UC Irvine
UC Irvine Previously Published Works

Title

Statistical Learning in Wasserstein Space

Permalink

Journal

ISSN

Authors

Karimi, Amirhossein
Ripani, Luigia
Georgiou, Tryphon T

Publication Date

DOI

Copyright Information

Peer reviewed

Statistical learning in Wasserstein space

Amirhossein Karimi, Luigia Ripani, and Tryphon T. Georgiou

Abstract

We seek a generalization of regression and principle component analysis (PCA) in a metric space where data points are distributions metrized by the Wasserstein metric. We recast these analyses as multimarginal optimal transport problems. The particular formulation allows efficient computation, ensures existence of optimal solutions, and admits a probabilistic interpretation over the space of paths (line segments). Application of the theory to the interpolation of empirical distributions, images, power spectra, as well as assessing uncertainty in experimental designs, is envisioned.

I. INTRODUCTION

The purpose of statistical learning is to identify structures in data, typically cast as relations between explanatory variables. Prime example is the case of principal components where affine manifolds are sought to approximate the distribution of data points in a Euclidean space. The data is then seen as a point-cloud clustering about the manifolds which, in turn, can be used to coordinatize the dataset. Such problems are central to system identification [1], [2].

Accounting for measurement noise and uncertainty has been an enduring problem from the dawn of statistics to the present [1]. In this work we view distributions as our principle objects which, besides representing point clouds, may also represent the uncertainty in localizing individual elements in a point set. Our aim is to develop a mathematical framework to allow for an analogue of principle components in the space of distributions. That is, we view distributions as points themselves in a suitable metric space (Wasserstein space) and seek structural features in collections of distributions.

From an applications point of view, these distributions may delineate density of particles, concentration of pollutants, fluctuation of stock prices, or the uncertainty in the glucose level in the blood of diabetic persons, indexed by a time-stamp, with the goal to capture the principal modes of underlying dynamics (flow). Another paradigm may involve intensity images in which case we seek suitable manifolds of distributions that images cluster about. Such manifolds would constitute the analogue of principal components and can be used to parametrize our image-dataset and capture directions with the highest variability. Such a setting could be applied in e.g., in longitudinal image study [3] where the brain development or tumor growth patterns need to be studied, power spectral tracking [4], traffic control [5] and so on.

There is a very recent and fast growing literature in the direction that we are envisioning, namely, to develop a theory of manifold learning in Wasserstein space. Most studies so far explore the tangent space of the Wasserstein manifold at a suitably selected reference point (e.g., the barycenter of the dataset [6]) so as to take advantage of the Hilbert structure of the tangent space and identify the dominant modes of variation. Such an approach typically leads to non-smooth and non-convex optimization [7].

Herein, we follow a different route in which we place a probability measure on the sought structures (manifolds, modes, interpolating paths that we wish to learn). To illustrate the effect of a probability law on paths, Figure 1 highlights the flow of Gaussian one-time marginals along linear paths that differ in the joint distribution between the corresponding end-point variables. Our approach has several advantages. First, it enjoys a multi-marginal optimal transport [8] formulation and the consequent computational benefits (solvable via linear programming). Second, it provides varying likelihood in the space of admissible paths/curves. When all the target distributions (observations) are Dirac measures, the problem reduces to interpolation in Euclidean space.

The first attempt to developing principal component analysis (PCA) for probability measures in) [9] utilized the pseudo-Riemannian structure of $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ space [10], labeled geodesic PCA (GPCA). Unfortunately, implementation can be challenging even in the simplest case of probability measures supported on a bounded subset of the real line due to non-differentiable and non-convex cost and constraints. Other studies introduce suitable relaxation which however prevents solutions from relating to geodesics in Wasserstein space [7], [11], [12]. Part of the ensuing problems

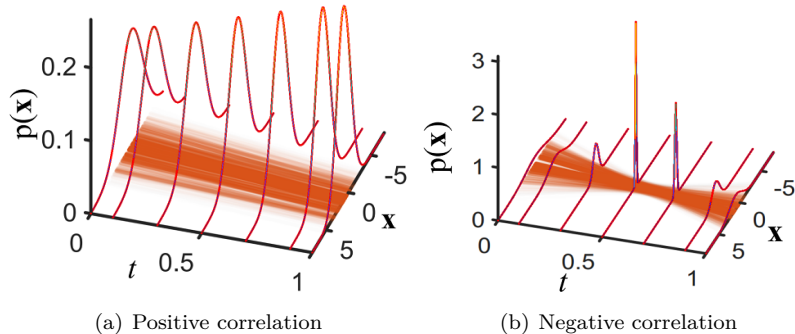(a) Positive correlation (b) Negative correlation

