Graphical Abstract

Distributional barycenter problem through data-driven flows

Esteban G. Tabak, Giulio Trigila, Wenjun Zhao

Highlights

Distributional barycenter problem through data-driven flows

Esteban G. Tabak, Giulio Trigila, Wenjun Zhao

- A new algorithm for the solution of the optimal transport barycenter problem on manifolds is proposed.
- The algorithm allows the adoption of very general, non necessarily pairwise, cost functions.
- The algorithm overcomes the adversarial nature of the barycenter problem.
- A new cost function penalizing non-isotropic maps is introduced.
- The efficacy of the method is illustrated on synthetic examples and on the MNIST data set.

Distributional barycenter problem through data-driven flows

Esteban G. Tabak[a], Giulio Trigila[b], Wenjun Zhao[a]

[a] *Courant Institute of Mathematical Sciences, 251 Mercer Street, New York, 10012, NY, USA*
[b] *Baruch College, CUNY, 55 Lexington avenue, New York, 10010, NY, USA*

Abstract

A new method is proposed for the solution of the data-driven optimal transport barycenter problem and of the more general distributional barycenter problem that the article introduces. The method improves on previous approaches based on adversarial games, by slaving the discriminator to the generator, minimizing the need for parameterizations and by allowing the adoption of general cost functions. It is applied to numerical examples, which include analyzing the MNIST data set with a new cost function that penalizes non-isometric maps.

Keywords: Optimal transport, barycenter problem, pattern visualization, filtering, adversarial optimization

1. Introduction

Optimal transport and the related Wasserstein barycenter problem have undergone rapid development during the last ten years, with a particular focus on applications to the analysis of data and machine learning [13], ranging from gene expression [14] to economics [15]. Procedures based on optimal transport have been used for density and conditional density estimation [1, 2], data augmentation [3], image classification [4, 5, 6], computer vision [7, 8, 9, 10], factor discovery [12] and data imputation [11].

Given two probability distributions ρ and μ, the optimal transport problem ([16, 17, 18]) seeks the map T with minimal cost among those satisfying the push forward condition $\mu = T_{\#}\rho$, with a cost function determined by the application

at hand. In the barycenter problem, a conditional distribution $\rho(x|z)$ is mapped to a single, unknown distribution $\mu(y)$, which minimizes the sum over z of the transportation cost from ρ to μ.

Some recent methodologies for the numerical solution of the data-based barycenter problem apply only to a canonical cost function, the squared Euclidean distance between points. The advantage of restricting attention to this or similar cost functions is that one can fully characterize the solution in terms of a convex potential, thus bypassing the need to actually perform a total cost minimization. However, a number of applications call for more general, field-specific cost functions. Consider for illustration the following instances:

1. The distributions $\rho(x|z)$ underlying real world data are often defined on high dimensional spaces, yet they concentrate near a manifold $\mathcal{M}$ of dimension m smaller than the dimension d of the ambient space. Exploiting this geometric property of the data reduces the complexity of the map, which should be a function of m rather than d. The geometry underlying the data encodes the nature of a system, so models consistent with it have a more meaningful data correlation structure. One way to carry out this program is to use a cost function that penalizes maps T moving data outside the manifold $\mathcal{M}$. Even in low dimensions, data often concentrates on or near a non-Euclidean sub-manifold, such as the Earth's surface for climate-related data.
2. The introduction of a new cost function is often dictated purely by properties that one wishes to impose on the barycenter. We introduce in section 6, in the context of an application to the MNIST data set, a cost function favoring isometric maps. This results in a smoother, more interpretable barycenter of hand-written digits, modeled as distributions in pixel space.

This last example goes beyond the realm of optimal transport, as the cost function does not adopt the form of the expect value of a pairwise cost $c(x, y)$. We call such extensions of the Wasserstein barycenter problem, *distributional barycenter problems*. They can be used to enforce problem-dependent desirable

conditions on the conditional maps, such as proximity to prescribed priors.

The methodology for the solution of the data-driven distributional barycenter problem proposed in this article can be used with general cost functions. It improves significantly over previous approaches to the barycenter problem based on adversarial games ([19, 2, 12]). The latter have two players: one that proposes cost-minimizing maps through time-evolving flows, and another that builds test functions to enforce the push-forward condition. The new approach slaves the test-functions to the flows, thus making the latter self-driven. Moreover, these flows are essentially non parametric, with a kernel's bandwidth as their single parameter.

1.1. Prior work

While optimal transport on Riemannian manifolds has been broadly studied from an analytical perspective ([20, 18]), few algorithms have been proposed for its numerical solution. Most are based on a regularization of optimal transport [7, 21] or are specific to particular manifolds [22]. The work in [23] finds a smooth interpolation of densities on discrete surfaces using the dynamical approach of Benamou and Brenier [24]. This approach, though grounded as ours on gradient flows, uses a different flow and requires the knowledge of the densities to be transported rather than samples thereof.

The optimal transport barycenter problem and its dual formulation were introduced in [25]. One of the first proposed methodologies for the numerical solution of the dual problem, a saddle point optimization problem, appeared in [26], where a modification of linear programming was adopted to compute the potentials associated to the optimal maps. Here we propose an alternative derivation of the formulation in [26], better suited for the discussion leading to the algorithm proposed in section 3.

1.2. Original contribution

The main contribution of this paper is an original methodology for the solution of the barycenter problem under general cost functions, through a time-dependent flow that pushes the marginal distributions $\rho(x|z)$ to their barycenter

$\mu(y)$. Its main novel aspects are that the maps require no parameterization and that the test function enforcing that all conditional distributions be mapped to the same barycenter is slaved to the maps. This yields a minimization problem with one constraint rather than a saddle point problem, equivalent to a minimization problem with infinitely many constraints. This reduction is achieved by proposing a specific –but sufficient– form for the test function F, therefore bypassing the adversarial formulation in which a Lagrangian is minimized over maps and maximized over test functions.

A numerical implementation based on a variation of the penalty method ([27]) results in a method that builds arbitrarily complex maps through flows and permits the adoption of very general cost functions. In particular, a new cost function is proposed that penalizes non-isometric conditional maps, a natural way to minimize data distortion.

1.3. Organization of the article

Section 2 reviews the barycenter problem and its dual. Section 3 proposes two specific test functions, yielding two alternative formulations, section 4 develops their data driven version and section 5 introduces a penalty method for their numerical solution. Section 6 contains numerical experiments on both synthetic data and the MNIST data set, using various test and cost functions. In particular, subsection 6.2 introduces a new cost penalizing maps far from isometric, and subsection 6.4 uses the barycenter problem to recover a hidden signal behind a time series defined on a sphere.

2. Data-driven distributional barycenter problem

Given a conditional probability distribution $\rho(x|z)$, the optimal transport barycenter problem seeks a target $\mu(y)$ and z-dependent maps $T(x,z)$ from ρ to μ with minimal total transportation cost:

$$\min_T \int\int c(x,T(x,z))\rho(x,z)dxdz, \quad s.t. \quad \forall z \; T\#\rho(\cdot|z) := \rho_T(\cdot|z) = \mu. \qquad (1)$$

Examples of cost functions are p-norms, as in the canonical cost $c(x,y) = \frac{1}{2}||x-y||_2^2$, and the squared geodesic distance on a manifold.

We will consider the more general distributional barycenter problem

$$\min_T C\left(T(x,z),\rho\right), \quad s.t. \quad \forall z\ T\#\rho(\cdot|z) := \rho_T(\cdot|z) = \mu, \tag{2}$$

where C can adopt forms different from the expected value of a pairwise cost function $c(x,T(x,z))$ of optimal transport. Examples of such more general costs include the Fermat distance introduced in [28] and a cost function introduced below to penalize deviations from isometry. For concreteness and to enable comparison with prior work, we describe below our methodology in the context of regular pairwise costs $c(x,y)$, explaining afterwards how it extends, quite straightforwardly, to the general case. The only constraint on $C(T,\rho)$ is that it must admit a data-based formulation, i.e. its dependence on ρ must be translatable into an expression involving only samples thereof. For the regular pairwise cost, such formulation simply replaces expected values by empirical means over the data points.

