ALPHA DIVERGENCES BASED MASS TRANSPORT MODELS FOR IMAGE MATCHING PROBLEMS

PENGWEN CHEN
Department of Mathematics
National Taiwan University
Taiwan

CHANGFENG GUI
School of Mathematics and Econometrics
Hunan University, Changsha 410082, China
and
Department of Mathematics, University of Connecticut
Storrs, CT 06269, USA

(Communicated by Hao-Min Zhou)

Abstract. Registration methods could be roughly divided into two groups: area-based methods and feature-based methods. In the literature, the Monge-Kantorovich (MK) mass transport problem has been applied to image registration as an area-based method. In this paper, we propose to use Monge-Kantorovich (MK) mass transport model as a feature-based method. This novel image matching model is a coupling of the MK problem with the well-known alpha divergence from the probability theory. The optimal matching scheme is the one which minimizes the weighted alpha divergence between two images. A primal-dual approach is employed to analyze the existence and uniqueness/non-uniqueness of the optimal matching scheme. A block coordinate method, analogous to the Sinkhorn matrix balancing method, can be used to compute the optimal matching scheme. We also derive a distance function for image morphing. Similar to elastic distances proposed by Younes, the geodesic under this distance function has an explicit expression.

1. Introduction. Image matching or image registration is one important task in medical imaging and computer vision. The goal is to align two or more pictures taken at different times, or from different sensors. Given two (or many) images of a set of objects, one needs to identify the corresponding objects in these images. Due to the existence of noises, or different imaging method, these corresponding objects may be located at different positions or have different intensities. A registration process is to find an optimal deformation from one image to the other in a certain sense. Over several decades, many registration methods are proposed to search for the optimal deformation over a certain class of transformations (e.g., affine, elastic) such that the similarity among the corresponding objects is maximized, for instance, see [47],[34],[16], and the references therein. Generally a registration...
method consists of two parts: one is the choice of the feature space, and the other is the selection of the similarity function. The feature space could include pixel intensity, edges, corners, or points, etc. The similarity function could be the sum of intensity differences, the sum of squared intensity differences, the cross-correlation function, the mutual information, etc. Due to the diversity of images to be registered, it is impossible to design a universal method applicable to all registration tasks. According to whether distinctive objects are detected, registration methods can be divided into two groups: area-based methods and feature-based methods. One survey of registration methods can be found in [61].

In this paper, we are interested in applying Monge-Kantorovich problem in the image registration, especially developing some feature-based methods. In the 1780’s, Monge formulated a problem of transporting a pile of soil with the least amount of work. In the 1940’s, Kantorovich [27] employed a dual variation principle to convert the original problem into a linear problem. A survey of theoretical works on this problem can be found in [13]. Recently, Monge-Kantorovich (MK) mass transport has been applied to image retrieval, image registration (matching) and image morphing, for instance in [55],[26],[41],[59],[37],[35]. In the image registration, the image intensity functions are regarded as piles of soil and image registration becomes the task of moving these piles of soil. The minimal transport cost is also called the earth mover’s distance [41] or the Kantorovich distance. The MK problem itself has several advantages: the minimizer is unique, and there is no other local minimizers, and the MK problem is also parameter free and symmetric. However, a well-known undesirable effect called the “fade-in and fade-out” effect might occur, i.e., the matching result could map a small high intensity region to a large low intensity region which is not related (see [21]). This intensity change effect mainly originates from the intrinsic measure-preserving constraint enforced in the MK problem. To alleviate this effect, Haker et. al. added an image intensity comparison term to the MK cost function to penalize the change of intensity [21]. On the other hand, causing by the measure-preserving constraint, numerically computing the Kantorovich distance for images is not an easy task. Their algorithm is a gradient descent method searching for a minimizer of the transport cost over the set of measure preserving mappings. The gradient direction is computed through the Helmholtz decomposition. To maintain the measure preserving property, the stepsize should not be too large. Each iteration of this algorithm has the computational complexity of order $O(N \log N)$, where $N$ is equal to the number of pixels in the two images. This is a significant improvement, compared with the computational complexity of previous numerical methods has order $O(N^2)$ (See, e.g. [26]). Since this method works directly with intensity values without detecting any salient object, it can be regarded as an area-based method. In practical applications, area based methods are sensitive to intensity changes which are usually introduced by noises or varying illumination [61].

As mentioned in [61], feature-based methods are recommended if the images contain enough distinctive and easily detectable objects. Traditionally, the feature objects were selected manually by experts. Currently, various feature detection methods are preferably employed to automatically extract salient structures in the images. The extracted features could be region features, line features or point features, such as the locations of corners and boundary points. For further processing, these features can be represented by their point representatives, which are called control points in the literature. Hence, point set representations
of image data are commonly used, and point-sets matching problem is of pivotal importance in many application domains, including medical image alignment[34], segmentation[28], tracking[44], shape recognition[43], and stereo matching[60].

The point pattern matching problem has attracted persistent interest in computer vision[50] and shape statistics[46],[11] communities for several decades. The problem could be separated into two main components, the spatial transformation and the point-to-point correspondence. The spatial transformation includes linear transforms (translation, rotation, scaling and affine transforms), and non-rigid transformations (e.g., a variety of splines[5],[51], and large deformation models[24],[19]). The correspondence problem can be tackled by a variety of methods, e.g., point assignment[42] and spectral methods[29]. The task of point-set registration is to establish a consistent point-to-point correspondence between two point sets or to recover the spatial transformation which yields the best alignment. A point set registration method usually should meet several desirable criteria: (1) it can align point sets accurately with a low computational complexity; (2) the alignment should be robust against the existence of outliers or missing corresponding points.

There are extensive studies on this problem. We shall give a brief review, focusing on those works directly related to this paper. The iterative closest point (ICP) algorithm is one common approach to feature-based image registration problem, because of its simplicity[4],[58]. One limitation of ICP is its local convergence problem. To avoid this problem, it requires sufficient overlap between the initial data-sets. Its vulnerability in performance also includes the proneness to outliers. To alleviate these difficulties, a variety of robust methods were developed. In these methods, multiple weighted matches are introduced to reduce the number of local minima. These models can be regarded as multiple-linked ICP methods. In[10], Chui and Rangarjan proposed a robust point matching (RPM) method, which estimates non-rigid transformation and correspondence simultaneously [10]. A technique called soft-assignment is used to establish correspondence. Later, the relation between the robust point matching method and the expectation maximization (EM) algorithm for Gaussian mixtures was studied in[9]. In[20], Granger and Pennec proposed the EM-ICP method, in which a maximal likelihood approach is used to estimate the transformation and the correspondence between point-sets. This probabilistic approach has been employed in many works, including Wells[54], Luo and Hancock[30], McNeil and Vijayakumar[33], and Tu et al.[49]. They viewed the point sets registration as a maximum likelihood estimation problem. The optimization algorithm is based on the Expectation Maximization principle: E-step is the computation of probabilities and M-step is the update of the transformation.

Recently, Tsin and Kanade[48] proposed a kernel correlation (KC) based point-set registration method to align point-sets without establishing the explicit point correspondence: Let $X := \{x_i : i = 1, \ldots, m\}$ and $Y := \{y_j : j = 1, \ldots, n\}$ be two point-sets belonging to a finite-dimensional vector space $\mathbb{R}^d$, $d = 2$ or 3. The Gaussian kernel correlation between any two points $x_i, y_j$ is given by

$$KC(x_i, y_j) = \exp\left(-\frac{\|x_i - y_j\|^2}{\sigma^2}\right),$$

and the kernel correlation between point-sets $X, Y$ is defined as

$$\left(\frac{mn}{mn}\right)^{-1} \sum_{i=1}^{m} \sum_{j=1}^{n} KC(x_i, y_j).$$
The matrix with entries \( \{ \exp(-\|x_i - y_j\|^2/\sigma^2) : i = 1, \ldots, m; j = 1, \ldots, n \} \) is known as the proximity matrix[42], and \( \sigma \) is the kernel scale. The transformation \( T \) is estimated through \( \min_{T=\mathbf{M}_{KC}(T(X), Y)} \{ \mathbf{KC}(Y, Y) - 2\mathbf{KC}(Y, T(X)) + \mathbf{KC}(T(X), T(X)) \} \), where \( T(X) := \{ T(x_i) : x_i \in X \} \). Their experiment shows that the multiple-linked registration is more robust than the ICP method, and also outperforms the EM-ICP method in terms of registration accuracy. The KC framework was further improved by Jian and Vemuri in [23]. They regarded points as realizations of some mixture of Gaussian distributions, which changes the registration of point-sets into the alignment of two Gaussian mixtures. In this framework, the transformation is estimated through minimizing the \( L^2 \) distance between Gaussian distributions. This framework was later applied in the multiple-point-sets registration[52].

In this paper, we would propose another multiple-linked ICP method, called the alpha divergence matching model, in which a kernel correlation is weighted by the alpha divergence\(^1\) of correspondence functions (see Eq. (7)). By maximizing this weighted kernel correlation, both the unknown transformation and correspondence are estimated. Given two point-sets \( X \) and \( Y \) as described above, we assign a unit weight \( \gamma_{i,j}^+ = 1 \) to each point \( x_i \), and a unit weight \( \gamma_{j,i}^- = 1 \) to each point \( y_j \). Represent a matching structure by two nonnegative \( m \times n \) matching (or correspondence) matrices \( \Gamma^+ := \{ \gamma_{i,j}^+ : i,j \} \), \( \Gamma^- := \{ \gamma_{j,i}^+ : i,j \} \) with the property

\[
\sum_{j=1}^{n} \gamma_{i,j}^+ = \gamma_{i}^+ = 1, \quad \sum_{i=1}^{m} \gamma_{j,i}^- = \gamma_{j}^- = 1 \quad \text{for all } i,j.
\]

In other words, \( \{ \gamma_{i,j}^+ : j \} \) is a partition of \( \gamma_{i}^+ = 1 \) and \( \{ \gamma_{j,i}^- : i \} \) is a partition of \( \gamma_{j}^- = 1 \). Let

\[
E_\alpha(\Gamma^+, \Gamma^-) = \sum_{i=1}^{m} \sum_{j=1}^{n} (\gamma_{i,j}^+)^\alpha (\gamma_{j,i}^-)^{1-\alpha} K(x_i, y_j),
\]

where \( 0 < \alpha < 1 \) and \( K(x,y) = K(\|x - y\|) \) with \( K : [0, \infty) \rightarrow (0, 1] \) satisfying

\[
(i) \ K(r) = 1 \text{ if and only if } r = 0, (ii) K(r) \text{ is a decreasing function of } r.
\]

Then we quantize the similarity between the point-sets be the maximum

\[
E_\alpha(X, Y) = \max_{\Gamma^+, \Gamma^-} E_\alpha(\Gamma^+, \Gamma^-),
\]

subject to \( m + n \) marginal constraints in Eq. (4). The matching structure is characterized by its maximizer.

Since \( \gamma_{i,j}^+, \gamma_{j,i}^- \) > 0 if and only if \( \gamma_{i,j}^+, \gamma_{j,i}^- \) > 0 for each pair \( (i,j) \) (see Lemma 4.3), then we may describe the matching structure by a set of matching pairs, \( \{(x_i, y_j) : \gamma_{i,j}^+, \gamma_{j,i}^- > 0 \} \), and each pair \( \gamma_{i,j}^+, \gamma_{j,i}^- > 0 \) indicates a link between \( x_i \) and \( y_j \). We call this maximization problem in Eq. (7) the alpha divergence matching (alpha-D)

\(^1\)The alpha-divergence[39] between two discrete distributions \( p = (p_1, \ldots, p_n), q = (q_1, \ldots, q_n) \) of fractional order \( \alpha \in (0, 1) \) is

\[
D_\alpha(p||q) := (\alpha - 1)^{-1} \log_2 \left( \sum_{i=1}^{n} p_i^\alpha q_i^{1-\alpha} \right).
\]
problem. In this discrete setting, the matching between point-sets is usually not one-to-one, i.e., a fuzzy matching is allowed.

Analogous to the KC method, the function

\[ M_\alpha(X, Y) := \alpha E_\alpha(X, X) - E_\alpha(X, Y) + (1 - \alpha) E_\alpha(Y, Y) \]

is always nonnegative and equals to zero if and only if \( X = Y \) (see Lemma B.1). Then using the closest point property (see Lemma A.1) we have \( E_\alpha(X, X) = \sum_{i=1}^{m} \gamma_i^+ \) and \( E_\alpha(Y, Y) = \sum_{j=1}^{n} \gamma_j^- \) are constants, we propose to estimate the transformation by

\[ \max_T E_\alpha(T(X), Y) + \text{some regularity term.} \]

We call the above framework the \textit{alpha divergence matching model} or the alpha-D model.

When \( K(x, y) = \exp(-\|x - y\|^2/\sigma^2) \), this approach can be regarded as a generalization of the KC method, hence it inherits the robustness of the KC method. This framework has several nice properties:

1. The alpha-D problem (Eq. (7)) has the closest point property (see Lemma A.1). When one point set differs from another point set by a sufficiently small perturbation, then the correspondence can be identified correctly;

2. The \( L^2 \) MK mass transport problem is an asymptotic limit of the alpha-D problem when the kernel scale \( \sigma \) approaches infinity (see Remark 1). The alpha-D problem also shares some measure-preserving property in the MK mass transport (see Theorem 3.7). Empirically, these properties would reduce the number of local optimal solutions through inhibiting a large number of matching links connecting to one single point.

3. The measure preserving condition does not appear as a constraint in the alpha-D problem. Each correspondence matrix is optimized only subject to row-sum or column-sum constraints. Numerically, this optimization problem is much easier than the MK problem, in which one correspondence matrix is optimized subject to both row-sum and column sum constraints. That is, one column stochastic matrix and one row stochastic matrix are computed to represent the matching structure in the alpha-D problem, instead of a doubly stochastic matrix in the MK problem. Besides, thanks to the partial decomposable structure in Eq. (9), the block-ascent coordinate method can be employed to solve the point-set matching problem as follows:

\[ \max_{\Gamma^+} \max_{\Gamma^-} \max_T E_\alpha(T(X), Y) + \text{some regularity term,} \]

subject to proper constraints. Keeping remaining variables fixed, the maximizer for \( \Gamma^+ \) or \( \Gamma^- \) is uniquely determined with explicit expressions. Thus, we may solve the problem by updating \( \Gamma^+, \Gamma^-, T \) cyclically.

In this paper, we aim at providing theoretical analysis of the continuous version of the alpha-D problem (Eq. (7)), including existence and uniqueness for the optimal matching scheme (the continuous counterpart of the correspondence matrices). In the MK problem, the uniqueness of the optimal measure preserving mapping follows from some absolute continuity condition called the \( N^{-1} \) property (see [7],[17]). Analogously, we list several conditions to ensure the uniqueness of the optimal matching scheme in the alpha-D problem, including the bijection condition and the absolute continuity condition. It is worth mentioning that Gaussian kernel...
functions satisfy the bijection condition. In other words, the continuous counterpart of the above fuzzy correspondence matrix could become a bijection mapping when the above proposed conditions are met. Numerically, an algorithm (with a resemblance to Sinkhorn balancing\[15\]) is proposed to compute the optimal correspondence matrices.

This paper is organized as follows. In section 2, we start with a brief review of the MK problem and then introduce the alpha-D problem. Several relations between the MK problem and the alpha-D problem are discussed. In section 3, we shall prove the existence and uniqueness of the solutions in the alpha-D problem and prove its convergence. Afterwards, we investigate its related image morphing problem for $\alpha = 1/2$. An algorithm is provided to compute a geodesic path between two images on its corresponding Riemannian metric. In section 5, we present several preliminary numerical results.

2. Background & problem formulation. The alpha divergence matching (alpha-D) problem proposed in this paper is to couple the Monge-Kantorovich (MK) mass transport problem with the alpha divergence. The main difference between the alpha-D problem and MK problem is the treatment on the measure preserving condition. In the MK problem, the measure preserving condition is strictly imposed; the condition is relaxed in the alpha-D problem. Next, we introduce these two problems formally.

2.1. MK mass transport. We list several related well-known properties in MK mass transport problem according to a survey \[12\].

Consider two nonnegative Radon measures $\nu_X, \nu_Y$, both absolutely continuous to Lebesgue measure, on $X := \mathbb{R}^d, Y := \mathbb{R}^d$ with their supports $\Omega_X := \text{spt}(\nu_X) \subset X, \Omega_Y := \text{spt}(\nu_Y) \subset Y$, satisfying the mass balance condition

\[ \nu_X(X) = \nu_Y(Y) < \infty. \]  

Given the cost density $c : X \times Y \to [0, \infty)$, we seek for an optimal measure-preserving mapping $s : X \to Y$, such that

\[ \text{the total cost } L_{MK}[s] := \int_X c(x, s(x))d\nu_X \text{ is minimized}, \]

where $s$ is a measure preserving mapping, i.e.,

\[ \int_X g(s(x))d\nu_X(x) = \int_Y g(y)d\nu_Y(y), \]

for all continuous functions $g \in C(Y)$. Then instead of searching for $s$, we seek for a product measure, which implicitly describes the mass transport scheme, in a class of product measures:

\[ \{ \text{Radon measures } \nu \text{ on } X \times Y: \text{proj}_{x \in X} \nu = \nu_X, \text{proj}_{y \in Y} \nu = \nu_Y \}. \]

Then the original problem in Eq. (12) becomes the minimization problem

\[ \min \{ J[\nu] := \int c(x, y)d\nu(x, y) \}, \]

2The support spt $\nu$ of a measure $\nu$ is the smallest closed subset $\mathcal{M}$ in $\mathbb{R}^d$ having total mass $\nu(\mathbb{R}^d) = \nu(\mathcal{M})$. 

Inverse Problems and Imaging Volume 5, No. 3 (2011), 551–590
subject to $\nu$ in the above class of product measures. This formulation circumvents the difficulty caused by the nonlinear structure of the constraint in Eq. (12). Then by introducing a Lagrangian function, one can derive its dual problem

$$
\max_{u \in C(X), v \in C(Y)} \left\{ M_{MK}[u, v] := \int_X u(x)d\nu_X(x) + \int_Y v(y)d\nu_Y(y) \right\},
$$

subject to $u(x) + v(y) \leq c(x, y)$, where $u, v$ are Lagrange multipliers.

In the $L^2$ case $c(x, y) = (x - y)\cdot(x - y)$, we introduce a pair of new dual variables $\phi(x) = x \cdot x - u(x), \psi(x) = y \cdot y - v(y)$, and then the dual problem is equivalent to the problem

$$
\min \int_X \phi(x)d\nu_X + \int_Y \psi(y)d\nu_Y, \text{ subject to } \phi(x) + \psi(y) \geq x \cdot y.
$$

Based on this formulation, one can show that the optimal mapping $s$ can be expressed as $s = \nabla \phi(x)$ a.e.\footnote{Here, a.e. refers to “almost everywhere”.} Besides, the new constraint in Eq. (17) yields that the optimal $(\phi, \psi)$ must be a convex conjugate pair. Thus, the mass transport problem is equivalent to finding a pair of convex conjugate functions, such that the mapping $s(x) := \nabla \phi(x)$ is measure-preserving. Thirdly, the optimal measure-preserving mapping $s$ is essentially a bijection from $X$ to $Y$. Readers are referred to $[7,17,32,38]$ for more discussion on the MK mass transport problem.

