbb{R}^k : v \geq 0 \text{ and } \sum_i v_i = 1 \right\}. \tag{2.8}$$

Our cost function becomes discrete as well and can be written as a matrix $C = (c_{ij})$. After simplification, the transport problem between $v \in \mathbb{S}_{k_1}$ and $w \in \mathbb{S}_{k_2}$ given cost matrix C becomes

$$\text{OT}(v, w; C) = \left\{ \begin{array}{ll} \min_{T \in \mathbb{R}^{k_1 \times k_2}} & \sum_{ij} T_{ij} c_{ij} \\ \text{s.t.} & T \geq 0 \\ & \sum_j T_{ij} = v_i \ \forall i \in \{1, \ldots, k_1\} \\ & \sum_i T_{ij} = w_j \ \forall j \in \{1, \ldots, k_2\}. \end{array} \right. \tag{2.9}$$

This linear program is solvable computationally and is the most obvious way to make optimal transport work in a discrete context. It was proposed in the computational literature as the "earth mover's distance" (EMD) [94]. When $k_1 = k_2 := k$ and C is symmetric, nonnegative, and satisfies the triangle inequality, one can check that $\mathrm{OT}(\cdot,\cdot;C)$ is a distance on $\mathbb{S}_k$; see [29] for a clear proof of this property.

Wasserstein distance. Next, suppose $X = Y = \mathbb{R}^n$, and take $c_{n,p}(x,y) := \|x-y\|_2^p$. Then, we recover the *Wasserstein distance* on $\mathrm{Prob}(\mathbb{R}^n)$, defined via

$$\mathcal{W}_p(\mu,\nu) := [\mathrm{OT}(\mu,\nu;c_{n,p})]^{1/p}. \tag{2.10}$$

$\mathcal{W}_p$ is a distance when $p \geq 1$, and $\mathcal{W}_p^p$ is a distance when $p \in [0,1)$ [115, §7.1.1]. In fact, the Wasserstein distance can be defined for probability measures over a surface, Riemannian manifold, or even a Polish space via the same formula.

The Wasserstein distance has drawn considerable application-oriented interest and aligns well with the basic motivation laid out in §1. Its basic role is to lift distances between points $\|x-y\|_2^p$ to distances between distributions in a compatible fashion: The Wasserstein distance between two δ-functions δ_x and δ_y is exactly the distance $\|x-y\|_2$. In §3, we will show how this basic property has strong bearing on several computational pipelines that need to lift geometric constructions to uncertain contexts.

2.4. One Value, Many Formulas. A remarkable property of the transport problem (2.7) is the sheer number of equivalent formulations that all lead to the same value, the cost of transporting mass from one measure onto another. These not only provide many interpretations of the transport problem but also suggest a diverse set of computational algorithms for transport, each of which tackles a different way of writing down the basic problem.

Duality. A basic idea in the world of convex optimization is that of *duality*, that every minimization problem admits a "dual" maximization problem whose optimal value lower-bounds that of the primal. As with most linear programs, optimal transport typically exhibits *strong duality*: The optimal values of the maximization and minimization problems coincide.

To motivate duality for transport, we will start with the finite-dimensional problem (2.9). We note two simple identities:

$$\max_{s\in\mathbb{R}} st = \begin{cases} 0 & \text{if } t = 0 \\ \infty & \text{otherwise} \end{cases} \qquad \max_{s\leq 0} st = \begin{cases} 0 & \text{if } t \geq 0 \\ \infty & \text{otherwise} \end{cases}$$

These allow us to write (2.9) as follows:

$$\min_T \max_{S\leq 0,\phi,\psi} \left[\sum_{ij} T_{ij}(c_{ij}+S_{ij}) + \sum_i \phi_i \left(v_i - \sum_j T_{ij} \right) + \sum_j \psi_j \left(w_j - \sum_i T_{ij} \right) \right].$$

The dual problem is derived by simply swapping the min and the max:

$$\max_{S\leq 0,\phi,\psi} \min_T \left[\sum_{ij} T_{ij}(c_{ij}+S_{ij}) + \sum_i \phi_i \left(v_i - \sum_j T_{ij} \right) + \sum_j \psi_j \left(w_j - \sum_i T_{ij} \right) \right]$$

$$= \max_{S\leq 0,\phi,\psi} \min_T \left[\sum_{ij} T_{ij}(c_{ij}+S_{ij}-\phi_i-\psi_j) + \sum_i \phi_i v_i + \sum_j \psi_j w_j \right] \text{ after refactoring.}$$

Since T is unbounded in the inner optimization problem of the dual, the solution of the inner minimization is $-\infty$ unless $S_{ij} = \phi_i + \psi_j - c_{ij}$ for all (i, j), that is, unless the coefficient of T_{ij} equals zero. Since the outer problem is a maximization, clearly we should avoid an optimal value of $-\infty$ for the inner minimization. Hence, we can safely add $S_{ij} = \phi_i + \psi_j - c_{ij}$ as a constraint to the dual problem. After some simplification, we arrive at the dual of (2.9):

$$\begin{array}{rl} \max_{\phi,\psi} & \sum_i [\phi_i v_i + \psi_i w_i] \\ \text{s.t.} & \phi_i + \psi_j \leq c_{ij} \ \forall (i,j). \end{array} \tag{2.11}$$

Although we have not justified that it is acceptable to swap a max and a min in this context, several techniques ranging from direct proof to the "sledgehammer" Slater duality condition [**102**] show that the optimal value of this maximization problem agrees with the optimal value of the minimization problem (2.9).

As is often the case in convex optimization, the dual (2.11) of the transport problem (2.9) has an intuitive interpretation. Suppose we change roles in optimal transport from the worker who wishes to minimize work to a company that wishes to maximize profit. The customer pays ϕ_i dollars per pound to drop off material v_i to ship from location i and ψ_j dollars per pound to pick up material w_j from location j. The dual problem (2.11) maximizes profit under the constraint that it is never cheaper for the customer to just drive from i to j and ignore the service completely: $\phi_i + \psi_j \leq c_{ij}$.

We pause here to note some rough trade-offs between the primal and dual transport problems. Since both yield the same optimal value, the designer of a computational system for solving optimal transport problems has a decision to make: whether to solve the primal problem, the dual problem, or both simultaneously (the latter aptly named a "primal–dual" algorithm). There are advantages and disadvantages to each approach. The primal problem (2.9) directly yields the matrix T, which tells not just the cost of transport but how much mass T_{ij} to move from source i to destination j; the only inequality constraint is that the entire matrix has nonnegative entries. On the other hand, the dual problem (2.11) has fewer variables, making it easier to store the output on the computer, but the "shadow price" variables (ϕ, ψ) are harder to interpret and are constrained by a quadratic number of inequalities. Currently there is little consensus as to which formulation leads to more successful algorithms or discretizations, and state-of-the-art techniques are divided among the two basic approaches.

As with many constructions in optimal transport, the dual of the measure-theoretic problem (2.7) resembles the discrete case up to a change of the notation. In particular, we can write

$$\boxed{\mathrm{OT}(\mu, \nu; c) := \left\{ \begin{array}{rl} \sup_{\substack{\phi \in L^1(d\mu), \\ \psi \in L^1(d\nu)}} & \int_X \phi(x)\, d\mu(x) + \int_Y \psi(y)\, d\nu(y) \\ \text{s.t.} & \phi(x) + \psi(y) \leq c(x,y) \\ & \quad \text{for } d\mu\text{-a.e. } x \in X,\ d\nu\text{-a.e. } y \in Y. \end{array} \right.} \tag{2.12}$$

It is worth noting a simplification that appears often in the transport world. Since μ and ν are positive measures and the overall problem in (2.12) is a maximization, we might as well choose ϕ and ψ as large as possible while satisfying the constraints. Suppose we fix the function $\phi(x)$ and *just* optimize for the function $\psi(x)$. Rearranging the constraint shows that for all $(x, y) \in X \times Y$ we must have $\psi(y) \leq c(x, y) - \phi(x)$. Equivalently, for all $y \in Y$ we must have

$\psi(y) \le \inf_{x\in X}[c(x,y) - \phi(x)]$. Define the *c-transform*

$$\phi^c(y) := \inf_{x\in X}[c(x,y) - \phi(x)]. \tag{2.13}$$

By the argument above we have

$$\mathrm{OT}(\mu,\nu;c) = \sup_{\phi\in L^1(d\mu)} \int_X \phi(x)\, d\mu(x) + \int_Y \phi^c(y)\, d\nu(y).$$

This problem is unconstrained, but the transformation from ϕ to ϕ^c is relatively complicated.

We finally note one special case of this dual formula, the 1-Wasserstein distance, which has gained recent interest in the machine learning world thanks to its application in generative adversarial networks (GANs) [4]. In this case, $X = Y = \mathbb{R}^n$ and $c(x,y) = \|x-y\|_2$. We can derive a bound as follows:

$$\begin{aligned}
|\phi^c(x) - \phi^c(y)| &= \left|\inf_z[\|x-z\|_2 - \phi(z)] - \inf_z[\|y-z\|_2 - \phi(z)]\right| \text{ by definition} \\
&\le \sup_z \big|\|x-z\|_2 - \|y-z\|_2\big| \\
&\qquad \text{by the identity } |\inf_x f(x) - \inf_x g(x)| \le \sup_x |f(x) - g(x)| \\
&\le \|x-y\|_2 \text{ by the reverse triangle inequality.} \qquad (2.14)
\end{aligned}$$

Furthermore, by definition of the c-transform (2.13) by taking $x = y$ we have $\phi^c(y) \le -\phi(y)$, or equivalently $\phi(y) \le -\phi^c(y)$. Hence,

$$\begin{aligned}
\mathcal{W}_1(\mu,\nu) &= \mathrm{OT}(\mu,\nu;c) \text{ through our choice } c(x,y) := \|x-y\|_2 \\
&= \sup_{\phi\in L^1(d\mu)} \int_{\mathbb{R}^n} \phi(x)\, d\mu(x) + \int_{\mathbb{R}^n} \phi^c(y)\, d\nu(y) \text{ by definition of the } c\text{-transform} \\
&\le \int_{\mathbb{R}^n} \phi^c(x)\,[d\nu(x) - d\mu(x)] \text{ since } \phi(y) \le -\phi^c(y)\ \forall y \in \mathbb{R}^n \\
&\le \sup_{\psi\in \mathrm{Lip}_1(\mathbb{R}^n)} \int_{\mathbb{R}^n} \psi(x)\,[d\nu(x) - d\mu(x)] \\
&\qquad \text{where } \mathrm{Lip}_1(\mathbb{R}^n) := \{f(x) : |f(x) - f(y)| \le \|x-y\|_2\ \forall x,y \in \mathbb{R}^n\}.
\end{aligned}$$

Lip_1 denotes the set of 1-Lipschitz functions; the last step is derived from (2.14), which shows that ψ^c is 1-Lipschitz.

In fact, this inequality is an equality. To prove this, take ψ to be any 1-Lipschitz function. Then,

$$\psi^c(y) = \inf_{x\in\mathbb{R}^n}[\|x-y\|_2 - \psi(x)] \ge \inf_{x\in\mathbb{R}^n}[\|x-y\|_2 - \|x-y\|_2 - \psi(y)] = -\psi(y). \tag{2.15}$$

where we have rearranged the Lipschitz property $\psi(x) - \psi(y) \le \|x-y\|_2$ to show $-\psi(x) \ge -\|x-y\|_2 - \psi(y)$. Hence,

$$\begin{aligned}
\sup_{\psi\in\mathrm{Lip}_1(\mathbb{R}^n)} \int_{\mathbb{R}^n} \psi(x)\,[d\nu(x) - d\mu(x)] &\le \sup_{\psi\in\mathrm{Lip}_1(\mathbb{R}^n)} \int_{\mathbb{R}^n} [\psi(x)\, d\nu(x) + \psi^c(y)]\, d\mu(y) \text{ by (2.15)} \\
&\le \sup_{\psi\in L^1(d\nu)} \int_{\mathbb{R}^n} [\psi(x)\, d\nu(x) + \psi^c(y)]\, d\mu(y) \\
&\qquad \text{since the constraints are loosened} \\
&= \mathcal{W}_1(\mu,\nu).
\end{aligned}$$

This finishes motivating our final formula

$$\boxed{\mathcal{W}_1(\mu,\nu) = \sup_{\psi\in\mathrm{Lip}_1(\mathbb{R}^n)} \int_{\mathbb{R}^n} \psi(x)\,[d\nu(x) - d\mu(x)].}$$

This convenient identity is used in computational contexts because the constraint that a function is 1-Lipschitz is fairly easy to enforce computationally; sadly, it does not extend to other Wasserstein $\mathcal{W}_p$ distances, which have nicer uniqueness and regularity properties when $p > 1$.