As the pushforward condition expresses the requirement that the random variable $y = T(x,z)$ be independent of z, it can be rewritten without explicit reference to the unknown barycenter μ. If z and $y = T(x,z)$ are independent, then $\rho_T(y,z) = \mu(y)\nu(z)$, so

$$\int F(y,z)\rho_T(y,z)dydz = \int F(T(x,z),z)\rho(x,z)dxdz = 0$$

for every test function F satisfying $\int F(y,z)\nu(z)dz = 0$. The converse is also true, leading to the minimax formulation of the barycenter problem:

$$\begin{cases} \min_T \max_F \int c(x,T(x,z))\rho(x,z)dxdz + \int F(T(x,z),z)\rho(x,z)dxdz \\ \forall y \ \int F(y,z)\nu(z)dz = E_z[F(y,\cdot)] = 0 \end{cases} \tag{3}$$

A comparison between (3) and the formulation in [25] reveals that the test function F is the Lagrange multiplier $\psi(y,z)$ of the dual Kantorovich problem.

The constraint in (3) can be satisfied automatically by subtracting from F

its expected value $E_z[F]$, which yields the unconstrained variational problem

$$\min_T \max_F L = \int c(x, T(x,z))\rho(x,z)dxdz + \int (F(y,z) - E_z[F])\rho_T(y,z)\, dydz. \quad (4)$$

The first integral in (4) corresponds to the cost function of optimal transport, to be extended below to far more general costs. For future reference, we will denote this integral as L_C, and the second integral, designed to test the fulfillment of the pushforward condition, as L_F:

$$L_C = \int c(x, T(x,z))\rho(x,z)dxdz, \quad L_F = \int (F(y,z) - E_z[F])\rho_T(y,z)\, dydz.$$

As noted in [1], the dual Kantorovich problem is a natural starting point for a data driven formulation of optimal transport. In particular, (4) has two main advantages over (1): the unknown barycenter μ does not appear explicitly, and the objective function is a sum of expected values, which can be replaced by their empirical counterpart

$$\min_T \max_F \frac{1}{N} \sum_i \left[c(x_i, T(x_i, z_i)) + F(T(x_i, z_i), z_i) - \frac{1}{N} \sum_j F(T(x_i, z_i), z_j) \right] \quad (5)$$

when only samples (x_i, z_i) of $\rho(x, z)$ are available.

3. Two choices for the test function $F(y, z)$

This section introduces a new algorithm for the numerical solution of the optimization problem in (5). We first define an the evolution equation for T through the gradient descent of L:

$$\dot{T} = -\left.\frac{\delta L}{\delta T}\right|_{x,z} = -\left[\nabla_y c(x,y) + \nabla_y F(y,z)\right]\rho(x,z), \quad y = T(x,z). \quad (6)$$

Notice that, for the canonical squared-distance cost, the first order optimality condition $\dot{T} = 0$ recovers the well-known relationship between the optimal map T^* and the optimal potential F^*, i.e. $x = T^*(x,z) - \nabla_y F^*(y,z)$.

Thus the evolution of the map T is defined in terms of the test function F. Since the role of F is to penalize any dependence of $\rho_T(y|z)$ on z, it is natural to

think that it should be able to resolve the family of distributions $\rho_T(y|z)$. The following two propositions clarify this point. We will use them to reformulate the problem in (5) so that the adversarial game played by F and T is reduced to a pure minimization algorithm over T.

Proposition 1. *If $F(y,z) = \rho_T(y|z)$ then the second term (L_F) in (4) is always strictly positive unless $\rho_T(y|z)$ is independent of z.*

Proof. We can rewrite L_F as

$$L_F = \int \left[F(y,z)\rho_T(y|z)\nu(z)dz - \int F(y,z)\bar{\rho}_T(y)\nu(z)dz \right] dy, \tag{7}$$

where $\bar{\rho}_T(y) = \int \rho_T(y|w)\nu(w)dw$. Substituting $F(y,z) = \rho_T(y|z)$ yields

$$L_F = \int \left(E_z[\rho_T^2(y|\cdot)] - E_z[\rho_T(y|\cdot)]^2 \right) dy. \tag{8}$$

By Jensen's inequality, the integrand is strictly positive for all values of y unless $\rho_T(y|z)$ does not depend on z. □

This result suggests adopting $F(y,z) = \lambda\rho_T(y|z)$, a test function that evolves as the conditional distributions $\rho(x|z)$ are pushed forward toward their barycenter μ. With this choice, the infinite dimensional maximization of (4) over F reduces to the maximization over the scalar λ:

Problem 1.

$$\min_T \max_\lambda \int c(x, T(x,z))\rho(x,z)dxdz + \\ + \lambda \int \left[\rho_T(x|z) - \int \rho_T(x|w)\nu(w)dw \right] \rho_T(x,z)dxdz.$$

Section 5 discusses in detail how to solve numerically Problem 1. Here we just point out that: 1) At all times, the information we have on $\rho_T(x,z)$ consists of samples thereof, i.e. the points $y^i = T(x^i, z^i)$ that have been transported by T, and 2) Since L_F is non-negative, the maximization over λ can be implemented through a penalty method, reducing Problem 1 to a pure minimization problem.

We show next that, alternatively, we can choose as test function $F(y,z)$ the product of two related functions, depending on y and z respectively:

Proposition 2. *If $F(y,z) = f(y)g(z)$ where $g(z) = \int f(y)\rho_T(y|z)dy$, then L_F in (4) is strictly positive unless the expected value of $f(y)$ under $\rho_T(y|z)$ is independent of z.*

Proof. It is not difficult to see that, with F given as above, one has

$$L_F = \int g(z)^2\nu(z)dz - \left(\int g(z)\nu(z)dz\right)^2. \tag{9}$$

By Jensen's inequality, (9) is always non-negative, vanishing only if g is independent of z. □

Under Proposition 2 we can relax (4) into

Problem 2.

$$\min_T \max_f \int c(x,T(x,z))\rho(x,z)dxdz + \int g(z)^2\nu(z)dz - \left(\int g(z)\nu(z)dz\right)^2$$

where $g(z) = \int f(y)\rho_T(y|z)dy$.

This formulation enforces the independence of $\rho_T(y|z)$ from z in a weak sense, with test function $f(y)$. For instance, restricting f to linear functions $f = \lambda y$ enforces that the conditional mean $\bar{y}(z)$ of $\rho_T(y|z)$ be independent of z. Notice that, in this case and under the canonical cost, the descent equation (6) implies that the map T is a z-dependent rigid translation, precisely the minimal family of maps able to remove conditional means.

These considerations suggest a preconditioning procedure whereby, rather than seeking the full barycenter from the start, one first limits the family of test functions and maps, yielding a less detailed but faster procedure that brings the $\rho(x|z)$ closer to each other. In particular, one can perform a preconditioning whereby only the conditional mean of $\rho(x|z)$ is removed, through a z-dependent rigid translation. Under the canonical cost, performing this preconditioning and subsequently computing the barycenter of the resulting push-forward distributions, results in the same barycenter that one would have found directly from the original ones. The proof, which extends arguments in [29] to the barycenter problem, is the content of the following proposition.

Proposition 3. *Consider the following, two-stage procedure for finding the barycenter of the conditional distributions $\rho(x|z)$ under the canonical cost $c(x,y) = \frac{1}{2}\|x-y\|^2$. First restrict the maps to the z-dependent rigid translations*

$$w = T_1(x,z) = x + \bar{x} - \bar{x}(z),$$

which make the conditional means of the resulting random variable W match. Then find the full barycenter of the resulting conditional distributions $\mu_1(w|z)$ through a map $y = T_2(w,z)$. Then the composition of the two maps,

$$y = T(x,z) = T_2\left(T_1(x,z), z\right)$$

solves the original barycenter problem.

Proof. Clearly the distribution $\mu(y)$ is independent of z, since μ is the barycenter of the $\mu_1(w|z)$. To prove optimality, it is enough [25, 30] to show that

1. T is the gradient of a convex function:

$$T(x,z) = \nabla_x \phi(x,z), \quad \phi(:,z) \text{ convex for all } z,$$

2. every point y is the geometrical barycenter of its pre-images under $T(x,z)$,

$$\forall y \; E_z\left[T^{-1}(y,z)\right] = y.$$

Since μ is the barycenter of the $\mu_1(w|z)$, both properties hold for T_2:

$$T_2(w,z) = \nabla_w \psi(w,z), \quad \psi(:,z) \text{ convex for all } z, \quad \forall y \; E_z\left[T_2^{-1}(y,z)\right] = y.$$

Then $T(x,z) = T_2\left(x + \bar{x} - \bar{x}(z), z\right) = \nabla_x \phi(x,z)$, where $\phi(x,z) = \psi\left(x + \bar{x} - \bar{x}(z), z\right)$ is convex in x for all values of z. Also $T^{-1}(y,z) = T_2^{-1}(y,z) + \bar{x}(z) - \bar{x}$, so

$$E_z\left[T^{-1}(y,z)\right] = E_z\left[T_2^{-1}(y,z)\right] + \bar{x} - \bar{x} = y,$$

concluding the proof.