Fig. 1. Marginals along linear paths (lines) with identical end-point distributions and different probability laws (positive and negative correlation between end-point marginals); the intensity of color is proportional to the likelihood of the path.

in developing PCA in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ are due to the fact that this is a positively curved space [13]. Earlier relevant papers on regression in Wasserstein space include [4], [14].

The model of geodesics in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ as the learning hypothesis comes with an important caveat. At times, it may lead to underfitting in that the hypothesis class, W_2-geodesics, i.e., $(\mathcal{P}_2(\mathbf{R}^d), W_2)$-line segments, as opposed to higher order $(\mathcal{P}_2(\mathbf{R}^d), W_2)$-curves [15], [16]) is too restrictive. This may result in temporally distant points to be more strongly correlated than adjacent ones.

In this paper the emphasis is placed on the path space of linear segments from $[0, 1]$ to $\mathbf{R}^d$ ($\mathbf{L}([0, 1], \mathbf{R}^d)$), and the problem of interpolating time-indexed target distributions which is cast as seeking a suitable probability law on line segments; it is seen as linear regression lifted to the space of probability distributions. We will discuss how the setting extends to the case where data/distributions are provided with no time-stamp. Interpolation with higher-order curves in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ that employs multi-marginal formulation is a possible future direction, as well as entropic regularization for computational purposes; these last topics will be noted in [17].

II. BACKGROUND AND NOTATION

A. Notation

We denote by $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ the Wasserstein space where $\mathcal{P}_2(\mathbf{R}^d)$ is the set of Borel probability measures with finite second moments, and W_2 is the Wasserstein distance (whose definition will be recalled later on). If a measure $\mu \in \mathcal{P}_2(\mathbf{R}^d)$ is absolutely continuous with respect to the Lebesgue measure on $\mathbf{R}^d$, we will represent by μ both the measure itself and its density. The push-forward of a measure μ by the measurable map $T : \mathbf{R}^d \to \mathbf{R}^d$ is denoted by $T_\# \mu = \nu \in \mathcal{P}(\mathbf{R}^d)$, meaning $\nu(B) = \mu(T^{-1}(B))$ for every Borel set B in $\mathbf{R}^d$. The Dirac measure at point $\mathbf{x}$ is represented by $\delta_{\mathbf{x}}$. We denote the set of linear paths from the time interval $[0, 1]$ to the state space $\mathbf{R}^d$ by $\Omega = \mathbf{L}([0, 1], \mathbf{R}^d)$. There is a bijective correspondence $(X_{0,1})$ between Ω and $\mathbf{R}^d \times \mathbf{R}^d$ using the values of each line at $t = 0$ and $t = 1$ such that any $\omega = (\omega_t)_{t \in [0,1]} \in \Omega$ corresponds to $X_{0,1}(\omega) := (\mathbf{x}_0, \mathbf{x}_1)$ for the endpoints $\mathbf{x}_0 = \omega_0, \mathbf{x}_1 = \omega_1 \in \mathbf{R}^d$. We equip Ω with the canonical σ-algebra generated by the projection map $X_{0,1}$. We often write $\mathbf{R}^{2d} := \mathbf{R}^d \times \mathbf{R}^d$ as a shorthand. For any $\pi(\mathbf{x}_0, \mathbf{x}_1) \in \mathcal{P}_2(\mathbf{R}^{2d})$, the one-time marginals are $\mu_t := ((1 - t)\mathbf{x}_0 + t\mathbf{x}_1)_\# \pi$, $t \in [0, 1]$.

B. Background on Optimal Transport Theory

We now provide a concise overview of optimal mass transportation theory based on [18], [19].

