Optimal Mass Transport: Signal processing and machine-learning applications

Article *in* IEEE Signal Processing Magazine · July 2017

DOI: 10.1109/MSP.2017.2695801

5 authors, including:

Soheil Kolouri
HRL Laboratories, LLC
70 PUBLICATIONS **1,149** CITATIONS
SEE PROFILE

Serim Park
Carnegie Mellon University
11 PUBLICATIONS **538** CITATIONS
SEE PROFILE

Transport-based analysis, modeling, and learning from signal and data distributions

Soheil Kolouri, Serim Park, Matthew Thorpe, Dejan Slepčev, Gustavo K. Rohde

***Abstract*—Transport-based techniques for signal and data analysis have received increased attention recently. Given their abilities to provide accurate generative models for signal intensities and other data distributions, they have been used in a variety of applications including content-based retrieval, cancer detection, image super-resolution, and statistical machine learning, to name a few, and shown to produce state of the art in several applications. Moreover, the geometric characteristics of transport-related metrics have inspired new kinds of algorithms for interpreting the meaning of data distributions. Here we provide an overview of the mathematical underpinnings of mass transport-related methods, including numerical implementation, as well as a review, with demonstrations, of several applications. Software accompanying this tutorial is available at [134].**

I. Introduction

A. Motivation and goals

Numerous applications in science and technology depend on effective modeling and information extraction from signal and image data. Examples include being able to distinguish between benign and malignant tumors from medical images, learning models (e.g. dictionaries) for solving inverse problems, identifying people from images of faces, voice profiles, or fingerprints, and many others. Techniques based on the mathematics of optimal mass transport have received significant attention recently (see Figure 1 and section II) given their ability to incorporate spatial (in addition to intensity) information when comparing signals, images, other data sources, thus giving rise to different geometric interpretation of data distributions. These techniques have been used to simplify and augment the accuracy of numerous pattern recognition-related problems. Some examples covered in this tutorial include image retrieval [136], [121], [98] , signal and image representation [165], [133], [145], [87], [89], inverse problems [126], [153], [95], [21], cancer detection [164], [11], [118], [156], texture and color modeling [42], [126], [127], [53], [123], [52], shape and image registration [70], [69], [71], [167], [159], [111], [103], [54], [93], [150], segmentation [113], [144], [125], watermarking [104], [105], and machine learning [148], [31], [60], [91], [110], [46], [81], [131], [116], [130], to name a few. This tutorial is meant to serve as an *introductory guide* to those wishing to familiarize themselves with these emerging techniques. Specifically we

- provide a brief overview on the mathematics of optimal mass transport
- describe recent advances in transport related methodology and theory
- provide a pratical overview of their applications in modern signal analysis, modeling, and learning problems.

Software accompanying this tutorial is available at [134].

B. Why transport?

In recent years numerous techniques for signal and image analysis have been developed to address important learning and estimation problems. Researchers working to find solutions to these problems have found it necessary to develop techniques to compare signal intensities across different signal/image coordinates. A common problem in medical imaging, for example, is the analysis of magnetic resonance images with the goal of learning brain morphology differences between healthy and diseased populations. Decades of research in this area have culminated with techniques such as voxel and deformation-based morphology [6], [7] which make use of nonlinear registration methods to understand differences in tissue density and locations [66], [149]. Likewise, the development of dynamic time warping techniques was necessary to enable the comparison of time series data more meaningfully, without confounds from commonly encountered variations in time [80]. Finally, researchers desiring to create realistic models of facial appearance have long understood that appearance models for eyes, lips, nose, etc. are significantly different and must thus be dependent on position relative to a fixed anatomy [30]. The pervasive success of these, as well as other techniques such as optical flow [10], level-set methods [33], deep neural networks [142], for example, have thus taught us that 1) nonlinearity and 2) modeling the location of pixel intensities are essential concepts to keep in mind when solving modern regression problems related to estimation and classification.

We note that the methodology developed above for modeling appearance and learning morphology, time series analysis and predictive modeling, deep neural networks for classification of sensor data, etc., is algorithmic in nature. The transport-related techniques reviewed below are nonlinear methods that, unlike linear methods such as Fourier, wavelets, and dictionary models, for example, explicitly model jointly signal intensities as well as their locations. Furthermore, they are often based on the theory of optimal mass transport from which fundamental principles can be put to use. Thus they hold the promise to ultimately play a significant role in the development of a theoretical foundation for certain subclasses of modern learning and estimation problems.

C. Overview and outline

As detailed below in section II the optimal mass transport problem first arose due to Monge [109]. It was later expanded

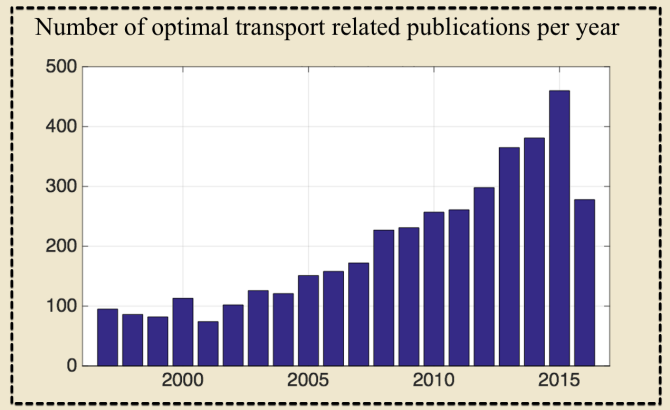

Fig. 1. The number of publications by year that contain any of the keywords: 'optimal transport', 'Wasserstein', 'Monge', 'Kantorovich', 'earth mover' according to webofknowledge.com. By comparison there were 819 in total pre 1997. The 2016 statistics is correct as of 29th August 2016.

by Kantarovich [74], [75], [77], [76] and found applications in operations research and economics. Section III provides an overview of the mathematical principles and formulation of optimal transport-related metrics, their geometric interpretation, and related embedding methods and signal transforms. We also explain Brenier's theorem [22], which helped pave the way for several practical numerical implementation algorithms, which are then explained in detail in section IV. Finally, in section V we review and demonstrate the application of transport-based techniques to numerous problems including: image retrieval, registration and morphing, color and texture analysis, image denoising and restoration, morphometry, super resolution, and machine learning. As mentioned above, software implementing the examples shown can be downloaded from [134].

II. A brief historical note

The optimal mass transport problem seeks the most efficient way of transforming one distribution of mass to another, relative to a given cost function. The problem was initially studied by the French Mathematician Gaspard Monge in his seminal work "Mémoire sur la théorie des déblais et des remblais" [109] in 1781. Between 1781 and 1942, several mathematicians made valuable contributions to the mathematics of the optimal mass transport problem among which are Arthur Cayley [27], Charles Duplin [44], and Paul Appell [5]. In 1942, Leonid V. Kantorovich, who at that time was unaware of Monge's work, proposed a general formulation of the problem by considering optimal mass transport plans, which as opposed to Monge's formulation allows for mass splitting [74]. Six years later and in 1948, Kantorovich became aware of Monge's work and made the connection between his work and that of the French mathematician in his short paper, "A Problem of Monge" [75]. Kantorovich shared the 1975 Nobel Prize in Economic Sciences with Tjalling Koopmans for his work in the optimal allocation of scarce resources. Kantorovich's contribution is considered as "the birth of the modern formulation of optimal transport" [163] and it made the optimal mass transport problem an active field of research in the following years [138], [96], [85], [128]. Among the published works in this era is the prominent work of Sudakhov [151] in 1979 on the existence of the optimal transport map. Later on, a gap was found in Sudakhov's work [151] which was eventually filled by Ambrosio in 2003 [1].

A significant portion of the theory of the optimal mass transport problem was developed in the Nineties. Starting with Brenier's seminal work on characterization, existence, and uniqueness of optimal transport maps [22], followed by Cafarrelli's work on regularity conditions of such mappings [25], Gangbo and McCann's work on geometric interpretation of the problem [62], [63], and Evans and Gangbo's work on formulating the problem through differential equations and specifically the p-Laplacian equation [50], [51]. A more thorough history and background on the optimal mass transport problem can be found in Rachev and Rüschendorf's book "Mass Transportation Problems" [129], Villani's book "Optimal Transport: Old and New" [163], and Santambrogio's book " Optimal transport for applied mathematicians" [139], or in shorter surveys such as that of Bogachev and Kolesnikov's [18] and Vershik's [162].

The significant contributions in mathematical foundations of the optimal transport problem together with recent advancements in numerical methods [36], [15], [14], [114], [160] have spurred the recent development of numerous data analysis techniques for modern estimation and detection (e.g. classification) problems. Figure 1 plots the number of papers related to the topic of optimal transport that can be found in the public domain per year demonstrating significant growth in these techniques.

III. Formulation of the problem and methodology

In this section we first review the two classical formulations of the optimal transport problem (i.e. Monge's and Kantorovich's formulations). Next, we review the geometrical characteristics of the problem, and finally review the transport based signal/image embeddings.

A. Optimal Transport: Formulation

*1) **Monge's formulation**:* The Monge optimal mass transportation problem is formulated as follows. Consider two probability measures μ and ν defined on measure spaces X and Y. In most applications X and Y are subsets of $\mathbb{R}^d$ and μ and ν have density functions which we denote by I_0 and I_1, $d\mu(x) = I_0(x)dx$ and $d\nu(x) = I_1(x)dx$, (originally representing the height of a pile of soil/sand and the depth of an excavation). Monge's optimal transportation problem is to find a measurable map $f : X \to Y$ that pushes μ onto ν and minimizes the following objective function,

$$M(\mu,\nu) = \inf_{f\in MP} \int_X c(x, f(x))d\mu(x) \tag{1}$$

where $c : X \times Y \to \mathbb{R}^+$ is the cost functional, and $MP := \{f : X \to Y \mid f_{\#}\mu = \nu\}$ where $f_{\#}\mu$ represents the pushforward of measure μ and is characterized as,

$$\int_{f^{-1}(A)} d\mu(x) = \int_A d\nu(y) \quad \text{for any measurable } A \subset Y. \tag{2}$$

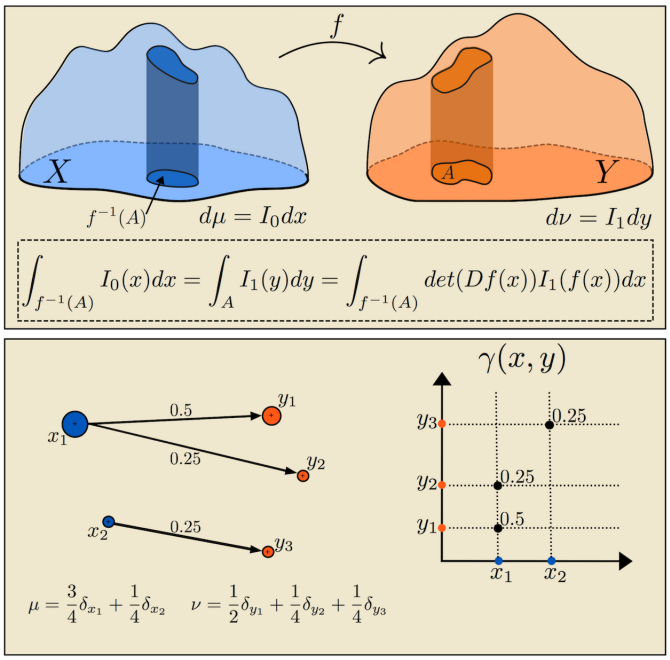

Fig. 2. Monge transport map (top panel) and Kantorovich's transport plan (bottom panel).

If μ and ν have densities and f is smooth and one to one then the equation above can be written in a differential form as

$$det(Df(x))I_1(f(x)) = I_0(x)$$

almost everywhere, where Df is the Jacobian of f (See Figure 2, top panel). Such measurable maps $f \in MP$ are sometimes called 'transport maps' or 'mass preserving maps'. Simply put, the Monge formulation of the problem seeks the best pushforward map that rearranges measure μ into measure ν while minimizing a specific cost function. Monge considered the Euclidean distance as the cost function in his original formulation, $c(x, f(x)) = |x - f(x)|$. Note that both the objective function and the constraint in Equation (1) are nonlinear with respect to $f(x)$. Hence, for over a century the answers to questions regarding existence and characterization of the Monge's problem remained unknown.

It should be mentioned that, for certain measures the Monge's formulation of the optimal transport problem is ill-posed; in the sense that there is no transport map to rearrange one measure to another. For instance, consider the case where μ is a Dirac mass while ν is not. Kantorovich's formulation alleviates this problem by finding the optimal transport plan as opposed to the transport map.

*2) **Kantorovich's formulation**:* Kantorovich formulated the transportation problem by optimizing over transportation plans, where a transport plan is a probability measure $\gamma \in P(X \times Y)$ with marginals μ and ν. One can think of γ as the joint distribution of I_0 and I_1 describing how much 'mass' is being moved to different coordinates. That is let A be a measurable subset of X and similarly $B \subseteq Y$. The quantity $\gamma(A, B)$ tells us how much 'mass' in set A is being moved to set B. Now let $\gamma \in P(X \times Y)$ be a plan with marginals μ and ν, i.e.,

$$(\Gamma_X)_{\#}\gamma = \mu \quad \text{and} \quad (\Gamma_Y)_{\#}\gamma = \nu$$

where $\Gamma_X : X \times Y \to X$ and $\Gamma_Y : X \times Y \to Y$ are the canonical projections. Let $\Gamma(\mu, \nu)$ be the set of all such plans. Kantorovich's formulation can then be written as,

$$K(\mu, \nu) = \min_{\gamma \in \Gamma(\mu,\nu)} \int_{X \times Y} c(x, y) d\gamma(x, y). \tag{3}$$

The minimizer of the optimization problem above, γ^*, is called the optimal transport plan. Note that unlike the Monge problem, in Kantorovich's formulation the objective function and the constraints are linear with respect to $\gamma(x, y)$. Moreover, Kantorovich's formulation is in the form of a convex optimization problem. Note that the Monge problem is more restrictive than the Kantorovich problem. That is, in Monge's version, mass from a single location in one domain is being sent to a single location in the domain of the other measure. Kantorovich's formulation considers transport plans which can deal with arbitrary measurable sets and has the ability to distribute mass from the one location in one density to multiple locations in another (See Figure 2, bottom panel). For any transport map $f : X \to Y$ there is an associated transport plan, given by

$$\gamma = (\mathrm{Id} \times f)_{\#}\mu, \tag{4}$$

which means that $\gamma(A, B) = \mu(\{x \in A : f(x) \in B\})$, or in other words for any integrable function $h(x, y)$ on measure space $X \times Y$

$$\int c(x, y) d\gamma(x, y) = \int c(x, f(x)) d\mu(x). \tag{5}$$

Furthermore when an optimal transport map f^* exists, with the optimal transport then γ^* given by (4) is an optimal transportation plan [163]. Note that the converse does not always hold. Finally, the following relationship holds between Equations (1) and (3) [18], [163],

$$K(\mu, \nu) \leq M(\mu, \nu). \tag{6}$$

The Kantorovich problem is especially interesting in a discrete setting, that is for probability measures of the form $\mu = \sum_{i=1}^{M} p_i \delta_{x_i}$ and $\nu = \sum_{j=1}^{N} q_j \delta_{y_j}$, where δ_{x_i} is a dirac measure centered at x_i, the Kantorovich problem can be written as,

$$\begin{aligned} K(\mu, \nu) \quad &= \quad \min_{\gamma} \sum_i \sum_j c(x_i, y_j)\gamma_{ij} \\ &s.t. \quad \sum_j \gamma_{ij} = p_i, \ \sum_i \gamma_{ij} = q_j \\ &\gamma_{ij} \geq 0, i = 1, ..., M, \ j = 1, ..., N \end{aligned} \tag{7}$$

where γ_{ij} identifies how much of the mass particle m_i at x_i needs to be moved to y_j (See Figure 2, bottom panel). Note that the optimization above has a linear objective function and linear constraints, therefore it is a linear programming problem. We note that the problem is convex, but not strictly so, and the constraint provides a polyhedral set of $M \times N$ matrices, $\Gamma(\mu, \nu)$.

In practice, a non-discrete measure is often approximated by a discrete measure and the Kantorovich problem is solved through the linear programming optimization expressed in equation (7). An important result that allows this discretization is the following. If μ^h and ν^h are sequences of measures that weakly converge to μ and ν respectively and γ^h is the optimal transport plan between μ^h and ν^h then, up to subsequences, γ^h converges weakly to the optimal transport plan between μ and ν, γ^* (see for example [163, Thm 5.20]). This result enables the application of discrete OT methods to more general measures.

Here we note that the Kantorovich problem can be extended to multi-marginal problems where instead of working in $X \times Y$ one defines $\chi := X_1 \times X_2 \times ... \times X_N$, the cost $c : \chi \to \mathbb{R}^+$, and marginals are defined as $(\Gamma_{X_i})_{\#}\gamma = \mu_i$. Multi-marginal transport problems and Martingale optimal transport problems [119], [139], [29] are two variations of the transport problem, which have recently attracted attention in economics [119], [72].