□

Two natural questions arise from proposition 3: can one perform pre-conditioning under more general cost functions, and can one implement richer pre-conditioners that bring the $\rho(x|z)$ closer to each other than merely translating them so that their conditional means match. To answer these questions, notice that proposition 3 allows one to start the follow-up barycenter problem directly from the $\mu_1(w|z)$ resulting from the pre-conditioning map, without any reference to the original random variable X. However, one does know the conditional pairing of X and W, i.e. the map $w = T_1(x, z)$ or, in the data-driven case, the point x_i that each w_i originated from. It follows that one can perform pre-conditioning under any cost function $C(T, \rho)$ and with any family of test functions F, provided that, in the subsequent full barycenter problem, though starting from the $W = T_1(X, z)$, one computes the cost C in terms of the original X:

$$C_2\left(T, \mu_1\right) = C\left(T * T_1, T_1^{-1} \# \mu_1\right).$$

In the data-driven setting developed below, this formula simply translates into using x_i in lieu of w_i in C.

4. Data-driven formulations

This section discusses the numerical representation of $\rho(y|z)$ and its use for implementing data-driven versions of Problems 1 and 2.

4.1. Data driven Problem 1

The map $y = T(x, z)$ is built from the composition of near-identity maps which yield, at each time-step of the algorithm, a current state of the map and a corresponding current conditional density $\rho_T(y|z)$. This conditional density, which evolves from $\rho(y|z)$ to $\mu(y)$, is known at all times through the points $y^i = T(x^i, z^i)$. A natural way to estimate $F(y, z) = \rho_T(y|z)$ from these samples is through a conditional kernel density estimation (CKDE) in the Nadaraya-Watson form ([31, 32]):

$$F(y, z_k) = \rho_T(y|z_k) \approx \frac{\sum_i \mathcal{K}_a(y, y_i)\mathcal{K}_b(z_k, z_i)}{\sum_j \mathcal{K}_b(z_k, z_j)} = \sum_{i=1}^{N} \mathcal{K}_a(y, y_i) Z_{ik}. \tag{10}$$

The kernel functions $\mathcal{K}_a(y, y_i)$, nonnegative and normalized so as to integrate to one, have centers y_i and bandwidth a –the algorithm's only free parameter, other than the choice of the kernels themselves, for which isotropic Gaussians were adopted in all the numerical examples below. The matrix $Z \in \mathbb{R}^{N\times N}$ is a normalized version of similar kernels in z-space:

$$Z_{ik} = \frac{\mathcal{K}_b(z_k, z_i)}{\sum_{j=1}^{N} \mathcal{K}_b(z_k, z_j)}. \tag{11}$$

With this choice for F, the empirical version of the term in square brackets in Problem 1 adopts the form

$$\rho_T(y_l|z_l) - \mathbb{E}_z\rho_T(y_l|z) \approx \sum_i \mathcal{K}_a(y_l, y_i)\left[Z_{il} - \frac{1}{N}\sum_k Z_{ik}\right] = \sum_i \mathcal{K}_a(y_l, y_i)C_{il}, \tag{12}$$

where the N by N matrix

$$C_{il} = Z_{il} - \frac{1}{N}\sum_k Z_{ik} \tag{13}$$

can be precomputed at the onset of the procedure, since the values of z_i remain unchanged throughout. Then the complete data driven formulation of Problem 1 adopts the simple form

$$\min_y \max_\lambda \sum_i c(x_i, y_i) + \lambda \sum_{i,l} \mathcal{K}_a(y_l, y_i)C_{il}. \tag{14}$$

4.2. Data driven Problem 2

In order to evaluate (9) from sample points, we rewrite $g(z)$ in the form

$$g(z) = \int f(y)\rho_T(y|z)dy = \int f(y)\frac{\rho_T(y,z)}{\nu(z)}dy = \int f(y)\frac{\rho_T(y,w)}{\nu(w)}\delta(w-z)dydw, \tag{15}$$

and propose the mollification

$$\delta(w-z) \approx \mathcal{K}_b(w,z), \quad \nu(w) = \frac{1}{n}\sum_j \mathcal{K}_b\left(w, z_j\right), \tag{16}$$

with a positive kernel $\mathcal{K}_b$ with bandwidth b that integrates to one. Then

$$g(z) \approx \sum_k f\left(y_k\right)\frac{\mathcal{K}_b(z, z_k)}{\sum_l \mathcal{K}_b\left(z_l, z_k\right)}, \tag{17}$$

which we can substitute in the test function $F(y, z)$ according to the proposal in Proposition 2, i.e. $F(y, z) = \lambda f(y)g(z)$. The resulting test component L_F of the Lagrangian is

$$L_F = \sum_i \left[F(T(x_i, z_i), z_i) - \frac{1}{N} \sum_j F(T(x_i, z_i), z_j) \right] =$$

$$\lambda \sum_i f(y_i) \sum_k f(y_k) \left[Z_{ki} - \frac{1}{N} \sum_j Z_{kj} \right] = \lambda \sum_{i,k} f(y_i) f(y_k) C_{ki}, \quad (18)$$

where $y_i = T(x_i, z_i)$ and the matrices Z_{ik} and C_{il} are those defined in (11) and (13). The overall data-driven version of Problem 2 with fixed test function f then becomes

$$\min_y \max_\lambda \sum_i c(x_i, y_i) + \lambda \sum_{i,k} f(y_i) f(y_k) C_{ki}. \quad (19)$$

The choice of the function f specifies a relaxation of the pushforward condition, with $f(y) = y^l$ (y^l here stands for the lth component of y) corresponding to moving each conditional mean $\bar{x^l}(z)$ to the mean $\bar{y^l}$ of the barycenter. To match the conditional means of all components y^l, as well as to enforce other moments, we can choose f to be a vectorial function whose entries can be chosen, for instance, as a polynomial basis: $f(y) = [f_1(y), f_2(y), .., f_m(y)]$, corresponding to the solution of

$$\min_y \max_\lambda \sum_i c(x_i, y_i) + \lambda \sum_{i,k,l} f_l(y_i) f_l(y_k) C_{ki}. \quad (20)$$

Notice that we do not need an independent factor λ_l for each f_l, as each term $\sum_{i,k} f_l(y_i) f_l(y_k) C_{ki}$ is independently non-negative, vanishing only when the expected value of f_l agrees for all values of z (We have proved this in proposition 2 for the problem posed in terms in distributions, and we will prove it below for the sample-based problem.) We may, however, weight each f_l differently if desired. For instance, we might want to start with most of the weight on the linear components of f, so as to enforce the agreement of the conditional means, then slowly increase the weight of the quadratic components, to match all conditional covariances, and add more terms, either higher order polynomials or

localized features, for a more detailed fulfillment of the pushforward condition. However, we have found empirically that simply pre-conditioning first with a linear f is enough to speed up the subsequent convergence of problem 1 in all of its generality, bypassing the need for the "continuous preconditioning" that the procedure just described would entail.

4.3. An alternative conditional density estimator

Formula (11) for the matrix Z_{ik} is not the only choice that makes (10) a robust conditional density estimator. The core requirements for Z are:

1. The entries Z_{ik} must be nonnegative and add up to zero row-wise:

$$Z_{ik} \geq 0, \quad \sum_i Z_{ik} = 1,$$

   to guarantee that the estimated $\rho_T(y|z_k)$ is positive and integrates to one.

2. Z_{ik} must be large when z_i and z_k are close to each other, and small when they are far away. This follows from conceptualizing (10) as a regular kernel density estimation which has the various centers y_i weighted by Z_{ik}. Then Z_{ik} must provide a measure of how relevant y_i is for an estimation of $\rho_T(y|z_k)$, i.e. how close the z_i associated with y_i is to z_k. The notion of closeness is, of course, problem dependent. For instance, for categorical factors z, a choice for Z_{ik} vanishes whenever $z_i \neq z_k$.