When the MK problem is applied to image matching tasks, the image intensity functions are regarded as piles of soil described by the measures $\nu_X$ and $\nu_Y$\footnote{Here, a.e. refers to “almost everywhere”.} and the image intensity functions are normalized so that the mass balance condition $\nu_X(\Omega_X) = \nu_Y(\Omega_Y)$ holds. The optimal matching scheme for two images is implicitly described by the optimal product measure $\nu$ with the least mass transport cost. Lastly, although a variety of cost density functions are discussed in the literature, e.g., $\|x - y\|^p$ with $1 \leq p < \infty$, the measure preserving constraint is strictly enforced in the MK problem.

### 2.2. The alpha-divergence (alpha-D) problem.

The setting of the alpha divergence matching (alpha-D) problem is given as follows. Consider two nonnegative Radon measures $\nu_X$ on $X := \mathbb{R}^d, \nu_Y$ on $Y := \mathbb{R}^d$ with supports $\Omega_X \subset X, \Omega_Y \subset Y$, i.e., $\nu_X(X) = \nu_Y(\Omega_X) < \infty$, and $\nu_Y(Y) = \nu_Y(\Omega_Y) < \infty$. Let the similarity function $K(x, y) : \Omega \rightarrow \mathbb{R}$ be a positive, smooth function (at least a twice differentiable function), which quantizes the similarity between two vectors $x \in X, y \in Y$.

Define the primal problem as follows. Assume that both $\nu_X, \nu_Y$ are absolutely continuous with respect to Lebesgue measures $\mu_X, \mu_Y$, i.e., $\nu_X \ll \mu_X, \nu_Y \ll \mu_Y$, where $\mu_X, \mu_Y$ denote Lebesgue measures on $\Omega_X$ and $\Omega_Y$ respectively. Let $\Omega := \Omega_X \times \Omega_Y$. Let $P = d\nu_X/d\mu_X$ and $Q = d\nu_Y/d\mu_Y$ be the densities of $\nu_X, \nu_Y$ with respect to $\mu_X, \mu_Y$. For simplicity, write $dx = d\mu_X$ and $dy = d\mu_Y$. Then $P = d\nu_X/dx$ and $Q = d\nu_Y/dy$. The primal problem is to find an optimal pair of correspondence functions $(p, q)$ (or an optimal matching scheme described by product measures $(\nu^X, \nu^Y)$ defined in Eq. (20)) for the maximization problem:

$$
\sup_{L_H(p, q)} := \int_{\Omega_Y} \int_{\Omega_X} p^a q^{1-a} K dxdy
$$

subject to

$$
\int_{\Omega_Y} p(x, y)dy = P(x), \text{ and } \int_{\Omega_X} q(x, y)dx = Q(y).
$$

Since the constraints are related to the marginal densities $P, Q$, we would call them marginal constraints.
In this paper, $\alpha$ is assumed to be a scalar in $(0, 1)$. In application, we use $\alpha = 1/2$. When $\alpha = 1/2$, the cost function is a symmetric function of $p, q$, and it could be viewed as a weighted Hellinger distance $^4$.

Unfortunately, the optimal solution $(p, q)$ of the primal problem does not exist in general. Loosely speaking, the optimal functions $p, q$ usually are the Dirac delta functions. To study the primal problem, we introduce their corresponding measures $\nu^X, \nu^Y$. For any measurable set $A = A_X \times A_Y$ with $A_X \subset X, A_Y \subset Y$, let

(20) \hspace{1cm} \nu^X(A) = \int_{A_Y} \int_{A_X} p(x, y) dx \, dy, \quad \nu^Y(A) = \int_{A_Y} \int_{A_X} q(x, y) dx \, dy.

Then the primal problem becomes

(21) \hspace{1cm} \max_{\nu^X, \nu^Y} \int_{\Omega} K(x, y) \left( \frac{d\nu^X}{d\omega} \right)^\alpha \left( \frac{d\nu^Y}{d\omega} \right)^{1-\alpha} d\omega,

subject to the marginal constraints

(22) \hspace{1cm} \text{proj}_{x \in X} \nu^X = \nu_X, \quad \text{proj}_{y \in Y} \nu^Y = \nu_Y.

Here $\omega$ is some dominating measure on $X \times Y$ with $\nu^X \ll \omega, \nu^Y \ll \omega$. The cost is invariant of the selection of $\omega$, then we may choose $\omega = \nu^X + \nu^Y$. The absolute continuity comes from the fact that for each measurable set $A$, if $\omega(A) = \nu^X(A) + \nu^Y(A) = 0$ then $\nu^X(A) = 0 = \nu^Y(A)$.

The dual problem is to find an optimal pair of positive functions $(a, b)$ for the minimization problem:

(23) \hspace{1cm} \inf \{M_H(a, b) := \alpha \int_{\Omega_X} ad\nu_X + (1-\alpha) \int_{\Omega_Y} b d\nu_Y \},

(24) \hspace{1cm} \text{subject to } K \leq a^\alpha b^{1-\alpha}, \ a > 0, \ b > 0.

The duality relation between these two problems would be shown in Theorem 3.2.

The $\alpha$-divergence has been used as a measure of dissimilarity between two densities. Hence, the $\alpha$-divergence can be applied to image registration, for instance [31]. In application of our model, consider two images having intensity functions $I_X, I_Y$ and represent features of these images by some weight functions $P$ and $Q$ respectively. For instance, in terms of object contours, we may choose the norm of the gradients $P(x) = |\nabla I_X(x)|$ and $Q(y) = |\nabla I_Y(y)|$; in terms of point-sets, we may choose various corner detectors (see references in [61]), and let $P = \sum_{i=1}^m \delta_{x_i}, Q = \sum_{j=1}^n \delta_{y_j}$, where $\delta_x$ is the Dirac delta mass and $m, n$ be the number of selected points in two images. The matching scheme is described by the optimal correspondence functions $(p, q)$ solved in the primal problem. The feature similarity function $K(x, y)$ is the Gaussian kernel function $\exp(-\|x-y\|^2/\sigma^2)$, where $x$ and $y$ refer to the spacial position of feature points. In applications of image morphing, we would choose cosine functions as feature similarity functions. Then the problem can be

$^4$The Hellinger distance is a well-known distance for probability functions, see e.g., [18]. Its definition is given as follows. Given densities $p, q$ of the measures $\nu_X, \nu_Y$ with respect to a dominating measure $\gamma$ on a measurable space $\Omega$, define

(19) \hspace{1cm} d_H(\nu_X, \nu_Y) \text{ or } d_H(p, q) := \left[ \int_{\Omega} (\sqrt{p} - \sqrt{q})^2 d\gamma \right]^{1/2} = \left[ 2(1 - \int_{\Omega} \sqrt{pq} d\gamma) \right]^{1/2}.

When $K \equiv 1$ and $\nu_X(X) = \nu_Y(Y) = 1$ in the primal problem, then $\sqrt{2 - L_H(p, q)}$ is the Hellinger distance between two densities $p, q$. 
regarded as a combination of the MK problem and the elastic distance proposed by Younes\cite{36} (See Remark 4).

2.3. Connection to the MK problem. We point out two relations between the MK problem and the alpha-D problem, when $K$ is the Gaussian kernel function. First, the $L^2$ case of the MK problem is an asymptotic limit of the Hellinger distance problem.

Remark 1. Let $\Omega = \Omega_X \times \Omega_Y$. Suppose $|\Omega| < \infty$, then

\begin{equation}
\int_\Omega \|x-y\|^4 \left(\frac{d\nu^X}{d\omega}\right)^\alpha \left(\frac{d\nu^Y}{d\omega}\right)^{1-\alpha} d\omega < \infty.
\end{equation}

As $\sigma \to \infty$, using the Taylor expansion of the exponential function, we can approximate the cost in the primal problem as follows,

\begin{equation}
\int_\Omega \exp \left(-\frac{(x-y)^2}{\sigma^2}\right) \left(\frac{d\nu^X}{d\omega}\right)^\alpha \left(\frac{d\nu^Y}{d\omega}\right)^{1-\alpha} d\omega \\
= \int_\Omega \frac{d\nu^X}{d\omega} \left(\frac{d\nu^Y}{d\omega}\right)^{1-\alpha} d\omega - \frac{1}{\sigma^2} \int_\Omega \|x-y\|^2 \left(\frac{d\nu^X}{d\omega}\right)^\alpha \left(\frac{d\nu^Y}{d\omega}\right)^{1-\alpha} d\omega + O\left(\frac{1}{\sigma^4}\right).
\end{equation}

By Hölder’s inequality\textsuperscript{5}, the first term achieves its maximum when the quotient $\frac{d\nu^Y}{d\omega}/\frac{d\nu^X}{d\omega} = k$ is constant. When two images are represented by two piles of equal mass, then

\begin{equation}
\int_\Omega \frac{d\nu^Y}{d\omega} d\omega = k \int_\Omega \frac{d\nu^X}{d\omega} d\omega = k \int_{\Omega^X} d\nu_X = k \int_{\Omega^Y} d\nu_Y = k \int_\Omega \frac{d\nu^Y}{d\omega} d\omega,
\end{equation}

which implies $k = 1$. Then $\nu^X = \nu^Y =: \nu$. Afterwards, maximizing the second term is equivalent to the problem

\begin{equation}
\min_{\nu} \left\{ \int_\Omega \|x-y\|^2 d\nu \right\}, \text{ subject to } proj_{x \in \Omega^X} \nu = \nu_X, \ proj_{y \in \Omega^Y} \nu = \nu_Y.
\end{equation}

This is exactly the $L^2$-MK mass transport problem.

Secondly, the alpha-D problem with $K$ being Gaussian kernel functions can be regarded as a robust problem modified from the MK problem. Reformulate the Gaussian kernel function in Eq. (9):

\begin{equation}
p^\alpha q^{1-\alpha} \exp(-\|T(x)-y\|^2/\sigma^2) = -\min_{r \geq 0} p^\alpha q^{1-\alpha} (r \log r - r + \|T(x)-y\|^2/\sigma^2),
\end{equation}

where the minimum is attained at $r = \exp(-\|T(x)-y\|^2/\sigma^2)$. The cost in the alpha-D problem can be regarded as an $L^2$ mass transport cost $p^\alpha q^{1-\alpha} \|T(x)-y\|^2$ weighted by a positive function $r$. Different weights are assigned according to the distances $\|T(x)-y\|$: the larger distance $\|T(x)-y\|$ has the smaller weight $r$. When $\|T(x)-y\| \gg \sigma$, then $r \approx 0$ and then this pair $(x, y)$ has less influence on the estimation $T$ than other pairs. Hence, roughly speaking the alpha-D problem tends to ignore these distant pairs. In contrast, due to the measure preserving condition, the image matching result based on the MK problem is easily affected by mass changes of the corresponding objects.

\textsuperscript{5}Hölder’s inequality (see page 622\textsuperscript{14}): Let scalars $u \in L^p, v \in L^p$ and $1 < p, q < \infty$ with $p^{-1} + q^{-1} = 1$, then $\int |uv| dx \leq \|u\|_{L^p} \|v\|_{L^q}$. Let $\alpha = p^{-1}, a = |u|^{1/\alpha}, b = |v|^{1/(1-\alpha)}$, then $0 < \alpha < 1$, and $\int a^\alpha b^{1-\alpha} dx \leq (\int a dx)^{\alpha} (\int b dx)^{1-\alpha}$. Note that the equality holds if and only if the quotient $[a/b]$ or $[b/a]$ is a constant function a.e. (This condition essentially comes from Young’s inequality).
3. Theoretical discussion: Existence, uniqueness, bijection.

3.1. Primal v.s. dual problems. In the following, we establish the existence of the optimal solution in the primal problem. First, we show the existence of an optimal solution in the dual problem and illustrate the duality relation between these two problems.

We shall prove that the optimal solution \((a, b)\) in the dual problem is a pair of convex functions (thus continuous functions) in Lemma 3.4. First, in the next theorem we first verify that \((a, b)\) exists in the space of measurable functions.

**Theorem 3.1 (Existence in the Dual Problem).** Let \(\Omega_x, \Omega_y, \nu_X, \nu_Y\), and \(K\) be defined as in the primal problem. There exist a \(\nu_X\)-measurable function \(a\) on \(\Omega_x\), and a \(\nu_Y\)-measurable function \(b\) on \(\Omega_y\), such that \((a, b)\) minimizes the problem

\[
\min_{(a,b)} \alpha a \nu_X(\Omega_x) + (1 - \alpha) b \nu_Y(\Omega_y)
\]

subject to \(d \nu_a = a d \nu_X, d \nu_b = b d \nu_Y\), \(a > 0, b > 0, 0 < K \leq a^n b(y)^{1-\alpha}\) a.e. w.r.t. \(\nu_X \times \nu_Y\).

**Proof.** Clearly this cost in (31) is bounded below by 0, then we take a minimizing sequence of Radon measures \(\{(\nu_{a_n}, \nu_{b_n}) : n\} \), given by \(d \nu_{a_n} = a_n d \nu_X\), \(d \nu_{b_n} = b_n d \nu_Y\), where \((a_n, b_n)\) is subject to \(K \leq a_n(x)^n b_n(y)^{1-\alpha}\) for all \((x, y) \in \Omega\). Then \(a_n\) is a positive \(\nu_X\)-measurable function, and \(b_n\) is a positive \(\nu_Y\)-measurable function.

By the weak compactness for measures [15], there exists a subsequence still denoted by \(\{(\nu_{a_n}, \nu_{b_n}) : n \in \mathbb{N}\}\) weakly converging to Radon measures \((\nu_a, \nu_b)\). For any Borel sets \(\Delta_1 \subset \Omega_x\), \(\Delta_2 \subset \Omega_y\) with \(\nu_X(\Delta_1) = 0, \nu_Y(\Delta_2) = 0\), we have \(\nu_{a_n}(\Delta_1) = \nu_{b_n}(\Delta_2) = 0\), then \(\nu_a(\Delta_1) = \nu_b(\Delta_2) = 0\), then \(\nu_a \ll \nu_X, \nu_b \ll \nu_Y\). Thus, by the Radon-Nikodym theorem, there exist some \(\nu_X\)-measurable function \(a\) and some \(\nu_Y\)-measurable function \(b\), such that \(d \nu_a = a d \nu_X, d \nu_b = b d \nu_Y\).

Next, we show that the functions \(a, b\) satisfy the constraints. Recall that \(a_n, b_n\) are the derivatives of \(\nu_{a_n}, \nu_{b_n}\) with respect to \(\nu_X, \nu_Y\) respectively. Since the function \(\log(\cdot)\) is locally summable, by Lebesgue-Besicovitch Differentiation theorem ([15], pp. 43),

\[
\log a_n = \lim_{r \to 0} \frac{\int_{B(x,r)} \log a_n}{\nu_X(B(x,r))}, \quad \log b_n = \lim_{r \to 0} \frac{\int_{B(y,r)} \log b_n}{\nu_Y(B(y,r))},
\]

where \(B(x, r)\) is an open ball with center \(x\) and radius \(r\). Since \(\log K \leq \alpha \log a_n + (1 - \alpha) \log b_n\) for each \(n\), then for each \((x, y)\) we have

\[
0 \leq \lim_{r \to 0} \frac{1}{\nu_X(B(x,r)) \nu_Y(B(y,r))} \int_{B(y,r)} \left( \alpha \log a_n + (1 - \alpha) \log b_n - \log K \right) d \nu_X d \nu_Y
\]

\[
= \lim_{n \to \infty} \left( \alpha \log a_n + (1 - \alpha) \log b_n - \log K \right) = \alpha \log a + (1 - \alpha) \log b - \log K.
\]

Since \(K > 0\) everywhere, then \((a, b)\) satisfies the constraints \(K \leq a^n b^{1-\alpha}\), \(a > 0, b > 0\) a.e. w.r.t. \(\nu_X \times \nu_Y\). \(\square\)

**Theorem 3.2 (Weak duality).** Let \(\Omega_X, \Omega_Y, \nu_X, \nu_Y\), and \(K\) be defined as in the primal problem. The cost of the dual problem (23) gives an upper bound for the primal problem (21). When the following two conditions hold, the cost function values of two problems are equal.
1. \( \nu^X(A) = \nu^Y(A) = 0 \), where \( A := \{(x, y) \in \Omega : K(x, y) < a(x)b(y)^{1-\alpha}\} \).

2. The weighted measure preserving condition:

\[
\int_{A_X \times A_Y} a(x) d\nu^X = \int_{A_X \times A_Y} b(y) d\nu^Y
\]

for each measurable set \( A_X \times A_Y \) with \( A_X \subset \Omega_X \), \( A_Y \subset \Omega_Y \).

Proof. Since \( K(x, y) \leq a(x)b(y)^{1-\alpha} \) and both \( a, b \) are positive, then

\[
\frac{d\nu^X}{d\omega} \leq a \frac{d\nu^X}{d\omega} + (1 - \alpha) \frac{d\nu^Y}{d\omega}
\]

where equality holds if and only if

\[
\min\{\phi(\omega), \psi(\omega)\} = \phi(\omega) \quad \text{or} \quad \psi(\omega)
\]

subject to the constraint

\[
H(x, y) \leq \phi(x) + \psi(y)
\]

Next, we shall examine the uniqueness of \((a, b)\).

**Theorem 3.3. [Uniqueness]**

The optimal solution \((\hat{\phi}, \hat{\psi})\) in Eq. (39) is unique, in the sense that if \((\phi, \psi)\) and \((\hat{\phi}, \hat{\psi})\) are both optimal solutions, then \(\phi = \hat{\phi}\) a.e. w.r.t. \(\nu^X\), and \(\psi = \hat{\psi}\) a.e. w.r.t. \(\nu^Y\). Thus, the optimal solution \((a, b)\) in the dual problem is unique a.e. w.r.t. \(\nu^X \times \nu^Y\).

---

6Young’s inequality(see page 622[14]): Let scalars \(x > 0, y > 0\) and \(1 < p, q < \infty\) with \(p^{-1} + q^{-1} = 1\), then \(xy \leq p^{-1}x^p + q^{-1}y^q\). Let \(\alpha = p^{-1}, a = x^{1/\alpha}, b = y^{1/(1-\alpha)}\), then \(0 < \alpha < 1\), and \(a^\alpha b^{1-\alpha} \leq \alpha a + (1 - \alpha)b\). By calculus, the equality holds if and only if \(a = b\).
Proof. Let \((\phi, \psi), (\hat{\phi}, \hat{\psi})\) be two different optimal solutions with
\[
M_H(\exp(\phi/\alpha), \exp(\psi/(1-\alpha))) = M_H(\exp(\hat{\phi}/\alpha), \exp(\hat{\psi}/(1-\alpha))).
\]
Since the exponential function is strictly convex, then \(((\phi + \hat{\phi})/2, (\psi + \hat{\psi})/2)\) also satisfies the constraint in Eq. (39), and has a lower cost. Here “two different solutions” means that their difference is not zero a.e. with respect to \(\nu_X, \nu_Y\).