Let μ_0 and μ_1 be two probability measures in $\mathcal{P}_2(\mathbf{R}^d)$. In the Monge formulation of optimal mass transportation, the cost

$$\int_{\mathbf{R}^d} ||T(\mathbf{x}) - \mathbf{x}||_2^2 \ \mu_0(d\mathbf{x})$$

is minimized over the space of maps

$$T : \quad \mathbf{R}^d \to \mathbf{R}^d$$
$$\mathbf{x} \mapsto T(\mathbf{x})$$

2

which "transport" mass $\mu_0(d\mathbf{x})$ at $\mathbf{x}$ so as to match the final distribution, namely, $T_{\#}\mu_0 = \mu_1$. If μ_0 and μ_1 are absolutely continuous, it is known that the optimal transport problem has a unique solution which is the gradient of a convex function ϕ, i.e., $T^* = \nabla\phi(\mathbf{x})$.

This problem is highly nonlinear and, moreover, a transport map may not exist as is the case when μ_0 is a discrete probability measure and μ_1 is absolutely continuous. To alleviate these issues, Kantorovich introduced a relaxation in which one seeks a joint distribution (coupling) π on $\mathbf{R}^d \times \mathbf{R}^d$, with the marginals μ_0 and μ_1 along the two coordinates, namely,

$$W_2^2(\mu_0, \mu_1) := \inf_{\pi \in \Pi(\mu_0, \mu_1)} \int_{\mathbf{R}^d \times \mathbf{R}^d} ||\mathbf{x} - \mathbf{y}||^2 \pi(d\mathbf{x}d\mathbf{y})$$

where $\Pi(\mu_0, \mu_1)$ is the space of couplings with marginals μ_0 and μ_1. In case the optimal transport map exists, we have $\pi = (\text{Id} \times T^*)_{\#}\mu_0$, where Id denotes the identity map.

The square root of the optimal cost, namely $W_2(\mu_0, \mu_1)$, defines a metric on $\mathcal{P}_2(\mathbf{R}^d)$ referred to as the Wasserstein metric [10], [20]. Further, assuming that T^* exists, the constant-speed geodesic between μ_0 and μ_1 is given by

$$\mu_t = \{(1-t)\mathbf{x} + tT^*(\mathbf{x})\}_{\#}\mu_0, \quad 0 \le t \le 1,$$

and is known as *displacement* or *McCann geodesic*.

We now discuss two cases of transport problems that are of interest in this paper.

1) Transport between Gaussian distributions: In case $\mu_0 \sim \mathcal{N}(\mathbf{m}_0, \boldsymbol{\Sigma}_0)$ and $\mu_1 \sim \mathcal{N}(\mathbf{m}_1, \boldsymbol{\Sigma}_1)$, i.e., both Gaussian, the transport problem has a closed-form solution [21]. Specifically,

$$T^* : \mathbf{x} \to \boldsymbol{\Sigma}_0^{-1/2}(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2})^{1/2}\boldsymbol{\Sigma}_0^{-1/2}\mathbf{x},$$

$$W_2(\mu_0, \mu_1) = \sqrt{||\mathbf{m}_0 - \mathbf{m}_1||^2 + \text{tr}(\boldsymbol{\Sigma}_0 + \boldsymbol{\Sigma}_1 - 2\mathbf{S})} \tag{1}$$

where tr(.) stands for trace and

$$\mathbf{S} = (\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1)^{1/2} = \boldsymbol{\Sigma}_0^{1/2}(\boldsymbol{\Sigma}_0^{1/2}\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0^{1/2})^{1/2}\boldsymbol{\Sigma}_0^{-1/2}. \tag{2}$$

The McCann geodesic μ_t corresponds to Gaussian distributions with mean $\mathbf{m}_t = (1-t)\mathbf{m}_0 + t\mathbf{m}_1$ and covariance

$$\boldsymbol{\Sigma}_t = (1-t)^2\boldsymbol{\Sigma}_0 + t^2\boldsymbol{\Sigma}_1 + t(1-t)(\boldsymbol{\Sigma}_0\boldsymbol{\Sigma}_1)^{1/2} \tag{3}$$
$$+ t(1-t)(\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_0)^{1/2}.$$

2) Multi-marginal optimal transportation: The multi-marginal problem is a generalization of optimal transport in which, for a set of given marginals, a correlating law is sought to minimize a cost function. This problem and its applications are surveyed in [8], [22]. The Kantorovich-version of this problem for marginals $\{\mu_i\}_{i=1}^N$ is to minimize the cost

$$\int_{\mathbf{R}^{Nd}} C(\mathbf{x}_1, ..., \mathbf{x}_N) d\gamma(\mathbf{x}_1, ..., \mathbf{x}_N)$$

where $\gamma(\mathbf{x}_1, ..., \mathbf{x}_N) \in \mathcal{P}_2(\mathbf{R}^{Nd})$ such that $\mathbf{x}_{i\#}\gamma = \mu_i$. It is a linear optimization problem over a weakly compact and convex set for which the numerical methods to solve it efficiently are well studied [22], [23].