The particular form (11) for Z_{ik} satisfies these properties, and it leads to robust and accurate numerical results in all examples that we have tried. Yet the resulting matrix Z_{ik} is asymmetric, as only its rows, not its columns, are normalized. For reason that the following subsection will clarify, we prefer a matrix Z that is symmetric and positive definite. Since a symmetric matrix Z with nonnegative entries whose rows add up to one is necessarily bi-stochastic, a natural candidate is the unique bi-stochastic matrix $\tilde{Z}$ that derives from the symmetric and positive Kernel matrix $K_{ik} = \mathcal{K}_b(z_k, z_i)$ through Sinkhorn's factorization: $\tilde{Z} = DKD$, where D is a diagonal matrix with positive diagonal entries. Since K is positive definite, so is $\tilde{Z}$, which also satisfies the required

properties for (10) to provide a consistent conditional density estimator, and is in fact better balanced than Z, in the sense that all points y_i have the same total weight (For points whose z_i is an outlier, this weight concentrates mostly in self-estimation, while for points with z_i in the core of the z-distribution, the weights are distributed among neighboring points in z, not necessarily y.)

We have found the numerical results with $\tilde{Z}$ and Z to be nearly indistinguishable. Since $\tilde{Z}$ comes with better theoretical guarantees, we use $\tilde{Z}$ in the remaining of the article and in the numerical examples, renaming it Z to avoid notational clumsiness.

The kernel-based matrix Z_i^j is suitable for continuous factors z with a notion of distance among points. Clearly, for categorical factors z, it should be replaced by the simpler

$$Z_i^j = \begin{cases} \frac{1}{N_i} & \text{for } z_i = z_j \\ 0 & \text{otherwise,} \end{cases} \qquad N_i = |\{z : z = z_i\}|,$$

also bi-stochastic, which simply discriminates among classes. To avoid repeating proofs and arguments, this can be considered as a particular case of the kernel-based Z with vanishing small bandwidth, so that different z_i do dot interact.

4.4. Positivity of L_F in the data-driven problem

We saw in section 3 that, for the two particular choices of the test function F corresponding to problems 1 and 2, the L_F in (4) is strictly positive unless all the marginals $\rho_T(y|z)$ agree. The positivity of L_F allows us to pre-multiply it by a positive scalar λ, replacing the minimax formulation by a penalized minimization. We show here that the positivity of L_F also holds in its data-driven version.

Proposition 4. *If the kernel matrices $\mathcal{K}_a$ and $\mathcal{K}_b$, with entries $\mathcal{K}_a(y_k, y_i)$ and $\mathcal{K}_b(z_k, z_i)$ respectively, are non-negative definite, then the test component L_F of both (14) and (19) is non negative.*

Proof. Notice that it is enough to show that the matrix C given by (13) is non-negative definite, i.e. that

$$\forall x,\ x^T C x \geq 0. \tag{21}$$

This sufficiency of (21) for (19) is obvious, as its test component is precisely a sum of terms of this form, but (21) is also sufficient for (14), since its test component is the inner product of C and $\mathcal{K}_a$, and the inner product of two non-negative definite matrices is a non-negative number, even when only one of them is symmetric (C, in general, is not.) To prove (21), write

$$\begin{aligned}
\sum_{i,j} x_i C_{ij} x_j &= \sum_{i,j} x_i Z_{ij} x_j - \frac{1}{N} \sum_{k,i,j} x_i Z_{ik} \left(x_j - x_k + x_k\right) \\
&= -\frac{1}{N} \sum_{k,i,j} x_i Z_{ik} \left(x_j - x_k\right) \\
&= \sum_{k,i} x_i Z_{ik} \left(x_k - \bar{x}\right) \\
&= \sum_{k,i} \left(x_i - \bar{x}\right) Z_{ik} \left(x_k - \bar{x}\right) \geq 0,
\end{aligned}$$

as the matrix Z (i.e. $\tilde{Z}$) is non-negative definite by construction.

□

4.5. Extension to general cost functions

We have so far restricted the cost component L_C of the objective function to the expected value of a pairwise cost $c(x, T(x, z))$, as pertains optimal transport. However, it is clear from the data-based formulations derived that the only requirement one must impose on $L_C = C(T, \rho)$ is that ρ should only appear through the expected value of functions, which can be replaced by their empirical counterpart when only samples (x_i, z_i) of ρ are known. Thus, for instance, in lieu of the pairwise cost $L_C = \int c(x, T(x, z))\rho(x, z)dxdz$, one may propose cost functions involving two points and their images under a common factor z,

$$L_C = \int c\left(x_1, T\left(x_1, z\right), x_2, T\left(x_2, z\right)\right) \rho(x_1|z)\rho(x_2|z)\nu(z)dx_1dx_2dz.$$

We will propose one such cost in an example below on hand-written digits, a cost that penalizes deviations of $T(:, z)$ from an isometry. A data-driven version of a cost of this form is

$$L_C = \frac{1}{N^2} \sum_{i,j} c\left(x_i, T\left(x_i, z_i\right), x_j, T\left(x_j, z_j\right)\right) Z(z_i, z_j),$$

involving the bi-stochastic matrix $Z_i^j = Z(z_i, z_j)$ introduced above.

Most formulas in this article are written, for concreteness, in terms of pairwise cost functions. In order to apply them to more general costs, it is enough to insert the corresponding expression for L_C and its derivatives, while all the formulas concerning L_F remain unaltered.

5. A penalty method

Both (14) and (19) are minimax problems of a special kind, where the maximization is carried out over a single positive scalar quantity λ whose optimal value is unbounded (as perteins the Lagrange multiplier of a single constraint requiring a non-negative quantity, L_F, to vanish). More effective than maximizing L over λ is to use a penalty method, whereby λ is externally increased at each iteration step to as to progressively enforce the constraint.

Among the possible strategies for controlling λ, we propose one guaranteeing that L_F decreases at every step, while not making λ grow so fast as to effectively make the minimization of L_C a secondary goal. The procedure applies to both (14) and (19), for concreteness we describe it here for (14):

1. Initialize $y_i = x_i$, $\lambda = \lambda_0 > 0$, and a maximum number of iteration $niter$. The iteration count starts from $n = 0$, and the learning rate from $\eta = \eta^0$. Precompute the matrix C as defined in (13).

   If λ is sufficiently small, L_C dominates the objective function, making the minimization problem convex (L_C is typically convex, at least near the identity map.) Therefore, we set $\lambda_0 = 1/\max(\text{abs}(\text{eigs}[-F_{xx}(x)]))$ resulting in a semi positive definite Hessian.

2. While $n < niter$ and y has not yet converged, tentatively evolve the learning rate through the formula $\eta^{n+1} = \min\{2.01\eta^n, \eta^0\}$.
   (a) Calculate the derivatives of the objective function (Appendix A.)
   (b) Compute λ^{n+1} according the criteria below, whith $\alpha > 0$:

$$\left\langle \nabla_y c(x,y) + \lambda \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{:l}, \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{:l} \right\rangle \geq \\ \geq \alpha \left\langle \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{il}, \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{il} \right\rangle, \quad (22)$$

   which implies the lower bound for λ:

$$\lambda \geq \alpha - \frac{\langle \nabla_y c(x,y), \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{:l} \rangle}{\langle \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{il}, \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{:l} \rangle} = \lambda_{min}. \quad (23)$$

   Here the inner product $\langle\rangle$ between two functions $f(x, y, :)$ and $g(x, y, :)$ is defined as $\langle f, g \rangle = \sum_i f(x_i, y_i, i) g(x_i, y_i, i)$. If λ_{min} is larger than λ^n and smaller than a threshold λ^{max}, set $\lambda^{n+1} = \lambda_{min}$. Otherwise, if $\lambda_{min} > \lambda^{max}$, set $\lambda^{n+1} = \lambda^{max}$, else set $\lambda^{n+1} = \lambda^n$. This guarantees that λ^{n+1} is not smaller than λ^n, so $L(y, \lambda^{n+1}) \geq L(y, \lambda^n)$ is satisfied for any y.
   (c) Update y using either gradient descent:

$$y^{n+1} = y^n - \eta \nabla_y L(y, \lambda^{n+1})|_{y=y^n}, \quad (24)$$

   or implicit gradient descent [33]:

$$y^{n+1} = y^n - \eta \left(I + \eta \nabla_{yy} L(y, \lambda^{n+1})|_{y=y^n} \right)^{-1} \nabla_y L(y, \lambda^n)|_{y=y^n}. \quad (25)$$

   These update rules couple the points $y_i \in \mathbb{R}^d$ in different ways, as discussed in the appendix.
   (d) Check that the objective function decreases after the step,

$$L(y^{n+1}, \lambda^{n+1}) \leq L(y^n, \lambda^{n+1}), \quad (26)$$

   where the kernel centers are evaluated at y^{n+1} on both sides of the inequality (see the appendix). Otherwise decrease η to $\eta/2$ and repeat (c) and (d) until it does.