Remark 2. In general, the uniqueness with respect to the measure \(\nu_X\) or \(\nu_Y\) does not imply the uniqueness with respect to Lebesgue measure. For instance, consider any set \(A_X \times A_Y \subset \Omega\) where \(A_X \subset \Omega_X\) with \(\nu_X(A_X) = 0\), and \(A_Y \subset \Omega_Y\) with \(\nu_Y(A_Y) = 0\), then the optimal solution \((a, b)\) can take arbitrary values on \(A_X \times A_Y\) as long as \((a, b)\) satisfies \(a^\alpha b^{1-\alpha} \geq K\), i.e., \((a, b)\) is not unique with respect to Lebesgue measure.

3.2.2. Convexity of \(\phi, \psi, a, b\). In the next two subsections, we would show that under the following conditions, an optimal matching scheme \((\nu^X, \nu^Y)\) exists uniquely:

**Assumption 1.** Let \(\Omega_X, \Omega_Y, X, Y, \nu_X, \nu_Y, \mu_X, \mu_Y\) and \(K\) be defined as in the primal problem. Define \(H\) as in Eq. (38).

- (Convexity condition) \(H(x, y)\) is convex in \(x\), and in \(y\).
- (Bijection condition) Fixing \(x \in X\), \(D_x H(x, y)\) is a bijection of \(y\), and fixing \(y\), \(D_y H(x, y)\) is a bijection of \(x\).
- (Absolute continuity condition) \(\nu_X \ll \mu_X\) and \(\nu_Y \ll \mu_Y\).

Here, the gradients \(D_x H := (\frac{\partial H}{\partial x_1}, \ldots, \frac{\partial H}{\partial x_d})\) for each \(x = (x_1, \ldots, x_d) \in X\) and the gradients \(D_y H := (\frac{\partial H}{\partial y_1}, \ldots, \frac{\partial H}{\partial y_d})\) for each \(y = (y_1, \ldots, y_d) \in Y\).

The convexity condition can be replaced by a weaker condition (see Remark 5 in appendix): for all \((x, y) \in \Omega_X \times \Omega_Y\),
\[
(40) \quad \text{the second derivatives of } H \text{ have a uniform bound.}
\]
The first two conditions make the function \(H(x, y)\) act like \(x \cdot y\) in the constraint Eq. (17) in the MK problem. The third condition is the absolute continuity condition used in the MK problem.

Before we show the construction of \((\nu^X, \nu^Y)\), we need to establish the convexity of \((\phi, \psi)\) first. Let us derive a necessary optimal condition for \((\phi, \psi)\) in the problem (39): \((\phi, \psi)\) must be a pair of convex conjugate functions with respect to \(\nu_X, \nu_Y\), i.e.,
\[
(41) \quad \psi(y) = \sup_{x \in X} (H(x, y) - \phi(x)), \quad \nu_Y\text{-a.e., } \phi(x) = \sup_{y \in Y} (H(x, y) - \psi(y)), \quad \nu_X\text{-a.e.}
\]
More precisely, let
\[
(42) \quad \phi^*(y) = \sup_{x \in X} (H(x, y) - \phi(x)), \quad \phi^{**}(x) = \sup_{y \in Y} (H(x, y) - \phi^*(y)),
\]
\[
(43) \quad \psi^*(x) = \sup_{y \in Y} (H(x, y) - \psi(y)), \quad \psi^{**}(y) = \sup_{x \in X} (H(x, y) - \psi^*(x)).
\]
Note that the exponential functions are increasing. If \((\phi, \psi)\) does not satisfy Eq. (41), then either \((\phi^*, \phi^{**})\) or \((\psi^*, \psi^{**})\) will improve the cost \(M\).

The mathematical property of \((\phi, \psi)\) depends on the choice of \(H(x, y)\). Assume that \(H(x, y)\) is both convex in \(x\) and \(y\) at this moment. We could make a weaker assumption (see Remark 5), but adopting the convexity assumption would simplify the discussion.
Lemma 3.4. Assume the convexity condition on $H(x,y)$. Then in the problem (39), the optimal functions $\phi(x)$ and $\psi(y)$ are convex with respect to $\nu_X, \nu_Y$. Then we may find a pair of convex functions $\phi, \psi$ both defined on $\mathbb{R}^d$ and $\phi = \phi \nu_X$-a.e. and $\psi = \psi \nu_Y$-a.e. Therefore, the minimizer $(a,b)$ of the dual problem, $a = \exp(\phi/\alpha), b = \exp(\psi/(1-\alpha))$, is a pair of convex functions with respect to $\nu_X$ and $\nu_Y$.

Proof. The optimal function $\psi$ satisfies $\psi(y) = \sup_{x \in X} (H(x,y) - \phi(x)), \nu_Y$-a.e.. If $\psi(y) = \sup_{x \in X} (H(x,y) - \phi(x))$ for some $y = y_1$ and $y_2$ in $Y$, then

\begin{align}
\psi(y_1) + \psi(y_2) & \geq \sup_{x \in X} (H(x,y_1) + H(x,y_2) - 2\phi(x)) \\
& \geq 2(\sup_{x \in X} H(x,(y_1+y_2)/2) - \phi(x)) = 2\psi((y_1+y_2)/2).
\end{align}

Hence, $\psi$ is convex with respect to $\nu_Y$. Similarly, $\phi$ is convex with respect to $\nu_X$. Since the exponential function is increasing and convex, and $\phi$ is convex, then $a$ is convex. Similar arguments apply to the convexity of $b$. \hfill \Box

The first optimal condition in Theorem 3.2 says that the masses of $(\nu_X, \nu_Y)$ should be concentrated on the set $\{(x,y) : K(x,y) = a(x)^\alpha b(y)^{1-\alpha}, \text{i.e., } H(x,y) = \phi(x) + \psi(y)\}$. According to Theorem 3.3, this set is uniquely determined with respect to $\nu_X \times \nu_Y$. This set would be called the set of matching pairs as follows.

Definition 3.5. Let $X$, $Y$, and $K$ be defined as in the primal problem and $(a,b)$ be an optimal solution of the dual problem. Let $(\phi, \psi, H)$ be computed from Eq. (38). We call any pair $(x,y) \in X \times Y$ a matching pair if

$$H(x,y) = \phi(x) + \psi(y),$$

or equivalently $K(x,y) = a(x)^\alpha b(y)^{1-\alpha}$. The matching relation $s : X \rightarrow Y$ is defined as the set of all the matching pairs $(x,y)$, i.e., $s(x) = \{y : (x,y)$ is a matching pair $\}$ for each $x \in X$.

Next, we would investigate the structure of the set of matching pairs and its associated matching relation. In convex analysis (e.g., [40]), given a convex function $f : \mathbb{R}^d \rightarrow \mathbb{R}$, a necessary and sufficient condition for unconstrained optimality of $x$ is $0 \in \partial f(x)$. Here we say that $y \in \mathbb{R}^d$ is a sub-gradient of $f$ at $x$ if $f(z) \geq y \cdot (z-x) + f(x)$ for all $z \in \mathbb{R}^d$. A sub-gradient $\partial f(x)$ exists whenever the convex function $f$ is finite in a neighborhood of $x$. Differentiability of $f$ at $x$ is equivalent to the existence of a unique sub-gradient $y$, in which case $y = \nabla f(x)$. It is known that differentiability of a convex function $f$ fails only on a Borel set of dimension of $d - 1$, thus the gradient of $f$ exists a.e. ([1] or see chapter 25 in [40]).

According to Lemma 3.4, an optimal solution of the problem (39) is given by a pair of convex functions $\phi, \psi$ defined on $X$ and $Y$. These two functions are continuous, and their sub-gradients both exist and form closed and convex sets. From Eq. (41), each matching pair $(x,y)$ satisfies

$$D_x H(x,y) \in \partial \phi(x), \text{ and } D_y H(x,y) \in \partial \psi(y).$$

Since the gradients of the optimal functions $\phi, \psi$ exist a.e. with respect to Lebesgue measure, then each matching pair $(x,y)$ satisfies

$$D_x H(x,y) = \nabla \phi(x) \text{ a.e., and } D_y H(x,y) = \nabla \psi(y) \text{ a.e.}$$

Now, we are ready to examine the bijection of the matching relation between $X$ and $Y$. Introduce notations $X_\phi := \{x \in X : \nabla \phi(x) \text{ exists}\}$ and $Y_\psi := \{y \in Y : \nabla \psi(y) \text{ exists}\}$.
\[ \nabla \psi(y) \text{ exists } \}. \text{ Let } \\
(49) \quad h_x(y) = D_x H(x, y) \text{ for each } x \in X, \quad \hat{h}_y(x) = D_y H(x, y) \text{ for each } y \in Y. \\

**Theorem 3.6.** Let \((\phi, \psi)\) be an optimal solution in the dual problem \((39)\). Let \\
h_x, \hat{h}_y, X_\phi, \text{ and } Y_\psi \text{ defined as above. Let } s^{-1}(Y_\psi) \text{ be the inverse image of } Y_\psi \\
under the relation } s \text{ in Definition 3.5, i.e.,} \\
(50) \quad s^{-1}(Y_\psi) = \{ x \in X : (x, y) \in s, y \in Y_\psi \} = \{ x : D_y H(x, y) = \nabla \psi(y), y \in Y_\psi \}. \\
Assume the convexity condition on } H. \text{ Although } s \text{ is not necessary a function, } s \\
\text{ and its inverse relation } s^{-1} \text{ have the following properties.} \\
1. \text{ If } h_x \text{ is bijective, then } s \text{ restricted to } X_\phi \text{ is a function. Likewise, if } \hat{h}_y \text{ is} \\
\text{ bijective, then } s^{-1} \text{ restricted to } Y_\psi \text{ is a function.} \\
2. \text{ Assume that both } h_x, \hat{h}_y \text{ are bijective. Then the range of } s \text{ is } Y \\
\text{ and the range of } s^{-1} \text{ is } X. \text{ Besides, } s \text{ restricted to the set } s^{-1}(Y_\psi) \cap X_\phi \text{ is} \\
\text{ injective and onto } Y_\psi \cap s(X_\phi). \\

**Proof.** For the first statement, let \(x_0 \in X_\phi\), then \(\nabla \phi(x_0) \text{ exists }\). Let \(h_{x_0}(y) = D_x H(x_0, y)\), then from Eq. \((48)\), \(s(x_0) = \{ y : h_{x_0}(y) = \nabla \phi(x_0) \}\). Since \(h_{x_0}\) is bijective, then the set \(s(x_0)\) has exactly one element. Thus \(s \text{ restricted to } X_\phi \) is a function. \(\Box\) \\

For the second statement, let \(y_0\) be any element in \(Y\). We need to show that there exists some \(x_0 \in X\), such that \(y_0 \in s(x_0)\), or equivalently from Eq. \((47)\), \\
\(\hat{h}_{y_0}(x_0) \in \nabla \psi(y_0)\). Since \(\hat{h}_{y_0}\) is surjective, then the existence of \(x_0\) is ensured. For the second part, suppose that for some \(y_0 \in Y_\psi\), there exist \(x_0, \hat{x}_0 \in X_\phi\) with \(y_0 \in s(x_0) \cap s(\hat{x}_0)\). Then \(y = s(x_0) = s(\hat{x}_0)\). And from Eq. \((48)\), \\
\[ \nabla \psi(y_0) = \hat{h}_{y_0}(x_0) = \hat{h}_{y_0}(\hat{x}_0). \]
Since \(\hat{h}_{y_0}\) is injective, then \(x_0 = \hat{x}_0\), which implies that \(s \text{ restricted to } s^{-1}(Y_\psi) \cap X_\phi \) is an injection. Moreover, since \(s^{-1}(Y_\psi \cap s(X_\phi)) = s^{-1}(Y_\psi) \cap X_\phi\), then the surjective part is verified. \(\Box\) \\

In some cases, the injectivity in \(h_x\) and \(\hat{h}_y\) is not easy to verify, we provide one sufficient condition based on the second derivatives of \(H\), see Lemma D.1 in appendix.

### 3.3. No duality gap

We are ready to verify no duality gap between the primal problem \((21)\) and the dual problem \((23)\). In this subsection, we assume the absolute continuity condition: \(\nu_X << \mu_X\) and \(\nu_Y << \mu_Y\). As mentioned in Theorem 3.2, it suffices to justify the two conditions in Eq. \((37)\). Here, we would present the following weighted measure-preserving property, which is related to the second condition.

**Theorem 3.7** (Weighted measure-preserving property). Consider the dual problem \((23)\). Let \(\nu_X, \mu_X, \nu_Y, \mu_Y, \Omega_X\) and \(\Omega_Y\) be defined as in the primal problem. Suppose that both \(h_x, \hat{h}_y\) defined in Eq. \((49)\) are bijections. Let \((a^*, b^*)\) be a solution of the dual problem. Denote measures \(d\nu_a = a^*d\nu_X, d\nu_b = b^*d\nu_Y\) with \(\nu_X << \mu_X\) and \(\nu_Y << \mu_Y\), then \(s \in \text{Definition 3.5}\) is a measure preserving map between \\
\((\Omega_X, \nu_a)\) and \((\Omega_Y, \nu_b)\), that is, \\
\[ \int_{\Omega_X} g(s(x))d\nu_a = \int_{\Omega_Y} g(y)d\nu_b, \]
for any \(g \in C(\Omega_Y) \cap L^\infty(\Omega_Y)\).
Proof. This proof is based on the approach used in [13,17]. Recall the dual problem minimizing the cost
\[ M(a, b) := \alpha \int_{\Omega_X} a(x) d\nu_X(x) + (1 - \alpha) \int_{\Omega_Y} b(y) d\nu_Y(y), \]
subject to \( K \leq a^\alpha b^{1-\alpha}, \)
and \( a > 0, b > 0. \) For a function \( g \in C(\Omega_Y) \cap L^\infty(\Omega_Y) \), fixing \( \tau > 0 \) with \( \tau < \|g\|_{L^\infty}^{-1}, \)
define the variations
\[ b_\tau(y) := b^*(y)(1 - \tau g(y)), \]
and
\[ a_\tau(x) := \max_{y \in \Omega_Y} \{ (K(x, y)/b_\tau(y))^{1-\alpha} \}. \]
Then \( M(a^*, b^*) \leq M(a_\tau, b_\tau) =: i(\tau). \) As the mapping \( \tau \to i(\tau) \) has a minimum at \( \tau = 0, \)
\[ 0 \leq \frac{M(a_\tau, b_\tau) - M(a^*, b^*)}{\tau} = \alpha \int_{\Omega_X} \frac{a_\tau(x) - a^*(x)}{\tau} d\nu_X + (1 - \alpha) \int_{\Omega_Y} \frac{b_\tau(y) - b^*(y)}{\tau} d\nu_Y. \]
For each \( x \in \Omega_X \cap X_\phi, \) if we choose \( y_\tau \in \Omega_Y \) to be one of maximizer in Eq. (55),
then \( a_\tau(x) = (K(x, y_\tau)/b_\tau(y_\tau))^{1-\alpha}/a^* \). Since \( (a^*, b^*) \) is a minimizer satisfying the constraint \( (K(x, y_\tau)/b^*(y_\tau))^{1-\alpha}/a^* \leq a^*(x) \),
then from Eq. (54,55)
\[ a_\tau(x) - a^*(x) = \frac{K(x, y_\tau)^{1/\alpha}}{(b^*(y_\tau)(1 - \tau g(y_\tau)))^{1-\alpha}/a^* - a^*(x) \leq a^*(x)(1 - \tau g(y_\tau))^{1-\alpha} - 1). \]
On the other hand, if we take \( y \in \Omega_Y \) with \( a^*(x)^\alpha = K(x, y)/b^*(y)^{1-\alpha}, \)
then from Eq. (54,55),
\[ a_\tau(x) - a^*(x) \geq \frac{K(x, y)^{1/\alpha}}{b^*(y)^{1-1/\alpha}} - a^*(x) = a^*(x)(1 - \tau g(y))^1 - 1 - 1). \]
Hence, we have
\[ ((1 - \tau g(y))^{1-1/\alpha} - 1)a^*(x) \leq a_\tau(x) - a^*(x) \leq ((1 - \tau g(y))^{1-1/\alpha} - 1)a^*(x). \]
By Theorem 3.6, \( s \) restricted to \( X_\phi \cap \Omega_X \) is a function. Then we have \( y = s(x) \). Also
as \( \tau \to 0, y_\tau \to s(x) \) and then by l’Hôpital’s rule, both sides in Eq. (59) converge
to the limit \((\alpha^{-1} - 1)g(s(x))a^*(x) \). Finally, applying the dominated convergence theorem to Eq. (56), we have
\[ \int_{\Omega_X} a^*(x)g(s(x)) d\nu_X = \int_{\Omega_X \cap X_\phi} a^*(x)g(s(x)) d\nu_X \geq \int_{\Omega_Y} b^*(y)g(y) d\nu_Y. \]
Replacing \( g \) by \(-g\), we deduce that equality holds. \( \square \)

Hence, although generally \( s \) is not a measure preserving mapping with respect to the measures \( \nu_X, \nu_Y, s \) preserves the measure preserving property with respect to the weighted measures \( \nu_\alpha, \nu_\beta. \)
Corollary 1. Here is another formulation of the weighted measure property. Assume that the absolute continuity condition holds. Under the assumption in Theorem 3.6, for each measurable set \( A \subset Y \),
\[
\nu_a(\{x : s(x) \in A\}) = \nu_b(\{y : y \in A\}).
\]

Proof. Consider a measurable set \( A \subset \mathbb{R}^d \) and its indicator function \( g_0(z) = 1 \) if \( z \in A \), \( g_0(z) = 0 \) otherwise. Let \( g_\epsilon(y) := \int_A \eta_\epsilon(y-z) g_0(z)dz \), where the standard mollifier \( \eta_\epsilon(y) \) is \( \epsilon^{-d} \eta(x/\epsilon) \) with
\[
\eta(y) = \begin{cases} 
  c \exp(1/(\|y\|^2-1)) & \text{if } |y| < 1, \\
  0 & \text{if } |y| \geq 1,
\end{cases}
\]
and the constant \( c \) adjusted so
\[
\int_{\mathbb{R}^d} \eta(y)dy = 1.
\]
Then as \( \epsilon \to 0 \), we have \( g_\epsilon(y) \in C^\infty(\Omega_Y) \) and \( g_\epsilon \to g_0 \) a.e. with respect to Lebesgue measure. From the above theorem, we have
\[
\int_X a^*(x) g_\epsilon(s(x))d\nu_X = \int_Y b^*(y) g_\epsilon(y)d\nu_Y,
\]
then as \( \epsilon \to 0 \),
\[
\int_X a^*(x) g_0(s(x))d\nu_X = \int_Y b^*(y) g_0(y)d\nu_Y.
\]
That is, for each measurable set \( A \subset Y \), we have
\[
\nu_a(\{x : s(x) \in A\}) = \nu_b(\{y : y \in A\}).
\]

Corollary 2. Assume that the absolute continuity condition holds. Under the assumption in Theorem 3.6, if both \( h_x \) and \( h_\phi \) are bijective from \( \mathbb{R}^d \) to \( \mathbb{R}^d \), then \( \nu_X(X_\phi \cap s^{-1}(Y_\psi)) = \nu_X(X_\phi) = \nu_X(X) \) and \( \nu_Y(Y_\phi \cap s(X_\phi)) = \nu_Y(Y) \). Hence, \( s \) is a bijection from \( X \) to \( Y \) a.e. with respect to \( \nu_X \) and \( \nu_Y \) respectively.