3) ***Dual formulation****:* The focus of the Kantorovich problem is to minimize the *cost* of transportation. Villani [163] describes the problem by an intuitive example using the transportation of bread between bakeries and cafés. Consider a company that owns bakeries throughout a city, which produce loaves that should be transported to cafés. If this company were to take delivery into its own hands, then it would like to minimize the transportation costs. Minimizing the cost of the transportation between bakeries and cafés is the focus of Kantorovich's problem. Now imagine that a second company comes into play, which offers to take care of all transportation (maybe they have a few tricks to do the transportation in a very efficient manner) offering competitive prices for buying bread at bakeries and selling them at cafés. Let $\phi(x)$ be the cost of bread bought at bakery x, and $\psi(y)$ be its selling price at café y. Before the second company showed up, the first company's selling price was equal to $\phi(x) + c(x, y)$, which is the price of the bread plus the delivery fee (transportation cost). Now if the second company wants to be competitive, it needs to sell bread to cafés with a price which is lower or equal to that which the first company offers. Meaning that,

$$\forall(x, y), \quad \psi(y) \leq c(x, y) + \phi(x) \tag{8}$$

At the same time, the second company of course wants to maximize its *profit*, which can be formulated as,

$$\begin{aligned} KD(\mu,\nu) = \quad & \sup_{\psi,\phi} \quad \int_Y \psi(y)d\nu(y) - \int_X \phi(x)d\mu(x) \\ & s.t. \qquad \psi(y) - \phi(x) \leq c(x, y), \quad \forall(x, y) \end{aligned} \tag{9}$$

Note that maximizing the profit is the dual problem [76], [129], [163], [139], [18] of the Kantorovich formulation. In addition, since the primal problem is convex we have,

$$K(\mu, \nu) = KD(\mu, \nu). \tag{10}$$

It is apparent from this scenario that the optimal prices for the dual problem must be tight meaning that, the second company cannot increase the selling price, $\psi(y)$, or decrease the buying price, $\phi(x)$. Formally,

$$\begin{aligned} \psi(y) &= \inf_x(\phi(x) + c(x, y)) \\ \phi(x) &= \sup_y(\psi(y) - c(x, y)) \end{aligned} \tag{11}$$

The equations above imply that the pair ψ and ϕ are related through a *c-transform* (see [163] for the definition and more detail), which coincides with the Legendre-Fenchel transform [20] when $c(x, y) = -\langle x, y\rangle$. Finally, for the quadratic transport cost, $c(x, y) = \frac{1}{2}|x - y|^2$, and when an optimal transport map exists, the optimal arguments of the dual problem and the optimal transport map are related through, $f^*(x) = \nabla(\frac{1}{2}|x|^2 - \phi^*(x))$.

4) ***Basic properties****:* The existence of a solution for the Kantorovich problem follows from Ulham's [158] and Prokhorov's [122] theorems. Ulham's theorem states that $\Gamma(\mu, \nu)$ is a tight set of probability measures defined on $X \times Y$. On the other hand, Prokhorov's theorem [122] states that $\Gamma(\mu, \nu)$ is relatively compact, and hence the limit of a sequence of transport plans $(\gamma_n) \subset \Gamma(\mu, \nu)$ where $\gamma_n \to \gamma$ is also in $\Gamma(\mu, \nu)$. Using these theorems one can show the following theorem [2], [163]: *For a lower semicontinuous and bounded from below cost function, $c(x, y)$ there exists a minimizer to the Kantorovich problem.*

In engineering applications the common cost functions almost always satisfy the existence of a transport plan. A further important question is regarding the existence of an optimal transport map instead of a plan. Brenier [22] addressed this problem for the special case where $c(x, y) = |x - y|^2$. Bernier's results was later relaxed to more general cases by Gangbo and McCann [63], which led to the following theorem:

Theorem Let μ and ν be two Borel probability measures on compact measurable supports X and Y, respectively. When $c(x, y) = h(x-y)$ for some strictly convex function h and μ is absolutely continuous with respect to the Lebesgue measure, then there exists a unique optimal transportation map $f^* : X \to Y$ such that $f^*_{\#}\mu = \nu$,

$$\int_X h(x - f^*(x))d\mu(x) = \min_{\gamma\in\Pi(\mu,\nu)} \int_{X\times Y} h(x - y)d\gamma(x, y). \tag{12}$$

In addition, the optimal transport plan is unique, and thus equal to $\gamma(x, y) = (\mathrm{Id} \times f^*)_{\#}\mu$ (See Eq. 4). Moreover, f^* is characterized by the gradient of a c-concave function $\phi : X \to \mathbb{R}$ as follows,

$$f^*(x) = x - \nabla h^{-1}(\nabla\phi(x)). \tag{13}$$

For a proof see [2], [63], [163]. Note that, $h(x) = \frac{1}{2}|x|^2$ is rather a special case, because the gradient of h is equal to identity,$\nabla h = \mathrm{Id}$, and the optimal transport map is simply characterized as $f^*(x) = x - \nabla\phi(x)$.

B. Optimal Mass Transport: Geometric properties

1) ***Wasserstein metric****:* Let Ω be a subset of $\mathbb{R}^d$ on which the measures we consider are defined. In most applications Ω is the domain where the signal is defined and thus bounded. Let $P_p(\Omega)$ be the set of Borel probability measures on Ω, with finite p'th moment, that is the set of probability measures

μ on $\mathbb{R}^d$ such that $\int_\Omega |x|^p d\mu(x) < \infty$. The p-Wasserstein metric, W_p, for $p \geq 1$ on $P_p(\Omega)$ is then defined as using the optimal transportation problem (3) with the cost function $c(x,y) = |x-y|^p$. For μ and ν in $P_p(\Omega)$,

$$W_p(\mu,\nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\Omega\times\Omega} |x-y|^p d\gamma(x,y) \right)^{\frac{1}{p}}. \quad (14)$$

For any $p \geq 1$, W_p is a metric on $P_p(\Omega)$. The metric space $(P_p(\Omega), W_p)$ is referred to as the p-Wasserstein space. If Ω is bounded then for any $p \geq 1$, W_p metrizes the weak convergence of measures on $P(\Omega)$. That is the convergence with respect to W_p is equivalent to weak convergence of measures.

Note that the p-Wasserstein metric can equivalently be defined using the dual Kantorovich problem,

$$W_p(\mu,\nu) = \left(\sup_{\phi} \left\{ \int_\Omega \phi(x) d\mu(x) - \int_\Omega \phi^c(y) d\nu(y) \right\} \right)^{\frac{1}{p}} \quad (15)$$

where $\phi^c(y) = \inf_x \{\phi(x) - |x-y|^p\}$.

For the specific case of $p = 1$ the p-Wasserstein metric is also known as the Monge–Rubinstein [163] metric, or the earth mover distance [136], which can also be expressed in the following form

$$W_1(\mu,\nu) = \sup_{Lip(\phi)\leq 1} \left\{ \int_\Omega \phi(x) d(\mu(x) - \nu(x)) \right\}, \quad (16)$$

where $Lip(\phi)$ is the Lipschitz constant for ϕ.

The p-Wasserstein metric for one-dimensional probability measures is specifically interesting due to its simple and unique characterization. For one-dimensional probability measures μ and ν on $\mathbb{R}$ the optimal transport map has a closed form solution. Let F_μ be the cumulative distribution function of a measure $\mu \in P_p(\mathbb{R})$

$$F_\mu(x) = \mu((-\infty, x)) \quad (17)$$

Note that this is a nondecreasing function going from 0 to 1, which is continuous if μ has a density. We define the *pseudoinverse* of F_μ as follows: for $z \in (0,1)$, $F^{-1}(z)$ is the smallest x for which $F_\mu(x) \geq z$, that is

$$F_\mu^{-1}(z) = \inf \{x \in \mathbb{R} \ : \ F_\mu(x) \geq z\} \quad (18)$$

If μ has positive density then F_μ is increasing (and thus invertible) and the inverse of the function F_μ is equal to the pseudoinverse we just defined. In other words the pseudoinverse is a generalization of the notion of the inverse of a function. The pseudoinverse (i.e. the inverse if the densities of μ and ν are positive) provides a closed form solution for the p-Wasserstein distance:

$$W_p(\mu,\nu) = \left(\int_0^1 |F_\mu^{-1}(z) - F_\nu^{-1}(z)|^p dz \right)^{\frac{1}{p}}. \quad (19)$$

The closed-form solution of the p-Wasserstein distance in one dimension is an attractive property, as it alleviates the need for optimization. This property was employed in the Sliced Wasserstein metrics as defined below.

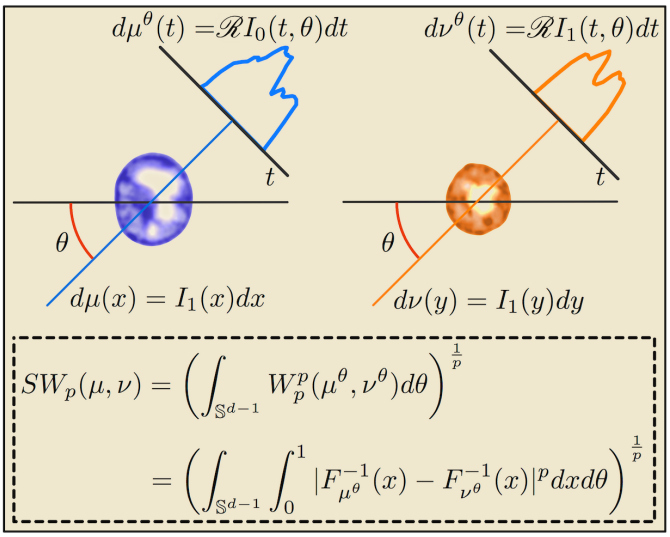

Fig. 3. Slicing measures μ and ν and the corresponding Sliced Wasserstein distance between them. The images represent segmented nuclei extracted from histopathology images.

*2) **Sliced-Wasserstein Metric**:* The idea behind the Sliced Wasserstein metric is to first obtain a family of one-dimensional representations for a higher-dimensional probability distribution through projections (slicing the measure), and then calculate the distance between two input distributions as a functional on the Wasserstein distance of their one-dimensional representations. In this sense, the distance is obtained by solving several one-dimensional optimal transport problems, which have closed-form solutions.

Slicing a measure is closely related to the well known Radon transform in the imaging and image processing community [19], [87]. The d-dimensional Radon transform $\mathscr{R}$ maps a function $I \in L_1(\mathbb{R}^d)$ where $L_1(\mathbb{R}^d) := \{I : \mathbb{R}^d \to \mathbb{R} | \int_{\mathbb{R}^d} |I(x)|dx \leq \infty\}$ into the set of its integrals over the hyperplanes of $\mathbb{R}^n$ and is defined as,

$$\mathscr{R}I(t,\theta) := \int_{\mathbb{R}} I(t\theta + s\theta^\perp) ds, \quad \forall t \in \mathbb{R}, \ \forall \theta \in \mathbb{S}^{d-1} \quad (20)$$

here $\theta^\perp$ is the subspace orthogonal to θ, and $\mathbb{S}^{d-1}$ is the unit sphere in $\mathbb{R}^d$. Note that $\mathscr{R} : L_1(\mathbb{R}^d) \to L_1(\mathbb{R}\times\mathbb{S}^{d-1})$. One can slice a probability measure μ defined on $\mathbb{R}\times\mathbb{S}^{d-1}$ into its conditional measures with respect to the uniform measure on $\mathbb{S}^{d-1}$ to obtain a measure μ^θ, which satisfies,

$$\int_{\mathbb{R}*\mathbb{S}^{d-1}} g(t,\theta) d\mu(t,\theta) = \int_{\mathbb{S}^{d-1}} \left(\int_{\mathbb{R}} g(t,\theta) d\mu^\theta(t) \right) d\theta. \quad (21)$$

for $\forall g \in L_1(\mathbb{R}\times\mathbb{S}^{d-1})$. The Sliced Wasserstein metric, for continuous probability measures μ and ν on $\mathbb{R}^d$ is then defined as,

$$SW_p(\mu,\nu) = \left(\int_{\mathbb{S}^{d-1}} W_p^p(\mu^\theta,\nu^\theta) d\theta \right)^{\frac{1}{p}} \quad (22)$$

where $p \geq 1$, and W_p is the Wasserstein metric, which for one dimensional measures μ^θ and ν^θ has a closed form solution (See Figure 3). For more details and definitions of the Sliced Wasserstein metric we refer the reader to [127], [93], [19], [87].

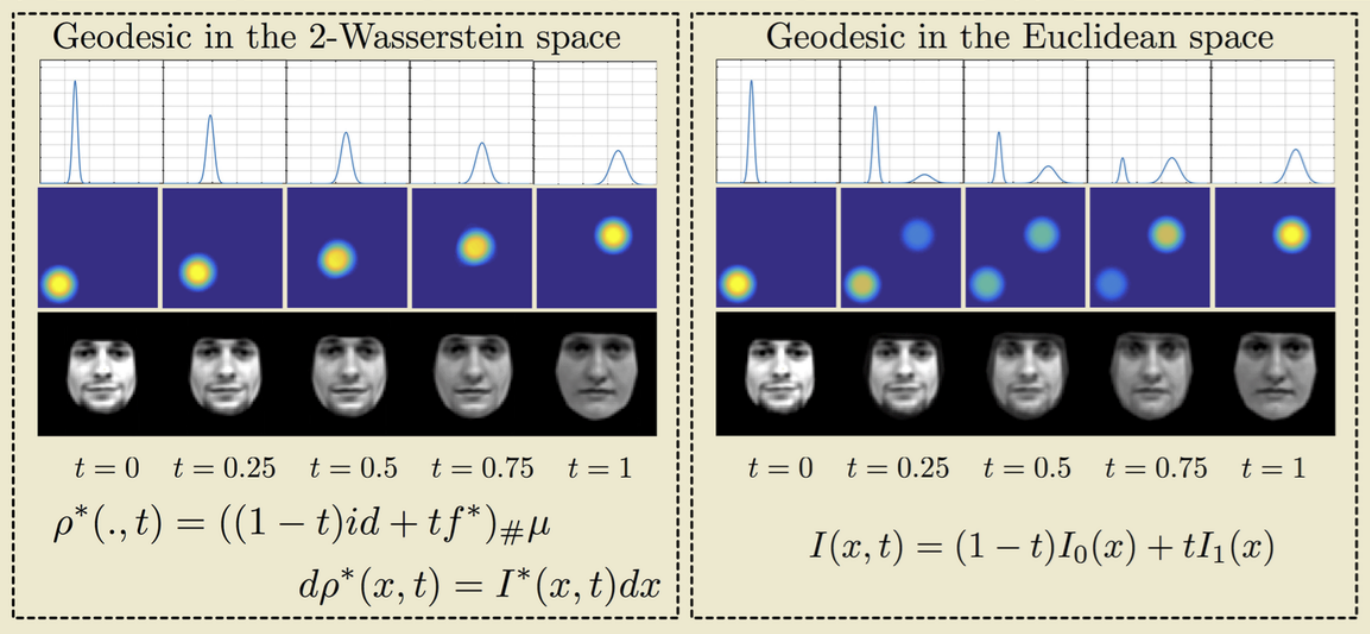

Fig. 4. Geodesic in the 2-Wasserstein space (left panel), and visualization of the geodesic, ρ^*, for one-dimensional signals and two-dimensional images (right panel). Note that the geodesic captures the nonlinear structure of the signals and images and provides a natural morphing.

*3) **Wasserstein spaces, geodesics, and Riemannian structure**:* In this section we assume that Ω is convex. Here we highlight that the p-Wasserstein space $(P_p(\Omega), W_p)$ is not just a metric space, but has additional geometric structure. In particular for any $p \geq 1$ and any $\mu_0, \mu_1 \in P_p(\Omega)$ there exists a continuous path between μ_0 and μ_1 whose length is the distance between μ_0 and μ_1.

Furthermore the space with $p = 2$ is special as it possesses a structure of an formal, infinite dimensional, Riemannian manifold. That structure was first noted by Otto [117] who developed the formal calculations for using this structure. The weak setting in which the notions of tangent space, Riemannian metric were made rigorous was developed by Ambrosio, Gigli and Savaré [3]. A slightly less general, but more approachable introduction is available in [2] and also in [139]. The notion of curvature was developed in [64], [101].

Here we review the two main notions, which have a wide use. Namely we characterize the geodesics in $(P_p(\Omega), W_p)$ and in the case $p = 2$ describe what is the local, Riemannian metric of $(P_2(\Omega), W_2)$. Finally we state the seminal result of Benamou and Brenier [14] who provided a characterization of geodesics via action minimization which is useful in computations and also gives an intuitive explanation of the Wasserstein metric.