The intuition behind the criteria for updating λ is the following. The two components of the objective function push the map $T(x,z)$ in opposite directions: while L_F decreases as ρ_T approaches the barycenter μ, L_C decreases as ρ_T returns to ρ, as the cost is typically minimal at $T(x,z) = x$. The two components are also different in nature: L_F represents the hard constraint that $y = T(x,z)$ be independent of z, while the minimization of L_C establishes a selection criterion among all maps satisfying $L_F = 0$. Because of this, one should always pick λ large enough that the direction of gradient descent of the full Lagrangian L is also a direction of descent for L_F. This condition reads

$$\left\langle \frac{\delta}{\delta T}\left[L_C + \lambda L_F\right], \frac{\delta}{\delta T} L_F \right\rangle \geq 0,$$

a requirement that we make more precise by establishing a threshold $\alpha > 0$:

$$\left\langle \frac{\delta}{\delta T}\left[L_C + \lambda L_F\right], \frac{\delta}{\delta T} L_F \right\rangle \geq \alpha \left\langle \frac{\delta}{\delta T} L_F, \frac{\delta}{\delta T} L_F \right\rangle,$$

which is the content of (23). Notice that, if the optimum is reached for the prior λ^n, then the first order condition yields:

$$\nabla_y c(x,y) + \lambda^n \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{il} = 0 \implies \nabla_y c(x,y) = -\lambda^n \nabla_y \sum_l \mathcal{K}_a(y_l, y) C_{il},$$

so (23) yields $\lambda_{min} = \alpha - (-\lambda^n) = \alpha + \lambda^n$. This suggests setting adaptively $\alpha = \omega \lambda^n$, whith $\omega \in (0,1)$.

6. Numerical examples

This section presents four representative numerical examples in different dimensions and with different types of covariates, to: (1) demonstrate the ability to work with cost functions different from the canonical L^2 and the effect that different choices for the cost have, and (2) show applicability to times series analysis with data distributed on a Riemannian manifold.

6.1. Barycenter of three ellipses under different costs

A toy example shows how the choice of a cost function affects the properties of the barycenter. The data points $x_i \in \mathbb{R}^2$ are sampled from 3 uniform densities

supported on 3 ellipses with different centers and shapes, and labelled by the discrete cofactor $z_i \in \{0, 1, 2\}$. The major axes of the ellipses are horizontal for $z = 1, 2$ (in green and yellow) and vertical for $z = 0$ (in blue). All ellipses have eccentricity $e = \frac{2\sqrt{2}}{3}$, with 100 points sampled from each.

We apply our algorithm with cost function induced by the p-norm:

$$c(x, y) = \sum_{i=1}^{2} |x_i - y_i|^p, \quad x, y \in \mathbb{R}^2, \quad p \in \mathbb{R}, \quad p \geq 1.$$

The test functions in x space are constructed by solving Problem 2, using as features f_j polynomials up to the second degree, i.e. y_1, y_2, y_1^2, $y_1 y_2$ and y_2^2. To address the issue that, for $p < 2$, the Hessian of the cost function is degenerate at the origin, we utilize the approximation $|x| \approx \sqrt{x^2 + \epsilon} - \sqrt{\epsilon}$,with $\epsilon = 0.01$. The data x_i and barycenter y_i (in purple), with $p \in \{1.2, 1.5, 2, 2.5, 3\}$ are shown in Figure 1.

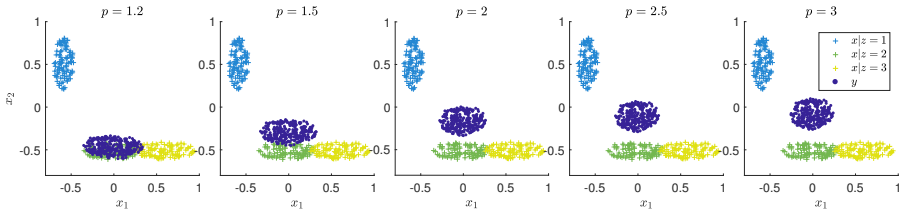


Figure 1: Barycenters with p-norm-based costs with different values of p. The color in the source data refers to the index z of the cluster, while the barycenter is displayed in purple.

Large values of p penalize outliers, i.e. distributions that are far from the barycenter. Thus the barycenter for p large must be such that no distribution is far from it. On the other hand, for p close to 1, majority rules: the average distance to the barycenter must be minimal. In our case, the "outlier" , both in shape and position, is the cluster $z = 0$ in blue. Thus in Figure 1, when p is small, the barycenter (in deep purple) is closer in shape and position to the ellipses with $z = 1, 2$, while, as p increases, the barycenter shifts gradually from the bottom to the middle of the figure and becomes nearly isotropic, so as not to be far from any cluster, including the outlier.

6.2. Handwritten digits

We use the MNIST dataset [34] to display the effect of the test functions chosen for Problem 2, contrast this with the non-parametric Problem 1, and illustrate how a non-pairwise cost function can help impose desired features on the barycenter. The MNIST dataset contains handwritten digits from 0 to 9. For each digit, we randomly select 6 images, which we randomly displace, and then compute their barycenter under various test functions and costs.

6.2.1. Effect of test function

We first demonstrate the effect of the richness of the test functions adopted, keeping as cost function the standard squared Euclidean distance. Two different sets of test functions are used: first order polynomials, which only detect the discrepancy in the conditional means, and polynomials up to 2nd order, testing both conditional mean and covariance. We then compare these results to the nonparametric algorithm (14), after preconditioning by subtracting the conditional mean. The results are displayed in Table 1. Qualitative improvements can be observed when the test function becomes richer. For example, the barycentric images are noisy when only the conditional mean is aligned, the edges are clearer when second order polynomials are adopted, and the nonparametric approach outperforms both choices.

6.2.2. Effect of the cost function

Looking at Table 1, one may think at first that the barycenter has not been fully resolved, as its contours are not well defined. Figure 2, displaying the push forward of each of the six marginals to the barycenter, shows that this is not quite the case, as all push forward measures agree, except when only the preconditioner is used, forcing each of the six marginals to keep its original shape. Small differences are due to the fact that each marginal contains a different number of sample points.

The fact that the digit six in Figure 2 (b) looks somewhat cloudy should not come as a surprize, since nothing in the objective function enforces the notion

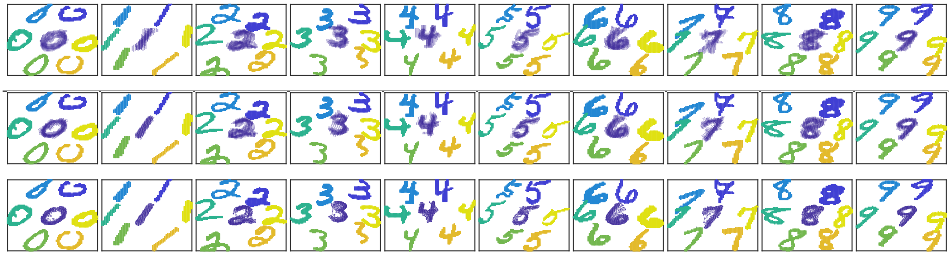
0	1	2	3	4	5	6	7	8	9
									

Table 1: Barycenters of 6 randomly chosen handwritten digits from MNIST dataset, solved numerically with different sets of test functions. The barycenter y is displayed in deep purple in the center, surrounded by the source data x in different colors indicating the index $z \in \{1, 2, 3, 4, 5, 6\}$. The first row contains the solution to Problem 2 using as test functions only polynomial of first degree, the second row uses polynomials of second degree, and the third row has the solution to Problem 1, where the test function, slaved to the map, evolves through the kernel density estimator in (10).

that the maps to the barycenter should not smear the original digits. A way to address this is to adopt a distortion-sensitive cost function in (2), namely

$$C(y(x,k),\rho) = \frac{1}{N^2} \sum_{1 \le i \ne j \le N} \left[\left(\frac{||y_i^k - y_j^k||^2}{||x_i^k - x_j^k||^2 + \epsilon^2} - 1 \right)^2 \right] + \omega \frac{1}{N} \sum_{i=1}^{N} ||y_i^k - x_i^k||^2. \tag{27}$$

The first term penalizes the deviation of the map from a conditional isometry, which would have equal pairwise distances in x and $y = T(x, z)$ space for each value of z (k in our discrete setting), with a small parameter ϵ to prevent division by zero. The second term is a remnant of a regular optimal transport cost, intended to anchor the barycenter in space, with small weight $\omega = 0.01$ in our numerical example. We compare the results obtained with the new cost and with the L^2 distance. The barycenters are displayed in Table 2, and the mapped samples from the digit six with different z values are shown in Figure 2. Clearly the adoption of the cost in (27) results in a barycenter with more defined contours, as the need to preserve pairwise distances prevents the points in the upper branch of the digit six to broaden up when mapped to the barycenter.