Proof. Based on the absolute continuity condition, \( \nu_X(X - X_\phi) = 0 = \nu_Y(Y - Y_\psi) \). From the definition of \( s^{-1}(Y_\psi) \) and the second statement in Theorem 3.6, \( s \) restricted to \( s^{-1}(Y_\psi) \cap X_\phi \) is injective and onto \( Y_\phi \cap s(X_\phi) \). Since \( \nu_Y(Y - Y_\psi) = 0 \), then from Corollary 1,
\[
0 = \nu_\phi(Y - Y_\psi) = \nu_\phi(s^{-1}(Y - Y_\psi)) = \nu_X(X - s^{-1}(Y_\psi)).
\]
Since \( a > 0 \) and \( \nu_X(X - X_\phi) = 0 \), then
\[
0 = \nu_X(X - s^{-1}(Y_\psi)) \geq \nu_X(X_\phi \cap (X - s^{-1}(Y_\psi))) = \nu_X(X_\phi - X_\phi \cap s^{-1}(Y_\psi)) \geq 0,
\]
which implies \( \nu_X(X_\phi \cap s^{-1}(Y_\psi)) = \nu_X(X_\phi) = \nu_X(X) \). Likewise, we can show that \( \nu_Y(Y_\psi \cap s(X_\phi)) = \nu_Y(Y) \). Done.

Now let us present the construction of an optimal matching scheme \((\nu_X^*, \nu_Y^*)\) in the primal problem. We have shown the existence and the uniqueness of \((a, b)\) and \((\phi, \psi)\) in the dual problems. Recall that from Theorem 3.2, \((\nu_X^*, \nu_Y^*)\) is optimal if and only if \((\nu_X^*, \nu_Y^*, a, b)\) satisfies those two conditions. According to the first condition in Theorem 3.2, the support of \((\nu_X^*, \nu_Y^*)\) must lie in the set of matching pairs \( \{x, y : H(x, y) = \phi(x) + \psi(y)\} \). From Theorem 3.6, this condition leads to the injectivity of \( s \) restricted to \( X_\phi \cap s^{-1}(Y_\psi) \). Since \((\nu_X^*, \nu_Y^*)\) must satisfy the marginal
constraint Eq. (22), then we define \((\nu^X, \nu^Y)\) as follows: Consider measurable sets \(A_X \subset X, A_Y \subset Y\). Then

\[
\nu^X(A_X \times A_Y) := \nu^X((x, y) \in (A_X \cap X_\phi) \times (A_Y \cap Y_\phi) : \phi(x) + \psi(y) = H(x, y))
\]

\[
= \nu_X((A_X \cap X_\phi) \cap s^{-1}(A_Y \cap Y_\phi))
\]

and likewise

\[
\nu^Y(A_X \times A_Y) := \nu_Y(s(A_X \cap X_\phi) \cap (A_Y \cap Y_\phi)).
\]

From Eq. (69),

\[
\int_{A_X \times A_Y} a d\nu^X = \nu_a(\{x : s(x) \in s(A_X \cap X_\phi) \cap (A_Y \cap Y_\phi)\}),
\]

and

\[
\int_{A_X \times A_Y} b d\nu^Y = \nu_b(\{y : y \in s(A_X \cap X_\phi) \cap (A_Y \cap Y_\phi)\}).
\]

By Corollary 2, \(\int_{A_X \times A_Y} a d\nu^X = \int_{A_X \times A_Y} b d\nu^Y\) for each measurable set \(A_X \times A_Y\), which verifies the second condition in Theorem 3.2. Therefore, no duality gap exists between the primal problem and the dual problem.

Now let us examine the uniqueness of \((\nu^X, \nu^Y)\). From Eq. (69), it suffices to show that the set \(\{(x, y) \in (A_X \times A_Y) \cap (X_\phi \times Y_\phi) : \phi(x) + \psi(y) = H(x, y)\}\) is uniquely determined with respect to \(\nu_X \times \nu_Y\). From Theorem 3.3, the optimal solution \((\phi, \psi)\) in the problem (39) is unique with respect to \(\nu_X, \nu_Y\) respectively. Hence, together with the absolute continuity condition \(\nu_X << \mu_X, \nu_Y << \mu_Y\), we deduce that both \(X_\phi \times Y_\psi\) and the set of the matching pairs are uniquely determined with respect to \(\nu_X \times \nu_Y\).

In summary, we have the uniqueness result:

**Theorem 3.8.** Suppose that Assumption 1 holds. Then the optimal matching scheme described by \((\nu^X, \nu^Y)\) in the primal problem is unique.

In general, the non-uniqueness of the optimal matching scheme \((\nu^X, \nu^Y)\) is caused by two factors: (a) \(H\) does not satisfy the bijection condition; (b) \(\nu_X, \nu_Y\) do not satisfy the absolute continuity condition, i.e., both \(\nu_X, \mu_X\) are mutually singular on some subset in \(X\) and \(\nu_Y, \mu_Y\) are mutually singular on some subset in \(Y\). For case (a), we provide a simple example: Let \(K(x, y)\) be a constant function 1 and \(X = Y = \mathbb{R}^d = \mathbb{R}^1, spt(\nu_X) = spt(\nu_Y) = [1, 2]\) with

\[
P(x) = Q(y) = 1 \text{ for } 1 \leq x \leq 2, 1 \leq y \leq 2, P(x) = 0 = Q(y), \text{ otherwise.}
\]

Since

\[
\int_{\Omega_Y} \int_{\Omega_X} p^a q^{1-a} dxdy \leq \int_{\Omega_Y} \int_{\Omega_X} apdxdy + \int_{\Omega_Y} \int_{\Omega_X} (1 - a)qdxdy
\]

then the first equality holds when \(p = q\). Thus every pair of correspondence functions \((p, q)\) satisfying the marginal constraints with \(p = q\) has the same cost. For case (b), one example is given as follows. Let \(d = 2\), and let the support \(\Omega_X = \{(x_1, x_1) : 1 \leq x_1 \leq 2\}\) and the support \(\Omega_Y = \{(y_1, 3 - y_1) : 1 \leq y_1 \leq 2\}\). Then for any fixed \(x \in \Omega_X\), \(K(x, y) = \exp(x \cdot y)\) is constant for each \(y \in \Omega_Y\), which
implies, any matching scheme \((\nu^X, \nu^Y)\) satisfying the marginal constraints is an optimal solution.

**Example 1 (Gaussian similarity function).** Assume the absolute continuity condition. Consider the log-similarity function \(H(x, y) = x \cdot y\). Clearly, \(H\) is convex in both \(x\) and \(y\). Also both \(D_xH = y, D_yH = x\) are bijections. Thus, the optimal matching relation \(s\) is a bijection almost everywhere. Also, Eq. (41) becomes the Legendre-Fenchel transform, and \(\{\phi(x), \psi(y)\}\) is a convex conjugate pair.

Consider the problem with the log-similarity function \(H = 2x \cdot y / \sigma^2\), marginal densities \((P, Q)\), and correspondence functions \((p, q)\). Clearly, the matching relation \(s\) is bijective a.e.. This problem is equivalent to the following problem with Gaussian kernel functions \(K(x, y) = \exp(-||x - y||^2 / \sigma^2)\), marginal densities \((\hat{P}, \hat{Q})\) and correspondence functions \((\hat{p}, \hat{q})\), provided that \((\hat{p}, \hat{q}), (P, Q)\) are chosen as follows. Let \(\hat{p} = p \exp(-||x||^2 / \alpha \sigma^2)\), \(\hat{q} = q \exp(-||y||^2 / (1 - \alpha) \sigma^2)\), and

\[
\begin{align*}
\hat{P}(x) &= \int_Y \hat{q} dy = P(x) \exp(-\frac{||x||^2}{\alpha \sigma^2}), \\
\hat{Q}(y) &= \int_X \hat{p} dx = Q(y) \exp(-\frac{||y||^2}{(1 - \alpha) \sigma^2}).
\end{align*}
\]

Then these two problems have the same cost value,

\[
\int_Y \int_X p^\alpha q^{1-\alpha} \exp(-\frac{||x - y||^2}{\sigma^2}) dx dy = \int_Y \int_X \hat{p}^\alpha \hat{q}^{1-\alpha} \exp(\frac{2x \cdot y}{\sigma^2}) dx dy.
\]

Therefore, the optimal matching relation \(s\) in the problem with Gaussian kernel functions is also bijective a.e..

4. Algorithms (discrete version). Here, we shall describe the numerical algorithm to compute the optimal matching scheme. When the alpha-D problem is applied to match objects, we take the following discretization. Each object is represented by one particle-set, which consists of a set of points at location \(\{x_i \in \mathbb{R}^d: i = 1, \ldots, m\}\) with weight (mass) \(\{\lambda_i > 0 : i = 1, \ldots, m\}\). In application, objects represented by weights would depend on user’s particular registration problem. For instance, weight could refer to area/length/the cardinality of points in images/shapes/pixel-sets matching problems respectively.

Here we shall focus on point-sets matching problems. A unit weight is assigned to each point and the following particle-sets matching problem is examined: Given two sets of particles located at \(\{x_i \in \mathbb{R}^d: i = 1, \ldots, m\}\) and \(\{y_j \in \mathbb{R}^d: j = 1, \ldots, n\}\) with weight (mass) \(\{\lambda_i^+: i = 1, \ldots, m\}, \{\lambda_j^-: j = 1, \ldots, n\}\) respectively, we aim to find a pair of optimal correspondence matrices \((\Lambda^+, \Lambda^-)\), whose entries are nonnegative, i.e., \(\lambda_{i,j}^+ := (\Lambda^+)_{i,j} \geq 0, \lambda_{i,j}^- := (\Lambda^-)_{i,j} \geq 0\)

\[
\begin{align*}
\{ & \text{to maximize } \sum_{i=1}^m \sum_{j=1}^n (\lambda_{i,j}^+)^\alpha (\lambda_{i,j}^-)^{1-\alpha} K(x_i, y_j), \\
& \text{subject to the marginal constraints } \sum_{j=1}^n \lambda_{i,j}^+ = \lambda_i^+, \sum_{i=1}^m \lambda_{i,j}^- = \lambda_j^- \text{ for each } i, j. 
\end{align*}
\]

As mentioned in section 2, \(\alpha\) is a scalar in \((0, 1)\), and in application, \(\alpha = 1/2\) is taken. The matching structure is described by a pair of \(m \times n\) matrices \(\Lambda^+\) and \(\Lambda^-\). Each nonzero pair \((\lambda_{i,j}^+, \lambda_{i,j}^-)\) indicates one matching link between points \(x_i\) and \(y_j\), and the pair \((x_i, y_j)\) is called a matching pair.

The problem in Eq. (77) can be regarded as the discrete counterpart of the primal problem given in Eq. (18): Consider \(P = \sum_{i=1}^m \lambda_i^+ \delta_{x_i}\), and \(Q = \sum_{j=1}^n \lambda_j^- \delta_{y_j}\), where \(\delta_x\) is the Dirac point mass. Then a pair of \((p, q)\) having the form \(p = \sum_{i,j} \lambda_{i,j}^+ \delta(x_i, y_j)\) and \(q = \sum_{i,j} \lambda_{i,j}^- \delta(x_i, y_j)\) is sought to maximize the problem in Eq. (18).
For the sake of simplicity, perform the normalization:

\[(78) \quad K_{i,j} := (\lambda^+_{i,j})^a (\lambda^-_{i,j})^{1-a} K(x_i, y_j), \gamma^+_{i,j} := \frac{\lambda^+_{i,j}}{\lambda^+_{i,j}}, \gamma^-_{i,j} := \frac{\lambda^-_{i,j}}{\lambda^-_{i,j}}.\]

Let the set of row/column stochastic matrices be

\[(79) \quad D^+ := \{\Gamma^+ \in \mathbb{R}^{m \times n} : \sum_{j=1}^n \gamma^+_{i,j} = 1, \gamma^+_{i,j} \geq 0\}, \quad D^- := \{\Gamma^- : \sum_{i=1}^m \gamma^-_{i,j} = 1, \gamma^-_{i,j} \geq 0\},\]

then we have an equivalent problem: Find an optimal scheme characterized by two matrices, \(\Gamma^+, \Gamma^-\) having entries \(\{\gamma^+_{i,j}, \gamma^-_{i,j} : 1 \leq i \leq m, 1 \leq j \leq n\}\), which

\[(80) \quad \left\{ \begin{array}{l}
\text{maximizes } L(\Gamma^+, \Gamma^-) := 2 \sum_{i=1}^m \sum_{j=1}^n (\gamma^+_{i,j})^a (\gamma^-_{i,j})^{1-a} K_{i,j}, \\
\text{subject to } \Gamma^+ \in D^+, \Gamma^- \in D^-.
\end{array} \right.\]

The normalization decouples \(\{\lambda^+_{i,j} : i\}, \{\lambda^-_{i,j} : j\}\) from the marginal constraints of the original matching problem. In the following, we would design an algorithm to compute the optimal solution \(\Gamma^+, \Gamma^-\) to Eq. \((80)\). Once \((\Gamma^+, \Gamma^-)\) is obtained, the optimal solution \((\Lambda^+, \Lambda^-)\) to Eq. \((77)\) can be computed from \((\Gamma^+, \Gamma^-)\) and \(\{\lambda^+_{i,j} : i\}, \{\lambda^-_{i,j} : j\}\), i.e., \(\lambda^+_{i,j} = \gamma^+_{i,j} \lambda^+_i, \lambda^-_{i,j} = \gamma^-_{i,j} \lambda^-_j\).

### 4.1. Algorithms

**Algorithm 1.** Given an \(m \times n\) matrix \(K\) with positive entries \(K_{i,j} > 0 : i, j\), then an optimal solution \((\Gamma^+, \Gamma^-)\) to Eq. \((80)\) can be computed by the following iterative rule.

Initialization: start with \((\Gamma^+(0), \Gamma^-(0))\) having nonzero entries \(\gamma^+_{i,j}(0), \gamma^-_{i,j}(0) : i, j\).

Generate a sequence of matrices \(\{\Gamma^+(l), \Gamma^-(l) : l \in \mathbb{N}\}\) (each of them has entries \(\gamma^+_{i,j}(l), \gamma^-_{i,j}(l) : i, j\) ) according to the rule:

\[\begin{align*}
\gamma^+_{i,j}(l+1) & := \frac{\gamma^+_{i,j}(l)}{\sum_{i=1}^m \gamma^+_{i,j}(l) K^2_{i,j}} (\sum_{i=1}^m \gamma^+_{i,j}(l) K^2_{i,j}), \\
\gamma^-_{i,j}(l+1) & := \frac{\gamma^-_{i,j}(l+1)}{\sum_{j=1}^n \gamma^-_{i,j}(l+1) K^2_{i,j}} (\sum_{j=1}^n \gamma^-_{i,j}(l+1) K^2_{i,j}).
\end{align*}\]

Then every limit point \((\Gamma^+, \Gamma^-)\) is a stationary point and in fact a maximizer of \(L\).

In this algorithm, clearly each column sum of \(\Gamma^-(l)\) is one, and each row sum of \(\Gamma^+(l)\) is one. To get a maximizer, it is crucial that the initial matrix \(\Gamma^+(0)\) is a positive matrix, which yields that \(\Gamma^+(l), \Gamma^-(l)\) are positive matrices for all \(l\).\(^7\)

Before we show the convergence of the algorithm, we need to study this problem in Eq. \((80)\) further.

**Lemma 4.1.** The cost function \(L\) in Eq. \((80)\) is concave, and the optimal solutions form a convex set. Given two optimal solutions \((\Gamma^+_1, \Gamma^-_1)\) with entries \(\gamma^+_{i,j,1}, \gamma^-_{i,j,1} : i, j\) and \((\Gamma^+_2, \Gamma^-_2)\) with entries \(\gamma^+_{i,j,2}, \gamma^-_{i,j,2} : i, j\), then

\[(81) \quad \gamma^+_{i,j,1} \gamma^-_{i,j,1} = \gamma^+_{i,j,2} \gamma^-_{i,j,2} \text{ for all } (i,j).\]

\(^7\) Otherwise, the algorithm might provide a stationary point which is not a maximizer. For instance, \(\Gamma^+(0)\) is the identity matrix then \(\Gamma^+(l) = \Gamma^+(0)\) for all \(l\), regardless of the selection \(K_{i,j}\).
Proof. It suffices to show the concavity of \( r^\alpha s^{1-\alpha} \), i.e.,
\[
(82) \quad r_1^\alpha s_1^{1-\alpha} + r_2^\alpha s_2^{1-\alpha} \leq 2 \left( \frac{r_1 + r_2}{2} \right)^\alpha \left( \frac{s_1 + s_2}{2} \right)^{1-\alpha} = (r_1 + r_2)^\alpha (s_1 + s_2)^{1-\alpha},
\]
for arbitrary four nonnegative scalars \( r_1, r_2, s_1, s_2 \). Equivalently,
\[
(83) \quad \hat{r}^\alpha \hat{s}^{1-\alpha} + (1-\hat{r})^\alpha (1-\hat{s})^{1-\alpha} \leq \alpha \hat{r} + (1-\alpha) \hat{s},
\]
where \( \hat{r} := r_1/(r_1 + r_2) \), \( \hat{s} := s_1/(s_1 + s_2) \).

By Young's inequality,
\[
\hat{r}^\alpha \hat{s}^{1-\alpha} + (1-\hat{r})^\alpha (1-\hat{s})^{1-\alpha} \leq \alpha \hat{r} + (1-\alpha) \hat{s} + (1-\alpha)(1-\hat{s}) = 1.
\]
The equality holds only when \( \hat{r} = \hat{s} \), i.e., \( r_1 s_2 = r_2 s_1 \), which implies Eq. (81). \( \square \)

**Definition 4.2.** Let \( E^+ := \{ \Gamma^+ \in D^+ : \sum_{i=1}^{m} \gamma_{i,j}^+ K_{i,j}^{1/\alpha} > 0 \text{ for } j = 1, \ldots, n \} \), and \( E^- := \{ \Gamma^- \in D^- : \sum_{j=1}^{n} \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)} > 0 \text{ for } i = 1, \ldots, m \} \).

Let \( E_1^+ := \{ \Gamma^+ \in D^+ : \gamma_{i,j}^+ > 0 \text{ for each } i, j \} \), i.e., the set of all interior points in \( D^+ \). Likewise, let \( E_1^- := \{ \Gamma^- \in D^- : \gamma_{i,j}^- > 0 \text{ for each } i, j \} \).

Let \( \phi(\Gamma^+) \) be a maximizer of the problem \( \max_{\Gamma^- \in D^-} L(\Gamma^+, \Gamma^-) \).

**Lemma 4.3.** Let \( E_1^+, E_1^-, D^+, D^- \), and \( \phi \) be defined as above. Consider the maximization problem for \( L \) given in Eq. (80). We have the following properties.