Eulerian transport. The language of fluid dynamics introduces two equivalent ways to understand the flow of a liquid or gas as it sloshes in a tank. In the *Lagrangian* framework, the fluid is thought of as a collection of particles whose path we trace as a function of time; the equations of motion roughly determine a map $\Phi_t(x)$ with $\Phi_0(x) = x$ determining the position at time $t \geq 0$ of the particle located at x when $t = 0$. Contrastingly, *Eulerian* fluid dynamics takes the point of view of a barnacle attached to a point in the tank of water counting the number of particles that flow past a point x; this formulation might seek a function $\rho_t(x)$ giving the density of the fluid at a non-moving point x as a function of time t.

So far, our formulation of transport has been Lagrangian: The transportation plan π explicitly determines how to match particles from the source distribution μ to the target distribution ν. Using a particularly clever change of variables, a landmark paper by Benamou & Brenier shows that the 2-Wasserstein distance from (2.10) over Euclidean space with cost $c(x,y) = \|x-y\|_2^2$ can be computed in an Eulerian fashion [**9**]:

$$\mathcal{W}_2^2(\rho_0,\rho_1) = \begin{cases} \min_{v(x,t),\rho(x,t)} & \frac{1}{2}\int_0^1 \int_{\mathbb{R}^n} \rho(x,t)\,\|v(x,t)\|_2^2\,dA(x)\,dt \\ \text{s.t.} & \rho(x,0) \equiv \rho_0(x)\ \forall x \in \mathbb{R}^n \\ & \rho(x,1) \equiv \rho_1(x)\ \forall x \in \mathbb{R}^n \\ & \frac{\partial\rho(x,t)}{\partial t} = -\nabla\cdot(\rho(x,t)v(x,t)) \\ & \forall x \in \mathbb{R}^n, t \in (0,1) \end{cases} \tag{2.16}$$

Here, we assume that we are computing the 2-Wasserstein distance between two distribution functions $\rho_0(x)$ and $\rho_1(x)$. This is often referred to as a *dynamical* model of transport and can be extended to spaces like Riemannian manifolds [**65**].

Formulation (2.16) comes with an intuitive physical interpretation. The time-varying function $\rho(x,t)$ gives the density of a gas as a function of time $t \in [0,1]$, which starts out in configuration ρ_0 and ends in configuration ρ_1. The constraint $\frac{\partial\rho}{\partial t} = -\nabla\cdot(\rho v)$ is the *continuity equation*, which states that the vector field $v(x,t)$ is the *velocity* of ρ as it moves as a function of time while preserving mass. Over all possible ways to "animate" the motion from ρ_0 to ρ_1, the objective function minimizes $\frac{1}{2}\rho\|v\|_2^2$ (mass times velocity squared): the total kinetic energy!

From a computational perspective, it can be convenient to replace velocity v with momentum $J := \rho\cdot v$ to obtain an equivalent formulation to (2.16):

$$\mathcal{W}_2^2(\rho_0,\rho_1) = \begin{cases} \min_{J(x,t),\rho(x,t)} & \frac{1}{2}\int_0^1 \int_{\mathbb{R}^n} \frac{\|J(x,t)\|_2^2}{\rho(x,t)}\,dA(x)\,dt \\ \text{s.t.} & \rho(x,0) \equiv \rho_0(x)\ \forall x \in \mathbb{R}^n \\ & \rho(x,1) \equiv \rho_1(x)\ \forall x \in \mathbb{R}^n \\ & \frac{\partial\rho(x,t)}{\partial t} = -\nabla\cdot J(x,t) \\ & \forall x \in \mathbb{R}^n, t \in (0,1) \end{cases} \tag{2.17}$$

This formulation is convex jointly in the unknowns (ρ, J).

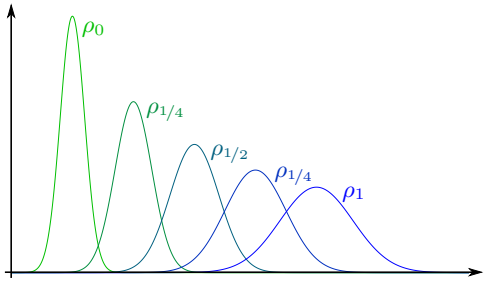


FIGURE 6. Displacement interpolation from ρ_0 to ρ_1 explains optimal transport between these two densities using a time-varying density function ρ_t, $t \in [0, 1]$.

Dynamical formulations of transport make explicit the phenomenon of *displacement interpolation* [**66, 64**], illustrated in Figure 6. Intuitively, the Wasserstein distance $\mathcal{W}_2$ between two distribution functions ρ_0 and ρ_1 is "explained" by a time-varying sequence of distributions ρ_t interpolating from one to the other. Unlike the trivial interpolation $\rho(t) := (1-t)\rho_0(x) + t\rho_1(x)$, optimal transport *slides* the distribution across the geometric domain similar to a geodesic shortest path between points on a curved manifold. Indeed, the intuitive connection to differential geometry is more than superficial: [**82, 61**] show how to interpret (2.16) as a geodesic in an infinite-dimensional manifold of probability distributions over a fixed domain.

Other p-Wasserstein distances $\mathcal{W}_p$ also admit Eulerian formulations. Most importantly, the 1-Wasserstein distance can be computed as follows:

$$\mathcal{W}_1(\rho_0, \rho_1) = \left\{ \begin{array}{ll} \min_{J(x)} & \int_{\mathbb{R}^n} \|J(x)\|_2 \, dA(x) \\ \text{s.t.} & \nabla \cdot J(x) = \rho_1(x) - \rho_0(x). \end{array} \right. \tag{2.18}$$

This problem, known as the *Beckmann problem*, has connections to traffic modeling and other tasks in geometry. From a computational perspective, it has the useful property that the vector field $J(x)$ has no time dependence, reducing the number of unknown variables in the optimization problem.

3. Motivating Applications

Having developed the basic definition and theoretical properties of the optimal transport problem, we can now divert from theoretical discussion to mention some concrete applications of transport in the computational world. These are just a few, chosen for their diversity (and no doubt biased toward areas adjacent to the author's research); in reality optimal transport is beginning to appear in a huge variety of computational pipelines. Our goal in this section is not to give the details of each problem and its resolution with transport, but just to give a flavor of how optimal transport can be applied as a powerful modeling tool in application-oriented disciplines as well as citations to more detailed treatments of each application.

Operations and logistics. Given its history and even its name, it comes as no surprise that a primary application of optimal transport is in the operations and logistics world, in which engineers are asked to find a minimum-cost routing of packages or materials to customers. The basic theory and algorithms for this case of optimal transport date back to World War II, in which optimal transport of soldiers, weapons, supplies, and the like were by no means theoretical problems.

A particular case of interest in this community is that of transport over a graph $G = (V, E)$. Here, shortest-path distances over the edges of G provide the costs for transport, leading to a problem known to computer scientists as *minimum-cost flow without edge capacities* [**93**]. This linear program is a classic algorithmic problem, with well-known algorithms including cycle canceling [**49**], network simplex [**81**], and the Ford–Fulkerson method [**40**]. A challenge for theoretical computer scientists is to design algorithms achieving lower-bound time complexity for solving this problem; recent progress includes [**100**], which achieves a near-linear runtime using an approach that almost resembles a numerical algorithm rather than a discrete method.

Histogram-based descriptors. Some of the earliest applications of optimal transport in the computational world come from computer vision [**94**]. Suppose we wish to perform *similarity search* on a database of photographs: Given one photograph, we wish to search the database for other photos that look similar. One reasonable way to do this is to describe each photograph as a histogram—or probability distribution—over the space of colors. Two photographs roughly look similar if they have similar color histograms as measured using optimal transport distances (known in this community as the "Earth Mover's Distance"), giving a simple technique for sorting and searching the dataset.

This basic approach comes up time and time again in the applied world. For images, rather than binning colors into a histogram one could bin the orientations and strengths of the gradients to capture the distribution of edge features [**85**]. Recent work has proposed an embedding of the words in an English dictionary into Euclidean space $\mathbb{R}^n$ [**72**], in which case the words present in a given document become a point cloud or superposition of δ-functions in $\mathbb{R}^n$; application of the Wasserstein ("Word-Movers") distance in this case is an effective technique for document retrieval [**53**].

Registration. Suppose we wish to use a medical imaging device such as the MRI to track the progress of a neurodegenerative disease. On a regular basis, we might ask the subject to return to the laboratory for a brain scan, each time measuring a signal over the volume of the MRI indicating the presence or absence of brain tissue. These signals can vary drastically from visit to visit, not just due to the progress of the disease but also due to more mundane issues like movement of the patient in the measurement device or nonrigid deformation of the brain itself.

Inspired by issues like those mentioned above, the task of computing a map from one scan to another is known as *registration*, and optimal transport has been proposed time and time again as a tool for this task. The basic idea of these tools is to use the transport map π as a natural way to transfer information from one scan to another [**43**]. One caveat is worth highlighting: Optimal transport does not penalize splitting mass or making non-elastic deformations in the optimal map, so long as points of mass individually do not move too far. A few recent methods attempt to cope with this final issue, e.g. by combining transport with an elastic deformation method more common in medical imaging [**39**] or by defining modified versions of optimal transport that are invariant to certain species of deformation [**24, 68, 106**].

Distance approximation. A predictable property of the p-Wasserstein distance $\mathcal{W}_p$ for distributions over a surface or manifold $\mathcal{M}$ is that the distance between δ-functions centered at two points $x_0, x_1 \in \mathcal{M}$ reproduces the geodesic (shortest-path) distance from x_0 to x_1. While distances in Euclidean space are computable

FIGURE 7. Level sets of geodesic distance to the front right toe of a 3D camel model approximated using the optimal transport technique [**107**].

FIGURE 8. A blue noise pattern generated using [**32**] (image courtesy F. de Goes, generated from photograph by F. Durand).

using a closed-form formula, distances along discretized surfaces can be challenging to compute algorithmically, requiring techniques like fast marching [**99**], the theoretically-justified but difficult-to-implement MMP algorithm [**75**], or diffusion-based approximation [**27**]. In this regime, fast algorithms for approximating optimal transport distances $\mathcal{W}_p$ restricted to δ-functions actually provide a way to approximate geodesic distances while preserving the triangle inequality [**107**]; the level sets of one such approximation are shown in Figure 7.

Blue noise and stippling. Certain laser printers and other devices can only print pages in black-and-white—no gray. The idea of *halftoning* is that gray values between black and white can be approximated in a perceptually reasonable fashion by patterns of black dots of varying radius or location over a white background; the halftoned image can be printed using the black-and-white printer and from a distance appears similar to the original.