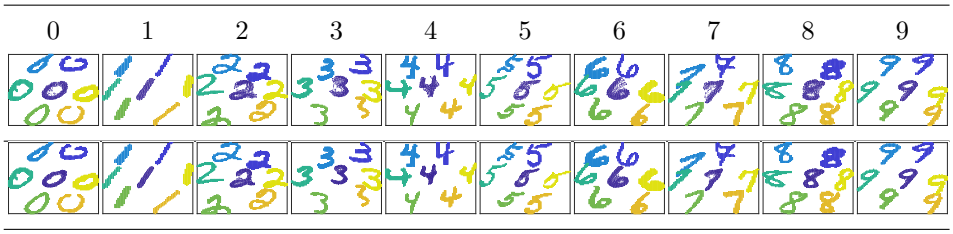
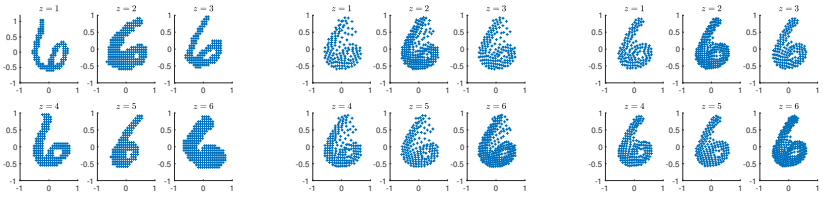


Table 2: Barycenters of digits solved non-parametrically, the first row under a squared distance cost and the second under the distortion-sensitive cost (27).



(a) Preconditioned (b) With L^2 cost (c) With cost in (27)

Figure 2: Samples of the digit 6 push-forwarded to the barycenter for each value of z.

6.3. Two patches on the unit sphere

This section performs a numerical experiment on the barycenter of two distributions, with samples shown in Figure 3, defined on the unit sphere $\mathbb{S}^2 = \{x \in \mathbb{R}^3 : ||x|| = 1\}$. Because the sample points are defined on the sphere, we can represent them in spherical coordinates,

$$x_1 = \cos\theta\cos\phi, \quad x_2 = \cos\theta\sin\phi, \quad x_3 = \sin\theta,$$

where $\theta \in [0, 2\pi)$ and $\phi \in [-\pi/2, \pi/2]$ represent longitude and latitude. Then the natural cost is not the canonical Euclidean L^2 distance $c(x, y) = ||x - y||^2$, but the geodesic distance between points:

$$\tilde{c}(x, y) = 2\arcsin\sqrt{\sin^2\left(\frac{|\theta_x - \theta_y|}{2}\right) + \cos\theta_x\cos\theta_y\sin^2\left(\frac{\phi_x - \phi_y|}{2}\right)}.$$

The example illustrates how the barycenters capture essential features of the manifold on which the data are defined. When the two distributions are

supported on the same hemisphere (left two panels of Figure 3, in red and black), the support of the barycenter (in blue) interpolates between them. By contrast, when the two distributions lie around the north and the south pole respectively, there is no preferred meridian on which the barycenter should lie, resulting in it being supported along the entire equator.

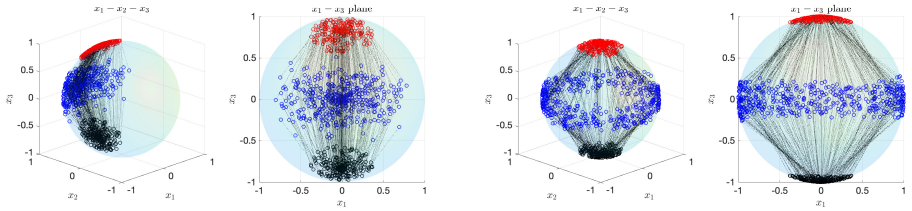


Figure 3: Barycenter (blue) of two distributions with $z = 0$ (red) and $z = 1$ (black), with 250 points sampled from each. The thin black lines indicate the one-to-one correspondence between the source data and its image under the map. Each distribution has $\theta \sim U_{[0,2\pi)}$, $\phi \sim U_{[\frac{3}{8}\pi,\frac{1}{2}\pi]}$ and $\phi \sim U_{[-\frac{1}{2}\pi,-\frac{3}{8}\pi]}$ respectively on the right two panels, while they are shifted to the same hemisphere on the left two.

6.4. Hidden variability recovery on the unit sphere

Time series are often modeled through a Markov model of the form

$$x^{n+1} = F(x^n, z^{n+1}_{known}, w^{n+1}, t^{n+1}), \tag{28}$$

where $\{x^n\}_{n=0}^T$ is the time series, t is the time, z_{known} represents known factors that influence x and w contains unknown sources of variability. In [35], the authors proposed a method to uncover the hidden variability w^n by removing from x^{n+1} the variability due to z^{n+1}, computing the barycenter of $\rho(x^{n+1}|z^{n+1})$ thorough the family of maps

$$y^n = T(x^n, z^n), \quad z^n = [x^{n-1}, t^n, z^n_{known}],$$

so that the "filtered" signal y^n is a function of only w^n. This section shows a synthetic example combining this idea with the algorithm described in section 5 to study time series defined on Riemannian manifolds. In particular, we consider the time series defined on the 3D unit sphere, generated as the sum of a deterministic dynamics and random noise:

- The deterministic dynamics in spherical coordinates is given by:

$$\begin{bmatrix} \tilde{\phi}^{n+1} \\ \tilde{\theta}^{n+1} \end{bmatrix} = \begin{bmatrix} \phi^n \\ \theta^n + \sin(\theta^n) + \frac{1}{2} \end{bmatrix},$$

  where θ^n and ϕ^n are the longitude and latitude at x^n respectively. In Cartesian coordinates, this becomes $\tilde{x}^{n+1} = Sph2Cart(R = 1, \tilde{\phi}^{n+1}, \tilde{\theta}^{n+1}))$.

- The hidden factor w^{n+1} is generated by first sampling a 2-dimensional uniform distribution in spherical coordinates, and then transforming the sampled points into Cartesian coordinates on the unit sphere:

$$w^{n+1} = Sph2Cart(1, \phi_w^{n+1}, \theta_w^{n+1}), \tag{29}$$

  where $\phi_w^{n+1} \sim U_{[\frac{\pi}{2}-0.45, \frac{\pi}{2}]}$ and $\theta_w^{n+1} \sim U_{[0,2\pi]}$. This results in the round patch centered at the north pole shown on the left panel of Figure 4. In order to add w^{n+1} to the deterministic part $\tilde{x}^{n+1}$, we define a one to one map between the tangent planes at the north pole and at $\tilde{x}^{n+1}$, through the reflection with respect to the axis $\tilde{x}_{1/2}^{n+1} = Sph2Cart\left(1, \frac{1}{2}(\tilde{\phi}^{n+1} + \frac{\pi}{2}), \tilde{\theta}^{n+1}\right)$ bisecting the angle between the north pole and $\tilde{x}^{n+1}$. Using Rodrigues' rotation formula, this yields

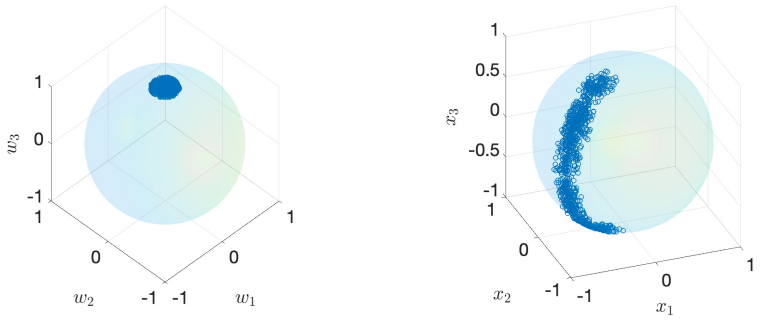
$$x^{n+1} = \left(I + 2K^2(\tilde{x}_{1/2}^{n+1})\right) w^{n+1}, \tag{30}$$

  where $K \in \mathbb{R}^{3\times 3}$ is the cross-product matrix:

$$K(x) = \begin{bmatrix} 0 & -x_3 & x_2 \\ x_3 & 0 & -x_1 \\ -x_2 & x_1 & 0 \end{bmatrix}.$$

Figure 4 shows a time series of 1000 steps, starting from the south pole, and the hidden signal w. Figure 5 compares the barycenters obtained with two different methods:

1. Filter x^{n+1} with $z^{n+1} = x^n$ in $\mathbb{R}^3$, using the Euclidean distance as cost.
2. Filter $[\phi^{n+1}, \theta^{n+1}]$ with $z^{n+1} = [\phi^n, \theta^n]$ and great-circle distance as the cost.