1. Fixing \( \Gamma^- \in E_1^- \), consider \( \max_{\Gamma^+ \in D^+} L(\Gamma^+, \Gamma^-) \). Then the optimal \( \Gamma^+ = \{ \gamma_{i,j}^+ : i, j \} \) is uniquely determined by
\[
(84) \quad \gamma_{i,j}^+ = \frac{\gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)}}{\left( \sum_{j=1}^{n} \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)} \right)}.
\]

2. Fixing \( \Gamma^+ \in E_1^+ \), consider \( \max_{\Gamma^- \in D^-} L(\Gamma^+, \Gamma^-) \). Then the optimal matrix \( \Gamma^- = \{ \gamma_{i,j}^- : i, j \} \) is uniquely determined by
\[
(85) \quad \gamma_{i,j}^- = \frac{\gamma_{i,j}^+ K_{i,j}^{1/\alpha}}{\left( \sum_{i=1}^{m} \gamma_{i,j}^+ K_{i,j}^{1/\alpha} \right)}.
\]

Then \( \phi(\Gamma^+) \) is unique for each \( \Gamma^+ \in E_1^+ \).

3. Suppose that \( \Gamma^+ = \{ \gamma_{i,j}^+ : i, j \} \in D^+ \) satisfies
\[
(86) \quad L(\hat{\Gamma}^+, \phi(\hat{\Gamma}^+)) \geq L(\Gamma^+, \phi(\Gamma^+)) \text{ for all } \Gamma^+ \in D^+.
\]

Then \( \sum_{i=1}^{m} \gamma_{i,j}^+ K_{i,j}^{1/\alpha} > 0 \text{ for all } j, \text{ i.e., } \Gamma^+ \in E_1^+ \).

**Proof.** For (1): The cost \( L \) can be viewed as the sum, from \( i = 1 \) to \( m \), of \( \sum_{j=1}^{n} (\gamma_{i,j}^+ \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)})^{1-\alpha} K_{i,j} \). For each term, using Hölder's inequality, we have
\[
(87) \quad \sum_{j=1}^{n} (\gamma_{i,j}^+ \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)})^{1-\alpha} \leq (\sum_{j=1}^{n} \gamma_{i,j}^+ \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)})^{1-\alpha} = (\sum_{j=1}^{n} \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)})^{1-\alpha}.
\]

Due to the constraint \( \sum_{j=1}^{n} \gamma_{i,j}^+ = 1 \), the maximal value of \( \sum_{j=1}^{n} (\gamma_{i,j}^+ \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)})^{1-\alpha} K_{i,j} \) occurs when the first equality holds, i.e.,
\[
\text{when } \gamma_{i,j}^+ = \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)}/(\sum_{j=1}^{n} \gamma_{i,j}^- K_{i,j}^{1/(1-\alpha)}).
\]
This is the optimal condition for $\gamma^+_{i,j}$. For (2): The arguments are similar to case (1).

For (3): Suppose $\Gamma^+ \notin E^+_1$. Without loss of generality, by permuting the index $j$, we may assume $\sum_{i=1}^m \bar{\gamma}^+_{i,j} K_1^{1/\alpha} = 0$, which implies $\bar{\gamma}^+_{i,j} = 0$ for all $i$. Then $\bar{\gamma}^+_{1,1} = 0$. Let $\phi(\Gamma^+) = \{ \gamma^-_{i,j} : i,j \}$. Since $\sum_{i=1}^m \bar{\gamma}^-_{i,1} = 1$, then by permuting the index $i$, without loss of generality, we may assume $\bar{\gamma}^-_{1,1} > 0$.

Using Hölder’s inequality, we have

$$
\sum_{j=1}^n (\bar{\gamma}^+_{1,j})^\alpha (\bar{\gamma}^-_{1,j})^{1-\alpha} K_{1,j} \leq (\sum_{j=1}^n \bar{\gamma}^+_{1,j})^\alpha (\sum_{j=1}^n \bar{\gamma}^-_{1,j} K_{1,j}^{1/(1-\alpha)})^{1-\alpha} = (\sum_{j=1}^n \bar{\gamma}^-_{1,j} K_{1,j}^{1/(1-\alpha)})^{1-\alpha}.
$$

From the assumption $\bar{\gamma}^-_{1,1} > 0$, we know $\sum_{j=1}^n \bar{\gamma}^+_{1,j} K_{1,j}^{1/(1-\alpha)} > 0$. Due to the constraint $\sum_{i=1}^m \bar{\gamma}^+_{i,j} = 1$ and the maximal value of $\sum_{j=1}^n (\bar{\gamma}^+_{1,j})^\alpha (\bar{\gamma}^-_{1,j})^{1-\alpha} K_{1,j}$ occurs only when the first equality in Eq. (88) holds, i.e., when

$$
\bar{\gamma}^+_{1,j} = \gamma^+_{1,j} K_{1,j}^{1/(1-\alpha)} / (\sum_{j=1}^n \bar{\gamma}^-_{1,j} K_{1,j}^{1/(1-\alpha)}).
$$

Then $\bar{\gamma}^-_{1,1} > 0$, which contradicts the assumption $\bar{\gamma}^+_{1,1} = 0$. \qed

From the above result, Alg. 1 can be viewed as one block coordinate ascent method,

$$
\Gamma^+(l + 1) = \arg \max_{\Gamma \in D^+} L(\Gamma, \Gamma^-(l)), \quad \Gamma^-(l + 1) = \arg \max_{\Gamma \in D^-} L(\Gamma^+(l + 1), \Gamma).
$$

The standard convergence result (a global maximizer of a concave function can be attained) of a block coordinate ascent method relies on a crucial assumption: the cost function must be continuously differentiable (See for instance[3],[53]). However, $\sqrt{xy}$ in Eq. (4.5) is not continuously differentiable at $(x, y) = (0, 0)$. Thus, a little extra effort is required to deduce the desired convergence result.

Make a few observations on Alg. 1. First, in the whole iterative process, $\{ \Gamma^+(l), \Gamma^-(l) : l \}$ are positive matrices. Then they are interior points of $D^+, D^-$ respectively, i.e., $\Gamma^+(l) \in E^+_2$, $\Gamma^-(l) \in E^-_2$. Secondly, the cost function $L$ is continuously differentiable at every interior point in $D^+ \times D^-$. Thirdly, for every $\Gamma^+ \in E^+_2$, define the gradient $\nabla L(\Gamma)$ with respect to the first component $\Gamma^+$ at $(\Gamma^+, \phi(\Gamma^+))$ to be a vector whose entries are $\partial L / \partial \gamma^+_{i,j}(\Gamma^+, \phi(\Gamma^+))$. Hence, the gradient $\nabla L(\Gamma^+)$ is well-defined for each $\Gamma^+ \in E^+_2$.

To get the possible largest domain of the gradient of $L$, we introduce a function $f$ to examine the behavior of $L$ near the boundary of $E^+_1 \times E^-_1$:

$$
f(\Gamma^+) := \max \{ L(\Gamma^+, \Gamma^-) : \Gamma^- \in D^- \} = L(\Gamma^+, \phi(\Gamma^+)) \text{ for each } \Gamma^+ \in E^+_1.
$$

From Lemma 4.3, the maximizer $\Gamma^- = \phi(\Gamma^+)$ has an explicit expression,

$$
\gamma^-_{i,j} = \gamma^+_{i,j} K_{i,j}^{1/\alpha} / \sum_{j=1}^n \gamma^+_{i,j} K_{i,j}^{1/\alpha}.
$$

Therefore, by eliminating $\{ \gamma^-_{i,j} : i,j \}$,

$$
f(\Gamma^+) = \sum_{j=1}^n \sum_{i=1}^m \gamma^+_{i,j} K_{i,j}^{1/\alpha} - \frac{\partial f}{\partial \gamma^+_{i,j}}(\Gamma^+) = \alpha K_{i,j}^{1/\alpha} \sum_{i=1}^m \gamma^+_{i,j} K_{i,j}^{1/\alpha} - 1.
$$
Then $f$ is concave and continuously differentiable in $E_1^+$. In the following lemma, we shall extend the domain of $\nabla^1 L(\Gamma^+)$ from $E_2^+$ to $E_1^+$.  

**Lemma 4.4.** Notations $E_1^+, E_2^+, \phi, f, L$ and its gradient $\nabla^1 L$ are defined as above and in Definition 4.2. At each interior point $\Gamma^+ \in E_2^+$, 

$$
\nabla f(\Gamma^+) = \frac{\partial L}{\partial \gamma_{i,j}^+(\Gamma^+, \phi(\Gamma^+))} = \alpha K_{i,j}^{1/\alpha} \left( \sum_{i=1}^{m} \gamma_{i,j}^+ K_{i,j}^{1/\alpha} \right)^{\alpha-1}.
$$

Clearly $\nabla f$ is continuous and finite in $E_1^+$.

For each $\Gamma^+ \in E_1^+$, define

$$
\nabla^1 L(\Gamma^+) := \lim_{x \in E_2^+ \rightarrow \Gamma^+} \nabla^1 L(x) = \lim_{x \in \mathbb{R}} \nabla f(x) = \nabla f(\Gamma^+).
$$

Then $\nabla^1 L(\Gamma^+)$ is well-defined and continuous in $E_1^+$.

**Proof.** Note that for any interior point $\Gamma^+ \in E_1^+$,

$$
\frac{\partial L}{\partial \gamma_{i,j}^+} = \alpha K_{i,j} \gamma_{i,j}^+ (\gamma_{i,j}^+)^{1-\alpha} = \alpha K_{i,j}^{1/\alpha} \left( \sum_{i=1}^{m} \gamma_{i,j}^+ K_{i,j}^{1/\alpha} \right)^{\alpha-1},
$$

where Eq. (91) is used. Comparing this result with the partial derivatives of $f$ in Eq. (92), we have

$$
\frac{\partial f}{\partial \gamma_{i,j}^+}(\Gamma^+) = \frac{\partial L}{\partial \gamma_{i,j}^+}(\Gamma^+, \phi(\Gamma^+)), \quad \forall i,j.
$$

Now we are ready to prove the convergence property.

**Theorem 4.5** (Proof of Alg. 1). Given that $\gamma_{i,j}^+(0) > 0, \gamma_{i,j}^-(0) > 0$ for all $i,j$, every limit point of $\{\Gamma^+(k), \Gamma^-(k)\}$ constructed in Alg. 1 is a stationary point, and a maximizer of $L$.

**Proof.** The proof consists of two parts, which is modified from Prop.2.7.1 in [3]. The arguments from Eq. (96) to Eq. (103) are provided to establish that every limit is a stationary point. This part is essentially the same proof of Prop. 2.7.1 in [3]. The remaining of this proof is given to verify that every limit point is a maximizer.

According to the algorithm,

$$
L(\Gamma^+(k), \Gamma^-(k)) \leq L(\Gamma^+(k+1), \Gamma^-(k)) \leq L(\Gamma^+(k+1), \Gamma^-(k+1)) \leq \ldots,
$$

and then

$$
f(\Gamma^+(k)) \leq f(\Gamma^+(k+1)) \leq \ldots.
$$

Let $\hat{\Gamma}^+$ be a limit point of the sequence $\{\Gamma^+(k) : k\}$. Then $\hat{\Gamma}^+ \in D^+$ because $D^+$ is closed. Also, $\lim_{k \rightarrow \infty} f(\Gamma^+(k)) = f(\hat{\Gamma}^+)$.

In the following, we show that $\Gamma^+$ maximizes $f$ over $D^+$.

Let $\{\Gamma^+(k_j) : j = 0, 1, 2, \ldots\}$ be a subsequence of $\{\Gamma^+(k)\}$ that converges to $\hat{\Gamma}^+$. We first show that $\{\Gamma^+(k_j+1) - \Gamma^+(k_j)\}$ converges to zero as $j \rightarrow \infty$. Suppose not. Let $r^{k_j} = ||\Gamma^+(k_j+1) - \Gamma^+(k_j)||$. By possibly restricting to a subsequence of $\{k_j\}$, we may assume that there exists some $\bar{r} > 0$ such that $r^{k_j} \geq \bar{r}$ for all $j$. Let $s^{k_j} = (\Gamma^+(k_j+1) - \Gamma^+(k_j))/r^{k_j}$, then $||s^{k_j}|| = 1$. Note that $s^{k_j}$ belongs to a compact set, therefore has a limit point $\bar{s}$. By restricting to a subsequence of $\{k_j\}$, we assume that $s^{k_j}$ converges to $\bar{s}$. 


Fix some $\epsilon \in (0, 1)$. Note that $0 \leq \epsilon r \leq r^{kj}$. Therefore, $\Gamma^+(k_j) + \epsilon r s^{kj}$ lies on the segment joining $\Gamma^+(k_j)$ and $\Gamma^+(k_j + 1)$, and belongs to $D^+$. Thus,
\begin{equation}
L(\Gamma^+(k_j + 1)) = L(\Gamma^+(k_j) + \epsilon r s^{kj}) \geq L(\Gamma^+(k_j), \phi(\Gamma^+(k_j))) \geq L(\Gamma^+(k_j), \phi(\Gamma^+(k_j))).
\end{equation}
Taking the limit as $j$ tends to infinity, we obtain
\begin{equation}
L(\Gamma^+, \phi(\Gamma^+)) \geq L(\Gamma^+, \phi(\Gamma^+)) \geq L(\Gamma^+, \phi(\Gamma^+)).
\end{equation}
Thus, $L(\Gamma^+, \phi(\Gamma^+)) = L(\Gamma^+, \phi(\Gamma^+))$. Since $\epsilon r s \neq 0$, this contradicts the fact that $L$ is uniquely maximized when viewed as a function of the first block-component. Hence, we must have that $\Gamma^+(k_j + 1) - \Gamma^+(k_j)$ converges to zero. In particular, $\Gamma^+(k_j)$ converges to $\bar{\Gamma}^+$.

Since $\Gamma^+(k_j + 1)$ is the maximizer in $\max_{\Gamma^+ \in D^+} L(\Gamma^+, \phi(\Gamma^+(k_j)))$, then
\begin{equation}
L(\Gamma^+(k_j + 1), \phi(\Gamma^+(k_j))) \geq L(\Gamma^+, \phi(\Gamma^+(k_j))), \forall \Gamma^+ \in D^+,
\end{equation}
which implies,
\begin{equation}
\nabla^1 L(\Gamma^+(k_j + 1), \phi(\Gamma^+(k_j))) \cdot (\Gamma^+(k_j + 1) - \Gamma^+) \geq 0, \forall \Gamma^+ \in D^+.
\end{equation}
Taking the limit as $j$ tends to infinity in the above two equations, we obtain
\begin{equation}
L(\Gamma^+, \phi(\Gamma^+)) \geq L(\Gamma^+, \phi(\Gamma^+)), \forall \Gamma^+ \in D^+,
\end{equation}
which implies that $\bar{\Gamma}^+ \in E^+_1$ (shown in Lemma 4.3), and
\begin{equation}
\nabla f(\bar{\Gamma}^+) \cdot (\bar{\Gamma}^+ - \Gamma^+) = \nabla^1 L(\bar{\Gamma}^+) \cdot (\bar{\Gamma}^+ - \Gamma^+) \geq 0, \forall \Gamma^+ \in D^+.
\end{equation}
Since $f$ is concave and Eq. (104), then $f(\bar{\Gamma}^+) - f(\Gamma^+) \leq \nabla f(\bar{\Gamma}^+) \cdot (\bar{\Gamma}^+ - \Gamma^+) \leq 0$ for any $\Gamma^+ \in D^+$. Thus, $\bar{\Gamma}^+$ maximizes $f$ over $D^+$.  \hfill \Box

4.2. The non-uniqueness in $\Gamma^+, \Gamma^-$. In general, $\Gamma^+, \Gamma^-$ is not unique because the discrete measures $\{\sum_{i=1}^n \lambda^+_i \delta_{x_i}, \sum_{j=1}^n \lambda^-_j \delta_{y_j}\}$ and Lebesgue measures are mutually singular. Here we shall show the non-uniqueness of $(\Gamma^+, \Gamma^-)$ by examining the following dual problems.

**Dual Problem I**: Given $K_{i,j} > 0$,
\begin{equation}
\begin{align*}
\text{minimize} & \quad \alpha \sum_{i=1}^m a_i + (1 - \alpha) \sum_{j=1}^n b_j, \\
\text{subject to constraints} & \quad K_{i,j} \leq a_i^+ b_j^{-\alpha}, a_i > 0, b_j > 0 \text{ for all } (i, j).
\end{align*}
\end{equation}

Dual problem I is a special case of the dual problem in Eq. (23) with $d\nu_X/d\mu_X$, $d\nu_Y/d\mu_Y$ being Dirac point masses. The uniqueness of $\{a_i, b_j\}$ directly comes from Theorem 3.3. The duality relation between this problem and the primal problem in Eq. (80) can be proved by similar arguments in the weak duality theorem: By Young’s inequality,
\begin{equation}
\sum_{i=1,j=1}^{m,n} K_{i,j} (\gamma^+_i)^\alpha (\gamma^-_j)^{1-\alpha} \leq \sum_{i=1,j=1}^{m,n} (a_i^+)^\alpha (b_j^-)^{1-\alpha}
\end{equation}
\begin{equation}
\sum_{i=1,j=1}^{m,n} a_i^+ b_j^- \leq \alpha \sum_{i=1,j=1}^{m,n} a_i^+ (1 - \alpha) \sum_{j=1}^n b_j = \alpha \sum_{i=1}^m a_i + (1 - \alpha) \sum_{j=1}^n b_j.
\end{equation}
Thus, the cost value in the dual problem is always not less than the cost value in the primal problem. Besides, the equality occurs if and only if we can find a set of $(\{\gamma^+_i\}, \{\gamma^-_j\}, \{a_i\}, \{b_j\})$ such that
\begin{equation}
a_i^+ b_j^- = \dot{b}_j^- \gamma^-_j, \ (K_{i,j} - a_i^+ b_j^-) (\gamma^+_i)^\alpha (\gamma^-_j)^{1-\alpha} = 0
\end{equation}
hold for all $i, j$. Indeed the minimal value achieved in the dual problem is equal to the maximal value achieved in the original problem. The proof is presented in Lemma E.1 in Appendix.

From Eq. (108), either $\gamma_i^+ = 0 = \gamma_i^-$ or

$$a_i \gamma_i^+ = b_j \gamma_i^-, \quad K_{i,j} = a_i^\alpha b_j^{1-\alpha}. \tag{109}$$

holds. Thus, when the matrix with entries $\{K_{i,j} : i, j\}$ is not a matrix product of a column vector with entries $\{a_i^\alpha\}$ and a row vector with entries $\{b_j^{1-\alpha}\}$, some entries of $\{\gamma_i^+ : i, j\}$ and $\{\gamma_i^- : i, j\}$ must be zero. In particular, when the matrix with entries $\{K_{i,j} : i, j\}$ has full rank, $\Gamma^+$ and $\Gamma^-$ would be highly sparse.