We first recall the definition of the length of a curve in a metric space. Let (X, d) be a metric space and $\mu : [a, b] \to X$. Then the length of μ, denoted by $L(\mu)$ is

$$L(\mu) = \sup_{m\in\mathbb{N},\, a=t_0<t_1<\cdots<t_m=b} \sum_{i=1}^{m} d(\mu(t_{i-1}), \mu(t_i)).$$

A metric space (X, d) is a *geodesic space* if for any μ_0 and μ_1 there exists a curve $\mu : [0, 1] \to X$ such that $\mu(0) = \mu_0$, $\mu(1) = \mu_1$ and for all $0 \leq s < t \leq 1$, $d(\mu(s), \mu(t)) = L(\mu|_{[s,t]})$. In particular the length of μ is equal to the distance from μ_0 to μ_1. Such curve μ is called a *geodesic*. The existence of geodesic is useful as it allows one to define the average of μ_0 and μ_1 as the midpoint of the geodesic connection them.

An important property of $(P_p(\Omega), W_p)$ is that it is a geodesic space and that geodesics are easy to characterize. Namely they are given by the *displacement interpolation* (a.k.a. McCann interpolation). We first describe them for μ_0 in $P_p(\Omega)$ which has a density: $d\mu_0 = I_0 dx$ and arbitrary $\mu_1 \in P_p(\Omega)$. Then there exists a unique transportation map $f^*_\sharp \mu_0 = \mu_1$ which minimizes the transportation cost for the given $p \geq 1$. The geodesic is obtained by moving the mass at constant speed from x to $f^*(x)$. More precisely, let $t \in [0, 1]$ and $x \in \Omega$ let

$$f^*_t(x) = (1 - t)x + tf^*(x)$$

be the position at time t of the mass initially at x. Note that f^*_0 is identity mapping and $f^*_1 = f^*$. Pushing forward the mass by f^*_t:

$$\mu^*(t) = f^*_{t\,\sharp}\mu_0$$

provides the desired geodesic from μ_0 to μ_1. We remark that the velocity of each particle $\partial_t f^*_t = f^*(x) - x$ which is nothing but the displacement of the optimal transportation map.

If μ_0 does not have a density, we need to use the optimal transportation plan $\gamma \in \Gamma(\mu_0, \mu_1)$ which minimizes the p-transportation cost. For $t \in [0, 1]$ let $F_t : \Omega \times \Omega \to \Omega$ be the convex interpolation: $F_t(x, y) = (1 - t)x + ty$. Then $\mu^*(t) = F_{t\,\sharp}\gamma$ is the desired geodesic. The intuitive explanation is that mass from x which ends up at y is being transported at constant speed, along the line segment from x to y. Figure 4 conceptualizes the geodesic between two measures in $P_2(\Omega)$, and visualizes it for three different pairs of measures.

An important fact regarding the 2-Wasserstein space is Otto's presentation of a formal Riemannian metric for this

space [117]. It involves shifting to Lagrangian point of view. To explain, consider path $\mu(t)$ in $P_2(\Omega)$ with smooth densities $I(x,t)$. Then $s(x,t) = \frac{\partial I}{\partial t}(x,t)$, the perturbation of $\mu(t)$, can be thought as a tangent vector. Instead of thinking of increasing/decreasing the density this perturbation can be viewed as resulting form moving the mass by a vector field. In other words consider vector fields $v(x,t)$ such that

$$s = -\nabla \cdot (Iv). \tag{23}$$

There are many such vector fields. Otto defined the size of $s(\,\cdot\,,t)$ as the (square root of) the minimal kinetic energy of the vector field that produces the perturbation to density s. That is

$$\langle s,s\rangle = \min_{v \text{ satisfies (23)}} \int |v|^2 I dx \tag{24}$$

Utilizing the Riemmanian manifold structure of $P_2(\Omega)$ together with the inner product presented in Equation (24) the 2-Wasserstein metric can be reformulated into finding the minimizer of the following action among all curves in $P_2(\Omega)$ connecting μ and ν [14],

$$\begin{aligned} W_2^2(\mu,\nu) = &\inf_{I,\mathrm{v}} \int_0^1 \int_\Omega I(x,t)|\mathrm{v}(x,t)|^2 dxdt \\ \text{such that } &\partial_t I + \nabla\cdot(I\mathrm{v}) = 0 \\ &I(\cdot,0) = I_0(\cdot),\ I(\cdot,1) = I_1(\cdot) \end{aligned} \tag{25}$$

where the first constraint is the continuity equation.

C. Optimal Transport: Embeddings and Transforms

The optimal transport problem and specifically the 2-Wasserstein metric and the Sliced 2-Wasserstein metric have been recently used to define bijective nonlinear transforms for signals and images [165], [133], [87], [89]. In contrast to commonly used linear signal transformation frameworks (e.g. Fourier and Wavelet transforms) which only employ signal intensities at fixed coordinate points, thus adopting an 'Eulerian' point of view, the idea behind the transport-based transforms is to consider the intensity variations together with the locations of the intensity variations in the signal. Therefore, such transforms adopt a 'Lagrangian' point of view for analyzing signals. Here we briefly describe these transforms and some of their prominent properties.

*1) **The linear optimal transportation framework**:* The linear optimal transportation (LOT) framework was proposed by Wang et al. [165]. The framework was successfully used in [11], [155], [118], [156] for pattern recognition in biomedical images and specifically histopathology and cytology images. Later, it was extended in [89] as a generic framework for pattern recognition and was used in [88] for single-frame super-resolution reconstruction of face images. The framework was modified by Seguy and Cuturi [145], where they studied the principal geodesics of probability measure. The LOT framework provides an invertible Lagrangian transform for images. It was initially proposed as a method to simultaneously amend the computationally expensive requirement of calculating pairwise 2-Wasserstein distances between N probability measures for pattern recognition purposes, and to allow for the construction of generative models for images involving

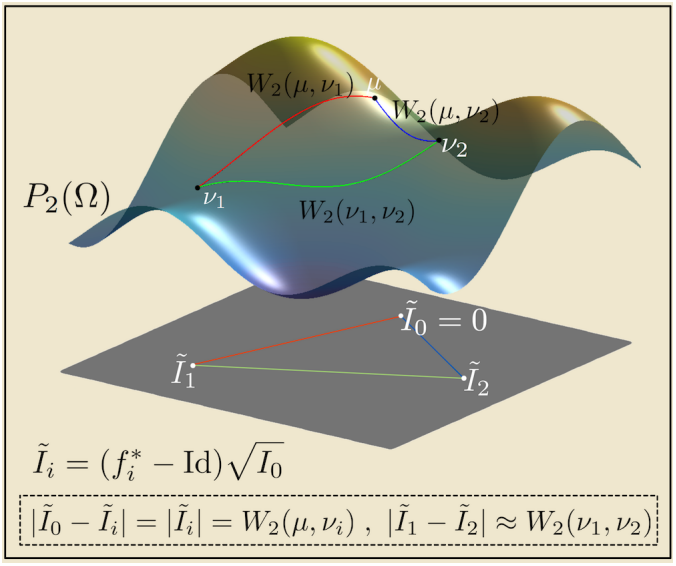

Fig. 5. Graphical representation of the LOT framework. The framework embeds the probability measures ν_i in the tangent space of $P_2(\Omega)$ at a fixed probability measure μ with density I_0. As a consequence, the Euclidean distance between the embedded functions $\tilde{I}_1$ and $\tilde{I}_2$ provides an approximation for the 2-Wasserstein distance, $W_2(\nu_1,\nu_2)$.

textures and shapes. For a given set of probability measures $\nu_i \in P_2(\Omega)$, for $i = 1,...,N$, and a fixed template probability measure μ, the transform projects the probability measures to the tangent space at μ, T_μ. The projections are acquired by finding the optimal velocity fields corresponding to the optimal transport plans between μ and each probability measure in the set.

Formally, the framework provides a linear embedding for $P_2(\Omega)$ with respect to a fixed measure $\mu \in P_2(\Omega)$. Meaning that, the Euclidean distance between the embedded measures, $\tilde{\nu}_i$, and the fixed measure, μ, is equal to $W_2(\mu,\nu_i)$ and the Euclidean distance between two embedded measures is, generally speaking, an approximation of their 2-Wasserstein distance. The geometric interpretation of the LOT framework is presented in Figure 5. The linear embedding then facilitates the application of linear techniques such as principal component analysis (PCA) and linear discriminant analysis (LDA) to probability measures.

*2) **The cumulative distribution transform**:* Park et al. [133] considered the LOT framework for one-dimensional probability measures, and since in dimension one the transport maps are explicit, they were able to characterize the properties of the transformed densities in the tangent space. Here, we briefly review their results. Similar to the LOT framework, let ν_i for $i = 1,...,N$, and μ be absolutely continuous probability measures defined on $\mathbb{R}$, with corresponding positive probability densities I_i and I_0. The framework first calculates the optimal transport maps between I_i and I_0 using $f_i(x) = F_{\nu_i}^{-1} \circ F_\mu(x)$ for all $i = 1,...,N$. Then the forward and inverse transport-based transform, denoted as the cumulative distribution transform (CDT) by Park et al. [133], for these density functions with respect to the fixed template μ is defined

Cumulative Distribution Transform pairs		
Property	Signal domain	CDT domain
	$I(x)$	$\tilde{I}(x)$
Translation	$I(x-\tau)$	$\tilde{I}(x)+\tau\sqrt{I_0(x)}$
Scaling	$aI(ax)$	$\frac{\tilde{I}(x)}{a}-x\frac{(a-1)}{a}\sqrt{I_0(x)}$
Composition	$g'(x)I(g(x))$	$(g^{-1}(\frac{\tilde{I}(x)}{\sqrt{I_0(x)}}+x)-x)\sqrt{I_0(x)}$

TABLE I
CUMULATIVE DISTRIBUTION TRANSFORM PAIRS. NOTE THAT THE COMPOSITION HOLDS FOR ALL STRICTLY MONOTONICALLY INCREASING FUNCTIONS g.

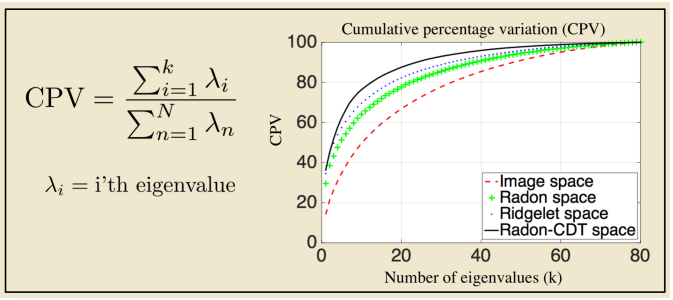

Fig. 7. The cumulative percentage of the face dataset in Figure 6 in the image space, the Radon transform space, the Ridgelet transform space, and the Radon-CDT transform space.

as,

$$\begin{cases} \tilde{I}_i = (f_i - \mathrm{Id})\sqrt{I_0} & \text{(Analysis)} \\ I_i = (f_i^{-1})'(I_0 \circ f_i^{-1}) & \text{(Synthesis)} \end{cases} \tag{26}$$

where $(I_0 \circ f_i^{-1})(x) = I_0(f_i^{-1}(x))$. Note that the L_2 norm of the transformed signals, $\tilde{I}_i$ corresponds to the 2-Wasserstein distance between μ and ν_i. In contrast to the higher-dimensional LOT, the Euclidean distance between two transformed (embedded) signals $\tilde{I}_i$ and $\tilde{I}_j$, however, is the exact 2-Wasserstein distance between ν_i and ν_j (See [133] for a proof) and not just an approximation. Hence, the transformation is isometric with respect to the 2-Wasserstein metric. This isometric nature of the CDT was utilized in [90] to provide positive definite kernels for n-dimensional measures.

From a signal processing point of view, the CDT is a nonlinear signal transformation that captures certain nonlinear variations in signals including translation and scaling. Specifically, it gives rise to the transformation pairs presented in Table I. From Table I one can observe that although $I(t-\tau)$ is nonlinear in τ, its CDT representation $\tilde{I}(t)+\tau\sqrt{I_0(t)}$ becomes affine in τ (similar effect is observed for scaling). In effect, the Lagrangian transformations (compositions) in original signal space are rendered into Eulerian perturbations in transform space, borrowing from the PDE parlance. Furthermore, Park et al. [133] demonstrated that the CDT facilitates certain pattern recognition problems. More precisely, the transformation turns certain not linearly separable and disjoint classes of signals into linearly separable ones. Formally, let C be a set of measurable maps and let $P, Q \subset P_2(\Omega)$ be sets of positive probability density functions born from two positive probability density functions $p_0, q_0 \in P_2(\Omega)$ (mother density functions) as follows,

$$\begin{aligned} P &= \{p | p = h'(p_0 \circ h), \forall h \in C\}, \\ Q &= \{q | q = h'(q_0 \circ h), \forall h \in C\}. \end{aligned} \tag{27}$$

The sets P and Q are disjoint but not necessarily linearly separable in the signal space. A main result of [133] states that the signal classes P and Q are guaranteed to be linearly separable in the transform space (regardless of the choice of the reference signal I_0) if C satisfies the following conditions,

i) $h \in C \iff h^{-1} \in C$
ii) $h_1, h_2 \in C \Rightarrow \rho h_1 + (1-\rho)h_2 \in C, \ \forall \rho \in [0,1]$
iii) $h_1, h_2 \in C \Rightarrow h_1(h_2), h_2(h_1) \in C$
iv) $h'(p_0 \circ h) \neq q_0, \ \forall h \in C$

The set of translations $C = \{f | f(x) = x + \tau, \tau \in \mathbb{R}\}$, and scaling $C = \{f | f(x) = ax, a \in \mathbb{R}^+\}$, for instance, satisfy above conditions. We refer the reader to [133] for further reading. The top panel in Figure 6 demonstrates the linear separation property of the CDT and demonstrate the capability of such nonlinear transformation. The signal classes P and Q are chosen to be the set of all translations of a single Gaussian and a Gaussian mixture including two Gaussian functions with a fixed mean difference, respectively. The discriminant subspace is calculated for these classes and it is shown that while the signal classes are not linearly separable in the signal domain, they become linearly separable in the transform domain.

*3) **The Radon cumulative distribution transform**:* The CDT framework was extended to higher dimensional density functions through the sliced-Wasserstein distance in [87], and was denoted as the Radon-CDT. It is shown in [87] that similar characteristics of the CDT, including the linear separation property, also hold for the Radon-CDT. Figure 6 clarifies the linear separation property of the Radon-CDT and demonstrate the capability of such transformations. Particularly, Figure 6 shows a facial expression dataset with two classes (i.e. neutral and smiling expressions) and its corresponding representations in the LDA discriminant subspace calculated from the images (bottom left panel), the Radon-CDT of the dataset and the corresponding representation of the transformed data in the LDA discriminant subspace (bottom right panel). It is clear that the image classes become more linearly separable in the transform space. In addition, the cumulative percentage variation of the dataset in the image space, the Radon transform space, the Ridgelet transform space, and the Radon-CDT space are shown in Figure 7. It can be seen that the variations in the dataset could be explained with fewer components in the nonlinear transform spaces as opposed to the linear ones.

IV. NUMERICAL METHODS

There exists a variety of fundamentally different approaches to finding optimal transportation maps and plans. Below we present the notable approaches. In the table II we summarize the expected computational complexity of the given algorithm.