(a) Hidden signal w (b) Time series x

Figure 4: (a) Hidden signal generated using (29). (b) Complete time series generated using (30).

The first approach ignores the fact that the time series is supported on a lower dimensional manifold of $\mathbb{R}^3$, resulting in a barycenter that is not on the surface of the sphere. The second approach respects the distance metric of the manifold, and the barycenter lays on the same lower dimensional manifold where the marginals are supported. To show that the filtered signal y^n is a surrogate for w^n, it is enough to establish a one to one map between y^n and w^n, a map that depends on the specific form of F in (28). In order to visualize this map, we align the (normalized) barycenter to the hidden noise by means of linear regression. Figure 6 shows the resulting smooth dependence between the hidden signal and the barycenter. Figure 7 dsiplayes a moving average of the barycenter and the hidden signal as a functions of time, providing further evidence that the two signal overlap.

7. Conclusions

This work introduces the distributional barycenter problem, an extension of the optimal transport barycenter problem where the cost needs not be the expected value of a pairwise function, allowing more general costs needed in applications, such as a new cost penalizing non-isometric maps.

A novel numerical algorithm is introduced for the solution of the barycenter

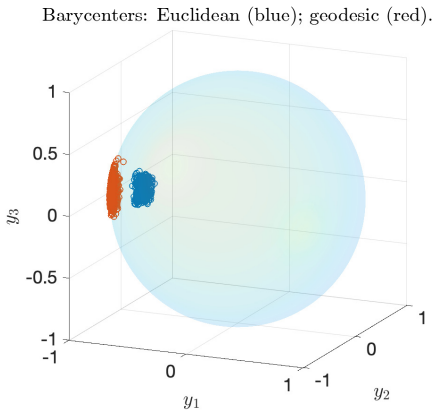


Figure 5: Barycenter solved by two different approaches, in 2D spherical coordinates (red) and 3D Cartesian coordinates (blue)

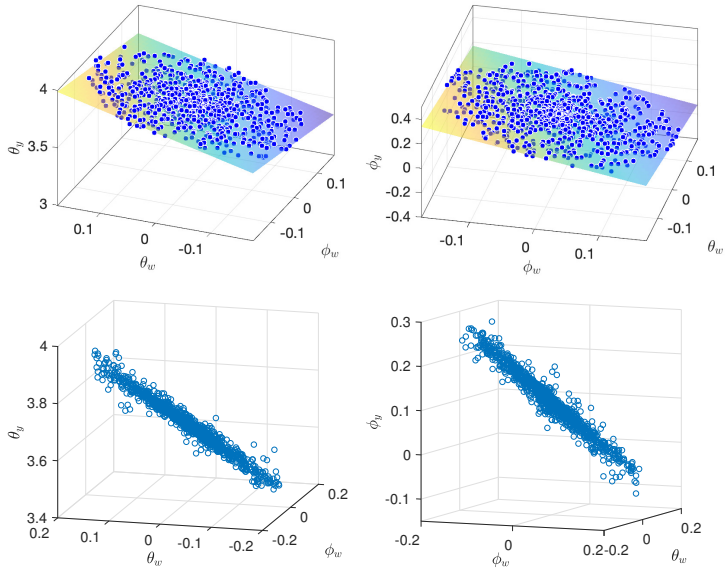


Figure 6: The spherical coordinates of the barycenter solved in 2D as functions of the hidden signal (also in spherical coordinates). Polynomial surfaces of order 5 are fitted to the data and visualized.

problem. The algorithm avoids the difficulties typical of adversarial approaches by slaving the discriminator to the generator. This results in a simpler approach that looks for a minimum rather than a saddle point of the objective function.

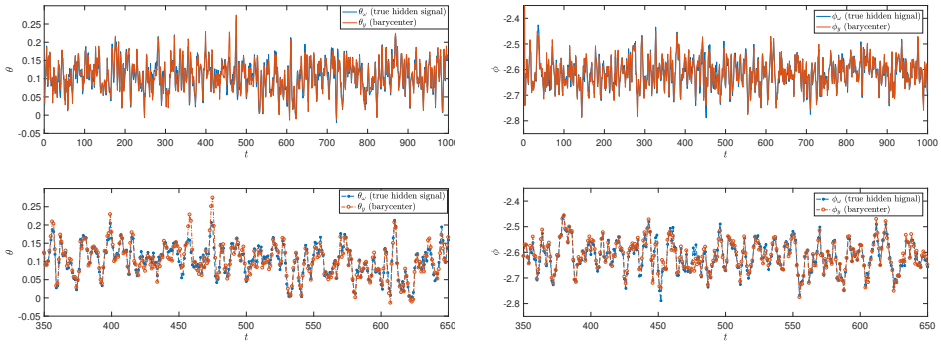


Figure 7: First row: moving average with window size 3 of the time series of the hidden signal ω and of the filtered signal y in spherical coordinates (left: longitude θ, right: lattitude ϕ). Second row: zoom of the first row.

The approach is essentially non-parametric, as the only parameter of the test functions and maps is the bandwidth of a kernel function.

Appendix A. Updating rules (24) and (25)

This appendix calculates the gradient and Hessian of the objective functions L in (14) and (19), used to implement the explicit (25) and implicit (26) schemes for updating the current position y_i^n of the original sample points x_i.

The update $y_i^n \to y_i^{n+1}$ is subtle, as both of the kernel $\mathcal{K}_a(y, w)$'s arguments are evaluated at the sample points y_i, yet they play very different roles in the Lagrangian L: while $y = y_i$ represents the map $T(x_i, z_i)$ that L is to be minimized over, the $w = y_i$ are the Kernels' centers, characterizing the test function $F = \rho(y|z)$ over which L was originally to be maximized! Our methodology replaced this maximization by a slaving of F to T, hence the appearance of the y_i in F, yet L must be minimized only over its first argument, not the second. Thus, for gradient descent, one must use terms such as

$$\frac{\partial \mathcal{K}_a(y, y_k^n)}{\partial y}\Big|_{y=y_i^n}$$

and, for implicit gradient descent,

$$\frac{\partial \mathcal{K}_a(y, y_k^{n+1})}{\partial y}\Big|_{y=y_i^{n+1}} \approx \frac{\partial \mathcal{K}_a(y, y_k^n)}{\partial y}\Big|_{y=y_i^n} + \frac{\partial^2 \mathcal{K}_a(y, w)}{\partial y^2}\Big|_{\substack{y=y_i^n \\ w=y_k^n}} \left(y_i^{n+1} - y_i^n\right) + \frac{\partial^2 \mathcal{K}_a(y, w)}{\partial y \partial w}\Big|_{\substack{y=y_i^n \\ w=y_k^n}} \left(y_k^{n+1} - y_k^n\right). \quad \text{(A.1)}$$

Though formulas below are developed for regular pairwise cost functions, their extension to the general case should be clear. The objective functions for problems 1 and 2 are:

$$\text{Kernel density estimation:} \qquad \mathcal{L}_1 = \sum_i c(x_i, y_i) + \lambda \sum_{i,k} \mathcal{K}_a(y_i, y_k) C_{ik}.$$

$$\text{Parametric:} \qquad \mathcal{L}_2 = \sum_i c(x_i, y_i) + \lambda \sum_{i,k} f(y_i) f(y_k) C_{ki}.$$

Explicit: Formula (25) is equivalent to forward Euler for ODEs. We can update the position of each point $y_i \in \mathbb{R}^d$ independently, through the update rule $y_i^{n+1} = y_i^n - \eta \nabla_y L|_{y=y_i^n}$, where

$$\nabla_y \mathcal{L}_1|_{y=y_i^n} = \left[\frac{\partial c(x_i, y)}{\partial y} + \lambda \sum_k \frac{\partial \mathcal{K}_a(y, y_k^n)}{\partial y} C_{ik}\right]_{y=y_i^n},$$

and

$$\nabla_y \mathcal{L}_2|_{y=y_i^n} = \left[\frac{\partial c(x_i, y)}{\partial y} + \lambda \sum_k \frac{\partial f(y_i)}{\partial y} f(y_k^n) C_{ki}\right]_{y=y_i^n}.$$