**Dual Problem II:** Thanks to Eq. (108), denote

$$\hat{\gamma}_{i,j} := a_i \gamma_i^+ = b_j \gamma_i^-. \tag{110}$$

Then the Dual problem I could be converted into the problem of finding the optimal matrix $\hat{\Gamma}$ with entries $\{\hat{\gamma}_{i,j} : \hat{\gamma}_{i,j} \geq 0\}$

$$\begin{align*}
\text{to minimize} & \quad \sum_{i=1}^m \sum_{j=1}^n \hat{\gamma}_{i,j}, \\
\text{subject to} & \quad -\alpha \log(\sum_{i=1}^m \hat{\gamma}_{i,j}) - (1-\alpha) \log(\sum_{j=1}^n \hat{\gamma}_{k,j}) \leq -\log K_{k,h} \text{ for all } k, h.
\end{align*} \tag{111}$$

Since $(-\log x)$ is a convex function, then the feasible domain for $\hat{\Gamma}$ is a convex set. Once the optimal matrix $\hat{\Gamma}$ is found, by summing over each row and each column, we can determine the original dual variables $\{a_i, b_j : i, j\}$ and the primal variables $\{\gamma_i^+, \gamma_i^- : i, j\}$.

After the introduction of $\hat{\gamma}_{i,j}$ in Eq. (110), the optimal condition in Eq. (108) reduces to the condition that $\hat{\gamma}_{i,j} = 0$ for every $(i, j)$ with $K_{i,j} < a_i^\alpha b_j^{1-\alpha}$. Hence the rest $\hat{\gamma}_{i,j}$ can take any value, provided that they satisfy the marginal constraints:

$$\sum_{j=1}^n \hat{\gamma}_{i,j} = a_i, \quad \sum_{i=1}^m \hat{\gamma}_{i,j} = b_j \text{ for all } i, j. \tag{112}$$

Therefore, in general $\hat{\Gamma}$ is not unique, which implies that the original primal variables $\Gamma^+, \Gamma^-$ are not uniquely determined.

The next example illustrates a common phenomena originated from the non-uniqueness: the loop structure — matching links indeed form a loop.

**Figure 1.** Matching links between $\{x_1, x_2, x_3\}$ and $\{y_1, y_2, y_3\}$ form a loop.
Example 2. Let
\[
\hat{\Gamma}^\delta := \begin{pmatrix}
\hat{\gamma}_{1,1} & \hat{\gamma}_{1,2} & \hat{\gamma}_{1,3} \\
\hat{\gamma}_{2,1} & \hat{\gamma}_{2,2} & \hat{\gamma}_{2,3} \\
\hat{\gamma}_{3,1} & \hat{\gamma}_{3,2} & \hat{\gamma}_{3,3}
\end{pmatrix}
= \begin{pmatrix}
0 & 1 - \delta & 1 + \delta \\
1 + \delta & 0 & 1 - \delta \\
1 - \delta & 1 + \delta & 0
\end{pmatrix}
\]
Here \((a_1, a_2, a_3) = (2, 2, 2) = (b_1, b_2, b_3). The matching links are shown in Fig. 1. Note that \(\hat{\Gamma}^\delta\) is a solution if and only if we have that \(K_{i,j} \leq a_i^* b_j^{1-a}\) for all \(i, j\), and \(K_{i,j} = a_i^* b_j^{1-a}\) for all \(i, j\) with \(\hat{\gamma}_{i,j} > 0\). Suppose that \(\hat{\Gamma}^\delta\) with \(\delta = 0\) is one solution of Dual problem II. Clearly, since the cost function is the same for each matrix \(\hat{\Gamma}^\delta\) with \(-1 < \delta < 1\), then \(\hat{\Gamma}^\delta\) is also a feasible solution for each \(-1 < \delta < 1\).

Last, in the coming remark, we point out a connection between our algorithm with the well-known matrix scaling problem proposed by Sinkhorn.

Remark 3 (Matrix scaling). In the special case that \(m = n\), \(K_{i,j} = 1\) for all \(i, j\) (then the dual variables \(a_i = 1 = b_j\) for all \(i, j\)), Alg. 1 is related to the well-known matrix scaling problem proposed by Sinkhorn: Any positive square matrix is diagonally equivalent to a unique doubly stochastic matrix, and the diagonal matrices which take part in the equivalence are unique to scalar factors[45],[6]. From Eq. (108), we have \(\gamma_{i,j}^+ = \gamma_{i,j}^-\) for all \(i, j\). Then Theorem 4.5 says that by repeating row and column normalization we can get two identical doubly stochastic matrices, which provides another proof of the Sinkhorn matrix scaling problem.

4.3. Image morphing. Various distances are proposed to measure the dissimilarity between two images. Sometimes we are interested in a distance function, which is generated by a Riemannian metric on an image space, namely \(S\). A Riemannian metric is a symmetric, positive definite, bilinear form defined on a manifold, called a Riemannian manifold. Given two images described by densities \(P, Q\) in \(S\), this Riemannian metric between \(P, Q\) is the distance along a path \(\{R_t : 0 \leq t \leq 1\}\) connecting \(P, Q\), and parameterized by \(t \in [0, 1]\), i.e., \(R_0 = P, R_1 = Q\). This distance \(T(P, Q)\) is given by the path integral
\[
T(P, Q) := \inf_{R(t) \in S} \int_0^1 \left( \sum_{i=1}^l \sum_{j=1}^l g_{i,j} \frac{dr_i}{dt} \frac{dr_j}{dt} \right)^{1/2} dt = \inf_{R(t) \in S} \left( \int_0^1 \sum_{i=1}^l \sum_{j=1}^l g_{i,j} \left( \frac{dr_i}{dt} \frac{dr_j}{dt} \right)^{1/2} dt \right)^{1/2}.
\]
Here \(\{r_i(t) : i = 1, \ldots, l\}\) is some particle-set representation of the image \(R_t\), and \(\{g_{i,j} : i = 1, \ldots, l,\text{ and } j = 1, \ldots, l\}\) is some metric tensor\(^8\). The path \(\{R_t : 0 \leq t \leq 1\}\) is called the geodesic connecting \(P, Q\) with respect to the metric tensor \(\{g_{i,j}\}\). In this section, we would study a special metric tensor related to the following weighted Hellinger distance (the alpha divergence with \(\alpha = 1/2\)):
\[
T(P, Q) = \inf_{p,q} \int_{\Omega_1} \int_{\Omega_0} \left( p + q - 2\sqrt{pq} \exp(-\|x-y\|^2/2\sigma^2) \right) dxdy,
\]
with both \(p, q\) satisfying their marginal constraints
\[
\int_{\Omega_1} pdy = P(x), \quad \int_{\Omega_0} qdy = Q(y).
\]
This desired metric tensor has the property that the corresponding geodesic \(\{R_t : 0 \leq t \leq 1\}\) in \(S\) has an explicit expression as the case of MK problem. Adopting
\footnote{\textsuperscript{8}The last equality is derived from the Cauchy-Schwartz inequality. The details can be found in section 2.1[25].}
this metric tensor, an imaging morphing could be easily constructed by displaying the interpolated images along the geodesic.

This idea is analogous to the $L^2$ MK problem. The cost of the $L^2$ MK problem could be formulated as a distance for the image space $S$. More precisely, the optimal cost in the $L^2$ MK problem is equal to the infimum of

$$
(116) \int_{\mathbb{R}^2} \int_0^1 R(t,x)|v(t,x)|^2 dt dx,
$$

over all time varying densities $R$ and velocity fields $v$ satisfying the continuity equation $\partial R/\partial t + \nabla \cdot (Rv) = 0$ for $0 < t < 1$, and the initial condition $R(0,\cdot) = dv_X/d\mu_X$ and the final condition $R(1,\cdot) = dv_Y/d\mu_Y$ (See [2],[21]). Let $(R_{\text{min}}, v_{\text{min}})$ be a minimizer, and $s_{\text{min}}$ be the optimal mapping of the $L^2$ MK problem. Then the flow $f(x,t)$ corresponding to the minimizing velocity field $v_{\text{min}}$ via $f(x,0) = x$, $\partial f/\partial t = v_{\text{min}} \circ f$ is given by the interpolation $f(x,t) = x + t(s_{\text{min}}(x) - x)$. Then the flow $f$ provides a continuous warping map between $\nu_X$ and $\nu_Y$. Based on this result, the task of computing the warping map boils down to solving the MK problem and applying the interpolation.

In this section, we would derive the corresponding result for the alpha-D problem with $\alpha = 1/2$. Let the support of the density $R_t$ be $\Omega_t$ for $t \in [0,1]$. First, we need to derive the corresponding Riemannian metric induced by this distance function $T$ such that $T(P,Q)^2$

$$
(117) \inf \left\{ \lim_{k \to \infty} k^{-1} \sum_{i=1}^k T(R_{(i-1)/k}, R_{i/k})k^2, \text{ subject to } \{R_{i/k} : i = 1, \ldots, k \} \subset S \right\},
$$

which is the Riemann sum expression of Eq. (113). As $k \to \infty$, $R_{(i-1)/k}$ and $R_{i/k}$ are close to each other, i.e., the distance $\|x-y\|$ of each matching pair $(x,y)$ between $R_{(i-1)/k}$ and $R_{i/k}$ is close to zero, using

$$
(118) \exp(-\|x-y\|^2/2\sigma^2) \approx 1 - \|x-y\|^2/2\sigma^2
$$

with higher order terms $O(\|x-y\|^4)$ dropped, $T(R_{(i-1)/k}, R_{i/k})$ can be approximated by

$$
(119) \inf_{p,q} \left\{ \int_{\Omega_{(i-1)/k}} \int_{\Omega_{(i-1)/k}} p + q - 2\sqrt{pq}(1 - \frac{\|x-y\|^2}{2\sigma^2}) dx dy \right\},
$$

subject to constraints

$$
(120) \int_{\Omega_{(i-1)/k}} pdy = R_{(i-1)/k}, \text{ and } \int_{\Omega_{i/k}} qdx = R_{i/k}.
$$

To obtain the metric tensor in Eq. (113), we proceed with the discrete form of Eq. (119). Let the optimal matching scheme between $R_{(i-1)/k}, R_{i/k}$ be a set of $l$ matching pairs $\{(x_i,y_i) : i = 1, \ldots, l\}$ with mass $\{(P_i, Q_i) : i = 1, \ldots, l\}$ respectively. Then

$$
(121) T(R_{(i-1)/k}, R_{i/k}) = \inf \sum_{i=1}^l \left( \sqrt{P_i} - \sqrt{Q_i} \right)^2 + \frac{\sqrt{P_iQ_i}\|x_i-y_i\|^2}{\sigma^2},
$$
where the infimum is taken over all the possible matching scheme. Substituting Eq. (121) into Eq. (117), as \( k \to \infty \), letting \( r_i(t) = (r_{i,1}(t), r_{i,2}(t)) \), we obtain the form of \( T(P, Q) \):

\[
T(P, Q)^2 = \inf_{\{r_i(t): i=1,\ldots,l\}} \int_0^1 \left( \frac{dr_{i,1}}{dt} \right)^2 + r_{i,2}^2 \left\| \frac{dr_{i,2}}{dt} \right\|^2 \, dt,
\]

where the first infimum is taken over all the possible matching schemes ( \( \{(P_i, Q_i) : i = 1, \ldots, l\} \) subject to the according marginal constraints, and the second infimum is taken subject to the boundary constraints

\[
r_i(0) = (\sqrt{P_i}, x_i/\sigma), \ r_i(1) = (\sqrt{Q_i}, y_i/\sigma), \text{ for each } i = 1, \ldots, l.
\]

To derive an efficient algorithm for the minimizers in Eq. (122), we would simply the integral by fixing one matching scheme and examining the optimality condition for the second infimum in Eq. (122). Since each \( r_i(t) \) in this cost function in Eq. (122) is decoupled, then we could optimize each \( r_i(t) \) individually. The result is listed in the following lemma. (Due to the same structure in each optimization problem, the subscript \( i \) is dropped.)

**Lemma 4.6.** Given that one particle \( P \) is located at \( x \in \mathbb{R}^d \) with mass \( \mathcal{P} \), and the other particle \( Q \) is located at \( y \in \mathbb{R}^d \) with mass \( \mathcal{Q} \), consider a particle path described by

\[
r(t) := (r_1(t), r_2(t)) \text{ for } 0 \leq t \leq 1,
\]

to minimize the distance:

\[
\int_0^1 \left( \frac{dr_1}{dt} \right)^2 + r_1^2 \left( \frac{dr_2}{dt} - \frac{dr_1}{dt} \right)^2 dt,
\]

with boundary conditions

\[
r(0) = (\sqrt{P}, x/\sigma), \ r(1) = (\sqrt{Q}, y/\sigma), \ \sigma > 0.
\]

Let

\[
K_c(x, y) = \cos(||x-y||/\sigma) \text{ if } ||x-y||/\sigma < \pi/2, \text{ and } K_c(x, y) = 0, \text{ otherwise.}
\]

Then \( r(t) \) can be computed explicitly through some coordinate transformation and the minimal distance in Eq. (125) can be simplified as

\[
(\sqrt{P})^2 + (\sqrt{Q})^2 - 2\sqrt{P} \sqrt{Q} K_c(x, y).
\]

**Proof.** The minimal path \( r(t) \) can be derived through calculus variations. Here, we present another approach based on the following observation: The above Riemannian metric in Eq. (125) has the same metric tensor components as the Euclidean distance in \( \mathbb{R}^2 \) in terms of polar coordinates (for instance, page 53 in [25]). Hence, the space for each particle path under a coordinate transformation is in fact locally isometric to the plane \( \mathbb{R}^2 \). That is, each path under a coordinate transformation becomes a line segment.

We outline three coordinate systems for the particle path used in the following discussion:

\[
(r_1, r_2) \leftrightarrow \text{polar coordinates}(r_1, \theta) \leftrightarrow \text{Cartesian coordinates}(\hat{r}_1, \hat{r}_2).
\]

To construct \( (r_1, \theta) \) from \( (r_1, r_2) \), consider the \( r_1 \)-\( \theta \)-plane in which the \( \theta \)-axis points in the direction of \( e_\theta = (y-x)/||y-x|| = (r_2(1) - r_2(0))/||r_2(1) - r_2(0)|| \). Let \( \theta(t) := (r_2(t) - r_2(0)) \cdot e_\theta \), then the coordinates \( (r_1, \theta) \) represent the projection
Then the Riemannian metric in Eq. (130) of a path \( \{r(t) : 0 \leq t \leq 1 \} \) on this \( r_1 \theta \)-plane, see Fig. 2. Then \( r(0), r(1) \) have coordinates \((r_1(0), 0)\) and \((r_1(1), \theta(1))\) respectively. Under this setting,

\[
\int_0^1 (\frac{dr_1}{dt})^2 + r_1^2 (\frac{dr_2}{dt})^2 dt \geq \int_0^1 (\frac{dr_1}{dt})^2 + r_1^2 (\frac{d\theta}{dt})^2 dt,
\]

and the equality holds only when vectors \( r_2(t) - r_2(0) \) parallels the vector \( e_\theta \) for all \( t \), which implies that the geodesic \( r(t) \) must lie in the \( r_1 \theta \) plane for \( t \in [0, 1] \). To handle the distance in the right-hand side of Eq. (129), we view \((r_1(t), \theta(t))\) as the polar coordinate expression of the geodesic, and let \( \hat{r}(t) := (\hat{r}_1(t), \hat{r}_2(t)) \) be the corresponding Cartesian coordinate expression, i.e.,

\[
\hat{r}(t) = (r_1(t) \cos(\theta(t)), r_1(t) \sin(\theta(t))) \quad \text{for} \quad t \in [0, 1].
\]

Then the Riemannian metric in Eq. (125) becomes

\[
\int_0^1 (\frac{d\hat{r}_1}{dt})^2 + (\frac{d\hat{r}_2}{dt})^2 dt.
\]

Thus the \( \hat{r}_1 \hat{r}_2 \) plane is locally isometric to \( \mathbb{R}^2 \), and the graph of \( \hat{r}(t) \) is a line segment connecting two points \( \hat{r}(0), \hat{r}(1) \), i.e.,

\[
\hat{r}(t) = \hat{r}(0) + t(\hat{r}(1) - \hat{r}(0)).
\]

Besides, the geodesic distance between particles \( P \) and \( Q \) is the Euclidean distance \( ||\hat{r}(0) - \hat{r}(1)|| \).

Another manner to explain Eq. (132) is through rewriting Eq. (121) as follows:

\[
P + Q - 2\sqrt{PQ} (1 - \frac{||x - y||^2}{2\sigma^2}) \approx (\sqrt{P})^2 + (\sqrt{Q})^2 - 2\sqrt{PQ} \cos(\frac{||x - y||}{\sigma}).
\]

Regard \( ||x - y||/\sigma \) as the angle between two vectors \( \hat{r}(0), \hat{r}(1) \). By the Law of Cosines, the distance in Eq. (133) is exactly the squared norm \( ||\hat{r}(0) - \hat{r}(1)||^2 \) (See the right in Fig. 2). Note that this distance is only valid for \( ||x - y||/\sigma \leq \pi/2 \)
according to the following reasons. First, the domain of the angular coordinate for a polar coordinate system is \((-\pi, \pi]\). Secondly, in case of \(\pi/2 < \|x - y\|/\sigma < \pi\), the distance with the link connected is greater than the distance without the link, \((\sqrt{P})^2 + (\sqrt{Q})^2\), then the matching link between these two particles does not exist. Therefore, based on Eq. (132), the squared geodesic distance between \(P, Q\) is given by Eq. (128).

Once \(\hat{r}(t)\) is obtained, we can compute the particle path \(r(t)\) backwards as follows:

- Use coordinate transformations to compute polar coordinates \((r_1(t), \theta(t))\) from Cartesian coordinates \(\hat{r}(t)\).
- To get \(r_1r_2\) coordinates from \(r_1\theta\) coordinates, let \(r_2(t) = x + \theta(t)e_\theta\) with \(e_\theta = (y - x)/\|x - y\|\).

\[\square\]

The result in Eq. (128) provides an explicit expression for the integral in Eq. (122). Then by substituting it into Eq. (122), we have

\[
T(P, Q)^2 = \inf \sum_{i=1}^{l} ((\sqrt{P_i})^2 + (\sqrt{Q_i})^2 - 2\sqrt{P_iQ_i}K_c(x_i, y_i))
\]

\[= \text{Total masses of } P \quad \text{and} \quad Q - 2 \sup \left( \sum_{i=1}^{l} \sqrt{P_iQ_i}K_c(x_i, y_i) \right),\]

where both \(\inf\), \(\sup\) are taken over all possible matching schemes. Hence, the optimal matching scheme is a maximizer of the discrete form of the cost function

\[
\text{subject to constraints}
\]

\[
\int_{\Omega_1} p \ dy = P, \quad \int_{\Omega_0} q \ dx = Q.
\]

Once the optimal matching scheme is found, then we construct a particle path \(r(t) = (r_1(t), r_2(t))\) for each matching pair. Then the representative point-set for each \(R_t\) is constructed by placing a particle with mass \(r_1(t)\) at the position \(r_2(t)\) for all the matching pairs.

We summarize the algorithm as follows. A numerical simulation is provided in the next section.