III. LEAST-SQUARES IN W_2

We start by highlighting the analogy with least squares in Euclidean setting. Here, given a dataset $\{(t_i, \mathbf{x}_i)\}_{i=1}^N$, we seek a linear model $\mathbf{x} = \mathbf{x}_0 + t\mathbf{v} + \boldsymbol{\epsilon}$, where $\mathbf{x}_0 \in \mathbf{R}^d$ and $\mathbf{v} \in \mathbf{R}^d$ serve as the initial point and tangent vector, respectively, and $\boldsymbol{\epsilon}$ represents $\mathbf{R}^d$-valued measurement error. Least squares estimation provides a choice for $\mathbf{x}_0$ and $\mathbf{v}$ with the least variance among all linear unbiased estimators, assuming that errors are uncorrelated having zero mean and finite variance.

In our setting, the data points lie in the probability space $\mathcal{P}_2(\mathbf{R}^d)$, and this entails a probabilistic representation of lines that generalizes linear regression to $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ as discussed below.

Problem 1. *Given $\{\mu_{t_i}\}_{i=1}^N \in \mathcal{P}_2(\mathbf{R}^d)$ a family of probability measures indexed by timestamps $\{t_i\}_{i=1}^N$, determine*

$$\inf_{(g_t)_{t\in[0,1]} \in \mathcal{G}_\mathcal{R}} \sum_{i=1}^N \frac{1}{N} W_2^2(g_{t_i}, \mu_{t_i}) \tag{4}$$

where $\mathcal{G}_\mathcal{R} = \{g_t = ((1-t)\mathbf{x}_0 + t\mathbf{x}_1)_{\#}\pi | \pi(\mathbf{x}_0, \mathbf{x}_1) \in \mathcal{P}_2(\mathbf{R}^{2d})\}$.

Naturally, g_t, for $t \in [0,1]$, represent one-time marginals. Also, $\mathcal{G}_\mathcal{R}$ includes as a subset the set of all geodesics on $\mathcal{P}_2(\mathbf{R}^d)$. Indeed, any $\pi(\mathbf{x}_0, \mathbf{x}_1)$ which is an optimal coupling between two arbitrary distributions on $\mathcal{P}_2(\mathbf{R}^d)$, matches a geodesic on $\mathcal{P}_2(\mathbf{R}^d)$. Therefore, (4) seeks the least sum of squared distances over a set that contains constant-speed geodesics.

The search space, $\mathcal{G}_\mathcal{R}$, is the building block of this study which comes with noteworthy properties. First, it leads to a multi-marginal transport problem (Section IV). Computationally, this is highly efficient as it amounts to a linear programming problem. Besides, entropic regularization can be used to considerable advantage (owing to the application of Sinkhorn's algorithm).

The problem amounts to finding a probability measure over the space of lines which defines a flux over linear paths for $t \in [0,1]$. It should be noted that t is taken to lie within $[0,1]$ for simplicity, however, we can extend the range of t to any arbitrary interval. The obtained curve in $\mathcal{P}_2(\mathbf{R}^d)$ can be used for prediction as one can extrapolate to values of t beyond the range in the observed data-set.

We first address the absolute continuity of the elements in $\mathcal{G}_\mathcal{R}$, a property of "smooth" curves in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ recalled below.

Definition 2 ([13]). *Let $\mu_t : (0,1) \to \mathcal{P}_2(\mathbf{R}^d)$ be a one-parameter family in $\mathcal{P}_2(\mathbf{R}^d)$; we say that $(\mu_t)_{t \in [0,1]}$ belongs to $AC(0,1; \mathcal{P}_2(\mathbf{R}^d))$, if there exists $m \in L^1(0,1)$ such that*

$$W_2(\mu_s, \mu_t) \leq \int_s^t m(r) dr, \quad \forall\, 0 < s \leq t < 1.$$

Absolutely continuous curves are important due to the boundedness of their metric derivative, namely μ_t', by $m(t)$ [13], i.e.,

$$\mu_t' := \lim_{s \to t} \frac{W_2(\mu_s, \mu_t)}{|s - t|} \leq m(t).$$

The next theorem establishes absolute continuity of elements in $\mathcal{G}_\mathcal{R}$.