A reasonable model for halftoning involves optimal transport. In particular, suppose we think of a grayscale image as a *distribution* of ink on a white page; that is, the image can be understood as a measure $\mu \in \text{Prob}([0,1]^2)$, where $[0,1]^2$ is the unit rectangle representing the image plane. Under the reasonable assumption that

ink is conserved, we might attempt to approximate μ with a set of dots of black inks, modeled using δ-functions centered at x_i. The intensity of the dot cannot be modulated (the printer only knows how to print in black-and-white), but the location can be moved, leading to an optimization problem to the effect:

$$\min_{x_1,\ldots,x_n} \mathcal{W}_2^2\left(\mu, \frac{1}{n}\sum_i \delta_{x_i}\right).$$

Here, the variables are the locations of the n dots approximating the image, and the Wasserstein 2-distance is used to measure how well the dots approximate μ. This basic idea is extended in [**32**] to a pipeline for computing *blue noise*; an example of their output is shown in Figure 8.

Political redistricting. A few recent attempts to propose political redistricting procedures have incorporated ideas from optimal transport to varying degrees of success. For example, optimal transport might provide one simplistic means of assigning voters to poling centers. The distribution of voters over a map is "transported" to a sparse set of polling places, where distributional constraints reflect the fact that each polling center can only handle so many voters; assigning each voter to his/her closest polling center might cause polling centers in city centers to become overloaded. A few papers have proposed variations on this idea to design compact voting districts e.g. for the US House of Representatives [**111, 73, 25, 1**]. Many confounding—but incredibly important—factors obscure the application of this simplistic mathematical model in practice, ranging from compliance with civil rights law to the simple decision of a transport cost (e.g. geographical versus road network versus public transportation versus travel time).

Statistical estimation. Parameter estimation is a key task in statistics that involves "explaining" a given dataset using a statistical model. For example, given the set of heights of people in a room $\{h_1, \ldots, h_n\}$, a simple parameter estimation task might be to estimate the mean h_0 and standard deviation σ of a normal (bell curve/Gaussian) distribution $g(h|h_0, \sigma)$ from which the data was likely sampled.

Principal among the techniques for parameter estimation is the *maximum likelihood estimator* (MLE). Continuing in our height data example, assuming the n heights are drawn independently, the joint probability of observing the given set of heights in the room is given by the product

$$P(h_1, \ldots, h_n|h_0, \sigma) = \prod_{i=1}^{n} g(h_i|h_0, \sigma).$$

The MLE of the data is the estimate of (h_0, σ) that maximizes this probability value:

$$(h_0, \sigma)_{\text{MLE}} := \arg\max_{h_0,\sigma} P(h_1, \ldots, h_n|h_0, \sigma).$$

For algebraic reasons it is often easier to maximize the *log likelihood* $\log P(\cdots)$, although this is obviously equivalent to the formulation above.

As an alternative to the MLE, however, the *minimum Kantorovich estimator* (MKE) [**8**] uses machinery from optimal transport. As the name suggests, the MKE estimates the parameters of a distribution by minimizing the transport distance between the parameterized distribution and the empirical distribution from data.

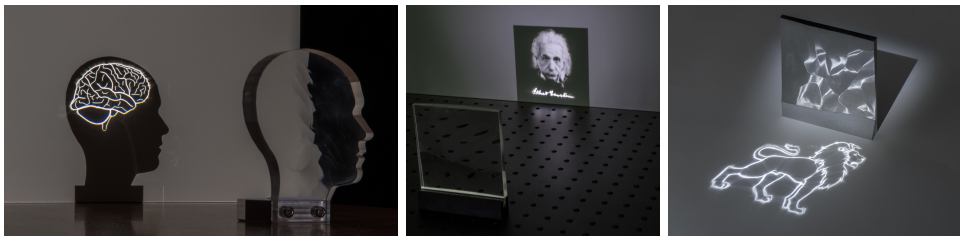


FIGURE 9. Optimal transport is used to design the shape of transparent or reflective material to show a particular caustic pattern (image courtesy of EPFL Computer Graphics and Geometry Laboratory and Rayform SA).

For our height problem, the optimization might look like

$$(h_0, \sigma)_{\text{MKE}} := \arg\min_{h_0,\sigma} \mathcal{W}_2^2 \left(\frac{1}{n} \sum_i \delta_{h_i}, g(\cdot|h_0, \sigma) \right)$$

The differences between MLE, MKE, and other alternatives can be subtle from the outside looking in, and the MKE is only recently being studied in applied environments in comparison to more conventional alternatives. Since it takes into account the distance measure of the geometric space on which the samples are defined, the MKE appears to be robust to geometric noise that can confound more traditional alternatives—at the price of increased computational expense. Recent applications have shown value of this estimator for training and inference in machine learning models [77, 14].

Domain adaptation. Many basic statistical and machine learning algorithms make a false assumption that the "training" and "test" data are distributed equally. As an example where this is not the case, suppose we wish to make an object recognition tool that learns how to label the contents of a photograph. As training data, we use the listings on an e-commerce site like Amazon.com, which contain not only a photographs of a given object but also text describing it. But, while this training data is extremely clean, it is not representative of possible test data, e.g. gathered by a robot navigating a shopping mall: Photographs collected by the latter likely contain clutter, a variety of lighting configurations, and countless other confounding factors. Algorithms designed to compensate for the difference between training and test data are known as *domain adaptation* techniques.

One possibility is to use optimal transport to design a stable domain adaptation tool. The basic idea is to view the training data as a point cloud in some Euclidean space $\mathbb{R}^d$. For instance, perhaps d could be the number of pixels in a photograph; the location of every point in the point cloud determines the contents of the photo, and as additional information each point is labeled with a text name. The test data is also a point cloud in $\mathbb{R}^d$, but thanks to the confounding factors listed above perhaps these two points clouds are not aligned. [26] proposes using optimal transport to align the training data to the test data and to carry the label information along, e.g. attempting to align the space of Amazon.com photos to the space of shopping mall photos. Once the training and test data are aligned, it makes sense to transfer information, classifiers, and the like from one to the other.

Engineering design. Optimal transport has found application in design tools for many engineering tasks, from reflector design [79, 117] to aerodynamics [92]. One intriguing paper uses optimal transport to design transparent objects made of materials like glass, which can focus light into *caustics* via refraction [98]. By minimizing the transport distance between the light rays by the glass and a desired black-and-white image, they can "shape" the distribution of light as it comes out of a window. An example caustic design computed using their method is shown in Figure 9.

4. One Problem, Many Discretizations

Computational optimal transport is a relatively new discipline, and techniques for solving the optimal transport problem and in particular computing Wasserstein distances are still a topic of active research. So far, it appears that no "one size fits all" approach has been discovered; rather, different applications and scenarios demand different numerical techniques for optimal transport.

Several desiderata inform the design of an algorithm for optimal transport:

- Efficiency: While L_1 distances and KL divergences are computable using closed-form formulas, optimal transport distance computation requires solving an optimization problem. The cost of solving this problem relative to the cost of direct computation of transport's simpler alternatives is largely the reason why optimal transport has not reached a higher level of popularity in the applied world. But this scenario is changing: New high-speed algorithms for large-scale transport are nearly competitive with more traditional alternatives while bringing to the table the geometric structure unique to transport world.
- Stability: A theme in the numerical analysis literature is stability, the resilience of a computation to small changes in the input. Stability of the minimal transport objective value and/or its accompanying transport map can be a challenging topic. Linear program discretizations of continuum optimal transport problems tend to resemble (2.9) above, a linear program whose optimal solution T *provably* has the sparsity of a permutation matrix; this implies that a small perturbation of v or w may result in a discrete change of T's sparsity structure.
- Structure preservation: Transport is well-studied theoretically, and one could reasonably expect that key properties of transport in the infinite-dimensional case are preserved either exactly or approximately when they are computed numerically. For instance, Wasserstein distances enjoy a triangle inequality, and Eulerian formulations of transport have connections to gradient flows and other PDE. Provable guarantees that these structures are preserved in discretizations of transport help assure that nothing critical is lost in the process of approximating transport distances algorithmically.

One reason why there are so many varied algorithms available for numerical OT is that the problem can be written in so many different ways (see §2.4). A basic recipe for designing a transport algorithm is to choose any one of many equivalent formulations of transport—all of which yield the same optimal value in theory—, discretize any variables that are otherwise infinite-dimensional, and design a bespoke optimization algorithm to solve the resulting problem, which now has a finite number of variables. The flexibility of choosing *which* version of transport to discretize usually is tuned to the geometry of a given application, desired properties of the resulting discretization, and ease of optimizing the discretized problem.

The reality of choosing a discretization to facilitate ease of computation reflects a tried-and-true maxim of engineering: "If a problem is difficult to solve, change the problem."

In this section, we roughly outline a few discretizations and accompanying optimization algorithms for numerical OT. Our goal is not to review all well-known techniques for computational transport thoroughly but rather to highlight the breadth of possible approaches and to give a few practical pointers for implementing state-of-the-art transport algorithms at home.

4.1. Regularized Transport. We will start by describing *entropically-regularized transport*, a technique that has piqued the interest of the machine learning community after its introduction there in 2013 [28]. This technique has an explicit trade-off between accuracy and computational efficiency and has shown particularly strong promise in the regime where a rough estimate of transport is sufficient. This regime aligns well with the demands of "big data" applications, in which individual data points are likely to be noisy—so obtaining an extremely accurate transport value would be overkill computationally.

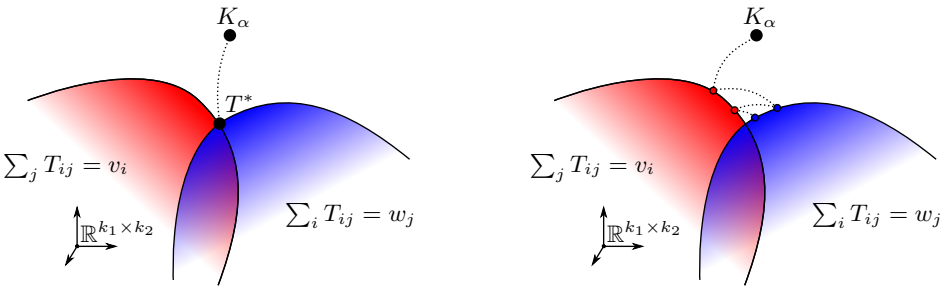
Regularization is a key technique in optimization and inverse problems in which an objective function is modified to encode additional assumptions and/or to make it easier to minimize. For example, suppose we wish to solve the least-squares problem $\min_x \|Ax-b\|_2^2$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. When A is rank-deficient or if $m < n$, an entire affine space of x's achieve the minimal value. To get around this, we could apply Tikhonov regularization (also known as ridge regression), in which we instead minimize $\|Ax - b\|_2^2 + \alpha\|x\|_2^2$ for some $\alpha > 0$. As $\alpha \to 0$ a solution of the original least-squares problem is recovered, while for any $\alpha > 0$ the regularized problem is guaranteed to have a unique minimizer; as $\alpha \to \infty$, we have $x \to 0$, a predictable but uninteresting value. From a high level, we can think of α as trading off between fidelity to the original problem $Ax \approx b$ and ease of solution: For small $\alpha > 0$ the problem is near-singular but close to the original least-squares formulation, while larger α makes the problem easier to solve.

The variables in the basic formulation of transport are nonnegative probability values, which do not appear to be amenable to standard least-squares style Tikhonov regularization. Instead, entropic regularization uses a regularizer from information theory: the entropy of a probability distribution. Suppose a probability measure has distribution function $\rho(x)$. The (differential) entropy of ρ is defined as

$$H[\rho] := -\int \rho(x)\log\rho(x)\,dx. \tag{4.1}$$

This definition makes two assumptions that are needed to work with entropy, that a probability measure admits a distribution and that it is nonzero everywhere—otherwise $\log\rho(x)$ is undefined. $H[\rho]$ is a concave function of ρ that roughly measures the "fuzziness" of a distribution. Low entropy indicates that a distribution is sharply peaked about a few points, while high entropy indicates that it is more uniformly distributed throughout space.