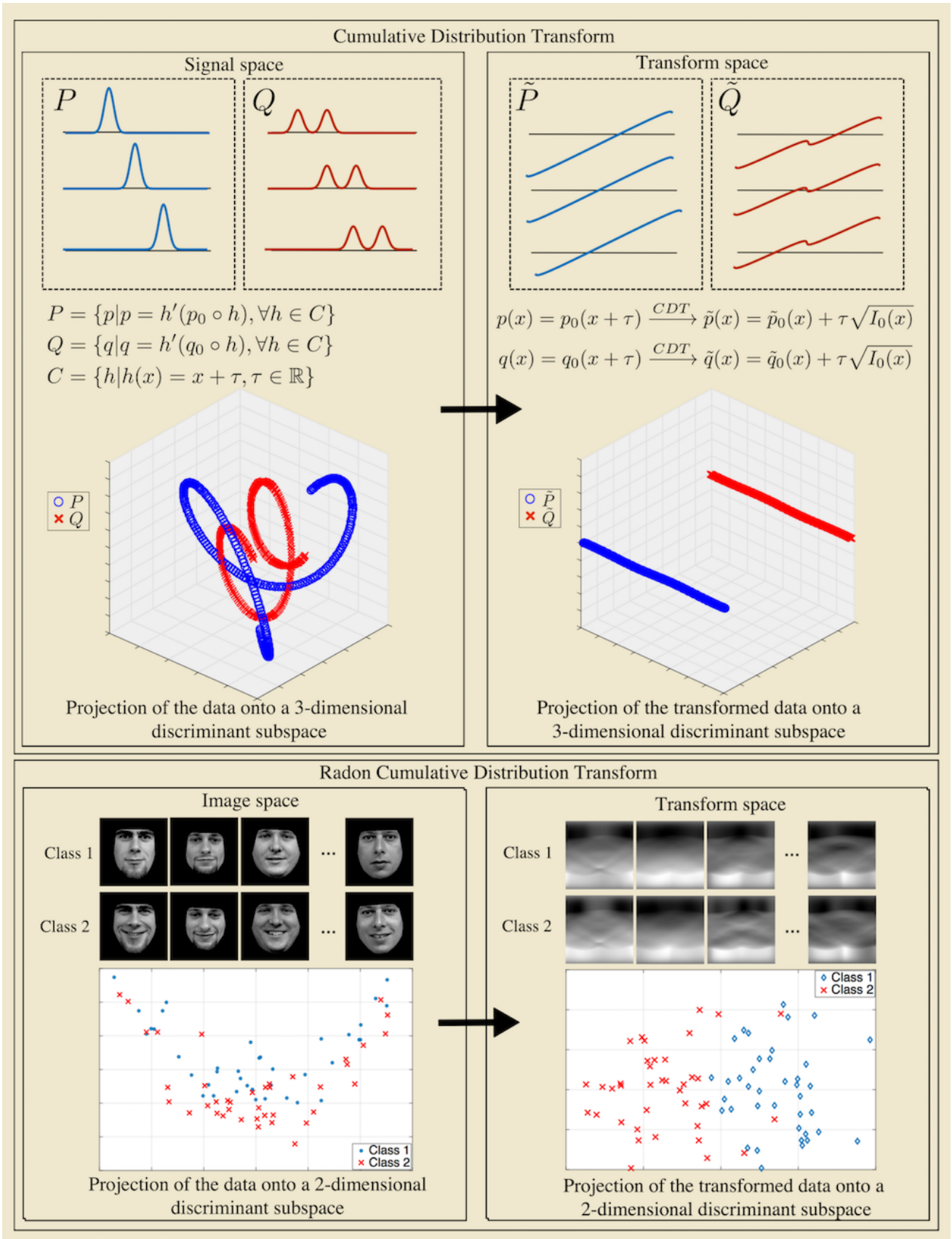

Fig. 6. Examples for the linear separability characteristic of the CDT and the Radon-CDT. The discriminant subspace for each case is calculated using the penalized-linear discriminant analysis (p-LDA). It can be seen that the nonlinear structure of the data is well captured in the transform spaces.

Comparison of Numerical Approaches		
Method	Compl.	Remark
linear programming	N^3	Applicable to general costs. Good approach if the measures are supported at very few sites
multi-scale linear programming	N	Applicable to general costs. Fast and robust method, though truncation involved can lead to imprecise distances.
Auction algorithm	N^2	Applicable only when the number of particles in the source and the target is equal and all of their masses are the same.
Entropy regularized linear programing	N	Applicable to general costs. Simple and performs very well in practice for moderately large problems. Difficult to obtain high accuracy.
Fluid mechanics	$N^{1+1/d}$	This approach can be adapted to generalizations of the quadratic cost, based on action along paths.
AHT minimization	N	Quadratic cost. Requires some smoothness and positivity of densities. Convergence is only guaranteed for infinitesimal stepsize.
Gradient descent on the dual problem	N	Quadratic cost. Convergence depends on the smoothness of the densities, hence a multi-scale approach is needed for non-smooth densities (i.e. normalized images).
Monge–Ampere solver	N	Quadratic cost. One in [16] is proved to be convergent. Accuracy is an issue due to wide stencil used.

TABLE II
THE NUMERICAL COMPLEXITY AND KEY PROPERTIES OF VARIOUS NUMERICAL APPROACHES. EACH MEASURE CONSISTS OF ABOUT N DELTA MASSES IN $\mathbb{R}^d$. (SO IF WE ARE DEALING WITH AN IMAGE THIS WOULD MEAN THAT N IS THE TOTAL NUMBER OF PIXELS AND $d = 2$). COMPLEXITIES ARE REPORTED WITH $\log N$ FACTORS NEGLECTED.

A. Linear programming

Leonid Kantorovich's extensive monologue on 'Mathematical Methods in the Organization and Planning of Production' in 1939 [76], George B Dantzig's pioneering work on the simplex algorithm in 1947 [38], John von Neumann's duality theory in 1947 [112], and Tjalling Koopmans' work on 'Activity analysis of production and allocation' in 1951 [55] were the founding pillars of the linear programming as we know it today. The term 'linear programming', was coined by Koopmans in [55]. The linear programming problem, is an optimization problem with a linear objective function and linear equality and inequality constraints. The set of feasible points to a linear programming problem forms a possibly unbounded convex set. The optimal solution can be thought of as the intersection of the hyper plane corresponding to the linear objective function with an extreme point (i.e. a corner point) of the feasible convex set of solutions.

Several numerical methods exist for solving linear programming problems [161], among which are the simplex method [38] and its variations [35], [39], [13] and the interior-point methods [78], [102]. The computational complexity of the mentioned numerical methods, however, scales at best cubically in the size of the domain. Hence, assuming the density measures have N particles the number of unknowns γ_{ij}s is N^2 and the computational complexities of the solvers are at best $\mathcal{O}(N^3 logN)$ [136], [36]. The computational complexity of the linear programming methods is a very important limiting factor for the applications of the Kantorovich problem.

We point out that in the special case where $M = N$ and the mass is equidistributed, the optimal transport problem simplifies to a one to one assignment problem, which can be solved by the auction algorithm of Bertsekas [17] or similar methods in $\mathcal{O}(N^2 logN)$. In addition, Pele and Werman showed that a thresholded version of the Kantorovich problem can be solved through the min-cost-flow algorithm in $\mathcal{O}(N^2)$ [121]. Several multiscale approaches and sparse approximation approaches have recently been introduced to improve the computational performance of the linear programming solvers, including Schmitzer [143],]. The work by Oberman and Ruan [114] provides an algorithm which is claimed to be of linear complexity for the linear programming problem. Here we provide a brief description of this method.

*1) **Multi-Scale algorithm**:* For a coarse grid of size h, the space $\Omega \subset \mathbb{R}^d$ is uniformly discretized into hypercubes with centers x_i presented with the set $\Omega_h = \{x_i\}_{i=1}^N$ and where the distance between neighboring points is h. Measures μ and ν are discretized by $\mu^h = \sum_{i=1}^N p_i \delta_{x_i}$ and $\nu^h = \sum_{j=1}^N q_i \delta_{x_i}$. One then performs the linear program (7) with inputs $p_i = p_i^h$, $q_i = q_i^h$, $x_i = x_i^h$ and $y_i = y_i^h$ to calculate the optimal map γ^h between μ^h and ν^h.

In the next step, the discretization is refined to $\frac{h}{2}$. Since γ^h is expected to converge to the optimal transport plan between μ and ν, γ^*, then $\gamma^{\frac{h}{2}}$ should be close to γ^h. Let S_h be the support of γ^h on the grid $\Omega_h \times \Omega_h$ and $\bar{S}_h$ be the set of neighbors (in space) of S_h. The Oberman and Ruan method assumes that

$$S_{\frac{h}{2}} \subseteq P_{h \mapsto \frac{h}{2}}(\bar{S}_h)$$

where $P_{h \mapsto \frac{h}{2}}$ is the projection onto the refined grid $\Omega^2_{\frac{h}{2}}$ (where one point in Ω_h is projected to 2^d points in $\Omega_{\frac{h}{2}}$). One then solves the linear program on the refined grid. For better resolution repeat the refinement procedure. The reason that the multi-scale procedure is advantageous is that the linear programming problem solved at the finest level is expected to have $O(N)$ variables as opposed to $O(N^2)$ variables that the original full linear program has.

B. Entropy regularized solution

Cuturi's work [36] provides a fast and easy to implement variation of the Kantorovich problem, which has attracted ample attention recently. Cuturi [36] proposed a variation of the Kantorovich problem by considering the transportation problem from a maximum-entropy perspective. He suggested to regularize the Wasserstein metric by the entropy of the transport plan. This modification simplifies the problem and enables much faster numerical schemes with complexity $\mathcal{O}(N^2)$ [36] or $\mathcal{O}(Nlog(N))$ using the convolutional Wasserstein distance

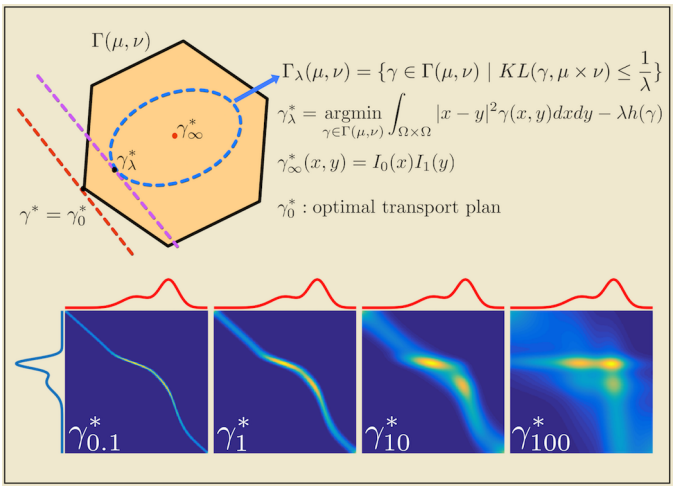

Fig. 8. The geometric interpretation of the entropy regularization of the Kantorovich problem. For $\lambda = 0$ the optimal transport plan is located at an extreme point (corner point) of the feasible set $\Gamma(\mu,\nu)$. Increasing λ corresponds to constraining the space to the convex subset of $\Gamma(\mu,\nu)$, for which the KL-divergence between the transport plans and $\gamma^*_\infty = \mu\times\nu$ is smaller than $\frac{1}{\lambda}$. The entropy regularized optimal transport plan, γ^*_λ is shown for a one-dimensional example with different values of λ.

presented in [147] (compared to $\mathcal{O}(N^3)$ of the linear programming methods), where N is the number of delta masses in each of the measures. The price one pays is that it is difficult to obtain high accuracy approximations of the optimal transport plan. Formally, the entropy regularized p-Wasserstein distance, otherwise known as the Sinkhorn distance, between probability measures μ and ν defined on the metric space (Ω, d) is defined as,

$$W^p_{p,\lambda}(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)}\int_{\Omega\times\Omega} d^p(x,y)\gamma(x,y)dxdy - \lambda h(\gamma) \quad (28)$$

where $h(\gamma)$ is the entropy of the plan and is defined as,

$$h(\gamma) = -\int_{\Omega\times\Omega}\gamma(x,y)\ln(\gamma(x,y))dxdy. \quad (29)$$

We note that this is not a true metric since $W^p_{p,\lambda}(\mu,\mu) > 0$. Since the entropy term is strictly concave, the overall optimization problem in (28) becomes strictly convex. It is straightforward to show that (see [36]) the entropy regularized p-Wasserstein distance in Equation (28) can be reformulated as,

$$W^p_{p,\lambda}(\mu,\nu) = \lambda \inf_{\gamma\in\Gamma(\mu,\nu)} \mathrm{KL}(\gamma|\mathcal{K}_\lambda) \quad (30)$$

where $\mathcal{K}_\lambda(x,y) = \exp(-\frac{d^p(x,y)}{\lambda})$ and $\mathrm{KL}(\gamma|\mathcal{K}_\lambda)$ is the Kullback-Leibler (KL) divergence between γ and $\mathcal{K}_\lambda$ and is defined as,

$$\mathrm{KL}(\gamma|\mathcal{K}_\lambda) = \int_{\Omega\times\Omega}\gamma(x,y)\ln(\frac{\gamma(x,y)}{\mathcal{K}_\lambda(x,y)})dxdy. \quad (31)$$

Figure 8 provides a geometric interpretation of the entropy regularization as discussed in [36]. It can be seen that, the regularizer enforces the plan to be within $\frac{1}{\lambda}$ radius in the KL-divergence sense from the transport plan $\gamma^*_\infty = \mu\times\nu$.

In the discrete setting, where $\mu = \sum_{i=1}^N p_i\delta_{x_i}$ and $\nu = \sum_{j=1}^N q_j\delta_{y_j}$, γ and $\mathcal{K}_\lambda$ are $N\times N$ matrices and hence the number of unknowns is N^2. However, using Equation (31) it can be shown that the optimal transport plan γ is of the form $D_v\mathcal{K}_\lambda D_w$ where D_v and D_w are diagonal matrices with diagonal entries $v, w \in \mathbb{R}^N$ [36]. Therefore, due to the new formulation the number of unknowns decreases from N^2 to $2N$. The new problem can then be solved through the iterative proportional fitting procedure (IPFP) [43], [137], otherwise known as the RAS algorithm [9], or alternatively through the Sinkhorn-Knopp algorithm [146], [84]. The Sinkhorn-Knopp algorithm is specifically interesting as it is computationally efficient and very simple to implement.

The entropy regularization of the transportation problem has recently attracted ample attention from both application and numerical analysis point of view. One can refer to Benamou et al. [15] and Solomon et al. [147] for further numerical analysis, and Cuturi and Doucet [37], Rabin et al. [125], and Rabin and Papadakis [124] and others for applications of the entropy regularized transportation problem. It should be mentioned, however, that the number of iterations for the entropy regularized solvers increases as the regularization parameter goes to zero [114]. In addition, these method become numerically unstable for small regularization parameters [15], [114]. This is due to the fact that the elements of matrix $\mathcal{K}_\lambda$ go to zero as λ goes to zero.

C. Fluid dynamics solution

Benamou and Brenier [14] presented a fluid dynamic formulation of the 2-Wassestein metric using the continuity equation as discussed in Section III-B3 and Equation (25). They provided a numerical scheme for optimizing the problem in (25), using the augmented Lagrangian method [56]. A brief overview of the method and its derivation is as follows. Let $\phi(x,t)$ be the space-time dependent Lagrange multiplier for constraints in Equation (25). Using $\phi(x,t)$ the Lagrangian is given by,

$$\begin{aligned} L(\phi,\rho,\mathrm{v}) &= \int_0^1\int_{\mathbb{R}^d} I|\mathrm{v}|^2 + \phi(\partial_t I + \nabla\cdot(I\mathrm{v}))dxdt \\ &= \int_0^1\int_{\mathbb{R}^d}\frac{|m|^2}{2I} - \partial_t\phi I - \nabla_x\phi\cdot m dxdt \\ &\quad - \int_{\mathbb{R}^d}\phi(0,\cdot)I_0 - \phi(1,\cdot)I_1 dx \end{aligned} \quad (32)$$

where $m = I\mathrm{v}$ and we used integration by parts together with the equality constraints in Equation (25) to obtain the second line. Note that,

$$W_2^2(\mu,\nu) = \inf_{\rho,m}\sup_{\phi} L(\phi,I,\mathrm{v}). \quad (33)$$

Using the Legendre transform [20] on $|m|^2/2I$ one can write,

$$\begin{aligned} \frac{|m(x,t)|^2}{2I(x,t)} &= \sup_{a,b}\ a(x,t)I(x,t) + b(x,t)m(x,t) \\ &\quad s.t.\quad a(x,t) + \frac{|b(x,t)|^2}{2} \leq 0,\ \forall x,t \end{aligned} \quad (34)$$

Define $\psi := \{I, m\}$, $q := \{a, b\}$, and their corresponding inner product to be,

$$\langle\psi, q\rangle = \int_0^1\int_{\mathbb{R}^d} I(x,t)a(x,t) + m(x,t)\cdot b(x,t)\ dxdt,$$

Also define ,

$$F(q) := \begin{cases} 0 & a(x,t) + \frac{|b(x,t)|^2}{2} \leq 0, \ \forall x,t \\ +\infty & O.W. \end{cases}$$

$$G(\phi) := \int_{\mathbb{R}^d} \phi(0,\cdot) I_0 - \phi(1,\cdot) I_1 dx$$

then it is straightforward to show that Equation (33) can be rewritten as,

$$W_2^2(\mu,\nu) = \sup_{\psi} \inf_{\phi,q} \ F(q) + G(\phi) + \langle \psi, \nabla_{t,x}\phi - q \rangle \quad (35)$$

where $\nabla_{t,x}\phi = \{\partial_t \phi, \nabla_x \phi\}$. In the formulation above, ψ can be treated as the Lagrange multiplier for a new constraint, namely $\nabla_{t,x}\phi = q$. Thus the augmented Lagrangian can be written as,

$$\begin{aligned} L_r(\phi,q,\psi) &= F(q) + G(\phi) + \langle \psi, \nabla_{t,x}\phi - q \rangle + \\ &\quad \frac{r}{2}\langle \nabla_{t,x}\phi - q, \nabla_{t,x}\phi - q \rangle \end{aligned} \quad (36)$$

and finally the corresponding saddle point problem is,

$$W_2^2 = \sup_{\psi} \inf_{\phi,q} L_r(\phi,q,\psi) \quad (37)$$

Benamou and Brenier [14] used a variation of the Uzawa algorithm [47] to solve the problem above. We note that recent methods based on Bregman iteration such as [166], [49] could also be used for solving the saddle point problem in (37).