Theorem 3. *For any $(g_t)_{t \in [0,1]} \in \mathcal{G}_\mathcal{R}$ where $g_t = ((1-t)\mathbf{x}_0 + t\mathbf{x}_1)_{\#}\pi$, $t \in [0,1]$, we have $(g_t)_{t \in [0,1]} \in AC(0,1; \mathcal{P}_2(\mathbf{R}^d))$.*

Proof. For $0 < s \leq t < 1$, $W_2^2(g_s, g_t)$ is bounded by $(t-s)^2 c$ where c is the bound derived from the finiteness of the second moments of π. Therefore, $W_2(g_s, g_t) \leq (t-s)\sqrt{c} = \int_s^t \sqrt{c}\, dr$. $\qquad\square$

A natural question arises as to whether

$$F(\pi) := \sum_{i=1}^N \frac{1}{N} W_2^2(((1-t_i)\mathbf{x}_0 + t_i\mathbf{x}_1)_{\#}\pi, \mu_{t_i}),$$

originating in (4), is strictly displacement convex with respect to π; a functional on $\mathcal{P}_2(\mathbf{R}^d)$ is said to be (strictly) displacement convex if it is (strictly) convex along geodesics on $\mathcal{P}_2(\mathbf{R}^d)$ in the classical sense [18], [24]. This notion of convexity inherits properties of classical convex analysis, in particular, strict displacement convexity guarantees uniqueness of the minimizer.

Unfortunately, the following counter-example reveals that $F(\pi)$ may not be displacement convex, in general. Consider $\pi_0 := \frac{1}{2}\delta_{(0,1)} + \frac{1}{2}\delta_{(3,2)}$ and $\pi_1 := \frac{1}{2}\delta_{(0,0)} + \frac{1}{2}\delta_{(-3,2)}$, and a single data point $\mu_{\{t=1\}} = \frac{1}{2}\delta_1 + \frac{1}{2}\delta_2$ at $t = 1$. The unique optimal map which pushes forward π_0 to π_1, maps $(0,1)$ into $(-3,2)$ and $(3,2)$ into $(0,0)$, and thus the displacement interpolation π_s between these two measures is $\pi_s = \frac{1}{2}\delta_{(-3s,s+1)} + \frac{1}{2}\delta_{(3-3s,2-2s)}$, $s \in [0,1]$. Computing $F(\pi_s)$ along $(\pi_s)_{0 \leq s \leq 1}$ gives

$$\begin{aligned} F(\pi_s) &= W_2^2(\mathbf{x}_{1\#}\pi_s, \mu_{\{t_i=1\}}) \\ &= \min(\frac{5}{2}s^2, \frac{5}{2}s^2 - 3s + 1), \end{aligned}$$

which fails to be convex in s.

Thus, in the next section, we provide a multi-marginal optimal transport formulation for (4) which helps circumvent issues with the inherent non-linearity and provides a handle on solutions.

IV. Multi-marginal reformulation

The following gives a multi-marginal formulation of (4) and establishes properties of the minimizer.

Theorem 4. *Problem (4) assumes the following multimarginal formulation*

$$\inf_\pi F(\pi) = \inf_\gamma \int \sum_{i=1}^N \frac{1}{N} ||(1-t_i)\mathbf{x}_0 + t_i\mathbf{x}_1 - \mathbf{y}_i||^2 d\gamma$$
$$subject\ to \quad \mathbf{y}_{i\#}\gamma = \mu_{t_i}(\mathbf{y}_i), \forall i=1,...,N. \tag{5}$$

where $\gamma(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_1, ..., \mathbf{y}_N) \in \mathcal{P}_2(\mathbf{R}^{(N+2)d})$ *and* $\pi(\mathbf{x}_0, \mathbf{x}_1) \in \mathcal{P}_2(\mathbf{R}^{2d})$. *Moreover, the minimizer for* $F(\cdot)$ *exists and is* $\pi^* = (\mathbf{x}_0, \mathbf{x}_1)_{\#}\gamma^*$ *where* γ^* *is the minimizers of the right-hand side in* (5).