The basic approach in entropically-regularized transport is to add a small multiple of $-H[\pi]$ to regularize the transport plan π in the OT problem. We will start by discussing the discrete problem (2.9), which after entropic regularization can be



(a) Projection (b) Alternating projection

FIGURE 10. (a) Intuition for the optimization problem (4.2) as a projection of K_α onto the prescribed row sum and column sum constraints with respect to KL divergence (4.3). (b) The Sinkhorn algorithm projects back and forth onto one set of constraints and then the other, converging to the transport matrix T^*.

written as follows:

$$\text{OT}_\alpha(v, w; C) := \left\{ \begin{array}{rl} \min_{T \in \mathbb{R}^{k_1 \times k_2}} & \sum_{ij} T_{ij} c_{ij} + \alpha \sum_{ij} T_{ij} \log T_{ij} \\ \text{s.t.} & \sum_j T_{ij} = v_i \ \forall i \in \{1, \dots, k_1\} \\ & \sum_i T_{ij} = w_j \ \forall j \in \{1, \dots, k_2\}. \end{array} \right. \tag{4.2}$$

We are able to drop the $T \geq 0$ constraint because $\log T_{ij}$ in the objective function prevents negative T values.

The objective function from (4.2) can be refactored slightly:

$$\begin{aligned} \sum_{ij} T_{ij} c_{ij} + \alpha \sum_{ij} T_{ij} \log T_{ij} &= \alpha \sum_{ij} T_{ij} \left(\frac{c_{ij}}{\alpha} + \log T_{ij} \right) \\ &= \alpha \sum_{ij} T_{ij} \log \frac{T_{ij}}{e^{-c_{ij}/\alpha}} \\ &= \alpha \text{KL}(T | K_\alpha). \end{aligned} \tag{4.3}$$

Here, we define a *kernel* K_α via $(K_\alpha)_{ij} := e^{-c_{ij}/\alpha}$. KL denotes the Kullback–Leibler divergence [52], a distance-like (but asymmetric) measure of the similarity between T and K from information theory; the definition of K_α is singular when $\alpha = 0$, indicating that the connection to KL is only possible in the $\alpha > 0$ regime.

Equation (4.3) gives an intuitive explanation for entropy-regularized transport illustrated in Figure 10(a). The matrix K does not satisfy the constraints of the regularized transport problem (4.2). Thinking of KL roughly as a distance measure, our job is to find the *closest projection* (with respect to KL) of K onto the set of T's satisfying the constraints $\sum_j T_{ij} = v_i$ and $\sum_i T_{ij} = w_j$. With this picture in mind, Figure 10(b) illustrates the Sinkhorn algorithm for entropy-regularized transport derived below, which alternates between projecting onto one of these sets and then the other.

Continuing in our derivation, we return to (4.2) to derive first-order optimality conditions. Since (4.2) is an equality-constrained differentiable minimization problem, it can be solved using a standard multi-variable calculus technique: the method of Lagrange multipliers. There are $k_1 + k_2$ constraints, so we need $k_1 + k_2$ Lagrange multipliers, which—following the derivation of (2.11)—we store in vectors

$\phi \in \mathbb{R}^{k_1}$ and $\psi \in \mathbb{R}^{k_2}$. The Lagrange multiplier function here is:

$$\begin{aligned}\Lambda(T;\phi,\psi) &:= \sum_{ij} T_{ij}c_{ij} + \alpha\sum_{ij} T_{ij}\log T_{ij} \\ &\quad + \sum_i \phi_i\left(v_i - \sum_j T_{ij}\right) + \sum_j \psi_j\left(w_j - \sum_i T_{ij}\right) \\ &= \langle T, C\rangle + \alpha\langle T, \log T\rangle + \phi^\top(v - T\mathbb{1}) + \psi^\top(w - T^\top\mathbb{1})\end{aligned}$$

Here, $\langle\cdot,\cdot\rangle$ indicates the element-wise inner product of matrices, the log is element-wise, and $\mathbb{1}$ indicates the vector of all ones. Taking the gradient with respect to T gives the following first-order optimality condition:[1]

$$\begin{aligned}0 &= \nabla_T\Lambda = C + \alpha\mathbb{1}\mathbb{1}^\top + \alpha\log T - \phi\mathbb{1}^\top - \mathbb{1}\psi^\top \\ &\implies \log T = \frac{(\phi - \alpha\mathbb{1})\mathbb{1}^\top}{\alpha} + \frac{\mathbb{1}\psi^\top}{\alpha} + \log K_\alpha \text{ where } K_\alpha := \exp[-C/\alpha] \\ &\implies \boxed{T = \text{diag}[p]K_\alpha\text{diag}[q]} \text{ where } p := \exp\left[\frac{\phi - \alpha\mathbb{1}}{\alpha}\right] \text{ and } q := \exp\left[\frac{\psi}{\alpha}\right].\end{aligned}$$

Here, diag$[v]$ indicates the diagonal matrix whose diagonal is v. The key result is the boxed equation, which gives a formula for the unknown transport matrix T in terms of two unknown vectors p and q derived by changing variables from the Lagrange multipliers ϕ and ψ. There are multiple choices of p and q in terms of ϕ and ψ that all give the same "diagonal rescaling" formula including some that are more symmetric, but this detail is not important.

Next we plug the new relationship $T = \text{diag}[p]K_\alpha\text{diag}[q]$ into the constraints of (4.2) to find

$$\begin{aligned}p \otimes (K_\alpha q) &= v \\ q \otimes (K_\alpha^\top p) &= w.\end{aligned} \tag{4.4}$$

Here, $\otimes$ denotes the elementwise (Hadamard) product of two vectors or matrices. These formulas determine the unknown vector p in terms of q and vice versa.

The formulas (4.4) directly suggest a state-of-the-art technique for entropy-regularized optimal transport, known as the *Sinkhorn (or Sinkhorn–Knopp) algorithm* and dating back to an early technique for matrix rescaling [**101**]. This extremely succinct algorithm successively updates estimates of p and q. Iteration k is given by the update formulas ($\oslash$ denotes elementwise division)

$$\begin{aligned}p^{k+1} &\leftarrow v \oslash (K_\alpha q^k) \\ q^{k+1} &\leftarrow w \oslash (K_\alpha^\top p^{k+1}).\end{aligned}$$

It can be implemented in fewer than ten lines of code! The basic approach is to update p in terms of q using the first relationship, then q in terms of p using the second relationship, then p again, and so on. Using essentially the geometric intuition provided in Figure 10(b) for this technique and explored in-depth in [**10**], one can prove that diag$[p]K_\alpha$diag$[q]$ converges asymptotically to the optimal T at a relatively efficient rate regardless of the initial guess.

[1] Readers uncomfortable with this sort of calculation are strongly encouraged to take a look at the useful "cheat sheet" document [**86**].

Several advantages distinguish the Sinkhorn method from its peers in the numerical optimization world. Most critically, beyond its ease of implementation, this algorithm is built from simple linear algebra operations—matrix-vector multiplies and elementwise arithmetic—that parallelize well and can be carried out extremely quickly on modern processing hardware. One modern spin on Sinkhorn shows how to shave off even more calculations while preserving its favorable convergence rate [**3**].

Beyond inspiring a huge body of follow-on work in machine learning and computer vision, the Sinkhorn rescaling algorithm provides a means to adapt optimal transport to discrete domains suggested in [**104**]. So far, our description of the Sinkhorn method has been generic to *any* cost matrix C. Adding geometric structure to the problem gives it a strong interpretation using heat flow and suggests a faster way to carry out Sinkhorn iterations on discrete domains.

Suppose that the transport cost C is given by squared pairwise distances along a discretized piece of geometry such as a triangulated surface, denoted Σ; this corresponds to computing a regularized version of the 2-Wasserstein distance (2.10). The dual variables p and q can be thought of as *functions* over Σ, discretized e.g. using one value per vertex. Then, the kernel K_α has elements

$$(K_\alpha)_{ij} = e^{-d(x_i,x_j)^2/\alpha},$$

where $d(x_i, x_j)$ denotes the shortest-path (geodesic) distance along the domain from vertex i to vertex j.

To start, if our domain is flat, or Euclidean, then $(K_\alpha)_{ij} = e^{-\|x_i-x_j\|_2^2/\alpha}$ for points $\{x_i\}_i \subseteq \mathbb{R}^n$. Considered as a function of the x_i's, we recognize K_α up to scale as a *Gaussian* (or normal distribution, or bell curve) in distance. Multiplication by K_α is then *Gaussian convolution*, an extremely simple operation that can be carried out algorithmically using methods like the Fast Fourier Transform (FFT). In other words, rather than explicitly computing and storing the matrix K_α as an initial step and computing matrix-vector products $K_\alpha p$ and $K_\alpha q$ (note K_α is symmetric in this case) in every iteration of the Sinkhorn algorithm, in this case we can replace these products with convolutions $g_\sigma * p$ and $g_\sigma * q$, where $*$ denotes convolution and g_σ is a Gaussian whose standard deviation is determined by the regularizer α. This is *completely equivalent* to the Sinkhorn method that explicitly computes the matrix-vector product, while eliminating the need to store K_α and improving algorithmic speed thanks to fast Gaussian convolution. Put more simply, in the Euclidean case **multiplication by K_α is more efficient than storing K_α** since we can carry out the former implicitly.

When Σ is curved, we can use a mathematical sleight of hand modifying the entropic regularizer to improve computational properties while maintaining convergence to the true optimal transport value as the regularizer goes to zero. We employ a well-known property of geodesic distances introduced in theory in [**113**] and applied to computing distances on discrete domains in [**27**]. This property, known as Varadhan's formula, states that geodesic distance $d(x, y)$ between two points x, y on a manifold can be recovered from heat diffusion over a short time:

$$d(x,y)^2 = \lim_{t\to 0}[-2t \ln \mathcal{H}_t(x,y)].$$

Recall that the heat kernel $\mathcal{H}_t(x, y)$ determines diffusion between $x, y \in \Sigma$ after time t. That is, if f_t satisfies the heat equation $\partial_t f_t = \Delta f_t$, where Δ denotes the

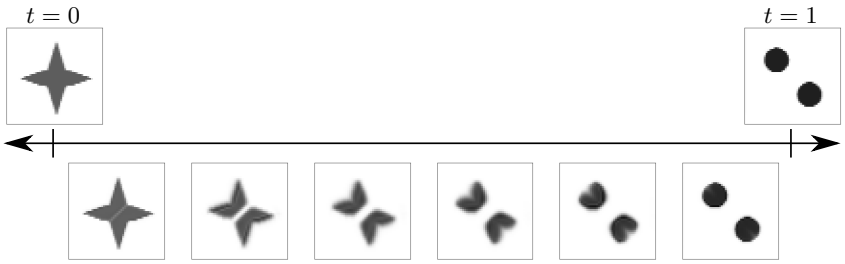


FIGURE 11. Output from an Eulerian algorithm for optimal transport extending [9] (image courtesy H. Lavenant); interpolation between the two distributions on the top is shown below the timeline. In addition to finding the transport cost, methods in this class also provide a sequence of distributions interpolating between the two inputs.

Laplacian operator, then

$$f_t(x) = \int_\Sigma f_0(y)\mathcal{H}_t(x,y)\,dy.$$

Connecting to the previous paragraph, the heat kernel in Euclidean space is exactly the Gaussian function! Hence, if we replace the kernel K_α with the heat kernel $\mathcal{H}_{\alpha/2}$ in Sinkhorn's method, in the Euclidean case nothing has changed. In the curved case, we get a new approximation of Wasserstein distances introduced as "convolutional Wasserstein distances."