Due to the space-time nature of the fluid dynamic formulation, the solvers require a space-time discretization (spatial-temporal grid), which increases the computational complexity of such solvers. However, the fluid dynamics solution enables one to handle situations for which there exist barriers in the domain. Take for instance transportation of a distribution through a maze, in which the mass can not be transported over the maze walls.

D. Flow minimization (AHT)

Angenent, Haker, and Tannenbaum (AHT) [4] proposed a flow minimization scheme to obtain the optimal transport map from the Monge problem. The method was used in several image registration applications [71], [167], pattern recognition [89], [165], and computer vision [88]. The method was polished in several follow up papers [71], [159]. The idea behind the method, is to first obtain an initial mass preserving transport map using the Knothe-Rosenblatt coupling [135], [163] and then update the initial map to obtain a curl free mass preserving transport map that minimizes the transport cost. A brief review of the method is provided here.

Let μ and ν be continuous probability measures defined on convex domains $X, Y \subseteq \mathbb{R}^d$ with corresponding positive densities I_0 and I_1. In order to find the optimal transport map, f^*, AHT starts with an initial transport map, $f_0 : X \to Y$ calculated from the Knothe-Rosenblatt coupling [135], [163], for which $(f_0)_{\#}\mu = \nu$. Then it updates f_0 to minimize the transport cost. The goal, however, is to update f_0 in a way that it remains a transport map from μ to ν. AHT defines $s(x,t)$, where for a fixed time, t_0, $s(x,t_0) : X \to X$ is a transport map from μ to itself. The initial transport map is then updated through $s(x,t)$, starting from $s(x,0) = x$, such that $f_0(s^{-1}(x,t))$ minimizes the transport cost. Following simple calculations, one can show (see [71]) that for $s(x,t)$ to be a MP mapping from I_0 to itself, $\frac{\partial s}{\partial t}$ should have the following form,

$$\frac{\partial s}{\partial t} = (\frac{1}{I_0}\xi) \circ s, \quad (38)$$

for some vector field ξ on X with $div(\xi) = 0$ and $\langle \xi, n \rangle = 0$ on the boundary, where n is the vector normal to the boundary. From (38), it follows that the time derivative of $f(x,t) = f_0(s^{-1}(x,t))$ satisfies,

$$\frac{\partial f}{\partial t} = -\frac{1}{I_0}(Df)\xi. \quad (39)$$

AHT then differentiate the Monge objective function,

$$M(f) = \int_X |f(x,t) - x|^2 I_0(x) dx, \quad (40)$$

with respect to t, which after rearranging will lead to,

$$\frac{\partial M}{\partial t} = -2 \int_X \langle f(x,t), I_0(x)(\frac{\partial s}{\partial t} \circ s^{-1}(x,t)) \rangle dx. \quad (41)$$

Substituting (38) in the above equation we get,

$$\frac{\partial M}{\partial t} = -2 \int_\Omega \langle f(x,t), \xi(x,t) \rangle dx. \quad (42)$$

Writing the Helmholtz decomposition of f as $f = \nabla\phi + \chi$ and using the divergence theorem we get,

$$\frac{\partial M}{\partial t} = -2 \int_\Omega \langle \chi(x,t), \xi(x,t) \rangle dx, \quad (43)$$

thus, $\xi = \chi$ decreases M the most. In order to find χ, AHT first find ϕ and then subtract its gradient from f. Given that $div(\chi) = 0$ and $\langle \chi, n \rangle = 0$ and taking the divergence of the Helmholtz decomposition of f leads to,

$$\begin{cases} \Delta(\phi) = div(f) \\ \langle \nabla\phi, n \rangle = \langle f, n \rangle \ \text{ on the boundary} \end{cases} \quad (44)$$

where Δ is the Laplace operator. Therefore, $\xi = f - \nabla(\Delta^{-1} div(f))$ where $\Delta^{-1} div(f)$ is the solution to (44). Substituting ξ back into (39) we have,

$$\frac{\partial f}{\partial t} = -\frac{1}{I_0} Df(f - \nabla(\Delta^{-1} div(f))), \quad (45)$$

which for $f \in \mathbb{R}^2$ can be even simplified further,

$$\frac{\partial f}{\partial t} = -\frac{1}{I_0} Df \nabla^\perp \Delta^{-1} div(f^\perp). \quad (46)$$

where $\perp$ indicates rotation by 90 degrees. Finally, the transport map f is updated with a gradient descent scheme,

$$f(x,t+1) = f(x,t) + \epsilon \frac{1}{I_0} Df(f - \nabla(\Delta^{-1} div(f))) \quad (47)$$

where ϵ is the gradient descent step size. AHT show that for infinitesimal step size, ϵ, $f(x,t)$ converges to the optimal transport map.

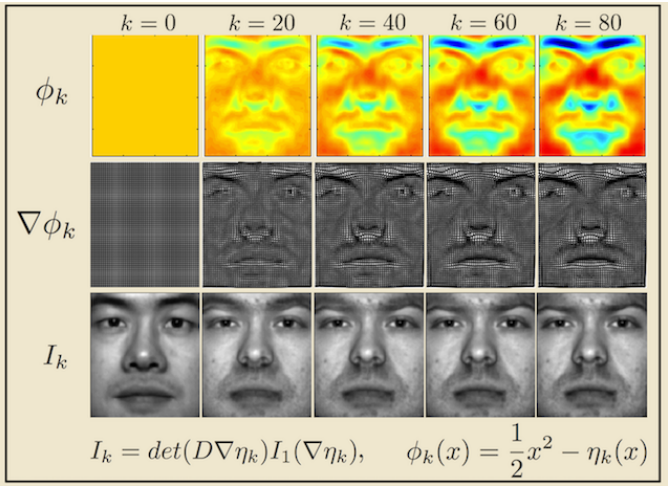

Fig. 9. Visualization of the iterative update of the transport potential and correspondingly the transport displacement map through CWVP iterations.

E. Gradient descent on the dual problem

Chartrand, Wohlberg, Vixie, and Bollt (CWVB) [28] provide a straight-forward algorithm for computing the optimal transport. It should be mentioned, however, that the method has not been studied in depth. The method reformulates the Kantorovich's dual formulation as described in Eq. 9 into minimization of a convex, continuous functional. More importantly, the derivative of the continuous functional can be explicitly found, and thus a gradient descent approach is developed for computing the optimal transport. A brief review of the method and its derivation is presented here.

For the strictly convex cost function, $c(x,y) = \frac{1}{2}|x-y|^2$, the dual problem can be written as,

$$KD(\mu,\nu) = \min_\eta \underbrace{\int_X \eta(x)d\mu(x) + \int_Y \eta^c(y)d\nu(y)}_{M(\eta)} \tag{48}$$

where $\eta^c(y) := \max_{x\in X}(x\cdot y - \eta(x))$ is the Legendre-Fenchel transform of $\eta(x)$. Note the relationship between η and ϕ in Section III-A3 is $\phi(x) = \frac{1}{2}x^2 - \eta(x)$.

The method utilizes the following property of the Legendre-Fenchel transform [20],

$$\nabla\eta^c(y) = \text{argmax}_{x\in X}(x\cdot y - \eta(x)) \tag{49}$$

and derives the functional derivative of M as,

$$\frac{dM(\eta)}{d\eta} = I_0 - \det(I - H\eta^{cc})I_1(id - \nabla\eta^{cc}) \tag{50}$$

where $H\eta^{cc}$ is the Hessian matrix of η^{cc}, I is the identity matrix, and $\eta^{cc}(x) = \max_{y\in Y}(y\cdot x - \eta^c(y))$. Note that if η is a concave function (respectively if ϕ is a convex function) then $\eta^{cc} = \eta$ (and $\phi^{cc} = \phi$). CWVB initializes $\eta_0(x) = \frac{1}{2}|x|^2$ and updates the potential field η through gradient descent,

$$\eta_{k+1} = \eta_k - \epsilon_k \frac{dM(\eta)}{d\eta} \tag{51}$$

where ϵ_k is the step size and is calculated with a line search method. We note that when μ and ν are absolutely continuous probability measures the optimal transport map is obtained from η^* through, $f^*(x) = \nabla\eta^*(x)$. Figure 9 visualizes the iterations of the CWVB method for two face images taken from YaleB face database (normalized to integrate to 1) [94]. We note that the number of iterations required for convergence of the CWVB method depends on the smoothness of the input densities. In [28] the authors proposed a multi-scale scheme (i.e. using Gaussian pyramids) to update η through different levels of smoothness and hence decrease the number of iterations needed for convergence.

F. Monge-Ampere equation

The Monge-Ampére partial differential equation (PDE) is defined as,

$$\det(H\phi) = h(x,\phi,D\phi) \tag{52}$$

for some functional h, and where $H\phi$ is the Hessian matrix of ϕ. The Monge-Ampére PDE is closely related to the Monge problem. According to Bernier's theorem (discussed in Section III-A4) when μ and ν are absolutely continuous probability measures on sets $X, Y \subset \mathbb{R}^n$, the optimal transport map than minimizes the 2-Wasserstein metric is uniquely characterized as the gradient of a convex function $\phi : X \to Y$. Moreover, we showed that the mass preserving constraint of the Monge problem can be written as $\det(Df)I_1(f) = I_0$. Combining these results one can write,

$$\det(D(\nabla\phi(x))) = \frac{I_0(x)}{I_1(\nabla\phi))} \tag{53}$$

where $D\nabla\phi = H\phi$, and therefore the equation shown above is in the form of the Monge-Ampére PDE. Now if ϕ is a convex function on X satisfying $\nabla\phi(X) = Y$ and solving the Equation (53) then $f^* = \nabla\phi$ is the optimal transportation map from μ to ν. The geometrical constraint on this problem is rather unusual in PDEs and is often referred to as the optimal transport boundary conditions. The equation (53) is a nonlinear elliptic equation and thus one hopes that methods achieving complexity $O(N)$ (as do the solvers for the Laplace equation) can be designed. Several authors have proposed numerical methods to obtain the optimal transport map through solving the Monge-Ampére PDE in Equation (53) [100], [157], [26], [140], [152], [16]. In particular the scheme in [16] is monotone, has compexity $O(N)$ (up to logarithms) and is provably convergent. We conclude by remarking that several regularity results on the optimal transport maps, including the pioneering work of Caffarreli [25], and the more recent results by De Philippis and Figalli [40] were established through the Monge-Ampére equation.

G. Other approaches

It is also worth pointing out the approach to optimal transport with quadratic cost using the *semi-discrete* approximation. Namely several works [8], [106], [83], [97] have considered the situation in which one of the measures considered has density $d\mu = I_0 dx$, while the other is a sum of delta masses, $\mu = \sum q_i \delta_{y_i}$. It turns out that there exists weights w_i such that the optimal transport map $x \mapsto y$ can be described via a

power diagram. More precisely the set of x mapping to y_i is the following cell of the power diagram:

$$PD_w(y_i) = \{x : \text{ for all } j \quad |x-y_i|^2 - w_i \leq |x-y_j|^2 - w_j\}.$$

The main observation (originating from [8] and stated in [106]) is that the weights w_i are minimizers of the following unconstrained convex functional

$$F(w) = \sum_i \left(p_i w_i - \int_{PD_w(y_i)} \|x-y_i\|^2 - w_i d\mu(x) \right).$$

Works of Kitagawa, Mérigot, and Thibert [83], Merigot [106], and Levy [97] use Newton based schemes and multiscale approaches to minimize the functional. The need to integrate over the power diagram makes the implementation somewhat geometrically delicate. Nevertheless recent implementation by Lévy [97] gives impressive results in terms of speed. We also note that this approach provides the transportation mapping (not just the approximation of a plan).

Finally, we mention a recent work of Aude, Cuturi, Peyré and Bach who investigated the possibility of using stochastic gradient descent on several formulations of OT problem with quadratic cost on large data sets.

Other notable numerical techniques that were not covered here but could be valuable in a number of instances include [108], [68], [59], [82], [92], [99].

V. Applications

In this section we review some recent applications of the optimal transport problem in signal and image processing, computer vision, and machine learning.

A. Image retrieval

One of the earliest applications of the optimal transport problem was in image retrieval. Rubner, Tomasi, and Guibas [136] employed the discrete Wasserstein metric, which they denoted as the Earth Mover's Distance (EMD), to measure the dissimilarity between image signatures. In image retrieval applications, it is common practice to first extract features (i.e. color features, texture feature, shape features, etc.) and then generate high dimensional histograms or signatures (histograms with dynamic/adaptive binning), to represent images. The retrieval task then simplifies to finding images with similar representations (i.e. small distance between their histograms/signatures). The Wasserstein metric is specifically suitable for such applications as it can compare histograms/signatures of different sizes (histograms with different binning). This unique capability turns the Wasserstein metric into an attractive candidate in image retrieval applications [136], [121], [98]. In [136], the Wasserstein metric was compared with common metrics such as the Jeffrey's divergence, the χ^2 statistics, the L_1 distance, and the L_2 distance in an image retrieval task; and it was shown that the Wasserstein metric achieves the highest precision/recall performance amongst all.

Speed of computation is an important practical consideration in image retrieval applications. For almost a decade, the high computational cost of the optimal transport problem overshadowed its practicality in large scale image retrieval applications. Recent advancements in numerical methods including the work of Pele and Werman [121], Merigot [106], and Cuturi [36], among many others, have reinvigorated optimal transport-based distances as a feasible and appealing candidate for large scale image retrieval problems.

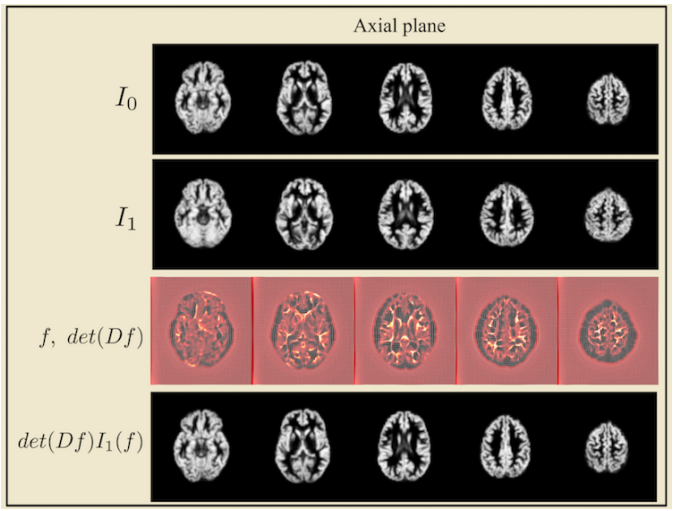

Fig. 10. An example of morphing a three dimensional image into another. The axial slices of images, and determinant of the Jacobian of the optimal transport map and the deformation field is shown. The last row shows the slices of the morphed image.

B. Registration and Morphing

Image registration deals with finding a common geometric reference frame between two or more images and it plays an important role in analyzing images obtained at different times or using different imaging modalities. Image registration and more specifically biomedical image registration is an active research area. Registration methods find a transformation f that maximizes the similarity between two or more image representations (e.g. image intensities, image features, etc.). Comprehensive recent surveys on image registration can be found in [149], [34], [115]. Among the plethora of registration methods, nonrigid registration methods are especially important given their numerous applications in biomedical problems. They can be used to quantify the morphology of different organs, correct for physiological motion, and allow for comparison of image intensities in a fixed coordinate space (atlas). Generally speaking, nonrigid registration is a non-convex and non-symmetric problem, with no guarantee on existence of a globally optimal transformation.