Implicit: This scheme, when applied to minimize the generic function $f(y, w)$ is obtained by the following approximation:

$$\begin{aligned} y^{n+1} &= y^n - \eta f_y(y^{n+1}, y^{n+1}) \\ &\approx y^n - \eta\{f_y(y^n, y^n) + (y^{n+1} - y^n)\left[f_{yy}(y^n, y^n) + f_{yw}(y^n, y^n)\right]\} \end{aligned} \quad \text{(A.2)}$$

that, once rearranged, results in the scheme in (26) [33]:

$$y^{n+1} = y^n - \eta \left[I + \eta(f_{yy}^n + f_{yw}^n)\right]^{-1} f_y^n \quad \text{(A.3)}$$

The Hessian matrix $\nabla_{yy} L_1$ in (26) is therefore given by

$$\nabla_{yy} \mathcal{L}_1 = \mathcal{L}_1^{yy} + \mathcal{L}_1^{yw}.$$

The matrix $\mathcal{L}_{1,yy}$ is diagonal, and we have:

$$\mathcal{L}^{yy}_{1,ii} = \left[\frac{\partial^2 c(x_i,y)}{\partial y^2} + \lambda \sum_k \frac{\partial^2 \mathcal{K}_a(y,y^n_k)}{\partial y^2} C_{ik}\right]_{y=y^n_i}, \quad \mathcal{L}^{yw}_{1,ik} = \lambda \left.\frac{\partial^2 \mathcal{K}_a(y,w)}{\partial y \partial w}\right|_{\substack{y=y^n_i \\ w=y^n_k}} C_{ik}.$$

This calculation applies to pairwise cost functions, where the only non diagonal $\mathbb{R}^{d\times d}$ blocks arise from the L_F in $\mathcal{L}_1$. One needs to adjust accordingly for more general costs.

Similarly for $\mathcal{L}_2$ we have $\nabla_{yy}\mathcal{L}_2 = \mathcal{L}^{yy}_2 + \mathcal{L}^{yw}_2$ with

$$\mathcal{L}^{yy}_{2,ii} = \left[\frac{\partial^2 c(x_i,y)}{\partial y^2} + \lambda \sum_k \frac{\partial^2 f(y)}{\partial y^2} f(y^n) C_{ki}\right]_{y=y^n_i}, \quad \mathcal{L}^{yw}_{2,ik} = \lambda \left.\frac{\partial f(y)}{\partial y}\right|_{y=y^n_i} \left.\frac{\partial f(y)}{\partial y}\right|_{y=y^n_k} C_{ki}.$$

When f is a vector, the gradient and Hessian have an additional sum over its components.

Acknowledgments

Tabak's work was partially supported by NSF grant DMS-1715753 and ONR grant N00014-15-1-2355.

References

[1] G. Trigila, E. G. Tabak, Data-driven optimal transport, Communications on Pure and Applied Mathematics 69 (4) (2016) 613–648.

[2] E. G. Tabak, G. Trigila, W. Zhao, Conditional density estimation and simulation through optimal transport, Machine Learning (2020) 1–24.

[3] M. Pavon, E. G. Tabak, G. Trigila, The data-driven schroedinger bridge, To appear in Communication of Pure and Applied Mathematics (2020).

[4] S. Kolouri, A. B. Tosun, J. A. Ozolek, G. K. Rohde, A continuous linear optimal transport approach for pattern analysis in image datasets, Pattern recognition 51 (2016) 453–462.

[5] W. Wang, J. A. Ozolek, D. Slepčev, A. B. Lee, C. Chen, G. K. Rohde, An optimal transportation approach for nuclear structure-based pathology, IEEE transactions on medical imaging 30 (3) (2010) 621–631.

[6] Y. Yang, Y.-F. Wu, D.-C. Zhan, Z.-B. Liu, Y. Jiang, Complex object classification: A multi-modal multi-instance multi-label deep network with optimal transport, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2594–2603.

[7] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, L. Guibas, Convolutional wasserstein distances: Efficient optimal transportation on geometric domains, ACM Transactions on Graphics (TOG) 34 (4) (2015) 66.

[8] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, G. K. Rohde, A linear optimal transportation framework for quantifying and visualizing variations in sets of images, International journal of computer vision 101 (2) (2013) 254–269.

[9] S. Angenent, S. Haker, A. Tannenbaum, Minimizing flows for the monge–kantorovich problem, SIAM journal on mathematical analysis 35 (1) (2003) 61–97.

[10] J. Rabin, S. Ferradans, N. Papadakis, Adaptive color transfer with relaxed optimal transport, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 4852–4856.

[11] E. G. Tabak, G. Trigila, Conditional expectation estimation through attributable components, Information and Inference: A Journal of the IMA 128 (00) (2018).

[12] H. Yang, E. G. Tabak, Conditional density estimation, latent variable discovery and optimal transport, arXiv preprint arXiv:1910.14090 (2019).

[13] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, G. K. Rohde, Optimal mass transport: Signal processing and machine-learning applications, IEEE signal processing magazine 34 (4) (2017) 43–59.

[14] G. Schiebinger, J. Shu, M. Tabaka, B. Cleary, V. Subramanian, A. Solomon, J. Gould, S. Liu, S. Lin, P. Berube, et al., Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming, Cell 176 (4) (2019) 928–943.

[15] A. Galichon, Optimal transport methods in economics, Princeton University Press, 2018.

[16] G. Monge, Mémoire sur la théorie des déblais et des remblais, De l'Imprimerie Royale, 1781.

[17] L. V. Kantorovich, On a problem of monge, Uspekhi Mat. Nauk, 3, No. 2 (1948) 225–226.

[18] F. Santambrogio, Optimal transport for applied mathematicians, Birkäuser, NY 55 (58-63) (2015) 94.

[19] M. Essid, D. Laefer, E. G. Tabak, Adaptive optimal transport, Submitted to Information and Inference (2018).

[20] M. Feldman, R. McCann, Monges transport problem on a riemannian manifold, Transactions of the American Mathematical Society 354 (4) (2002) 1667–1697.

[21] E. Tenetov, G. Wolansky, R. Kimmel, Fast entropic regularized optimal transport using semidiscrete cost approximation, SIAM Journal on Scientific Computing 40 (5) (2018) A3400–A3422.

[22] O. Yair, F. Dietrich, R. Talmon, I. G. Kevrekidis, Optimal transport on the manifold of spd matrices for domain adaptation, arXiv preprint arXiv:1906.00616 (2019).

[23] H. Lavenant, S. Claici, E. Chien, J. Solomon, Dynamical optimal transport on discrete surfaces, ACM Transactions on Graphics (TOG) 37 (6) (2018) 1–16.

[24] J.-D. Benamou, Y. Brenier, A computational fluid mechanics solution to the monge-kantorovich mass transfer problem, Numerische Mathematik 84 (3) (2000) 375–393.

[25] M. Agueh, G. Carlier, Barycenter in the Wasserstein space, SIAM J. MATH. ANAL. 43 (2) (2011) 094–924.

[26] G. Carlier, A. Oberman, E. Oudet, Numerical methods for matching for teams and wasserstein barycenters, ESAIM: Mathematical Modelling and Numerical Analysis 49 (6) (2015) 1621–1642.

[27] J. Nocedal, S. Wright, Numerical optimization, Springer Science & Business Media, 2006.

[28] F. Sapienza, P. Groisman, M. Jonckheere, Weighted geodesic distance following fermat's principle, 6th International Conference on Learning Representations (2018).
URL https://openreview.net/forum?id=BJfaMIJwG

[29] M. Kuang, E. G. Tabak, Preconditioning of optimal transport, SIAM Journal on Scientific Computing 39 (4) (2017) A1793–A1810.

[30] M. Kuang, E. G. Tabak, Sample-based optimal transport and barycenter problems, Communications on Pure and Applied Mathematics 72 (8) (2019) 1581–1630.

[31] M. Rosenblatt, Conditional probability density and regression estimators, Multivariate analysis II 25 (1969) 31.

[32] J. G. De Gooijer, D. Zerom, On conditional density estimation, Statistica Neerlandica 57 (2) (2003) 159–176.

[33] M. Essid, E. Tabak, G. Trigila, An implicit gradient-descent procedure for minimax problems, Submitted to Machine Learning (Springer) (2019).

[34] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[35] E. G. Tabak, G. Trigila, Explanation of variability and removal of confounding factors from data through optimal transport, Communications on Pure and Applied Mathematics 71 (1) (2018) 163–199.