All that remains is to convince ourselves that we can compute matrix-vector products $\mathcal{H}_t \cdot p$ when $\mathcal{H}_t$ is the heat kernel of a discretized domain Σ that is not Euclidean. Thankfully, armed with material from other chapters in this tutorial, this is quite straightforward in the context of discrete differential geometry. In particular, the well-known cotangent approximation of the Laplacian Δ can be combined with standard ordinary differential equation (ODE) solution techniques to carry out heat diffusion in this case using sparse linear algebra. We refer the reader for [104] for details of one implementation that uses DDG tools extensively.

4.2. Eulerian Algorithms. Entropically-regularized transport works with the Kantorovich formulation (2.7). This may be one of the earliest and most intuitive definitions of optimal transport, but this in itself is not a strong argument in favor of tackling this formulation numerically. As a point of contrast, we now explore a *completely different* approximation of Wasserstein distances that can be useful in low-dimensional settings, built from the Eulerian (fluid mechanics) formulation of the 2-Wasserstein distance $\mathcal{W}_2^2$ (2.16). Historically, this method pre-dates the popularity of entropically-regularized transport and has distinct advanges and disadvantages: It explicitly computes a time-varying displacement interpolation of a density "explaining" the transport (see Figure 11) but in the process must solve a difficult boundary-value PDE problem. Beyond the original paper [9], we recommend the excellent tutorial [87] that steps through an implementation of this technique in practice.

We make a few more simplifications to the continuum formulation before discretizing it. We start by making a quick observation: for any vector $J \in \mathbb{R}^n$ and

$\rho > 0$ we have

$$\frac{\|J\|_2^2}{2\rho} = \left\{ \begin{array}{ll} \sup_{a\in\mathbb{R}, b\in\mathbb{R}^n} & a\rho + b^\top J \\ \text{s.t.} & a + \frac{\|b\|_2^2}{2} \leq 0. \end{array} \right.$$

This convex program not only justifies that the quotient $\|J\|_2^2/2\rho$ is convex jointly in J and ρ, but also it shows we can write the optimization problem (2.17) with additional variables as

$$\begin{array}{rl} \inf_{J,\rho} \sup_{a,b} & \int_0^1 \int_{\mathbb{R}^n} [a(x,t)\rho(x,t) + b(x,t)^\top J(x,t)]\, dA(x)\, dt \\ \text{s.t.} & \rho(x,0) \equiv \rho_0(x)\ \forall x \in \mathbb{R}^n \\ & \rho(x,1) \equiv \rho_1(x)\ \forall x \in \mathbb{R}^n \\ & \frac{\partial \rho(x,t)}{\partial t} = -\nabla \cdot J(x,t)\ \forall x \in \mathbb{R}^n, t \in (0,1) \\ & a(x,t) + \frac{\|b(x,t)\|_2^2}{2} \leq 0\ \forall x \in \mathbb{R}^n, t \in (0,1). \end{array}$$

Next, we introduce a dual potential function $\phi(x,t)$ similarly to the derivation of (2.12) to take care of all but the last constraint:

$$\begin{array}{rl} \inf_{J,\rho} \sup_{a,b,\phi} & \int_0^1 \int_{\mathbb{R}^n} \Big[a(x,t)\rho(x,t) + b(x,t)^\top J(x,t) \\ & \qquad + \phi(x,t) \left(\frac{\partial \rho(x,t)}{\partial t} + \nabla \cdot J(x,t) \right) \Big]\, dA(x)\, dt \\ & + \int_{\mathbb{R}^n} [\phi(x,1)(\rho_1(x) - \rho(x,1)) - \phi(x,0)(\rho_0(x) - \rho(x,0))]\, dA(x) \\ \text{s.t.} & a(x,t) + \frac{\|b(x,t)\|_2^2}{2} \leq 0\ \forall x \in \mathbb{R}^n, t \in (0,1). \end{array} \tag{4.5}$$

We can simplify some terms in this expression. First, using integration by parts we have

$$\int_0^1 \phi(x,t) \frac{\partial \rho(x,t)}{\partial t}\, dt = [\rho(x,1)\phi(x,1) - \rho(x,0)\phi(x,0)] - \int_0^1 \rho(x,t) \frac{\partial \phi(x,t)}{\partial t}\, dt$$

We also can integrate by parts in x to show

$$\int_{\mathbb{R}^n} \phi(x,t) \nabla \cdot J(x,t)\, dA(x) = -\int_{\mathbb{R}^n} J(x,t)^\top \nabla \phi(x,t)\, dA(x).$$

This simplification works equally well if we replace $\mathbb{R}^n$ with the box $[0,1]^n$ with periodic boundary conditions. Incorporating these two integration by parts formulae into our objective function yields a new one:

$$\int_{\mathbb{R}^n} \left\{ \int_0^1 \left(\rho(x,t) \left[a(x,t) - \frac{\partial \phi(x,t)}{\partial t} \right] + J(x,t)^\top [b(x,t) - \nabla \phi(x,t)] \right) dt \right.$$
$$\left. - \phi(x,0)\rho_0(x) + \phi(x,1)\rho_1(x)) \right\} dA(x)$$

We now make some notational simplifications. Define $z := \{\rho, J\}$ and $q := \{a, b\}$ with inner product

$$\langle z, q \rangle := \int_{\mathbb{R}^n} \int_0^1 (a(x,t)\rho(x,t) + b(x,t)^\top J(x,t))\, dt\, dA(x).$$

Furthermore, define

$$F(q) := \left\{ \begin{array}{ll} 0 & \text{if } a(x,t) + \frac{\|b(x,t)\|_2^2}{2} \leq 0\ \forall x \in \mathbb{R}^n, t \in (0,1) \\ \infty & \text{otherwise.} \end{array} \right.$$

$$G(\phi) := \int_{\mathbb{R}^n} (\phi(x,0)\rho_0(x) - \phi(x,1)\rho_1(x))\, dA(x)$$

These functions are both convex. These functions, plus our simplifications and a sign change, allow us to write (4.5) in a compact fashion as:

$$-\sup_{z}\inf_{q,\phi}\left[F(q)+G(\phi)+\langle z,\nabla_{x,t}\phi-q\rangle\right], \tag{4.6}$$

where $\nabla_{x,t}\phi := \{\partial\phi/\partial t, \nabla_x\phi\}$.

Blithely assuming strong duality, namely that we can swap the supremum and the infimum, we arrive at an alternative interpretation of (4.6). In particular, we can view z as a Lagrange multiplier corresponding to a constraint $q = \nabla_{x,t}\phi$. From this perspective, we actually can find a saddle point (max in z, minimum in (q,ϕ)) of the *augmented Lagrangian* L_r for any $r \geq 0$:

$$L_r(\phi,q,z) := F(q)+G(\phi)+\langle z,\nabla_{x,t}\phi-q\rangle+\frac{r}{2}\langle\nabla_{x,t}\phi-q,\nabla_{x,t}\phi-q\rangle.$$

The extra term here effectively adds zero to the objective function, assuming the constraint is satisfied.

The algorithm proposed in [**9**] iteratively updates estimates $(\phi^\ell, q^\ell, z^\ell)$ by cycling through the following three steps:

$$\begin{aligned}\phi^{\ell+1} &\leftarrow \arg\min_{\phi} L_r(\phi,q^\ell,z^\ell)\\ q^{\ell+1} &\leftarrow \arg\min_{q} L_r(\phi^{\ell+1},q,z^\ell)\\ z^{\ell+1} &\leftarrow z^\ell + r(q^{\ell+1}-\nabla_{x,t}\phi^{\ell+1}).\end{aligned}$$

The first two steps update some variables while holding the rest fixed to the best possible value. The third step is gradient step for z. This cycling algorithm and equivalent formulations has many names in the literature—including ADMM [**17**], the Douglas–Rachford algorithm [**37, 59**], and the Uzawa algorithm [**112**]—and is known to converge under weak assumptions.

The advantage of this algorithm is that the individual update formulae are straightforward. In particular, the ϕ update is equivalent to solving a Laplace equation

$$\Delta_{x,t}\phi^{\ell+1} = \nabla_{x,t}\cdot(z^\ell - rq^\ell),$$

where $\Delta_{x,t}$ is the Laplacian operator in time and space. The q update decouples over x and t, amounting to projecting $\nabla_{x,t}\phi^{\ell+1}+{}^{z^\ell}\!/_r$ onto the constraints in the definition of $F(q)$ with respect to L_2, a one-dimensional problem solvable analytically. And, the z update is already in closed-form.

So far, we have described the Benamou–Brenier algorithm using continuum variables, but of course at the end of the day we must discretize the problem for computational purposes. The most straightforward discretization assumes ρ_0 and ρ_1 are supported in the unit square $[0,1]^n$, which is broken up into a $m\times m\times\cdots\times m$ grid, and further discretizes the time variable $t\in[0,1]$ into p steps. Then, all degrees of freedom (ϕ,q,z) can be put on the grid vertices and interpolated in between using multilinear basis functions; this leads to a finite element (FEM) discretization of the problem that can be approached using techniques discussed in earlier chapters. An alternative grid-based discretization and accompanying optimization algorithm is also given in [**83**].

The use of PDE language makes this dynamical formulation of transport seem attractive as potentially compatible with machinery like discrete exterior calculus (DEC) [**44**], which could define a discrete notion of transport on simplicial complexes like triangle meshes that discretize curved surfaces. This remains an open

problem for challenging technical reasons.[2] Principally, discretizing the objective function $\|J\|_2^2/\rho$ on a triangle mesh is challenging because scalar quantities like ρ typically are discretized on vertices or faces while vectorial quantities like J are better suited for edges. Evaluating $\|J\|^2/2\rho$ then requires averaging J or ρ so that the two end up on the same simplices. If this problem is overcome, it still remains to prove a triangle inequality for discretizations of the Wasserstein distance resulting from such an approach. Some recent papers with analogous constructions on graphs [**62, 109, 38**] suggest that such an approach may be possible.

While the Benamou–Brenier dynamical formulation of transport is the best known, it is worth noting that the Beckmann problem (2.18) for the 1-Wasserstein distance $\mathcal{W}_1$ more readily admits discretization using the finite element method (FEM) while preserving a triangle inequality. Details of such a formulation as well as an efficient optimization algorithm are provided in [**107**]. The reason (2.18) is easier to discretize is that the time-varying aspect of transport is lost in this formulation: All that is needed is a single vector $J(x)$ per point x. What makes this problem easy to discretize and optimize is its downfall application-wise: Interpolation with respect to $\mathcal{W}_1$ between two densities μ_0 and μ_1 is given by the uninteresting solution $\mu_t = (1-t)\mu_0 + t\mu_1$, which does not displace mass but rather "teleports" it from the source to the target.

Another PDE-based approach to optimal transport is worth noting and has strong connections to the theory of transport without connecting to fluid flow. Recall the Monge formulation of optimal transport on $\mathbb{R}^n$ in equation (2.3), which seeks a map $\phi(x)$ that pushes forward one distribution function $\rho_0(x)$ onto another $\rho_1(x)$. A famous result by Brenier [**18**] shows that ϕ can be written as the gradient of a convex potential $\Psi(x)$: $\phi(x) = \nabla\Psi(x)$. Using H to denote the Hessian operator, this potential satisfies the Monge-Ampère PDE

$$\det(H\Psi(x))\rho_1(\nabla\Psi(x)) = \rho_0(x), \tag{4.7}$$

a second-order nonlinear elliptic equation that is extremely challenging to solve in practice. A few algorithms, e.g. [**80, 60, 12, 41, 13**], tackle this nonlinear system head-on, discretizing the variables involved and solving for Ψ.

4.3. Semidiscrete Transport. Our final example from the computational transport world uses yet another formulation of the transport problem. This time, our inspiration is the one-dimensional semidiscrete problem, whose solution is motivated from the formulation in equation (2.6). Our exposition of this problem closely follows the excellent tutorial [**57**].