Haker et al. [70] proposed to use the Monge problem for image warping and elastic registration. Utilizing the Monge problem in an image warping/registration setting has a number of advantages. First, the existence and uniqueness of the global transformation (the optimal transport map) is known. Second, the problem is symmetric, meaning that the optimal transport map for warping I_0 to I_1 is the inverse of the optimal transport map for warping I_1 to I_0. Lastly, it provides a landmark-free and parameter-free registration scheme with a built-in mass

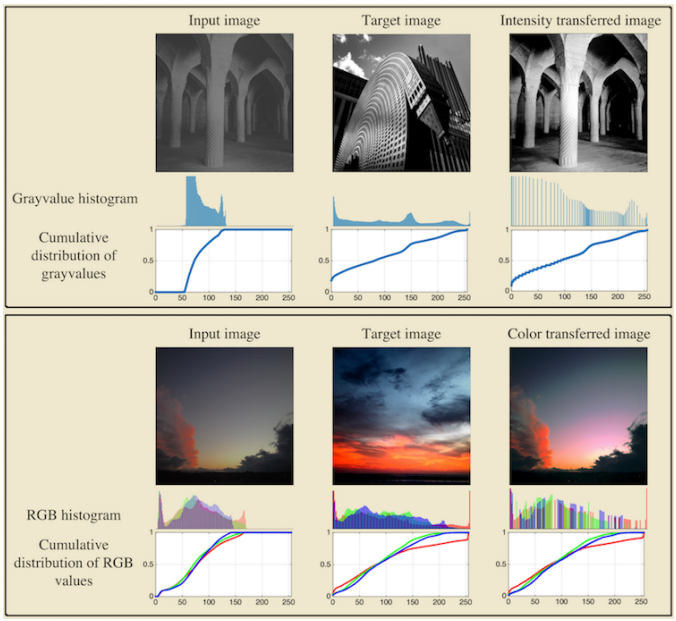

Fig. 11. Grayvalue and color transfer via optimal transportation.

preservation constraint. These advantages motivated several follow-up work to investigate the application of the Monge problem in image registration and warping [69], [71], [167], [159], [111], [68]. Figure 10 shows an example of morphing a Magnetic Resonance Imaging (MRI) image of a brain into another. The axial slices of three dimensional images are shown, as well as slices of the determinant of the Jacobian of the optimal transport map and the deformation field.

In addition to images, the optimal mass transport problem has also been used in point cloud and mesh registration [103], [54], [93], [150], which have various applications in shape analysis and graphics. In these applications, shape images (2D or 3D binary images) are first represented whether with sets of weighted points (i.e. point clouds), using clustering techniques such as K-means or Fuzzy C-means, or with meshes. Then, a regularized variation of the optimal transport problem is solved to match such representations. The regularization on the transportation problem is often imposed to enforce the neighboring points (or vertices) to remain near to each other after the transformation [54]. As another example, Su et al. [150] proposed a composition of conformal maps and optimal transport maps and introduced the *Conformal Wasserstein Shape Space*, which enables efficient shape matching and comparison.

C. Color transfer and texture synthesis

Texture mixing and color transfer are among appealing applications of the optimal transport framework in image analysis, graphics, and computer vision. Here we briefly discuss these applications.

1) **Color transfer:** The purpose of color transfer [132] is to change the color palette of an image to impose the feel and look of another image. The color transfer is generally performed through finding a map, which morphs the color distribution of the first image into the second one. For grayscale images, the color transfer problem simplifies to a histogram matching problem, which is solved through the one-dimensional optimal transport formulation [42]. In fact, the classic problem of histogram equalization is indeed a one-dimensional transport problem [42]. The color transfer problem, on the other hand, is concerned with pushing the three-dimensional color distribution of the first image into the second one. This problem can also be formulated as an optimal transport problem as demonstrated in [126], [123], [52].

A complication that occurs in the color transfer on real images, however, is that a perfect match between color distributions of the images is often not satisfying. This is due to the fact that a color transfer map may not transfer the colors of neighboring pixels in a coherent manner, and may introduce artifacts in the color transferred image. Therefore, the color transfer map is often regularized to make the transfer map spatially coherent [123]. Figure 11 shows a simple example of grayvalue and color transfer via optimal transport framework. It can be seen that the cumulative distribution of the grayvalue and color transferred images are similar to that of the input image.

2) **Texture synthesis and mixing:** Texture synthesis is the problem of synthesizing a texture image that is visually similar to an exemplar input texture image, and has various applications in computer graphics and image processing [45], [61]. Many methods have been proposed for texture synthesis, among which are *synthesis by recopy* and *synthesis by statistical modeling*. Texture mixing, on the other hand, considers the problem of synthesizing a texture image from a collection of input texture images in a way that the synthesized texture provides a meaningful integration of the colors and textures of the input texture images. Metamorphosis is one of the successful approaches in texture mixing, which performs the mixing via identifying correspondences between elementary features (i.e. textons) among input textures and progressively morphing between the shapes of elements. In other approaches, texture images are first parametrized through a tight frame (often steerable wavelets) and statistical modeling is performed on the parameters. Rabin et al. [127], for instance, used the optimal transport framework, specifically the sliced Wasserstein barycenters, on the coefficients of a steerable wavelet tight frame for texture mixing.

Other successful approaches include the random phase and spot noise texture modeling [53], which model textures as stationary Gaussian random fields [61]. Briefly, these methods are based on the assumption that the visual texture perception is based on the spectral magnitude of the texture image. Therefore, utilizing the spectral magnitude of an input image and randomizing its phase will lead to a new synthetic texture image which is visually similar to the input image. Ferradans et al. [53] utilized this assumption together with the Wasserstein geodesics to interpolate between spectral magnitude of texture images, and provide synthetic mixed texture images. Figure 12 shows an example of texture missing via the Wasserstein geodesic between the spectral magnitudes of the input texture images. The in-between images are synthetically generated using the random phase technique.

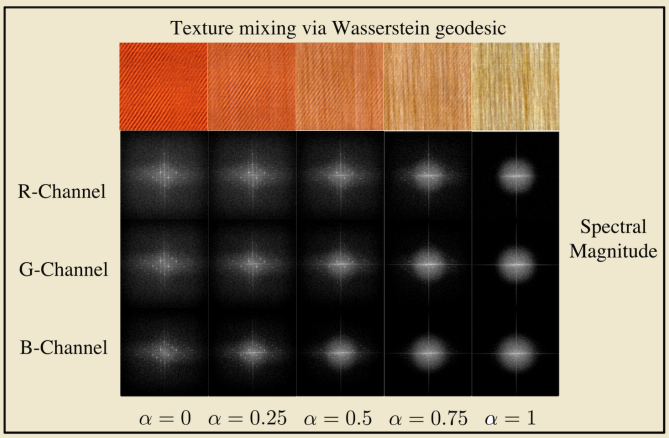

Fig. 12. An example of texture mixing via optimal transport using the method presented in Ferradans et al. [53]

D. Image denoising and restoration

The optimal transport problem has also been used in several image denoising and restoration problems [95], [126], [153]. The goal in these applications is to restore or reconstruct an image from noisy or incomplete observation. Rabin and Peyré [126], for instance, introduced an optimal transport-based framework to enforce statistical priors for constrained functional minimization. More precisely, they utilized the Sliced Wasserstein metric to measure the adequacy of the statistics of the reconstructed signal/image with the given constraints. In a similar approach, Swoboda and Schörr proposed a variational optimization problem with the Total Variation (TV) as the regularizer and the Wasserstein distance (dual formulation) to enforce prior statistics on reconstructed signal/image. They relaxed the optimization into a convex/concave saddle point problem and solved it via a proximal gradient descent.

In another approach, Lellmann et al. [95] utilized the Kantorovich-Rubinsten discrepancy term together with a Total Variation term in the context of image denoising. They called their method Kantorovich-Rubinstein-TV (KR-TV) denoising. It should be noted that, the Kantorovich-Rubinstein metric is closely related to the 1-Wasserstein metric (for one dimensional signals they are equivalent). The KR term in their proposed functional provides a fidelity term for denoising while the TV term enforces a piecewise constant reconstruction.

E. Transport based morphometry

Given their suitability for comparing mass distributions, transport-based approaches for performing pattern recognition of morphometry encoded in image intensity values have also recently emerged. Recently described approaches for transport-based morphometry (TBM) [165], [11], [89] work by computing transport maps or plans between a set of images and a reference or template image. The transport plans/maps are then utilized as an invertible feature/transform onto which pattern recognition algorithms such as principal component analysis (PCA) or linear discriminant analysis (LDA) can be applied. In effect, it utilizes the LOT framework described earlier in Section III-C1. These techniques has been recently employed to decode differences in cell and nuclear morphology for drug screening [11], and cancer detection histopathology [164], [118] and cytology [156], amongst other applications including the analysis of galaxy morphologies [89], for example.

We note the strong similarity between deformation-based methods which have long been used to analyzed radiological images [66], [73], for example. The difference being that TBM allows for numerically exact, uniquely defined solutions for the transport plans or maps used. That is, images can be matched with little perceptible error. The same is not true in methods that rely on registration via the computation of deformations, given the significant topology differences commonly found in medical images. Moreover, TBM allows for comparison of the entire intensity information present in the images (shapes and textures), while deformation-based methods are usually employed to deal with shape differences. Figure 13 shows a schematic of the TBM steps applied to a cell nuclei dataset. It can be seen that the TBM is capable of modeling the variation in the dataset. In addition, it enables one to visualize the classifier, which discriminates between image classes (in this case malignant versus benign).

F. Super-Resolution

Super-resolution is the process of reconstructing a high-resolution image from one or several corresponding low-resolution images. Super-resolution algorithms can be broadly categorized into two major classes namely “multi-frame” super resolution and “single-frame” super resolution, based on the number of low-resolution images they require to reconstruct the corresponding high-resolution image. The transport-based morphometry approach was used for single frame super resolution in [88] to reconstruct high-resolution faces from very low resolution input face images. The authors utilized the transport-based morphometry in combination with subspace learning techniques to learn a nonlinear model for the high-resolution face images in the training set.

In short, the method consists of a training and a testing phase. In the training phase, it uses high resolution face images and morphs them to a template high-resolution face through optimal transport maps. Next, it learns a subspace for the calculated optimal transport maps. A transport map in this subspace can then be applied to the template image to synthesize a high-resolution face image. In the testing phase, the goal is to reconstruct a high-resolution image from the low-resolution input image. The method searches for a synthetic high-resolution face image (generated from the transport subspace) that provides a corresponding low-resolution image which is similar to the input low-resolution image. Figure 14 shows the steps used in this method and demonstrates reconstruction results.

G. Machine-Learning and Statistics

The optimal transport framework has recently attracted ample attention from the machine learning and statistics communities [148], [31], [60], [87], [91], [110]. Some applications of the optimal transport in these arenas include various

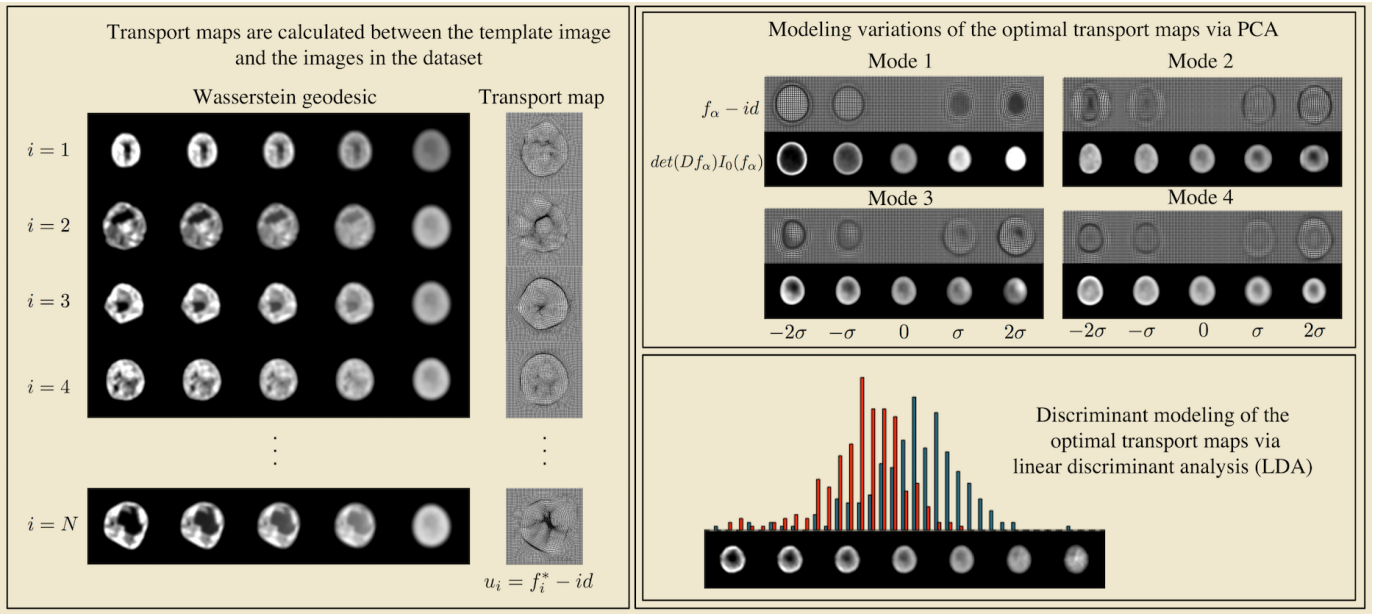

Fig. 13. The schematic of the TBM framework. The optimal transport maps between input images $I_1, ..., I_N$ and a template image I_0 is calculated. Next, linear statistical modeling such as principal component analysis (PCA), linear discriminant analysis (LDA), and canonical correlation analysis (CCA) is performed on the optimal transport maps. The resulting transport maps obtained from the statistical modeling step are then applied to the template image to visualize the results of the analysis in the image space.

transport-based learning methods [110], [60], [91], [148], domain adaptation, Bayesian inference [54], [31], [32], and hypothesis testing [41], [130] among others. Here we provide a brief overview of the recent developments of transport-based methods in machine learning and statistics.

1) **Learning:** Transport-based distances have been recently used in several works as a loss function for regression, classification, etc. Montavon, Müller, and Cuturi [110] for instance utilized the dual formulation of the entropy regularized Wasserstein distance to train restricted Boltzmann machines (RBMs). Boltzmann machines are probabilistic graphical models (Markov random fields) that can be categorized as stochastic neural networks and are capable of extracting hierarchical features at multiple scales. RBMs are bipartite graphs which are special cases of Boltzmann machines, which define parameterized probability distributions over a set of d-binary input variables (observations) whose states are represented by h binary output variables (hidden variables). RBMs' parameters are often learned through information theoretic divergences such as KL-Divergence. Montavon et al. [110] proposed an alternative approach through a scalable entropy regularized Wasserstein distance estimator for RBMs, and showed the practical advantages of this distance over the commonly used information divergence-based loss functions.

In another approach, Frogner et al. [60] used the entropy regularized Wasserstein loss for multi-label classification. They proposed a relaxation of the transport problem to deal with unnormalized measures by replacing the equality constraints in Equation (7) with soft penalties with respect to KL-divergence. In addition, Frogner et al. [60] provided statistical bounds on the expected semantic distance between the prediction and the groundtruth. In yet another approach, Kolouri et al. [91] utilized the sliced Wasserstein metric and provided a family of positive definite kernels, denoted as Sliced-Wasserstein Kernels, and showed the advantages of learning with such kernels. The Sliced-Wasserstein Kernels were shown to be effective in various machine learning tasks including classification, clustering, and regression.

Solomon et al. [148] considered the problem of graph-based semi-supervised learning, in which graph nodes are partially labeled and the task is to propagate the labels throughout the nodes. Specifically, they considered a problem in which the labels are histograms. This problem arises for example in traffic density prediction, in which the traffic density is observed for few stop lights over 24 hours in a city and the city is interested in predicting the traffic density in the un-observed stop lights. They pose the problem as an optimization of a Dirichlet energy for distribution-valued maps based on the 2-Wasserstein distance, and present a Wasserstein propagation scheme for semi-supervised distribution propagation along graphs.