Proof. We sketch the main steps of the proof and we refer to [17] for details. First, suppose $\pi(\mathbf{x}_0, \mathbf{x}_1) \in \mathcal{P}_2(\mathbf{R}^{2d})$ and $\mu(\mathbf{y}) \in \mathcal{P}_2(\mathbf{R}^d)$ are given such that $g_t = \{(1-t)\mathbf{x}_0 + t\mathbf{x}_1\}_{\#}\pi,\ t \in [0,1]$ and $\eta(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}) \in \Pi(\pi, \mu)$, namely, a coupling between π and μ. We define

$$W_\eta(g_t, \mu) := \int_{\mathbf{R}^{3d}} ||(1-t)\mathbf{x}_0 + t\mathbf{x}_1 - \mathbf{y}||^2 d\eta.$$

It is easy to check that $W_2^2(g_t, \mu) \leq W_\eta^2(g_t, \mu),\ \forall t \in [0, 1]$. We can show the tightness of this inequality for some η, namely, $W_2^2(g_t, \mu) = \min_{\eta(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}) \in \Pi(\pi, \mu)} W_\eta^2(g_t, \mu)$. This can be shown by the application of Gluing Lemma (see Proposition 7.3.1 in [13]). Using the Disintegration Theorem [13], we can extend this result to a family of measures $\{\mu_{t_i}(\mathbf{y}_i)\}_{i=1,...,N} \in \mathcal{P}_2(\mathbf{R}^d)$ to show that

$$\sum_{i=1}^N W_2^2(g_{t_i}, \mu_{t_i}) = \min_{\gamma \in \mathcal{P}_2(\mathbf{R}^{(N+2)d})} \sum_{i=1}^N W_\gamma^2(g_{t_i}, \mu_{t_i})$$
$$subject\ to \quad \mathbf{y}_{i\#}\gamma = \mu_{t_i}(\mathbf{y}_i), (\mathbf{x}_0, \mathbf{x}_1)_{\#}\gamma = \pi.$$

Let γ^* be any minimizer of the right-hand side term in (5) which exists due to the existence of solution to the multi-marginal problem with quadratic cost [15]. It follows that $\pi^* = (\mathbf{x}_0, \mathbf{x}_1)\#\gamma^*$ minimizes left-hand side of (5). □

Remark 5. In Lagrangian particle tracking, one can track the Lagranigian particles resident in a fluid to measure the desired properties of them at several time steps. This tracking technique results in not only distributions of particles but also *correlation between different points in time*. In the proposed approach, the correlation of data can be incorporated by adding the linear constraints $(\mathbf{y}_i, \mathbf{y}_j)_{\#}\gamma = \pi_{t_i t_j}$, where $\pi_{t_i t_j}$ represents the joint at time steps t_i and t_j.

Remark 6. The linear problem in multi-marginal formulation (5) is computationally burdensome. A well-known strategy to alleviate this is to regularize the cost by entropy which makes the problem strictly convex. Then, we can employ iterative Bergman projections [25] to solve the regularized problem.

The following proposition accounts for consistency of the proposed method with linear regression in a Euclidean space when the target distributions are Dirac measures.

Proposition 7. *If all data points are Dirac measures, i.e.,* $\mu_{t_i}(\mathbf{y}_i) = \delta_{\mathbf{v}_i}(\mathbf{y}_i),\ i = 1, ..., N$ *where* $\mathbf{v}_i \in \mathbf{R}^d$, *the multi-marginal formulation in (5) reduces to Euclidean regression.*

Proof. Any γ that satisfies the marginal constraints in (5) can be written as $\gamma = \pi(\mathbf{x}_0, \mathbf{x}_1) \otimes \delta_{\mathbf{v}_1}(\mathbf{y}_1) \otimes ... \otimes \delta_{\mathbf{v}_N}(\mathbf{y}_N)$, where $\pi(\mathbf{x}_0, \mathbf{x}_1) \in \mathcal{P}_2(\mathbf{R}^{2d})$ and $\otimes$ represents the product of two measures. Using this fact, we can show that

$$\inf_\gamma \int_{\mathbf{R}^{(N+2)d}} \sum_{i=1}^N \frac{1}{N} ||(1-t_i)\mathbf{x}_0 + t_i\mathbf{x}_1 - \mathbf{y}_i||^2 d\gamma$$
$$= \inf_{\mathbf{w}_0, \mathbf{w}_1 \in \mathbf{R}^d} \sum_{i=1}^N \frac{1}{N} ||(1-t_i)\mathbf{w}_0 + t_i\mathbf{w}_1 - \mathbf{v}_i||^2.$$