**Algorithm 2.** Choose a kernel size \(\sigma > 0\). Given two images represented by two point-sets \(R_0, R_1\) on \(\Omega_0, \Omega_1\) respectively, our goal is to produce a sequence of images represented by point-sets \(R_t\) on \(\Omega_t\) for \(0 < t < 1\) along the geodesic path (the geodesic distance is given in Eq. (117)).

- Use Alg. 1 to find an optimal matching scheme described by matching pairs \(\{(x_i, y_i) : i = 1, \ldots, l\}\) maximizing \(T(R_0, R_1)\) in Eq. (135).\(^9\)
- For each matching pair \((x_i, y_i)\) with mass \((P, Q)\), denote \(r(0) := (P, x_i/\sigma)\), \(r(1) := (Q, y_i/\sigma)\), and

\[
\hat{r}(0) := (P, 0), \quad \hat{r}(1) := (Q \cos(\|y_i - x_i\|/\sigma), Q \sin(\|y_i - x_i\|/\sigma)).
\]

\(^9\)To meet the positivity condition in Alg. 1, one might use a positive kernel function like \(\max(K_c(x, y), \epsilon)\) instead of \(K_c(x, y)\) itself, where \(\epsilon\) is a positive small scalar.
The path in the \( \hat{r}_1 \hat{r}_2 \) plane is
\[
\hat{r}(t) := \hat{r}(0) + t(\hat{r}(1) - \hat{r}(0)), \quad \text{for } 0 < t < 1.
\]

The mass/position information of the path between \( x_i, y_i \) is given by
\[
(139) \quad r_1(t; x_i, y_i) := \|\hat{r}(t)\|, \quad r_2(t; x_i, y_i) = x_i/\sigma + (y_i - x_i)\theta(t)/\|x_i - y_i\|,
\]
where \( \theta(t) := \arctan(\hat{r}_2(t)/\hat{r}_1(t)) \). Then each particle-set \( R_t \) is described by \( \{r_1(t; x_i, y_i), r_2(t; x_i, y_i) : i = 1, \ldots, l\} \).

- To display images, we need to have the intensity values at specific grid points (pixels). Since \( r_2(t; x_i, y_i) \) is not necessarily at those grid points, to display an image corresponding to each point-set \( R_t \), we convolute the mass function of the particle-set with a Gaussian kernel function:
\[
(140) \quad R_t(u) = \sum_{i=1}^{l} r_1(t; x_i, y_i) \exp\left(-\frac{(u - r_2(t; x_i, y_i))^2}{2\sigma'^2}\right),
\]
for each grid point \( u \in \Omega_t \). Here \( \sigma' \) is another parameter, which is empirically chosen as a number larger than the grid size of the displayed image.

Remark 4. The speciality of adopting square root functions in this Riemannian metric was studied in [56]. Younes gave several distances (e.g., Eq. (21), (22)) between two shapes (curves) based on the form of a 1/2-divergence weighted by cosines of the angle differences. For instance, one geodesic distance is given by Eq. (22) in [56]:
\[
(141) \quad \left( l_1 + l_2 - 2\sup_g \int_0^1 \sqrt{l_1 l_2 g''(t)} \left| \cos \frac{\theta_1(g(t)) - \theta_2(t)}{2} \right| \, dt \right)^{1/2}
\]
\[
(142) \quad = \left( E_Y(C_1, C_1) + E_Y(C_2, C_2) - 2E_Y(C_1, C_2) \right)^{1/2},
\]
with \( E_Y(C_i, C_j) := \sup_{h_i, h_j} \int_0^1 \sqrt{h_i'(t) h_j'(t)} \left| \cos \frac{\theta_i(h_i(t)) - \theta_j(h_j(t))}{2} \right| \, dt \).

The notations are briefly explained here. Consider two curves \( C_1, C_2 \) in the \( xy \) plane with arc-lengths \( l_1, l_2 \), and angle functions \( \theta_1, \theta_2 \) (between the tangent and the axis \( y = 0 \)). Parameterize two curves by \( t \) and \( g(t) \) respectively, where \( g \) is some increasing diffeomorphism of \([0, 1]\) such that the distance is minimized. The correspondence between two curves is given by \( t \leftrightarrow g(t), \ t \in [0, 1], \) or equivalently, could be constructed through the parameterizations \( h_1(t) \) and \( h_2(t) \) on curves \( C_1, C_2 \). Here \( h_i \) is some increasing differentiable functions from \([0, 1]\) to \([0, l_i]\) for \( i = 1, 2 \). In case of \( l_1 = 1 = l_2 \), the derivatives \( h'_1, h'_2 \) could be regarded as the densities of the distribution functions \( h_1, h_2 \), then the elastic distance could be regarded as a 1/2-divergence weighted by cosine functions. Since this distance is not convex in \( g \), the optimal correspondence is usually computed in the product space \([0, l_1] \times [0, l_2]\) by dynamical programming. Although this distance can be used to measure the distance between two closed curves, as mentioned in page 583 [56] the curves along the geodesic are not closed in general. Later, Younes et. al. proposed another Riemannian metric based on Jordan angles in Grassmann manifolds [36] in which the curves constructed along the geodesic path are closed [57]. Surely the cosine-weighted alpha-D problem in Eq. (135) can be applied to the curve matching problem. In some sense, Eq. (135) could be viewed as a generalization of Younes’ elastic distance in Eq. (141) from \( \mathbb{R}^1 \) to \( \mathbb{R}^d \).
5. **Numerical results.** The algorithms in this paper have order \(O(N^2)\) where \(N\) is the total number of points. Thus we do not suggest to use the alpha-D model as an area-based method, in which \(N\) becomes the number of image pixels. In this section, we provide some experiment results in which the alpha-D model (\(\alpha = 1/2\)) is employed as a feature-based method and the size of \(N\) is between 100 to 1000.

First, we provide a few numerical simulation results to reveal several properties of the alpha-D model. Then we show two matching results of real photos based on Eq. (143). Empirically, matching results vary depending on the kernel scale \(\sigma\) in \(K\), especially when extreme values are taken. Empirically, when \(\sigma^2\) is too small, the cost in Eq. (143) has many local optimal solutions and the algorithm tends to get stuck at some local optimal solution; when \(\sigma^2\) is too large, the estimated transformation \(T\) is not robust, i.e., very sensitive to noise. For simplicity, in our experiments a kernel scale \(\sigma\) is chosen between 0.1 and 0.4, if images are scaled to fit in a unit square.

**[Experiment I:]** In Fig. 3, the first sub-figure (from the left to the right) shows image 1 consisting of 20 \(\times\) 20 pixels. The second sub-figure shows image 2 consisting of 20 \(\times\) 20 pixels. Each pixel with lighter marker · has mass 0.5 and the pixels with darker makers • has mass 1. The third sub-figure shows the matching result based on the alpha-D model. Clearly we can see that these two blocks are correctly matched. Here the similarity function \(K(x, y)\) is given by \(\exp(-\|x - y\|^2/0.3^2)\).

The fourth sub-figure shows the result of the MK problem. A normalization on the total mass of each image is employed to meet the mass balance condition. Since the upper block in image 1 has more mass than the upper block in image 2, a few pixels (about 4 pixels with marker •) merge into the lower block shown in ▼.

**[Experiment II:]** In Fig. 4, given two particle sets – two letters “i” and “j”, we use Alg. 2. to construct five interpolated particle-sets at \(t = 1/6, 2/6, 3/6, 4/6,\) and \(5/6\) respectively. The kernel function is \(\exp(-\|x - y\|^2/0.3^2)\). Both two particle-sets consist of 60 particles (shown in ·), and each particle has mass 1. The first row shows the interpolated particle-sets, in which each dot (·) refers to the trace of the geodesic path between each matching pair in \(\Omega_t\), and each dot does not necessarily have equal mass. To visualize both the mass and spatial position information, we construct interpolated images by convolving each interpolated particle-set (shown in the first row) with \(\exp(-\|x - y\|^2/0.02^2)\). (Here the value 0.02 is larger then the grid size of the image \(5 \times 10^{-3}\)). Notice that there is no “fade-in” and “fade-out” effect in this result under this kernel scale.

**[Experiment III:]** In applications, sometimes the estimated transformation is expected to own some smoothness. However, the cost functions of the MK problem or the alpha-D problem have no control over the smoothness of the matching relation. Introducing splines is a common approach in describing non-rigid deformations. Hence, we incorporate the alpha-D problem with splines to estimate both the correspondence and the nonrigid deformation function. Consider two point-sets \(\{x_i : i = 1, \ldots, m\}, \{y_j : j = 1, \ldots, n\}\) in \(\mathbb{R}^d\). Given a linear bounded operator \(\mathcal{L}\), we estimate corresponding matrices \(\Gamma^+, \Gamma^-\) with entries \(\gamma_{i,j}^+, \gamma_{i,j}^- : i, j\) and a

\[10\] Another approach (but more time-consuming) is that one first estimates transformations and correspondences under a large value of \(\sigma\), and improves these estimations by letting \(\sigma\) gradually decreasing to 0. This approach has been used in the RPM method[10].
smooth function $f$ through maximizing the energy function,

$$E_S(\Gamma^+, \Gamma^-, f) := -\lambda \|L f\|^2 + \sum_{i,j=1}^{m,n} \sqrt{\gamma_{i,j}^+ \gamma_{i,j}^-} \exp(-\|y_j - x_i - f(x_i)\|^2/\sigma^2),$$

subject to the marginal constraints $\sum_{i=1}^m \gamma_{i,j}^- = 1$, $\sum_{j=1}^n \gamma_{i,j}^+ = 1$ where $\lambda > 0$ is a parameter. A reproducing kernel Hilbert space with its associated reproducing kernel $U$ is introduced by a bounded linear operator $L$. The smooth function $f$ in this Hilbert space could be represented by linear combination of eigenfunctions of the reproducing kernels $U$. The details about the relation between $U$ and $L$ can be found in the reference[51]. Two popular splines are radical-basis-function (RBF) splines or thin-plate-splines.

In this experiment, the alpha-D problem incorporated with splines is employed to register two artificial shapes, which are represented by two particle-sets in Fig. 5. Each particle-set consists of $n = 30$ particles, $\{x_i : i = 1, \ldots, 30\}$ and $\{y_i : i = 1, \ldots, 30\}$, placed along the shapes clockwise and with masses given by the arc-lengths $\{||y_{i+1} - x_{i-1}||/2, ||y_{i+1} - y_{i-1}||/2 : i = 1, \ldots, 30\}$, with $x_{31} := x_1, y_{31} := y_1, x_0 := x_{30}, y_0 := y_{30}$}. Then two shapes are matched through maximizing the energy function in Eq. (143). Consider the reproducing kernel $U(x, y) = \exp(-\|x - y\|^2/0.05)$, and construct $f$ by the RBF splines: $f(x) = \sum_{i=1}^n \alpha_i U(x_i, x)$ with vectors $\alpha_i \in \mathbb{R}^2$ to be determined. Then the regularity term of RBF splines $\|L f\|^2$ in Eq. (143) becomes

$$\|L f\|^2 = \sum_{i,j=1}^n (\alpha_i \cdot \alpha_j U(x_i, x_j)).$$

By formulating the exponential function as shown in Eq. (30), $\{\alpha_i : i\}$ is computed by solving a weighted least square problem. Numerically, the problem given in Eq. (143) could be solved by a block coordinate ascent method. The whole algorithm boils down to updating $\Gamma^+$, $\Gamma^-$, and $\{\alpha_i : i\}$ cyclically. The forward/backward matching results are shown in the third and fourth sub-figures.

Here we provide two image matching experiments on photos. Fig. 6 shows the matching result between two point-sets extracted from two photos of one building. Here, each set of feature points is extracted by applying the edge detection on images $I$, $\{x : |\nabla I(x)| > \epsilon, \text{ for some positive } \epsilon\}$. The affine transformation and the correspondence between point-sets are estimated through maximizing

$$E_S(\Gamma^+, \Gamma^-, A) := \sum_{i,j=1}^{m,n} \sqrt{\gamma_{i,j}^+ \gamma_{i,j}^-} \exp(-\|y_j - x_i - A(x_i)\|^2/\sigma^2),$$

where $A(x)$ is an affine transformation of $x$. Similarly, Fig. 7 shows the matching result between two photos of one baby. In this experiment, the RBF spline and the correspondence between point-sets are estimated through minimizing Eq. (143).

Empirically, the alpha-D model has a similar matching performance as the RPM method[10] or the KC method[48]. The main difference is the computational cost. Compare with the RPM method. Because the correspondence matrix in the RPM method is subject to some row-sum and column-sum constraints, similar to doubly stochastic constraints, they proposed to use the soft-assignment to obtain a doubly-stochastic-like matrix after each iteration of the transformation. When the size of correspondence matrices is large, the soft-assignment might take a large number of iterations to get an acceptable doubly-stochastic-like matrix. Hence the RPM
method generally converges slower than the block coordinate ascent method in the αD model. Now compare the αD model with the KC method. Correspondence is not estimated in the KC model, thus the computational cost in the KC method is more economical, especially in the case of rigid motions. However, in the case of nonrigid deformations, the cost in Eq. (3) cannot be handled by a block coordinate descent method. Instead, a gradient descent method with backtracking is required. Under this circumstance, the computational cost in the KC method could become more expensive. Several comparison experiments can be found in [8].

Appendix A. Proof of closest point property.

Lemma A.1. Consider two point-sets \( X, Y \) with the same cardinality \( n \). Suppose for each \( i \), \( x_i \) is the closest point of \( y_i \) among the point set \( X \), and \( y_i \) is the closest point of \( x_i \) among the point set \( Y \). Then \( (\Gamma^+, \Gamma^-) \) being a pair of identity matrices is a maximizer of \( E_\alpha(\Gamma^+, \Gamma^-) \) in Eq. (7).

Proof. Let \( \{\gamma^+_{i,j}, \gamma^-_{i,j} : i,j\} \) be the entries of \( \Gamma^+, \Gamma^- \) respectively. Let \( \delta_{i,j} \) be the Kronecker delta. Since \( \|x_i - y_i\| \leq \min\{\|x_i - y_j\|, \|x_j - y_i\| : j \} \) for each \( i \), then

\[
K(x_i, y_i)^\alpha K(x_j, y_j)^{1-\alpha} \geq K(x_i, y_j),
\]

and then

\[
E_\alpha(\Gamma^+, \Gamma^-) = \sum_{i,j=1}^{n,n} (\gamma^+_{i,j})^\alpha (\gamma^-_{i,j})^{1-\alpha} K(x_i, y_j)
\]

\[
\leq \sum_{i,j=1}^{n,n} (\gamma^+_{i,j} K(x_i, y_i))^\alpha (\gamma^-_{i,j} K(x_j, y_j))^{1-\alpha},
\]

\[
\leq (\sum_{i,j=1}^{n} \gamma^+_{i,j} K(x_i, y_i))^\alpha (\sum_{i,j=1}^{n} \gamma^-_{i,j} K(x_j, y_j))^{1-\alpha},
\]

\[
= (\sum_{i=1}^{n} K(x_i, y_i))^\alpha (\sum_{j=1}^{n} K(x_j, y_j))^{1-\alpha}
\]

\[
= \sum_{i,j=1}^{n,n} (\delta_{i,j})^\alpha (\delta_{i,j})^{1-\alpha} K(x_i, y_j) = E_\alpha(\{\delta_{i,j}\}, \{\delta_{i,j}\}).
\]

The second inequality comes from Hölder’s inequality. Thus, \( (\Gamma^+, \Gamma^-) \) with \( \gamma^+_{i,j} = \delta_{i,j} = \gamma^-_{i,j} \) for all \( i, j \) is a maximizer. Done.

Appendix B. Proof of nonnegativity.

Lemma B.1. Consider two point-sets \( X = \{x_i : i = 1, \ldots, m\} \) and \( Y = \{y_i : i = 1, \ldots, n\} \) and \( M_\alpha \) defined in Eq. (8). If \( M_\alpha(X, Y) = 0 \), then \( X = Y \).

Proof. Since \( K(x, y) \leq 1 \), then

\[
M_\alpha(X, Y) := \alpha E_\alpha(X, X) + (1 - \alpha) E_\alpha(Y, Y) - E_\alpha(X, Y)
\]

(147)

\[
\geq \alpha \sum_i \gamma^+_{i} + (1 - \alpha) \sum_j \gamma^-_{j} - \sum_{i=1,j=1}^{m,n} (\gamma^+_{i,j})^\alpha (\gamma^-_{i,j})^{1-\alpha}
\]

(148)

\[
= \sum_{i=1,j=1}^{m,n} (\alpha \gamma^+_{i,j} + (1 - \alpha) \gamma^-_{i,j} - (\gamma^+_{i,j})^\alpha (\gamma^-_{i,j})^{1-\alpha}) \geq 0.
\]

(149)
where the first inequality comes from the assumption $0 < K \leq 1$ and the second inequality comes from Young’s inequality. Hence $M_\alpha(X, Y) = 0$ implies that

$$
\gamma_{i,j}^+ - \gamma_{i,j}^- = (\gamma_{i,j}^+ - \gamma_{i,j}^-) \alpha (K(x_i, y_j) - 1) = 0
$$

for all $i, j$, then $\gamma_{i,j}^+ (K(x_i, y_j) - 1) = 0$ for all $i, j$, which implies $K(x_i, y_j) = 1$ (i.e., $x_i = y_j$) for each pair $(i, j)$ with $\gamma_{i,j}^+ = \gamma_{i,j}^- > 0$. This statement in fact is equivalent to $X = Y$. In summary, $M_\alpha(X, Y) = 0$ implies $X = Y$.  

\[\square\]

**Appendix C. A substitution for the convexity condition.**

**Remark 5.** The convexity condition can be replaced by the weaker condition in Eq. (40). Denote two $d \times d$ matrices whose entries are partial derivatives

$$
\{\partial^2 H / \partial x_i \partial x_j(x, y), \partial^2 H / \partial y_i \partial y_j(x, y) : 1 \leq i \leq d, 1 \leq j \leq d\}
$$

by $\mathcal{H}_1(x, y), \mathcal{H}_2(x, y)$. Let

$$
c_i = \inf \{\text{ the smallest eigenvalue of } \mathcal{H}_i(x, y) : (x, y) \in \Omega\} \text{ for } i = 1, 2.
$$

Thanks to the uniform bound condition on the entries of $\mathcal{H}_1, \mathcal{H}_2$, their $l^\infty$ matrix norms have uniform bounds as well. Then from the fact that the spectral radius of a matrix can be bounded by its $l^\infty$ matrix norms (See Theorem 5.6.9 in [22]), the existence of $c_1, c_2$ is verified. Let $\hat{\phi} : \Omega_X \to \mathbb{R}, \hat{\psi} : \Omega_Y \to \mathbb{R},$

$$
\hat{\phi} = \sum_{i=1}^d \max(-c_1/2, 0)x_i^2, \hat{\psi} = \sum_{i=1}^d \max(-c_2/2, 0)y_i^2.
$$

Let $\hat{H}(x, y) = H(x, y) + \hat{\phi}(x) + \hat{\psi}(y)$, and let $\hat{\mathcal{H}}_1(x, y), \hat{\mathcal{H}}_2(x, y)$ be the according $d \times d$ matrices whose entries are partial derivatives of $\hat{H}$. Then the matrices $\hat{\mathcal{H}}_1(x, y), \hat{\mathcal{H}}_2(x, y)$ are both positive semi-definite for each $(x, y)$ in $\Omega$, which implies that $\hat{H}(x, y)$ is convex in $x$ and convex in $y$. The original constraint $H(x, y) \leq \phi(x) + \psi(y)$ is replaced by $\hat{H}(x, y) \leq \hat{\phi}(x) + \hat{\psi}(y)$, where $\hat{\phi}(x) := \phi(x) + \hat{\phi}(x), \hat{\psi}(y) := \psi(y) + \hat{\psi}(y)$ are both convex functions with respect to $\nu_X, \nu_Y$ (Lemma 3.4). Note that

$$
D_x \hat{H} = D_x H + \nabla \hat{\phi}(x) = D_x H + \max(-c_1, 0), D_y \hat{H} = D_y H + \max(-c_2, 0),
$$

then the bijective property of the gradients of $\hat{H}$ is inherited from the bijective property of the gradients of $H$. Thus under the conditions of Corollary 2, one can use the same arguments to conclude that $s$ is a bijection $a.e.$.