In this setting, optimal transport is computed from a distribution whose mass is concentrated at a finite set of isolated points to a distribution with a known but potentially smooth density function. Recall that in the one-dimensional case, we learned that each point of mass in the source is mapped to an *interval* in the target. That is, the domain of the target density is partitioned into contiguous cells whose mass is assigned to a single source point. We will find that the higher-dimensional analog is spiritually identical: Each point of mass in the source density is assigned to a convex region of space in the target. This observation will suggest algorithms constructed from ideas in discrete geometry extending Voronoi diagrams and similar constructions.

[2]Interested readers are encouraged to contact the author of this tutorial for preliminary results on this problem!

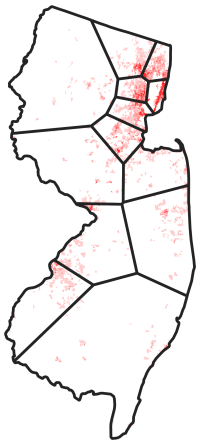


FIGURE 12. Power diagram from a semidiscrete transport problem (image courtesy R. Barnes). Here, semidiscrete transport is used to partition the state of New Jersey into cells with equal population; population density is shaded in red.

As in (2.6), suppose we are computing the 2-Wasserstein distance from a discrete measure $\mu := \sum_{i=1}^{k} a_i \delta_{x_i}$, whose mass is concentrated at points $x_i \in \mathbb{R}^n$ with weights $a_i > 0$, to an absolutely continuous measure ν with distribution function $\rho(x)$. The dual formulation of transport (2.12) in this case can be written

$$\begin{array}{rl} \sup_{\phi,\psi} & \sum_{i=1}^{k} a_i \phi(x_i) + \int_{\mathbb{R}^n} \psi(y)\rho(y)\, dA(y) \\ \text{s.t.} & \phi(x) + \psi(y) \leq c(x,y)\ \forall x, y \in \mathbb{R}^n. \end{array}$$

The objective in this case "does not care" about values of $\phi(x)$ for $x \notin \{x_i\}_{i=1}^k$. Define $\phi_i := \phi(x_i)$. By this observation, we can write a problem with only one continuum variable:

$$\begin{array}{rl} \sup_{\phi,\psi} & \sum_i a_i \phi_i + \int_{\mathbb{R}^n} \psi(y)\rho(y)\, dA(y) \\ \text{s.t.} & \phi_i + \psi(y) \leq c(x_i,y)\ \forall y \in \mathbb{R}^n, i \in \{1,\dots,k\}. \end{array}$$

In a slight abuse of notation, for the rest of this section we will think of ϕ as a vector $\phi \in \mathbb{R}^k$ rather than a function $\phi(x)$. Given the supremum, we might as well choose the largest ψ possible that satisfies the constraints. Hence,

$$\psi(y) = \inf_{i \in \{1,\dots,k\}} [c(x_i, y) - \phi_i].$$

This leads to a final optimization problem in a *finite* set of variables $\phi_1, \dots, \phi_k$:

$$\begin{aligned} \mathcal{W}_2^2(\mu,\nu) &= \sup_{\phi \in \mathbb{R}^k} \sum_i a_i \phi_i + \int_{\mathbb{R}^n} \rho(y) \left(\inf_{i \in \{1,\dots,k\}} [c(x_i,y) - \phi_i] \right) dA(y) \\ &= \sup_{\phi \in \mathbb{R}^k} \sum_i \left[a_i \phi_i + \int_{\mathrm{Lag}_\phi^c(x_i)} \rho(y)[c(x_i,y) - \phi_i]\, dA(y) \right] \end{aligned} \tag{4.8}$$

Here, $\mathrm{Lag}_\phi^c(x_i)$ indicates the *Laguerre cell* corresponding to x_i:

$$\mathrm{Lag}_\phi^c(x_i) := \{y \in \mathbb{R}^n : c(x_i,y) - \phi_i \leq c(x_j,y) - \phi_j\ \forall j \neq i\}. \tag{4.9}$$

The set of Laguerre cells yields the *Laguerre diagram*, a partition of $\mathbb{R}^n$ determined by the cost function c and the vector ϕ; the ϕ_i's effectively control the sizes of the Laguerre cells in the diagram. When $c(x,y) = \|x-y\|_2$ is a distance function and $\phi = 0$, the Laguerre diagram equals the well-known Voronoi diagram of the x_i's that partitions $\mathbb{R}^n$ into loci of points S_i corresponding to those closer to x_i than to the other x_j's [6]. More importantly for the 2-Wasserstein distance, when $c(x,y) = 1/2\|x-y\|_2^2$, the Laguerre diagram is known as the *power diagram*, an object studied since the early days of computational geometry [5]; an example is shown in Figure 12.

Since (4.8) comes from a simplification of the dual of the transport problem, it is concave in ϕ; a direct proof can be found in [7]. This implies that a simple gradient ascent procedure starting from any initial estimate of ϕ will reach a global optimum. Define the objective function

$$F(\phi) := \sum_i \left[a_i\phi_i + \int_{\mathrm{Lag}^c_\phi(x_i)} \rho(y)[c(x_i,y) - \phi_i]\, dA(y) \right].$$

The gradient can be computed using the partial derivative expression

$$\frac{\partial F}{\partial \phi_i} = a_i - \int_{\mathrm{Lag}^c_\phi(x_i)} \rho(y)\, dA(y). \tag{4.10}$$

This expression is predictable from the definition of $F(\phi)$; a similar formula exists for the second derivatives of F. Setting the gradient (4.10) to zero formalizes an intuition for the optimization problem (4.8), that it resizes the Laguerre cells by modifying the ϕ_i's until the cell corresponding to each x_i contains mass a_i:

$$a_i = \int_{\mathrm{Lag}^c_\phi(x_i)} \rho(y)\, dA(y).$$

The main ingredient needed to compute the derivatives of F is an algorithm for integrating ρ over Laguerre cells. Hence, gradient ascent and Newton's method applied to optimizing for ϕ cycle between updating the Laguerre diagram for the current ϕ estimate, recomputing the gradient and/or Hessian, assembling these into a search direction, and updating the current estimate of ϕ. For squared Euclidean costs, these algorithms are facilitated by fast algorithms for computing power diagrams, e.g. [16, 118]. While convergence of gradient descent with line search follows directly from concavity, [48] proves that under certain assumptions a damped version of Newton's algorithm—which employs the Hessian in addition to the gradient to accelerate convergence—exhibits global convergence.

Example techniques following this template include [20], which proposed an early technique for 2D problems; [70], for semi-discrete transport to piecewise-linear distribution functions in 2D supported on triangle meshes improved using a multiscale approximation; and [56], which proposes semi-discrete transport to distributions in 3D that are piecewise-linear on tetrahedral meshes. [32] provides an early example of a Newton solver for 2D semidiscrete transport using power diagrams and additionally uses derivatives of transport in the support points x_i and weights a_i for assorted approximation problems.

Beyond providing fast algorithms for transport in the semidiscrete case, this formulation is also valuable for applications incorporating transport terms. [33] employs semidiscrete transport to a collection of distributions concentrated on line segments to reconstruct line drawings from point samples; [35] proposes a similar

technique for reconstructing triangulated surfaces from point clouds in $\mathbb{R}^3$. [42] defines a version of semi-discrete transport intrinsic to a triangulated surface, which can be used for tasks like parameterizing the set of per-vertex area weights in terms of the values ϕ_i.

5. Beyond Transport

Beyond improving tools for solving the basic optimal transport problem, some of the most exciting recent work in computational transport involves using transport as a single term in a larger model. In a recent tutorial for the machine learning community, we termed this new trend *"Wassersteinization"* [31]: using Wasserstein distances to improve geometric properties of variational models in statistics, learning, applied geometry, and other disciplines. Further extending the scope of applied transport, variations of the basic problem have been proposed to apply OT to objects other than probability distributions.

While a complete survey of these creative new applications and extensions is far beyond the scope of this tutorial, we highlight a few interesting pointers into the literature:

- Unbalanced transport: One limitation of the basic model for optimal transport is that it is a distance between histograms or probability distributions, rather than a distance between functions or vectors in $\mathbb{R}^n$—which may not integrate to 1 or may contain negative values. This leads to the problem of *unbalanced transport*, in which mass conservation and/or positivity must be relaxed. Models for this problem range from augmenting the transport problem with a "trash can" that can add or remove mass from distributions [85] to extensions of dynamical transport to this case [22]. Making transport work for functions rather than distributions while preserving the triangle inequality and other basic properties is challenging both theoretically and from a numerical perspective.
- Barycenters: The idea of displacement interpolation we motivated using (2.17) suggests a generalization to more than two distributions, known as the *Wasserstein barycenter* problem [2]. Given k distributions $\mu_1, \ldots, \mu_k$, the Wasserstein barycenter $\mu_{\text{barycenter}}$ is defined as the minimizer of the following optimization problem

$$\mu_{\text{barycenter}} := \arg\min_{\mu} \sum_{i=1}^{k} \mathcal{W}_2^2(\mu, \mu_i). \tag{5.1}$$

  The Wasserstein barycenter gives some notion of *averaging* a set of probability distributions, motivated by the observation that the average $\frac{1}{k}\sum_{i=1}^{k} x_i$ of a set of vectors $x_i \in \mathbb{R}^n$ is the minimizer $\arg\min_x \sum_i \|x - x_i\|_2^2$. Barycenter algorithms range from extensions of the Sinkhorn algorithm [10, 104] to methods that perform gradient descent on μ by differentiating the distance $\mathcal{W}_2$ in its argument [30] and stochastic techniques requiring only samples from the distributions μ_i [110, 23]. Other algorithms are inspired by a connection to multi-marginal transport [84], a generalization of optimal transport involving a distribution over the product of more than two measures. The optimization problem (5.1) is also one of the earliest examples of "Wassersteinization," in the sense that it is an optimization problem for an unknown distribution μ including Wasserstein distance terms, contrasting somewhat from the optimization problems we considered in §4 in which the unknown is the transport distance itself.

Further generalizing the barycenter problem leads to a notion of the Dirichlet energy of a map from points in one space to distributions over another [19, 54], with applications in machine learning [108] and shape matching [105, 63]. An intriguing recent paper also proposes an inverse problem for barycentric coordinates seeking weights for (5.1) that "explain" an input distribution as a transport barycenter of others [15].

- Quadratic assignment: The basic optimization problem for transport has an objective function that is *linear* in the unknown transport matrix, expressing a preference for transport maps that do not move any single particle of probabilistic mass very far. This model, however, does not necessarily extract *smooth* maps, wherein distance traveled by any single particle is less important than making sure that nearby particles in the source are mapped to nearby locations in the target. Such a smoothness term leads to a *quadratic* term in the transport problem and allows it to be extended to a distance between metric-measure spaces known as the Gromov–Wasserstein distance [68, 69], inspired by the better-known but more rigid Gromov–Hausdorff distance. From an optimization perspective, Gromov–Wasserstein computation leads to a "quadratic assignment" problem, known in the most general case to be NP-hard [95]; practical instances of the problem in shape matching, however, can be tackled using spectral [67] or entropy-based [106] approximations and have shown promise for applications in shape matching. [91] proposes a method for averaging metric spaces using a barycenter formulation similar to (5.1).
- Capacity-constrained transport: Yet another extension of the transport problem comes from introducing *capacity constraints* limiting the amount of mass that can travel between assorted pairs of source and target points; in the measure-theoretic formulation, this amounts to constraining transport plan to be dominated by another input plan [51]. This constraint makes sense in many operations-type applications and has intriguing theoretical properties, but design of algorithms and discretizations for capacity-constrained transport remains largely open although [9] provides one approach again extending Sinkhorn's algorithm.
- Gradient flows and PDE: Given a function $f : M \to \mathbb{R}$ defined over a geometric space M like a manifold, a *gradient flow* of f starting at some $x_0 \in M$ attempts to minimize f via "gradient descent" from $x(0) := x_0$ expressed as an ordinary differential equation (ODE) $x'(t) = -\nabla f(x(t))$. Since OT puts a geometry on the space of distributions $\text{Prob}(\mathbb{R}^n)$ over $\mathbb{R}^n$, we can define an analogous procedure that flows probability distributions to reduce certain functionals [46, 97]. For instance, gradient flow on the entropy functional (4.1) in the Wasserstein metric leads to the heat diffusion equation $\partial\rho/\partial t = -\Delta\rho$, where Δ is the Laplacian operator; that is, performing gradient descent on entropy in the Wasserstein metric is exactly the same as diffusing the initial probability distribution like an unevenly-heated metal plate. Beyond giving a variational motivation for certain PDE, this mathematical idea inspired numerical methods for solving PDE that can be written as gradient flows [88, 11]. Recent work has even incorporated transport into numerical methods for PDE that cannot easily be written as gradient flows in Wasserstein space, such as those governing incompressible fluid flow [58, 74, 34, 71]. Gradient flow properties can also be leveraged as structure to be preserved in discrete models of transport; for instance, [62] proposes a

model for dynamical optimal transport on a graph and checks that the gradient flow of entropy—now an ODE rather than a PDE—agrees with a discrete heat equation.