□

A. Measure-valued PCA

Classical PCA produces subspaces of a given dimension that approximate a data point cloud. Their projection onto these subspaces can serve as approximate coordinates for the data points. Our aim is to extend the concept so as to approximate and parametrize families of probability measures $\{\mu_i\}_{i=1}^N$. In this context, the first principal component corresponds to a line segment in $\mathcal{P}_2(\mathbf{R}^d)$ that is at a minimal distance from the data points which are now probability distributions. To this end, it is natural to consider the following problem

$$\inf_{\pi \in \mathcal{P}_2(\mathbf{R}^{2d})} \sum_{i=1}^N \frac{1}{N} \min_{t \in \mathbf{R}} W_2^2(((1-t)\mathbf{x}_0 + t\mathbf{x})_{\#}\pi, \mu_i).$$

An equivalent multi-marginal formulation is

$$\inf_{\gamma} \sum_{i=1}^{N} \frac{1}{N} \min_{t \in \mathbf{R}} \int_{\mathbf{R}^{(N+2)d}} ||(1-t)\mathbf{x}_0 + t\mathbf{x}_1 - \mathbf{y}_i||^2 d\gamma$$

$$\text{subject to} \quad \mathbf{y}_{i\#}\gamma = \mu_i(\mathbf{y}_i), \forall i = 1, ..., N \tag{6}$$

where $\gamma(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_1, ..., \mathbf{y}_N) \in \mathcal{P}_2(\mathbf{R}^{(N+2)d})$. However, in the present setting of $\mathcal{P}_2(\mathbf{R}^d)$, the problem is non-convex. We employ coordinate descent method [26] in seeking a minimizer, although global optimality is not guaranteed.

The key idea is to successively minimize for t and $\gamma(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_1, ..., \mathbf{y}_N)$ to find the minimum of (6). To do so, at each iteration, we fix either γ or the parameters t, and optimize in the other direction. With respect to t the problem is quadratic assuming the closed-form solution

$$t_i^* = \frac{\int_{\mathbf{R}^{3d}} <\mathbf{y}_i - \mathbf{x}_0, \mathbf{x}_1 - \mathbf{x}_0> d\eta_i(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_i)}{\int_{\mathbf{R}^{2d}} ||\mathbf{x}_1 - \mathbf{x}_0||^2 d\pi_i(\mathbf{x}_0, \mathbf{x}_1)} \tag{7}$$

where $\eta_i(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_i) = (\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_i)_{\#}\gamma$ and $< ., . >$ denotes a corresponding inner product. Minimization over γ is a linear programming problem.

V. Results for Gaussian measures

In this section, we consider Gaussian target distributions, i.e., $\mu_{t_i}(\mathbf{y}_i) \sim N(\mathbf{m}_{\mathbf{y}_i}, \mathbf{C}_{\mathbf{y}_i})$. By applying the proposed method to this problem, we will show that the optimal interpolating curve remains Gaussian at each time step. Regarding (1), we observe that in regression analysis for Gaussian measures, the means can be treated separately by linear regression in Euclidean space. Thereby, the means for the optimal curve in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ can be readily found. Therefore, without loss of generality, we assume the means of μ_{t_i}'s to be zero. The following proposition recast the problem as an SDP.

Proposition 8. *Consider Gaussian-points* $\mu_i \sim N(0, \mathbf{C}_{\mathbf{y}_i})$. *The minimizing* $\gamma^*(\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_1, ... \mathbf{y}_N)$ *in (6) is a Gaussian distribution with mean zero and covariance that solves*

$$\min_{\mathbf{C}_{\gamma} \succeq 0} tr((1-2\bar{t}+\overline{t^2})\mathbf{C}_{\mathbf{x}_0} + \overline{t^2}\mathbf{C}_{\mathbf{x}_1} + 2(\bar{t}-\overline{t^2})\mathbf{S}_{\mathbf{x}_0\mathbf{x}_1})$$
$$-\frac{2}{N}\sum_{i=1}^{N} tr((1-t_i)\mathbf{S}_{\mathbf{x}_0\mathbf{y}_i} + t_i\mathbf{S}_{\mathbf{x}_1\mathbf{y}_i}) \tag{8}$$