In application, when both domains $\Omega_X, \Omega_Y$ are bounded, one can easily see that the above weaker condition (Eq. (40)) is satisfied.

**Appendix D. Proofs of the injectivity induced by $D_x H, D_y H$.**

**Lemma D.1.** Assume that $F$ is continuously differentiable from $X = \mathbb{R}^d$ to $Y = \mathbb{R}^d$. Write the vector field $F$ as $F = (F_1, \ldots, F_d)$ with each scalar field $F_j : X \to \mathbb{R}$.

Let $\mathcal{J}(x)$ be its Jacobian matrix with entries $\{\partial F_i / \partial x_j : 1 \leq i \leq d, 1 \leq j \leq d\}$.

1. If $\mathcal{J}(x)$ is positive definite for each $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, then $F$ is an injection from $X$ to $Y$.
2. If $\mathcal{J}$ is a nonsingular constant matrix, then $F$ is a bijection from $X$ to $Y$. 

**Inverse Problems and Imaging**, Volume 5, No. 3 (2011), 551–590
Proof. Part 1: Suppose not. Then we may find \( \bar{x} \neq x \) in \( X \) such that \( F(x) = F(\bar{x}) \). Consider the problem
\[
\max_{t \in [0,1]} \{ F(t) := u \cdot F(x + t(\bar{x} - x)) \},
\]
for some nonzero vector \( u \in \mathbb{R}^d \), say \( u = \bar{x} - x \). Since \( F(0) = F(1) \), then by Rolle's theorem there exists some \( t_0 \in [0,1] \) such that \( F'(t_0) = 0 \), i.e.,
\[
u \cdot (F(x + t_0(\bar{x} - x))(\bar{x} - x)) = 0.
\]
Hence, when \( u = \bar{x} - x \), then Eq. (153) indicates that \( J \) is not positive definite at \( x + t_0(\bar{x} - x) \). Contradiction. Hence, we have the desired injection.

Part 2: This proof is suggested by one reviewer. Since the gradient of each component \( F_j \) is constant, then \( F \) in this case is affine, i.e., \( F(x) = Jx + b \) where \( b \) is some constant vector. Hence, \( F \) is a bijection. \( \square \)

According to this lemma, we can get the desired injectivity for the gradients \( D_xH, D_yH \) in Theorem 3.6. Fix some \( x_0, y_0 \in \mathbb{R}^d \). Consider a matrix function \( H : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}^{d \times d} \) given by \( H = \{ \partial^2 H/\partial x_i \partial y_j : 1 \leq i \leq d, 1 \leq j \leq d \} \). If \( H(x_0,y) \) is positive definite for all \( y \in \mathbb{R}^d \), then \( D_xH(x_0,y) \) is an injection of \( y \). Likewise, if \( H(x,y_0) \) is positive definite for all \( x \in \mathbb{R}^d \), then \( D_yH(x,y_0) \) is an injection of \( x \).

Appendix E. Proof of the discrete counterpart of the duality. In the proof of no duality gap for the continuous version of the \( \alpha \)D model, we assume the absolute continuity condition. Hence, it is necessary to provide another proof to establish the duality for the discrete version of the \( \alpha \)D model.

Lemma E.1. Let \( D^+, D^- \) be the sets of row-stochastic and column stochastic matrices respectively. Consider the primal problem (80) and the dual problem (105). There exists a set \( \{ \Gamma^+ := \{ \gamma_{i,j}^+ \} \in D^+, \Gamma^- := \{ \gamma_{i,j}^- \} \in D^-, \{ a_i \} \in \mathbb{R}^m, \{ b_j \} \in \mathbb{R}^n \} \) satisfying the condition in Eq. (108).

Proof. Consider the problem \( \max_{\Gamma^+ \in D^+} f(\Gamma^+) \), where \( f \) is defined in Eq. (90). Form the Lagrangian function
\[
f(\Gamma^+) = \sum_{i=1}^m \hat{a}_i \left( \sum_{j=1}^n \gamma_{i,j}^+ \right) - 1,
\]
where \( \{ \hat{a}_i \geq 0 : i = 1, \ldots, m \} \) are Lagrangian multipliers. From Lemma 4.3, all maximizers exist in \( E^+_1 \). Suppose that \( \Gamma^+ \) with entries \( \{ \hat{a}_{i,j} \} \) is one maximizer. By the Kuhn-Tucker condition, for all \( j \), we have
\[
\alpha K_{i,j}^{1/\alpha} \left( \sum_{i=1}^m \hat{a}_{i,j} K_{i,j}^{1/\alpha} \right)^{\alpha-1} = \hat{a}_i, \quad \text{if } \hat{a}_{i,j} > 0,
\]
and
\[
\alpha K_{i,j}^{1/\alpha} \left( \sum_{i=1}^m \gamma_{i,j}^+ K_{i,j}^{1/\alpha} \right)^{\alpha-1} \leq \hat{a}_i, \quad \text{if } \hat{a}_{i,j} = 0.
\]
Let \( b_j = (\sum_{i=1}^m \gamma_{i,j}^+ K_{i,j}^{1/\alpha})^{\alpha} \) for each \( j \), then
\[
K_{i,j} = \left( a_i / \alpha \right) b_j^{1-\alpha} \text{ for all } (i,j) \text{ with } \gamma_{i,j}^+ > 0,
\]
Figure 3. Illustration of images matching. The first two sub-figures (from left to right) represent image 1 and image 2. The third sub-figure shows the matching result from image 1 to image 2 using the αD problem. The fourth sub-figure is the matching result using the MK problem. The grid is plotted to reveal the fade-in and fade-out distortion.

Figure 4. Illustration of image morphing. The third row represents two given particle-sets \( P \) (left) and \( Q \) (right). The first row is the intermediate images at \( t = 1/6, 2/6, 3/6, 4/6, 5/6 \) (from left to right). The second row is the intermediate images convolved with a Gaussian kernel function at \( t = 1/6, 2/6, 3/6, 4/6, 5/6 \) (from left to right).

and

\[
(158) \quad K_{i,j} \leq (\hat{a}_i/\alpha)b_j^{1-\alpha} \text{ for all } (i,j) \text{ with } \tilde{\gamma}_{i,j}^+=0.
\]

On the other hand, let \( \tilde{\Gamma}^- = \phi(\tilde{\Gamma}^+) \) with entries \( \{\tilde{\gamma}_{i,j}^-\} \) given in Eq. (91). Using Eq. (157) and the definition of \( b_j \) above, Eq. (91) becomes \( b_j^{1/\alpha}\tilde{\gamma}_{i,j}^- = K_{i,j}^{1/\alpha}\tilde{\gamma}_{i,j}^- = \hat{a}_i b_j^{(1-\alpha)/\alpha}\tilde{\gamma}_{i,j}^- \), then \( b_j\tilde{\gamma}_{i,j}^- = \tilde{\gamma}_{i,j}^+(\hat{a}_i/\alpha) \). Hence, this set \( (\{\tilde{\gamma}_{i,j}^+, \{\tilde{\gamma}_{i,j}^-\}, \{\hat{a}_i/\alpha\}, \{b_j\}) \) satisfies the condition in Eq. (108).

REFERENCES

[1] R. D. Anderson and V. L. Klee, Jr., *Convex functions and upper semi-continuous collecti on*, Duke Math. J., 19 (1952), 349-357.

[2] Jean-David Benamou and Yann Brenier, *A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem*, Numer. Math., 84 (2000), 375–393.
Figure 5. Matching two artificial shapes shown in the first two sub-figures (from the left to the right). The third sub-figure shows the matching result from the first shape to the second shape. The fourth sub-figure shows the matching result from the second shape to the first shape. The parameter $\lambda = 0.01$ and $\sigma^2 = 0.4^2$ are used in Eq. (143).

[3] D. P. Bertsekas, “Nonlinear Programming,” Athena Scientific, 2003.
[4] P. J. Besl and N. D. McKay, A method for registration of 3-d shapes, IEEE Trans. Pattern Anal. Mach. Intell., 14 (1992), 239–256.
[5] F. L. Bookstein, Principal warps: Thin-plate splines and the decomposition of deformations, IEEE Trans. Pattern Anal. Mach. Intell., 11 (1989), 567–585.
[6] Alberto Borobia and Rafael Cantó, Matrix scaling: A geometric proof of Sinkhorn’s theorem, Linear Algebra Appl., 268 (1998), 1–8.
[7] Yann Brenier, Polar factorization and monotone rearrangement of vector-valued functions, Comm. Pure Appl. Math., 44 (1991), 375–417.
[8] P. Chen, A novel kernel correlation model with the correspondence estimation, JMIV, 39 (2011), 100–120.
[9] H. Chui and A. Rangarajan, A feature registration framework using mixture models, IEEE workshop on MMBIA, (2000), 190–197.
[10] ______, A new algorithm for non-rigid point matching, CVIU, 89 (2003), 114–141.
[11] I. L. Dryden and K. V. Mardia, “Statistical Shape Analysis,” Wiley Series in Probability and Statistics: Probability and Statistics, John Wiley & Sons, Ltd., Chichester, 1998.
[12] L. C. Evans and W. Gangbo, Differential equations methods for the Monge-Kantorovich mass transfer problem, Mem. Amer. Math. Soc., 137 (1999), viii+66 pp.
[13] Lawrence C. Evans, Partial differential equations and Monge-Kantorovich mass transfer, Current developments in mathematics, 1997, (Cambridge, MA), Int. Press, Boston, MA, (1999), 65–126.
[14] ______, “Partial Differential Equations,” Second edition, Graduate Studies in Mathematics, 19, American Mathematical Society, Providence, RI, 2010.
[15] Lawrence C. Evans and Ronald F. Gariepy, “Measure Theory and Fine Properties of Functions,” Studies in Advanced Mathematics, CRC Press, Boca Raton, FL, 1992.
[16] Olivier Faugeras and Gerardo Hermosillo, Well-posedness of two nonrigid multimodal image registration methods, SIAM J. Appl. Math., 64 (2004), 1550–1587.
[17] Wilfrid Gangbo, An elementary proof of the polar factorization of vector-valued functions, Arch. Rational Mech. Anal., 128 (1994), 381–399.
[18] A. L. Gibbs and F. E. Su, On choosing and bounding probability metrics, Intl. Stat. Rev., 70 (2002), 419–435.
[19] J. Glaunes, A. Trouve and L. Younes, Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching, CVPR, 2 (2004), 712–718.
[20] S. Granger and X. Pennec, Multi-scale EM-ICP: A fast and robust approach for surface registration, ECCV, 4 (2002), 418–432.
[21] S. Haker, L. Zhu, A. Tannenbaum and S. Angenent, Optimal mass transport for registration and warping, International Journal of Computer Vision, 60 (2004), 225–240.
[22] Roger A. Horn and Charles R. Johnson, “Matrix Analysis,” Cambridge University Press, Cambridge, 1985.
[23] B. Jian and B. C. Vemuri, A robust algorithm for point set registration using mixture of Gaussians, ICCV, 2 (2005), 1246–1251.
Figure 6. Matching two photos of a building (Top left). Pointsets (marked in red) with cardinality 290 and 660 are selected by the edge detector (Top right). The bottom left shows 20 randomly selected matching pairs. The bottom right shows the deformation of the matching result after 10 iterations in the block coordinate ascent method. The parameter $\sigma^2 = 0.2^2$ is used in Eq. (145).

[24] S. C. Joshi and M. I. Miller, *Landmark matching via large deformation diffeomorphism*, IEEE Image Proc., 9 (2000), 1357–1370.
[25] Jürgen Jost and Xianqing Li-Jost, “Calculus of Variations,” Cambridge Studies in Advanced Mathematics, 64. Cambridge University Press, Cambridge, 1998.
[26] Thomas Kaijser, *Computing the Kantorovich distance for images*, J. Math. Imaging Vision, 9 (1998), 173–191.
[27] L. V. Kantorovich, *On the transfer of masses*, Dokl. Akad. Nauk. SSSR, 37 (1942), 227–229.
[28] M. Kass, A. Witkin and D. Terzopoulos, *Snake: Active contour models*, International Journal of Computer Vision, 1 (1988), 321–331.
[29] M. Leordeanu and M. Hebert, *A spectral technique for correspondence problems using pairwise constraints*, ICCV, (2005), 1482–1489.
Figure 7. Matching two photos of a baby (Top left). Point-sets (marked in red) with cardinality 285 and 342 are selected by the edge detector (Top right). The bottom left shows 15 randomly selected matching pairs. The bottom right shows the deformation of the matching result after 15 iterations in the block coordinate ascent method. Here 16 grid points are selected as control points. The parameters $\lambda = 0.01$, $\sigma^2 = 0.2^2$ (for $K$), and $\sigma^2 = 0.3$ (for $U$) are used in Eq. (143).

[30] B. Luo and E. R. Hancock, A unified framework for alignment and correspondence, Computer Vision and Image Understanding, 92 (2003), 26–55.
[31] B. Ma, R. Narayanan, H. Park, A. O. Hero, P. H. Bland and C. R. Meyer, Comparing pairwise and simultaneous joint registrations of decorrelating interval exams using entropic graphs, Information Processing in Medical Imaging, 4584 (2008), 270–282.
[32] Robert J. McCann, Existence and uniqueness of monotone measure-preserving maps, Duke Math. J., 80 (1995), 309–323.
[33] G. McNeil and S. Vijayakumar, A probabilistic approach to robust shape matching, IEEE ICIP, (2006), 937–940.
[34] M. I. Miller and L. Younes, Group actions, homeomorphism, matching: A general framework, Int. J. Comput. Vis., 41 (2001), 61–84.
[35] O. Museyko, M. Stiglmayr, K. Klamroth and G. Leugering, On the application of the Monge-Kantorovich problem to image registration, SIAM J. Imaging Sci., 2 (2009), 1068–1097.
[36] Yurii A. Neretin, On Jordan angles and the triangle inequality in Grassmann manifolds, Geom. Dedicata, 86 (2001), 81–92.
[37] J. Rabin, J. Delon and Y. Gousseau, A statistical approach to the matching of local features, SIAM J. Imaging Sci., 2 (2009), 931–958.
[38] Svetlozar T. Rachev, “Probability Metrics and the Stability of Stochastic Models,” Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons, Ltd., Chichester, 1991.
[39] A. Rényi, On measures of entropy and information, Proc. 4th Berkeley Symp. Math. Stat. and Prob., Vol. 1, Univ. California Press, Berkeley, Calif., (1961), 547–561.
[40] R. Tyrrell Rockafellar, “Convex Analysis,” Princeton Mathematical Series, No. 28, Princeton University Press, Princeton, N.J., 1970.
[41] Y. Rubner, C. Tomasi and L. J. Guibas, The earth mover’s distance as a metric for image retrieval, Int. J. Comput. Vis., 40 (2000), 99–121.
[42] G. L. Scott and H. C. Longuet-Higgins, An algorithm for associating the features of two images, Proceedings of the Royal Society London: Biological Sciences, 244 (1991), 21–26.
[43] T. Sebastian, P. Klein and B. Kimia, On aligning curves, IEEE Trans. Pattern Anal. Mach. Intell., 5 (2003), 116–125.
[44] I. K. Sethi and R. Jain, Finding trajectories of feature points in a monocular image sequence, IEEE Trans. Pattern Anal. Mach. Intell., 9 (1987), 56–73.
[45] Richard Sinkhorn, A relationship between arbitrary positive matrices and doubly stochastic matrices, Ann. Math. Statist., 35 (1964), 876–879.
[46] D. W. Thompson, “On Growth and Form,” Cambridge University Press, Cambridge, 1917.
[47] A. Trouvé, Diffeomorphism groups and pattern matching in image analysis, Int. J. Comput. Vis., 28 (1998), 213–221.
[48] Y. Tsin and T. Kanade, A correlation-based approach to robust point set registration, ECCV, 3 (2004), 558–569.
[49] Zhuowen Tu, Songfeng Zheng and Alan Yuille, Shape matching and registration by data-driven EM, Computer Vision and Image Understanding, 109 (2008), 290–304.
[50] S. Ullman, “The Interpretation of Visual Motion,” MIT Press, Cambridge, MA, 1979.
[51] Grace Wahba, “Spline Models for Observational Data,” CBMS-NSF Regional Conference Series in Applied Mathematics, 59, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.
[52] F. Wang, B. C. Vemuri, A. Rangarajan and S. J. Eisenschenk, Simultaneous nonrigid registration of multiple point sets and atlas construction, IEEE PAMI, 30 (2008), 2011–2022.
[53] J. Warga, Minimizing certain convex functions, J. Soc. Industr. Appl. Math., 11 (1963), 588–593.
[54] W. M. Wells, Statistical approaches to feature-based object recognition, International Journal of Computer Vision, 22 (1997), 63–98.
[55] M. Werman, S. Peleg and A. Rosenfeld, A distance metric for multi-dimensional histograms, Comp. Vis. Graphics Image Proc., 32 (1985), 328–336.
[56] Laurent Younes, Computable elastic distances between shapes, SIAM J. Appl. Math., 58 (1998), 565–586 (electronic).
[57] Laurent Younes, Peter W. Michor, Jayant Shah and David Mumford, A metric on shape space with explicit geodesics, Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl., 19 (2008), 25–57.
[58] Z. Zhang, Iterative point matching for registration of free-form curves and surfaces, International Journal of Computer Vision, 13 (1994), 119–152.
[59] L. Zhu, Y. Yang, S. Haker and A. Tannenbaum, An image morphing technique based on optimal mass preserving mapping, IEEE Trans. Image Processing, 16 (2007), 1481–1495.
[60] C. L. Zitnick and T. Kanade, A cooperative algorithm for stereo matching and occlusion detection, IEEE Trans. Pattern Anal. Mach. Intell., 22 (2000), 675–684.
[61] B. Zitova and J. Flusser, Image registration methods: A survey, Image and Vis. Compu., 21 (2003), 977–1000.

Received May 2009; revised June 2010.

E-mail address: pengwen@math.ntu.edu.tw
E-mail address: gui@math.uconn.edu