- MATRIX FIELDS AND VECTOR MEASURES: *Vector measures* generalize probability measures by replacing scalar-valued probability values $\mu(S) \in [0, 1]$ with values in other cones $\mathcal{C}$. For instance, a tensor-valued measure μ assigns measurable sets S to $d \times d$ postive semidefinite matrices $\mu(S) \in \mathcal{S}_+^d$ while satisfying analogous axioms to those laid out for probability measures in §2.1. These tensor fields find application in diffusion tensor imaging (DTI), which measures diffusivity of molecules like water in the interior of the human brain as a proxy for directionality of white matter fibers; OT extended to this setting can be used to align multiple such images. A few recent models extend OT to this case and propose related numerical methods [**78**, **21**, **89**].

6. Conclusion

The techniques covered in this tutorial are just a few of many ways to approach discrete optimal transport. New algorithms are proposed every month, and there is considerable room for mathematical, algorithmic, and application-oriented researchers to improve existing methods or make their own for different types of data or geometry. Furthermore, mathematical properties such as convergence and approximation quality are still being established for new techniques. Many questions also remain in linking to other branches of discrete differential geometry, e.g. at the most fundamental level defining a *purely discrete* notion of optimal transport compatible with polyhedral meshes or simplicial complexes without requiring regularization and while preserving structure from the smooth case.

These challenges aside, discrete optimal transport is demonstrating that OT holds interest far beyond mathematical analysis. New discretizations and algorithms bring down OT's complexity to the point where it can be incorporated into practical engineering pipelines and into larger models without incurring a huge computational expense. Further research into this new discipline holds unique potential to improve both theory and practice and eventually to bring insight into other branches of discrete and smooth geometry.

Acknowledgments. The author acknowledges the generous support of Army Research Office grant W911NF-12-R-0011 ("Smooth Modeling of Flows on Graphs"), from the MIT Research Support Committee ("Structured Optimization for Geometric Problems"), and from the MIT–IBM Watson AI Lab ("Large-Scale Optimal Transport for Machine Learning").

Many thanks to MIT Geometric Data Processing Group members Mikhail Bessmeltsev, Edward Chien, Sebastian Claici, David Palmer, and Dima Smirnov for proofreading this document.

References

1. *OptimalDistricts.org*, http://www.optimaldistricts.org/.
2. Martial Agueh and Guillaume Carlier, *Barycenters in the Wasserstein space*, SIAM Journal on Mathematical Analysis **43** (2011), no. 2, 904–924.
3. Jason Altschuler, Jonathan Weed, and Philippe Rigollet, *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration*, Proc. NIPS, 2017, pp. 1961–1971.
4. Martin Arjovsky, Soumith Chintala, and Léon Bottou, *Wasserstein generative adversarial networks*, International Conference on Machine Learning, 2017, pp. 214–223.

5. Franz Aurenhammer, *Power diagrams: properties, algorithms and applications*, SIAM Journal on Computing **16** (1987), no. 1, 78–96.
6. ———, *Voronoi diagrams—a survey of a fundamental geometric data structure*, ACM Computing Surveys (CSUR) **23** (1991), no. 3, 345–405.
7. Franz Aurenhammer, Friedrich Hoffmann, and Boris Aronov, *Minkowski-type theorems and least-squares partitioning*, Proceedings of the Eighth Annual Symposium on Computational Geometry, ACM, 1992, pp. 350–357.
8. Federico Bassetti, Antonella Bodini, and Eugenio Regazzini, *On minimum Kantorovich distance estimators*, Statistics & probability letters **76** (2006), no. 12, 1298–1302.
9. Jean-David Benamou and Yann Brenier, *A computational fluid mechanics solution to the Monge–Kantorovich mass transfer problem*, Numerische Mathematik **84** (2000), no. 3, 375–393.
10. Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré, *Iterative Bregman projections for regularized transportation problems*, SIAM Journal on Scientific Computing **37** (2015), no. 2, A1111–A1138.
11. Jean-David Benamou, Guillaume Carlier, and Maxime Laborde, *An augmented Lagrangian approach to Wasserstein gradient flows and applications*, ESAIM: Proceedings and Surveys **54** (2016), 1–17.
12. Jean-David Benamou, Brittany D Froese, and Adam M Oberman, *Two numerical methods for the elliptic Monge-Ampère equation*, ESAIM: Mathematical Modelling and Numerical Analysis **44** (2010), no. 4, 737–758.
13. ———, *Numerical solution of the optimal transportation problem using the Monge–Ampère equation*, Journal of Computational Physics **260** (2014), 107–126.
14. Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert, *Inference in generative models using the Wasserstein distance*, arXiv:1701.05146 (2017).
15. Nicolas Bonneel, Gabriel Peyré, and Marco Cuturi, *Wasserstein barycentric coordinates: histogram regression using optimal transport*, ACM Transactions on Graphics **35** (2016), no. 4, 71–1.
16. Adrian Bowyer, *Computing Dirichlet tessellations*, The Computer Journal **24** (1981), no. 2, 162–166.
17. Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein, *Distributed optimization and statistical learning via the alternating direction method of multipliers*, Foundations and Trends in Machine Learning **3** (2011), no. 1, 1–122.
18. Yann Brenier, *Polar factorization and monotone rearrangement of vector-valued functions*, Communications on Pure and Applied Mathematics **44** (1991), no. 4, 375–417.
19. ———, *Extended Monge–Kantorovich theory*, Lecture Notes in Mathematics (2003), 91–122.
20. Guillaume Carlier, Alfred Galichon, and Filippo Santambrogio, *From Knothe's transport to Brenier's map and a continuation method for optimal transport*, SIAM Journal on Mathematical Analysis **41** (2010), no. 6, 2554–2576.
21. Yongxin Chen, Tryphon T Georgiou, and Allen Tannenbaum, *Matrix optimal mass transport: a quantum mechanical approach*, IEEE Transactions on Automatic Control (2017).
22. Lénaïc Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard, *An interpolating distance between optimal transport and Fisher–Rao metrics*, Foundations of Computational Mathematics (2016), 1–44.
23. Sebastian Claici, Edward Chien, and Justin Solomon, *Stochastic Wasserstein barycenters*, arXiv:1802.05757 (2018).
24. Scott Cohen and Leonidas Guibas, *The earth mover's distance under transformation sets*, Proc. ICCV, vol. 2, IEEE, 1999, pp. 1076–1083.
25. Vincent Cohen-Addad, Philip N Klein, and Neal E Young, *Balanced power diagrams for redistricting*, arXiv:1710.03358 (2017).
26. Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy, *Optimal transport for domain adaptation*, PAMI **39** (2017), no. 9, 1853–1865.
27. Keenan Crane, Clarisse Weischedel, and Max Wardetzky, *Geodesics in heat: A new approach to computing distance based on heat flow*, ACM Transactions on Graphics (TOG) **32** (2013), no. 5, 152.
28. Marco Cuturi, *Sinkhorn distances: Lightspeed computation of optimal transport*, Advances in Neural Information Processing Systems, 2013, pp. 2292–2300.

29. Marco Cuturi and David Avis, *Ground metric learning*, Journal of Machine Learning Research **15** (2014), no. 1, 533–564.
30. Marco Cuturi and Arnaud Doucet, *Fast computation of Wasserstein barycenters*, International Conference on Machine Learning, 2014, pp. 685–693.
31. Marco Cuturi and Justin Solomon, *A primer on optimal transport*, NIPS Tutorial, 2017.
32. Fernando De Goes, Katherine Breeden, Victor Ostromoukhov, and Mathieu Desbrun, *Blue noise through optimal transport*, ACM Transactions on Graphics (TOG) **31** (2012), no. 6, 171.
33. Fernando De Goes, David Cohen-Steiner, Pierre Alliez, and Mathieu Desbrun, *An optimal transport approach to robust reconstruction and simplification of 2d shapes*, Computer Graphics Forum, vol. 30, Wiley Online Library, 2011, pp. 1593–1602.
34. Fernando de Goes, Corentin Wallez, Jin Huang, Dmitry Pavlov, and Mathieu Desbrun, *Power particles: an incompressible fluid solver based on power diagrams*, ACM Transactions on Graphics **34** (2015), no. 4, 50–1.
35. Julie Digne, David Cohen-Steiner, Pierre Alliez, Fernando De Goes, and Mathieu Desbrun, *Feature-preserving surface reconstruction and simplification from defect-laden point sets*, Journal of Mathematical Imaging and Vision **48** (2014), no. 2, 369–382.
36. Roland L'vovich Dobrushin, *Definition of random variables by conditional distributions*, Teoriya Veroyatnostei i ee Primeneniya **15** (1970), no. 3, 469–497.
37. Jim Douglas and Henry H Rachford, *On the numerical solution of heat conduction problems in two and three space variables*, Transactions of the American Mathematical Society **82** (1956), no. 2, 421–439.
38. Matthias Erbar, Martin Rumpf, Bernhard Schmitzer, and Stefan Simon, *Computation of optimal transport on discrete metric measure spaces*, arXiv:1707.06859 (2017).
39. Jean Feydy, Benjamin Charlier, François-Xavier Vialard, and Gabriel Peyré, *Optimal transport for diffeomorphic registration*, MICCAI 2017, 2017.
40. Lester Randolph Ford Jr. and Delbert Ray Fulkerson, *Solving the transportation problem*, Management Science **3** (1956), no. 1, 24–32.
41. Brittany D Froese and Adam M Oberman, *Convergent finite difference solvers for viscosity solutions of the elliptic Monge–Ampère equation in dimensions two and higher*, SIAM Journal on Numerical Analysis **49** (2011), no. 4, 1692–1714.
42. Fernando de Goes, Pooran Memari, Patrick Mullen, and Mathieu Desbrun, *Weighted triangulations for geometry processing*, ACM Transactions on Graphics (TOG) **33** (2014), no. 3, 28.
43. Steven Haker, Lei Zhu, Allen Tannenbaum, and Sigurd Angenent, *Optimal mass transport for registration and warping*, International Journal of Computer Vision **60** (2004), no. 3, 225–240.
44. Anil Nirmal Hirani, *Discrete exterior calculus*, Ph.D. thesis, California Institute of Technology, 2003.
45. Frank L Hitchcock, *The distribution of a product from several sources to numerous localities*, Studies in Applied Mathematics **20** (1941), no. 1-4, 224–230.
46. Richard Jordan, David Kinderlehrer, and Felix Otto, *The variational formulation of the Fokker–Planck equation*, SIAM Journal on Mathematical Analysis **29** (1998), no. 1, 1–17.
47. Leonid Vitalievich Kantorovich, *On the translocation of masses*, Dokl. Akad. Nauk SSSR, vol. 37, 1942, pp. 199–201.
48. Jun Kitagawa, Quentin Mérigot, and Boris Thibert, *A Newton algorithm for semi-discrete optimal transport*, arXiv:1603.05579 (2016).
49. Morton Klein, *A primal method for minimal cost flows with applications to the assignment and transportation problems*, Management Science **14** (1967), no. 3, 205–220.
50. Tjalling C Koopmans, *Exchange ratios between cargoes on various routes*, (1941).
51. Jonathan Korman and Robert McCann, *Optimal transportation with capacity constraints*, Transactions of the American Mathematical Society **367** (2015), no. 3, 1501–1521.
52. Solomon Kullback and Richard A Leibler, *On information and sufficiency*, The Annals of Mathematical Statistics **22** (1951), no. 1, 79–86.
53. Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger, *From word embeddings to document distances*, International Conference on Machine Learning, 2015, pp. 957–966.
54. Hugo Lavenant, *Harmonic mappings valued in the Wasserstein space*, arXiv:1712.07528 (2017).