2) **Domain adaptation :** Domain adaptation is one of the fundamental problems in machine learning which has gained proper attention from the machine learning research community in the past decade [120]. Domain adaptation is the task of transferring knowledge from classifiers trained on available labeled data to unlabeled test domains with data distributions that differ from that of the training data. The optimal transport framework is recently presented as a potential major player in domain adaptation problems [54], [31], [32]. Courty, Flamary, and Davis [31], for instance, assumed that there exists a non-rigid transformation between the source and target distributions and find this transformation using an entropy regularized optimal transport problem. They also proposed a label-aware version of the problem in which the transport plan is regularized so a given target point (testing exemplar) is only

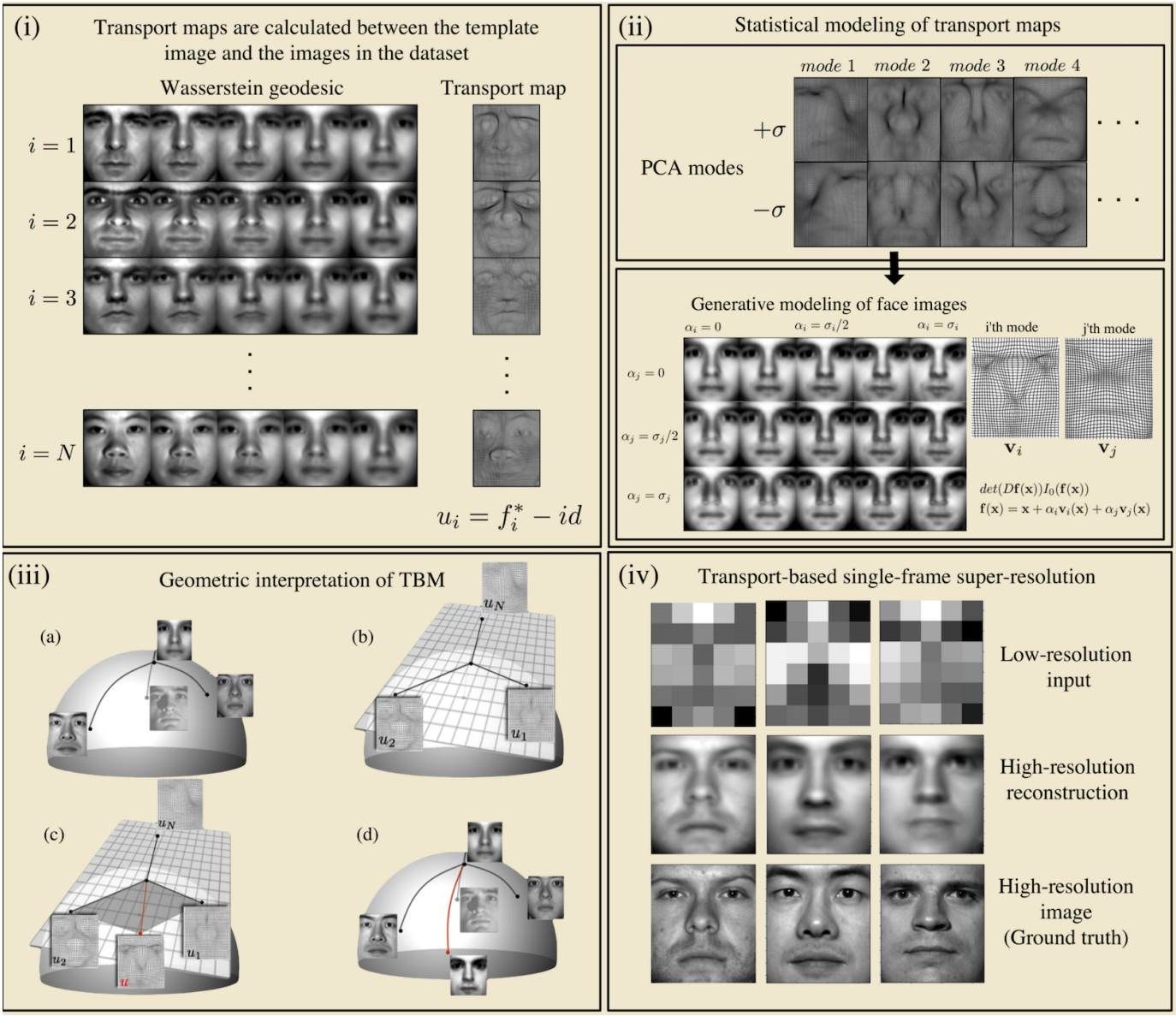

Fig. 14. In the training phase, optimal transport maps that morph the template image to high-resolution training face images are calculated (i). Principal component analysis (PCA) is used to learn a linear subspace for transport maps for which a linear combination of obtained eigenmaps can be applied to the template image to obtain synthetic face images (ii). A geometric interpretation of the problem is depicted in panel (iii), and reconstruction results are shown in panel (iv).

associated with source points (training exemplars) belonging to the same class. Courty et al. [31] showed that domain adaptation via regularized optimal transport outperform the state-of-the-art results in several challenging domain adaptation problems.

3) **Bayesian inference:** Another interesting and emerging application of the optimal transport problem is in Bayesian inference [46], [81], [131], [116]. In Bayesian inference, one critical step is the evaluation of expectations with respect to a posterior probability function, which leads to complex multidimensional integrals. These integrals are commonly solved through the Monte Carlo numerical integration, which requires independent sampling from the posterior distribution. In practice, sampling from a general posterior distribution might be difficult, therefore the sampling is performed via a Markov Chain which converges to the posterior probability after certain number of steps. This leads to the celebrated Markov Chain Monte Carlo (MCMC) method. The downside of MCMC is that the samples are not independent and hence the convergence of the empirical expectation is slow. El Moselhy and Marzouk [46] proposed a transport-based method that evades the need for Markov chain simulation by allowing direct sampling from the posterior distribution. The core idea in their work is to find a transport map (via a regularized Monge formulation), which pushes forward the prior measure to the posterior measure. Then, sampling the prior distribution and applying the transport map to the samples, will lead to a sampling scheme from the posterior distribution. Similar ideas were used in [81], [116]. Figure 15 shows the basic idea behind these methods.

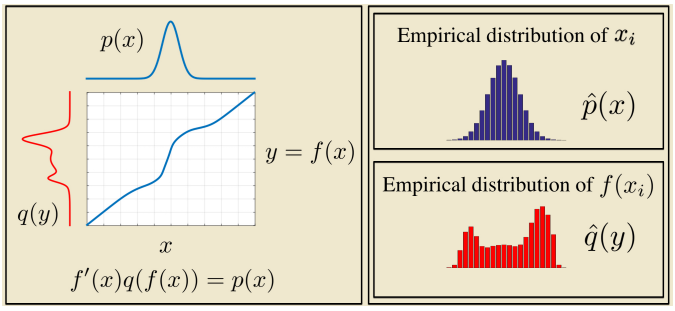

Fig. 15. Left panel shows the prior distribution p and the posterior distribution q and the corresponding transport map f that pushes p into q. One million samples, x_i, were generated from distribution p and the top-right panel shows the empirical distribution of these samples denoted as $\hat{p}$. The bottom-right panel shows the empirical distribution of transformed samples, $y_i = f(x_i)$, denoted as $\hat{q}$.

4) **Hypothesis testing:** Wasserstein distance is used for goodness of fit testing in [41] and for two sample testing in [130]. Ramdas et al. [130] presented connections between the entropy regularized Wasserstein distance, multivariate Energy distance, and the kernel maximum mean discrepancy (MMD) [67], and provided a "distribution free" univariate Wasserstein test statistic. These and other applications of transport-related concepts show the promise of the mathematical modeling technique in the design of statistical data analysis methods to tackle modern learning problems.

Finally, we note that in the interest of brevity, a number of other important applications of transport-related techniques were not discussed above, but are certainly interesting on their own right. These include astronomical sciences [57], [23], [58], meteorology sciences [24], optics [65], [12], particle image velocimetry [141], information theory [154], optical flow estimation [86], [79], and geophysics [107], [48] among others.

VI. Summary and Conclusions

Transport-related methods and applications have come a long way. While earlier applications focused primarily in civil engineering and economics problems, they have recently begun to be employed in a wide variety of problems related to signal and image analysis, and pattern recognition. In this tutorial, seven main areas of application were reviewed: image retrieval V-A, registration and morphing V-B, color transfer and texture analysis V-C, image restoration V-D, transport-based morphometry V-E, image super-resolution V-F, and machine learning and statistics V-G. As evidenced by the number of papers published (see Fig 1) per year, transport and related techniques have received increased attention in recent years. Overall, researchers have found that the application of transport-related concepts can be helpful to solve problems in diverse applications. Given recent trends, it seems safe to expect that the the number of different application areas will continue to grow.

In its most general form, the transport-related techniques reviewed in this tutorial can be thought as mathematical models for signals, images, and in general data distributions. Transport-related metrics involve calculating differences not only of pixel or distribution intensities, but also "where" they are located in the corresponding coordinate space (a pixel coordinate in an image, or a particular axis in some arbitrary feature space). As such, the geometry (e.g. geodesics) induced by such metrics can give rise to dramatically different algorithms and data interpretation results. The interesting performance improvements recently obtained could motivate the search for a more rigorous mathematical understanding of transport-related metrics and applications.

We note that the emergence of numerically precise and efficient ways of computing transport-related metrics and geodesics, presented in section IV also serves as an enabling mechanism. Coupled with the fact that several mathematical properties of transport-based metrics have been extensively studied, we believe that the ground is set of their increased use as foundational tools or building blocks based on which complex computational systems can be built. The confluence of these emerging ideas may spur a significant amount of innovation in a world where sensor and other data is becoming abundant, and computational intelligence to analyze these is in high demand. We believe transport-based models while become an important component of the ever expanding tool set available to modern signal processing and data science experts.

VII. Acknowledgements

Authors gratefully acknowledge funding from the NSF (CCF 1421502) and the NIH (GM090033, CA188938) in contributing to a portion of this work. DS also acknowledges funding by NSF (DMS DMS-1516677)

References

[1] L. Ambrosio. *Lecture notes on optimal transport problems*. Springer, 2003.
[2] L. Ambrosio and N. Gigli. A user's guide to optimal transport. In *Modelling and optimisation of flows on networks*, pages 1–155. Springer, 2013.
[3] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.
[4] S. Angenent, S. Haker, and A. Tannenbaum. Minimizing flows for the Monge–Kantorovich problem. *SIAM journal on mathematical analysis*, 35(1):61–97, 2003.
[5] P. Appell. Le probleme geometrique des déblais et remblais. *Mémorial des sciences mathématiques*, 27:1–34, 1928.
[6] J. Ashburner and K. J. Friston. Voxel-based morphometry the methods. *Neuroimage*, 11(6):805–821, 2000.
[7] J. Ashburner, C. Hutton, R. Frackowiak, I. Johnsrude, C. Price, K. Friston, et al. Identifying global anatomical differences: deformation-based morphometry. *Human brain mapping*, 6(5-6):348–357, 1998.
[8] F. Aurenhammer, F. Hoffmann, and B. Aronov. Minkowski-type theorems and least-squares clustering. *Algorithmica*, 20(1):61–76, 1998.
[9] M. Bacharach. Estimating nonnegative matrices from marginal data. *International Economic Review*, 6(3):294–310, 1965.
[10] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.
[11] S. Basu, S. Kolouri, and G. K. Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014.
[12] A. Bäuerle, A. Bruneton, R. Wester, J. Stollenwerk, and P. Loosen. Algorithm for irradiance tailoring using multiple freeform optical surfaces. *Optics express*, 20(13):14477–14485, 2012.
[13] M. S. Bazaraa, J. J. Jarvis, and H. D. Sherali. *Linear programming and network flows*. John Wiley & Sons, 2011.

[14] J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[15] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[16] J.-D. Benamou, B. D. Froese, and A. M. Oberman. Numerical solution of the optimal transportation problem using the Monge-Ampére equation. *Journal of Computational Physics*, 260:107–126, 2014.

[17] D. P. Bertsekas. The auction algorithm: A distributed relaxation method for the assignment problem. *Annals of operations research*, 14(1):105–123, 1988.

[18] V. I. Bogachev and A. V. Kolesnikov. The Monge-Kantorovich problem: achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67(5):785, 2012.

[19] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and Radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[20] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[21] C. Brauer and D. Lorenz. Cartoon-texture-noise decomposition with transport norms. In *Scale Space and Variational Methods in Computer Vision*, pages 142–153. Springer, 2015.

[22] Y. Brenier. Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics*, 44(4):375–417, 1991.

[23] Y. Brenier, U. Frisch, M. Hénon, G. Loeper, S. Matarrese, R. Mohayaee, and A. Sobolevski. Reconstruction of the early universe as a convex optimization problem. *Monthly Notices of the Royal Astronomical Society*, 346(2):501–524, 2003.

[24] C. J. Budd, M. Cullen, and E. Walsh. Monge–ampére based moving mesh methods for numerical weather prediction, with applications to the Eady problem. *Journal of Computational Physics*, 236:247–270, 2013.

[25] L. A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, pages 99–104, 1992.

[26] L. A. Caffarelli and R. J. McCann. Free boundaries in optimal transport and Monge-Ampére obstacle problems. *Annals of mathematics*, pages 673–730, 2010.

[27] A. Cayley. On Monge's "mémoire sur la théorie des déblais et des remblais.". *Proceedings of the London Mathematical Society*, 14:139–142, 1882.

[28] R. Chartrand, K. Vixie, B. Wohlberg, and E. Bollt. A gradient descent solution to the Monge-Kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009.

[29] M. Colombo, L. De Pascale, and S. Di Marino. Multimarginal optimal transport maps for 1-dimensional repulsive costs. *Canad. J. Math*, 2013.

[30] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

[31] N. Courty, R. Flamary, and D. Tuia. Domain adaptation with regularized optimal transport. In *Machine Learning and Knowledge Discovery in Databases*, pages 274–289. Springer, 2014.

[32] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. Optimal transport for domain adaptation. *arXiv preprint arXiv:1507.00504*, 2015.

[33] D. Cremers, M. Rousson, and R. Deriche. A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape. *International journal of computer vision*, 72(2):195–215, 2007.

[34] W. R. Crum, T. Hartkens, and D. Hill. Non-rigid image registration: theory and practice. *The British Journal of Radiology*, 2014.

[35] W. H. Cunningham. A network simplex method. *Mathematical Programming*, 11(1):105–116, 1976.

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

[37] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. *arXiv preprint arXiv:1310.4375*, 2013.

[38] G. B. Dantzig. Programming in a linear structure. *Washington, DC*, 1948.

[39] G. B. Dantzig and M. N. Thapa. *Linear programming 2: theory and extensions*. Springer Science & Business Media, 2006.

[40] G. De Philippis and A. Figalli. W 2, 1 regularity for solutions of the Monge–ampère equation. *Inventiones mathematicae*, 192(1):55–69, 2013.

[41] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, et al. Tests of goodness of fit based on the l_2-Wasserstein distance. *The Annals of Statistics*, 27(4):1230–1239, 1999.

[42] J. Delon. Midway image equalization. *Journal of Mathematical Imaging and Vision*, 21(2):119–134, 2004.

[43] W. E. Deming and F. F. Stephan. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, 11(4):427–444, 1940.

[44] C. Dupin. Applications de la géométrie et de la mécanique. 1822. re-edition by Bachelier.

[45] A. A. Efros and W. T. Freeman. Image quilting for texture synthesis and transfer. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 341–346. ACM, 2001.

[46] T. A. El Moselhy and Y. M. Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[47] H. C. Elman and G. H. Golub. Inexact and preconditioned Uzawa algorithms for saddle point problems. *SIAM Journal on Numerical Analysis*, 31(6):1645–1661, 1994.

[48] B. Engquist, B. D. Froese, and Y. Yang. Optimal transport for seismic full waveform inversion. *arXiv preprint arXiv:1602.01540*, 2016.

[49] E. Esser. Applications of Lagrangian-based alternating direction methods and connections to split Bregman. *CAM report*, 9:31, 2009.

[50] L. C. Evans. Partial differential equations and Monge-Kantorovich mass transfer. *Current developments in mathematics*, pages 65–126, 1997.

[51] L. C. Evans and W. Gangbo. *Differential equations methods for the Monge-Kantorovich mass transfer problem*. American Mathematical Soc., 1999.

[52] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[53] S. Ferradans, G.-S. Xia, G. Peyré, and J.-F. Aujol. *Static and dynamic texture mixing using optimal transport*. Springer, 2013.

[54] R. Flamary, N. Courty, D. Tuia, and A. Rakotomamonjy. Optimal transport with Laplacian regularization: Applications to domain adaptation and shape matching. In *NIPS Workshop on Optimal Transport and Machine Learning OTML*, 2014.

[55] C. C. for Research in Economics and T. Koopmans. *Activity analysis of production and allocation*. Monograph. Wiley, 1951.

[56] M. Fortin and R. Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*. Elsevier, 2000.

[57] U. Frisch, S. Matarrese, R. Mohayaee, and A. Sobolevski. A reconstruction of the initial conditions of the universe by optimal mass transportation. *Nature*, 417(6886):260–262, 2002.

[58] U. Frisch and A. Sobolevskii. Application of optimal transportation theory to the reconstruction of the early universe. *Journal of Mathematical Sciences (New York)*, 133(1):303–309, 2004.

[59] B. D. Froese and A. M. Oberman. Convergent filtered schemes for the Monge-Ampére partial differential equation. *SIAM Journal on Numerical Analysis*, 51(1):423–444, 2013.

[60] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2044–2052, 2015.

[61] B. Galerne, Y. Gousseau, and J.-M. Morel. Random phase textures: Theory and synthesis. *Image Processing, IEEE Transactions on*, 20(1):257–267, 2011.

[62] W. Gangbo and R. J. McCann. Optimal maps in Monge's mass transport problem. *Comptes Rendus de l'Academie des Sciences-Serie I-Mathematique*, 321(12):1653, 1995.

[63] W. Gangbo and R. J. McCann. The geometry of optimal transportation. *Acta Mathematica*, 177(2):113–161, 1996.

[64] N. Gigli. Second order analysis on $(P_2(M), W_2)$. *Mem. Amer. Math. Soc.*, 216(1018):xii+154, 2012.

[65] T. Graf and V. I. Oliker. An optimal mass transport approach to the near-field reflector problem in optical design. *Inverse Problems*, 28(2):025001, 2012.