55. Elizaveta Levina and Peter Bickel, *The earth mover's distance is the Mallows distance: Some insights from statistics*, Proc. ICCV, vol. 2, IEEE, 2001, pp. 251–256.
56. Bruno Lévy, *A numerical algorithm for L_2 semi-discrete optimal transport in 3D*, ESAIM: Mathematical Modelling and Numerical Analysis **49** (2015), no. 6, 1693–1715.
57. Bruno Lévy and Erica Schwindt, *Notions of optimal transport theory and how to implement them on a computer*, Computers and Graphics **72** (2018), 135–148.
58. Bo Li, Feras Habbal, and Michael Ortiz, *Optimal transportation meshfree approximation schemes for fluid and plastic flows*, International Journal for Numerical Methods in Engineering **83** (2010), no. 12, 1541–1579.
59. Pierre-Louis Lions and Bertrand Mercier, *Splitting algorithms for the sum of two nonlinear operators*, SIAM Journal on Numerical Analysis **16** (1979), no. 6, 964–979.
60. Grégoire Loeper and Francesca Rapetti, *Numerical solution of the Monge–Ampère equation by a Newton's algorithm*, Comptes Rendus Mathematique **340** (2005), no. 4, 319–324.
61. John Lott, *Some geometric calculations on Wasserstein space*, Communications in Mathematical Physics **277** (2008), no. 2, 423–437.
62. Jan Maas, *Gradient flows of the entropy for finite Markov chains*, Journal of Functional Analysis **261** (2011), no. 8, 2250–2292.
63. Manish Mandad, David Cohen-Steiner, Leif Kobbelt, Pierre Alliez, and Mathieu Desbrun, *Variance-minimizing transport plans for inter-surface mapping*, ACM Transactions on Graphics **36** (2017), 14.
64. Robert J McCann, *A convexity principle for interacting gases*, Advances in Mathematics **128** (1997), no. 1, 153–179.
65. ———, *Polar factorization of maps on Riemannian manifolds*, Geometric and Functional Analysis **11** (2001), no. 3, 589–608.
66. Robert John McCann, *A convexity theory for interacting gases and equilibrium crystals*, Ph.D. thesis, Princeton University, 1994.
67. Facundo Mémoli, *Spectral Gromov–Wasserstein distances for shape matching*, Proc. ICCV Workshops, IEEE, 2009, pp. 256–263.
68. ———, *Gromov–Wasserstein distances and the metric approach to object matching*, Foundations of Computational Mathematics **11** (2011), no. 4, 417–487.
69. ———, *The Gromov–Wasserstein distance: A brief overview*, Axioms **3** (2014), no. 3, 335–341.
70. Quentin Mérigot, *A multiscale approach to optimal transport*, Computer Graphics Forum, vol. 30, Wiley Online Library, 2011, pp. 1583–1592.
71. Quentin Mérigot and Jean-Marie Mirebeau, *Minimal geodesics along volume-preserving maps, through semidiscrete optimal transport*, SIAM Journal on Numerical Analysis **54** (2016), no. 6, 3465–3492.
72. Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean, *Efficient estimation of word representations in vector space*, arXiv:1301.3781 (2013).
73. Stacy Miller, *The problem of redistricting: the use of centroidal Voronoi diagrams to build unbiased congressional districts*, Senior project, Whitman College (2007).
74. Jean-Marie Mirebeau, *Numerical resolution of Euler equations, through semi-discrete optimal transport*, Journées Équations aux Dérivées Partielles (2015), 1–16.
75. Joseph SB Mitchell, David M Mount, and Christos H Papadimitriou, *The discrete geodesic problem*, SIAM Journal on Computing **16** (1987), no. 4, 647–668.
76. Gaspard Monge, *Mémoire sur la théorie des déblais et des remblais*, Histoire de l'Académie Royale des Sciences de Paris (1781).
77. Grégoire Montavon, Klaus-Robert Müller, and Marco Cuturi, *Wasserstein training of restricted Boltzmann machines*, Advances in Neural Information Processing Systems, 2016, pp. 3718–3726.
78. Lipeng Ning, Tryphon T Georgiou, and Allen Tannenbaum, *On matrix-valued Monge–Kantorovich optimal mass transport*, IEEE Transactions on Automatic Control **60** (2015), no. 2, 373–382.
79. Vladimir I. Oliker, *Near radially symmetric solutions of an inverse problem in geometric optics*, Inverse Problems **3** (1987), no. 4, 743.

80. Vladimir I. Oliker and Laird D. Prussner, *On the numerical solution of the equation* $\frac{\partial^2 z}{\partial x^2}\frac{\partial^2 z}{\partial y^2} - \left(\frac{\partial^2 z}{\partial x \partial y}\right)^2 = f$ *and its discretizations, I*, Numerische Mathematik **54** (1989), no. 3, 271–293.
81. James B Orlin, *A polynomial time primal network simplex algorithm for minimum cost flows*, Mathematical Programming **78** (1997), no. 2, 109–129.
82. Felix Otto, *The geometry of dissipative evolution equations: the porous medium equation*, (2001).
83. Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet, *Optimal transport with proximal splitting*, SIAM Journal on Imaging Sciences **7** (2014), no. 1, 212–238.
84. Brendan Pass, *Multi-marginal optimal transport: theory and applications*, ESAIM: Mathematical Modelling and Numerical Analysis **49** (2015), no. 6, 1771–1790.
85. Ofir Pele and Michael Werman, *Fast and robust earth mover's distances*, Proc. ICCV, IEEE, 2009, pp. 460–467.
86. Kaare Brandt Petersen and Michael Syskind Pedersen, *The matrix cookbook*, Technical University of Denmark **7** (2008), 15.
87. Gabriel Peyré, *Optimal transport with Benamou–Brenier algorithm*, http://www.numerical-tours.com/matlab/optimaltransp_2_benamou_brenier/, 2010.
88. Gabriel Peyré, *Entropic approximation of Wasserstein gradient flows*, SIAM Journal on Imaging Sciences **8** (2015), no. 4, 2323–2351.
89. Gabriel Peyré, Lénaïc Chizat, François-Xavier Vialard, and Justin Solomon, *Quantum entropic regularization of matrix-valued optimal transport*, European Journal of Applied Mathematics (2017), 1–24.
90. Gabriel Peyré and Marco Cuturi, *Computational optimal transport*, Submitted, 2017.
91. Gabriel Peyré, Marco Cuturi, and Justin Solomon, *Gromov–Wasserstein averaging of kernel and distance matrices*, International Conference on Machine Learning, 2016, pp. 2664–2672.
92. Alexander Plakhov, *Billiards, optimal mass transport and problems of optimal aerodynamic resistance*, Preprint (2012).
93. K. Ahuja Ravindra, Thomas L Magnanti, and James B. Orlin, *Network flows: theory, algorithms, and applications*, 1993.
94. Yossi Rubner, Carlo Tomasi, and Leonidas J Guibas, *The earth mover's distance as a metric for image retrieval*, International journal of computer vision **40** (2000), no. 2, 99–121.
95. Sartaj Sahni and Teofilo Gonzalez, *P-complete approximation problems*, Journal of the ACM (JACM) **23** (1976), no. 3, 555–565.
96. Filippo Santambrogio, *Optimal transport for applied mathematicians*, Springer, 2015.
97. ———, {*Euclidean, metric, and Wasserstein*} *gradient flows: an overview*, Bulletin of Mathematical Sciences **7** (2017), no. 1, 87–154.
98. Yuliy Schwartzburg, Romain Testuz, Andrea Tagliasacchi, and Mark Pauly, *High-contrast computational caustic design*, ACM Transactions on Graphics (TOG) **33** (2014), no. 4, 74.
99. James A Sethian, *Fast marching methods*, SIAM review **41** (1999), no. 2, 199–235.
100. Jonah Sherman, *Generalized preconditioning and undirected minimum-cost flow*, Proc. SODA, SIAM, 2017, pp. 772–780.
101. Richard Sinkhorn and Paul Knopp, *Concerning nonnegative matrices and doubly stochastic matrices*, Pacific Journal of Mathematics **21** (1967), no. 2, 343–348.
102. Morton Slater, *Lagrange multipliers revisited*, Cowles Commission Discussion Paper (1950), no. 403, 1–13.
103. Justin Solomon, *Computational optimal transport*, Snapshots of Modern Mathematics from Oberwolfach (2017), no. 8, 1–15.
104. Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas, *Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains*, ACM Transactions on Graphics (TOG) **34** (2015), no. 4, 66.
105. Justin Solomon, Leonidas Guibas, and Adrian Butscher, *Dirichlet energy for analysis and synthesis of soft maps*, Computer Graphics Forum, vol. 32, Wiley Online Library, 2013, pp. 197–206.
106. Justin Solomon, Gabriel Peyré, Vladimir G Kim, and Suvrit Sra, *Entropic metric alignment for correspondence problems*, ACM Transactions on Graphics (TOG) **35** (2016), no. 4, 72.

107. Justin Solomon, Raif Rustamov, Leonidas Guibas, and Adrian Butscher, *Earth mover's distances on discrete surfaces*, ACM Transactions on Graphics (TOG) **33** (2014), no. 4, 67.
108. ———, *Wasserstein propagation for semi-supervised learning*, International Conference on Machine Learning, 2014, pp. 306–314.
109. ———, *Continuous-flow graph transportation distances*, arXiv:1603.06927 (2016).
110. Matthew Staib, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka, *Parallel streaming Wasserstein barycenters*, Advances in Neural Information Processing Systems, 2017, pp. 2644–2655.
111. Lukas Svec, Sam Burden, and Aaron Dilley, *Applying Voronoi diagrams to the redistricting problem*, The UMAP Journal **28** (2007), no. 3, 313–329.
112. Hirofumi Uzawa, *Iterative methods for concave programming*, Studies in Linear and Non-Linear Programming **2** (1968), 154.
113. Sathamangalam R.S̃rinivasa Varadhan, *On the behavior of the fundamental solution of the heat equation with variable coefficients*, Communications on Pure and Applied Mathematics **20** (1967), no. 2, 431–455.
114. Leonid Nisonovich Vaserštei̇n, *Markov processes over denumerable products of spaces, describing large systems of automata*, Problemy Peredachi Informatsii **5** (1969), no. 3, 64–72.
115. Cédric Villani, *Topics in optimal transportation*, no. 58, American Mathematical Soc., 2003.
116. ———, *Optimal transport: old and new*, vol. 338, Springer Science & Business Media, 2008.
117. Xu-Jia Wang, *On the design of a reflector antenna*, Inverse problems **12** (1996), no. 3, 351.
118. David F Watson, *Computing the n-dimensional Delaunay tessellation with application to Voronoi polytopes*, The Computer Journal **24** (1981), no. 2, 167–172.

MIT Department of Electrical Engineering and Computer Science
E-mail address: jsolomon@mit.edu