[66] U. Grenander and M. I. Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56(4):617–694, 1998.

[67] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[68] E. Haber, T. Rehman, and A. Tannenbaum. An efficient numerical method for the solution of the l_2 optimal mass transfer problem. *SIAM Journal on Scientific Computing*, 32(1):197–211, 2010.

[69] S. Haker and A. Tannenbaum. On the Monge-Kantorovich problem and image warping. *IMA Volumes in Mathematics and its Applications*, 133:65–86, 2003.

[70] S. Haker, A. Tannenbaum, and R. Kikinis. Mass preserving mappings and image registration. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2001*, pages 120–127. Springer, 2001.

[71] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal of*

Computer Vision, 60(3):225–240, 2004.
[72] D. Hobson and M. Klimmek. Robust price bounds for the forward starting straddle. *Finance and Stochastics*, 19(1):189–214, 2015.
[73] S. C. Joshi and M. I. Miller. Landmark matching via large deformation diffeomorphisms. *Image Processing, IEEE Transactions on*, 9(8):1357–1370, 2000.
[74] L. V. Kantorovich. On translation of mass (in Russian), C R. Doklady. *Acad. Sci. USSR*, 37:199–201, 1942.
[75] L. V. Kantorovich. A problem of Monge. *Uspekhi Mat. Nauk*, 3(24):225–226, 1948.
[76] L. V. Kantorovich. Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422, 1960.
[77] L. V. Kantorovitch. On the translocation of masses. *Management Science*, 5(1):1–4, 1958.
[78] N. Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM, 1984.
[79] T. Kato, H. Itoh, and A. Imiya. Optical flow computation with locally quadratic assumption. In *Computer Analysis of Images and Patterns*, pages 223–234. Springer, 2015.
[80] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7(3):358–386, 2005.
[81] S. Kim, R. Ma, D. Mesa, and T. P. Coleman. Efficient bayesian inference methods via convex optimization and optimal transport. In *Information Theory Proceedings (ISIT), 2013 IEEE International Symposium on*, pages 2259–2263. IEEE, 2013.
[82] J. Kitagawa. An iterative scheme for solving the optimal transportation problem. *Calc. Var. Partial Differential Equations*, 51(1-2):243–263, 2014.
[83] J. Kitagawa, Q. Mérigot, and B. Thibert. Convergence of a newton algorithm for semi-discrete optimal transport. *arXiv preprint arXiv:1603.05579*, 2016.
[84] P. A. Knight. The Sinkhorn-Knopp algorithm: Convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1):261–275, 2008.
[85] M. Knott and C. Smith. On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43(1):39–49, 1984.
[86] I. Kolesov, P. Karasev, A. Tannenbaum, and E. Haber. Fire and smoke detection in video with optimal mass transport based optical flow and neural networks. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*, pages 761–764. IEEE, 2010.
[87] S. Kolouri, S. R. Park, and G. K. Rohde. The Radon cumulative distribution transform and its application to image classification. *Image Processing, IEEE Transactions on*, 25(2):920–934, 2016.
[88] S. Kolouri and G. K. Rohde. Transport-based single frame super resolution of very low resolution face images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2015.
[89] S. Kolouri, A. B. Tosun, J. A. Ozolek, and G. K. Rohde. A continuous linear optimal transport approach for pattern analysis in image datasets. *Pattern Recognition*, 51:453–462, 2016.
[90] S. Kolouri, Y. Zou, and G. K. Rohde. Sliced Wasserstein kernels for probability distributions. *arXiv preprint arXiv:1511.03198*, 2015.
[91] S. Kolouri, Y. Zou, and G. K. Rohde. Sliced-Wasserstein kernels for probability distributions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4876–4884, 2016.
[92] M. Kuang and E. Tabak. Preconditioner for optimal transport. preprint, 2016.
[93] R. Lai and H. Zhao. Multi-scale non-rigid point cloud registration using robust sliced-Wasserstein distance via Laplace-Beltrami eigenmap. *arXiv preprint arXiv:1406.3758*, 2014.
[94] K. Lee, J. Ho, and D. Kriegman. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Trans. Pattern Anal. Mach. Intelligence*, 27(5):684–698, 2005.
[95] J. Lellmann, D. A. Lorenz, C. Schnlieb, and T. Valkonen. Imaging with Kantorovich–Rubinstein discrepancy. *SIAM Journal on Imaging Sciences*, 7(4):2833–2859, 2014.
[96] V. L. Levin. Duality theorems in the Monge-Kantorovich problem. *Uspekhi Matematicheskikh Nauk*, 32(3):171–172, 1977.
[97] B. Lévy. A numerical algorithm for L_2 semi-discrete optimal transport in 3D. *ESAIM Math. Model. Numer. Anal.*, 49(6):1693–1715, 2015.
[98] P. Li, Q. Wang, and L. Zhang. A novel earth mover's distance methodology for image matching with gaussian mixture models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1689–1696, 2013.
[99] M. Lindsey and Y. A. Rubinstein. Optimal transport via a monge-ampre optimization problem. *arXiv preprint arXiv:1603.07435*, 2016.
[100] G. Loeper and F. Rapetti. Numerical solution of the Monge–ampère equation by a newton's algorithm. *Comptes Rendus Mathematique*, 340(4):319–324, 2005.
[101] J. Lott. Some geometric calculations on Wasserstein space. *Communications in Mathematical Physics*, 277(2):423–437, 2008.
[102] I. J. Lustig, R. E. Marsten, and D. F. Shanno. Interior point methods for linear programming: Computational state of the art. *ORSA Journal on Computing*, 6(1):1–14, 1994.
[103] Y. Makihara and Y. Yagi. Earth mover's morphing: Topology-free shape morphing using cluster-based EMD flows. In *Computer Vision–ACCV 2010*, pages 202–215. Springer, 2010.
[104] B. Mathon, P. Bas, F. Cayre, and B. Macq. Optimization of natural watermarking using transportation theory. In *Proceedings of the 11th ACM workshop on Multimedia and security*, pages 33–38. ACM, 2009.
[105] B. Mathon, F. Cayre, P. Bas, and B. Macq. Optimal transport for secure spread-spectrum watermarking of still images. *Image Processing, IEEE Transactions on*, 23(4):1694–1705, 2014.
[106] Q. Mérigot. A multiscale approach to optimal transport. *Computer Graphics Forum*, 30(5):1583–1592, 2011.
[107] L. Métivier, R. Brossier, Q. Mérigot, E. Oudet, and J. Virieux. Measuring the misfit between seismograms using an optimal transport distance: application to full waveform inversion. *Geophysical Journal International*, 205(1):345–377, 2016.
[108] T. Mikami and M. Thieullen. Optimal transportation problem by stochastic optimal control. *SIAM Journal on Control and Optimization*, 47(3):1127–1139, 2008.
[109] G. Monge. *Mémoire sur la théorie des déblais et des remblais*. De l'Imprimerie Royale, 1781.
[110] G. Montavon, K.-R. Müller, and M. Cuturi. Wasserstein training of Boltzmann machines. *arXiv preprint arXiv:1507.01972*, 2015.
[111] O. Museyko, M. Stiglmayr, K. Klamroth, and G. Leugering. On the application of the Monge-Kantorovich problem to image registration. *SIAM Journal on Imaging Sciences*, 2(4):1068–1097, 2009.
[112] L. J. Neumann and O. Morgenstern. *Theory of games and economic behavior*. Princeton university press Princeton, 1947.
[113] K. Ni, X. Bresson, T. Chan, and S. Esedoglu. Local histogram based segmentation using the Wasserstein distance. *International journal of computer vision*, 84(1):97–111, 2009.
[114] A. M. Oberman and Y. Ruan. An efficient linear programming method for optimal transportation. *arXiv preprint arXiv:1509.03668*, 2015.
[115] F. P. Oliveira and J. M. R. Tavares. Medical image registration: A review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93, 2014.
[116] D. S. Oliver. Minimization for conditional simulation: Relationship to optimal transport. *Journal of Computational Physics*, 265:1–15, 2014.
[117] F. Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26(1-2):101–174, 2001.
[118] J. A. Ozolek, A. B. Tosun, W. Wang, C. Chen, S. Kolouri, S. Basu, H. Huang, and G. K. Rohde. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical image analysis*, 18(5):772–780, 2014.
[119] B. Pass. Multi-marginal optimal transport: theory and applications. *arXiv preprint arXiv:1406.0026*, 2014.
[120] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *Signal Processing Magazine, IEEE*, 32(3):53–69, 2015.
[121] O. Pele and M. Werman. Fast and robust earth mover's distances. In *Computer vision, 2009 IEEE 12th international conference on*, pages 460–467. IEEE, 2009.
[122] Y. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.
[123] J. Rabin, S. Ferradans, and N. Papadakis. Adaptive color transfer with relaxed optimal transport. In *Image Processing (ICIP), 2014 IEEE International Conference on*, pages 4852–4856. IEEE, 2014.
[124] J. Rabin and N. Papadakis. Non-convex relaxation of optimal transport for color transfer. In *NIPS Workshop on Optimal Transport for Machine Learning*, 2014.
[125] J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer, 2015.
[126] J. Rabin and G. Peyré. Wasserstein regularization of imaging problem. In *Image Processing (ICIP), 2011 18th IEEE International Conference on*, pages 1541–1544. IEEE, 2011.
[127] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2012.
[128] S. T. Rachev. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985.
[129] S. T. Rachev and L. Rüschendorf. *Mass Transportation Problems: Volume I: Theory, Volume II: Appications*, volume 1. Springer, New

York, 1998.
[130] A. Ramdas, N. Garcia, and M. Cuturi. On Wasserstein two sample testing and related families of nonparametric tests. *arXiv preprint arXiv:1509.02237*, 2015.
[131] S. Reich. A nonparametric ensemble transform method for Bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.
[132] E. Reinhard, M. Ashikhmin, B. Gooch, and P. Shirley. Color transfer between images. *IEEE Computer graphics and applications*, 21(5):34–41, 2001.
[133] S. Rim Park, S. Kolouri, S. Kundu, and G. Rohde. The Cumulative Distribution Transform and Linear Pattern Classification. *ArXiv e-prints 1507.05936*, July 2015.
[134] G. K. Rohde and Others. Transport and other Lagrangian transforms for signal analysis and discrimination. http://faculty.virginia.edu/rohde/transport.
[135] M. Rosenblatt. Remarks on a multivariate transformation. *The annals of mathematical statistics*, 23(3):470–472, 1952.
[136] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.
[137] L. Ruschendorf. Convergence of the iterative proportional fitting procedure. *The Annals of Statistics*, pages 1160–1174, 1995.
[138] E. J. Russell. Letters to the editor-extension of Dantzig's algorithm to finding an initial near-optimal basis for the transportation problem. *Operations Research*, 17(1):187–191, 1969.
[139] F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY (due in September 2015)*, 2015.
[140] L.-P. Saumier, M. Agueh, and B. Khouider. An efficient numerical algorithm for the l2 optimal transport problem with applications to image processing. *arXiv preprint arXiv:1009.6039*, 2010.
[141] L.-P. Saumier, B. Khouider, and M. Agueh. Optimal transport for particle image velocimetry. *Communications in Mathematical Sciences*, 13(1):269–296, 2015.
[142] J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.
[143] B. Schmitzer. A sparse algorithm for dense optimal transport. In *Scale Space and Variational Methods in Computer Vision*, pages 629–641. Springer, 2015.
[144] B. Schmitzer and C. Schnörr. Object segmentation by shape matching with Wasserstein modes. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 123–136. Springer, 2013.
[145] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pages 3294–3302, 2015.
[146] R. Sinkhorn and P. Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific Journal of Mathematics*, 21(2):343–348, 1967.
[147] J. Solomon, F. de Goes, P. A. Studios, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (Proc. SIGGRAPH 2015), to appear*, 2015.
[148] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of The 31st International Conference on Machine Learning*, pages 306–314, 2014.
[149] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *Medical Imaging, IEEE Transactions on*, 32(7):1153–1190, 2013.
[150] Z. Su, Y. Wang, R. Shi, W. Zeng, J. Sun, F. Luo, and X. Gu. Optimal mass transport for shape matching and comparison. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 37(11):2246–2259, 2015.
[151] V. Sudakov. *Geometric Problems in the Theory of Infinite-dimensional Probability Distributions*. Number no. 141 in Geometric Problems in the Theory of Infinite-dimensional Probability Distributions. American Mathematical Society, 1979.
[152] M. M. Sulman, J. Williams, and R. D. Russell. An efficient approach for the numerical solution of the Monge–Ampère equation. *Applied Numerical Mathematics*, 61(3):298–307, 2011.
[153] P. Swoboda and C. Schnorr. Convex variational image restoration with histogram priors. *SIAM Journal on Imaging Sciences*, 6(3):1719–1735, 2013.
[154] E. Tannenbaum, T. Georgiou, and A. Tannenbaum. Signals and control aspects of optimal mass transport and the Boltzmann entropy. In *49th IEEE Conference on Decision and Control (CDC)*, pages 1885–1890. IEEE, 2010.
[155] A. B. Tosun, O. Yergiyev, S. Kolouri, J. F. Silverman, and G. K. Rohde. Novel computer-aided diagnosis of mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. In *SPIE Medical Imaging*, pages 90410Z–90410Z. International Society for Optics and Photonics, 2014.
[156] A. B. Tosun, O. Yergiyev, S. Kolouri, J. F. Silverman, and G. K. Rohde. Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. *Cytometry Part A*, 2015.
[157] N. S. Trudinger and X.-J. Wang. On the second boundary value problem for Monge-Ampére type equations and optimal transportation. *arXiv preprint math/0601086*, 2006.
[158] S. M. Ulam. *A collection of mathematical problems*, volume 8. Interscience Publishers, 1960.
[159] T. ur Rehman, E. Haber, G. Pryor, J. Melonakos, and A. Tannenbaum. 3D nonrigid registration via optimal mass transport on the GPU. *Medical image analysis*, 13(6):931–940, 2009.
[160] T. ur Rehman, G. Pryor, and A. Tannenbaum. Fast multigrid optimal mass transport for image registration and morphing. In *British Machine Vision Conference*, 2007.
[161] R. J. Vanderbei. *Linear programming*. Springer, 2014.
[162] A. Vershik. Long history of the Monge-Kantorovich transportation problem. *The Mathematical Intelligencer*, 35(4):1–9, 2013.
[163] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[164] W. Wang, J. A. Ozolek, D. Slepcev, A. B. Lee, C. Chen, and G. K. Rohde. An optimal transportation approach for nuclear structure-based pathology. *Medical Imaging, IEEE Transactions on*, 30(3):621–631, 2011.
[165] W. Wang, D. Slepcev, S. Basu, J. A. Ozolek, and G. K. Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.
[166] X. Zhang, M. Burger, and S. Osher. A unified primal-dual algorithm framework based on Bregman iteration. *Journal of Scientific Computing*, 46(1):20–46, 2011.
[167] L. Zhu, Y. Yang, S. Haker, and A. Tannenbaum. An image morphing technique based on optimal mass preserving mapping. *Image Processing, IEEE Transactions on*, 16(6):1481–1495, 2007.

Soheil Kolouri received his B.S. degree in electrical engineering from Sharif University of Technology, Tehran, Iran, in 2010, and his M.S. degree in electrical engineering in 2012 from Colorado State University, Fort Collins, Colorado. He earned his doctorate degree in biomedical engineering from Carnegie Mellon University in 2015. His thesis, titled, "Transport-based pattern recognition and image modeling", won the best thesis award. He is currently at HRL Laboratories, Malibu, California, United States.

Serim Park received her B.S. degree in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea in 2011 and is currently a doctoral candidate in the Electrical and Computer Engineering Department at Carnegie Mellon University, Pittsburgh, United States. She is mainly interested in signal processing and machine learning, especially designing new signal and image transforms and developing novel systems for pattern recognition.

Matthew Thorpe received his BSc, MSc and PhD in Mathematics from the University of Warwick, UK in 2009, 2012 and 2015 respectively and his MScTech in Mathematics from the University of New South Wales, Australia, in 2010. He is currently a postdoctoral associate within the mathematics department at Carnegie Mellon University.

Dejan Slepčev earned B.S degree in mathematics from the University of Novi Sad in 1995, M.A. degree in mathematics from University of Wisconsin Madison in 2000 and Ph.D. in mathematics from University of Texas at Austin in 2002. He is currently associate professor at the Department of Mathematical Sciences at Carnegie Mellon University.

Gustavo K. Rohde earned B.S. degrees in physics and mathematics in 1999, and the M.S. degree in electrical engineering in 2001 from Vanderbilt University. He received a doctorate in applied mathematics and scientific computation in 2005 from the University of Maryland. He is currently an associate professor of Biomedical Engineering, and Electrical and Computer Engineering at the University of Virginia. Contact: gustavo@virginia.edu.