where $\bar{t} = 1/N \sum_{i=1}^{N} t_i$, $\overline{t^2} = 1/N \sum_{i=1}^{N} t_i^2$ *and*

$$\mathbf{C}_{\gamma} = \begin{bmatrix} \mathbf{C}_{\mathbf{x}_0} & \mathbf{S}_{\mathbf{x}_0\mathbf{x}_1} & \mathbf{S}_{\mathbf{x}_0\mathbf{y}_1} & \cdots & \mathbf{S}_{\mathbf{x}_0\mathbf{y}_N} \\ \mathbf{S}_{\mathbf{x}_0\mathbf{x}_1}^T & \mathbf{C}_{\mathbf{x}_1} & \mathbf{S}_{\mathbf{x}_1\mathbf{y}_1} & \cdots & \mathbf{S}_{\mathbf{x}_1\mathbf{y}_N} \\ \mathbf{S}_{\mathbf{x}_0\mathbf{y}_1}^T & \mathbf{S}_{\mathbf{x}_1\mathbf{y}_1}^T & \mathbf{C}_{\mathbf{y}_1} & \cdots & \mathbf{S}_{\mathbf{y}_1\mathbf{y}_N} \\ \vdots & \vdots & \vdots & \ddots & \\ \mathbf{S}_{\mathbf{x}_0\mathbf{y}_N}^T & \mathbf{S}_{\mathbf{x}_1\mathbf{y}_N}^T & \mathbf{S}_{\mathbf{y}_1\mathbf{y}_N} & \cdots & \mathbf{C}_{\mathbf{y}_N} \end{bmatrix}. \tag{9}$$

Proof. When the marginals $(\{\mu_{t_i}\}_{i=1}^{N})$ of the joint measure γ in (5) are Gaussian, we can conclude that γ^* in (5) is also Gaussian since the marginal constraints act only on the second moments of γ, namely, no constraint is imposed onto higher-order moments, and the cost function is quadratic in $\mathbf{x}_0, \mathbf{x}_1, \mathbf{y}_1, ..., \mathbf{y}_N$. Simple calculation shows the cost function in (5) for any Gaussian γ with covariance matrix given in (9) reads

$$\int_{\mathbf{R}^{(N+2)d}} \sum_{i=1}^{N} \frac{1}{N} ||(1-t_i)\mathbf{x}_0 + t_i\mathbf{x}_1 - \mathbf{y}_i||^2 d\gamma =$$
$$(1-2\bar{t}+\overline{t^2})tr(\mathbf{C}_{\mathbf{x}_0}) + \overline{t^2}tr(\mathbf{C}_{\mathbf{x}_1}) + 2(\bar{t}-\overline{t^2})tr(\mathbf{S}_{\mathbf{x}_0\mathbf{x}_1})$$
$$-\frac{2}{N}\sum_{i=1}^{N} tr((1-t_i)\mathbf{S}_{\mathbf{x}_0\mathbf{y}_i} + t_i\mathbf{S}_{\mathbf{x}_1\mathbf{y}_i}).$$

From this, we can readily conclude that the multi-marginal formulation in (5) is equivalent to the semi-definite programming in (8). □

Noticing that $(\mathbf{x}_0, \mathbf{x}_1)_{\#}\gamma = \pi$, one can write the interpolating curve as

$$\mu_t \sim N(0, (1-t)^2 \mathbf{C}_{\mathbf{x}_0}^* + t^2 \mathbf{C}_{\mathbf{x}_1}^* + t(1-t)(\mathbf{S}_{\mathbf{x}_0\mathbf{x}}^* + \mathbf{S}_{\mathbf{x}_0\mathbf{x}}^{*T})) \tag{10}$$

where asterisk in the equation above denotes the optimal values obtained in (8). One can compare the result above to (3) by noticing that in (10) the term $\mathbf{S}_{\mathbf{x}_0\mathbf{x}_1}^*$ is not equal to $(\mathbf{C}_{\mathbf{x}_0}^* \mathbf{C}_{\mathbf{x}_1}^*)^{1/2}$. Therefore, the interpolating curve (μ_t) in (10) doesn't characterize a geodesic in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$, which may underfit the dataset as discussed earlier.

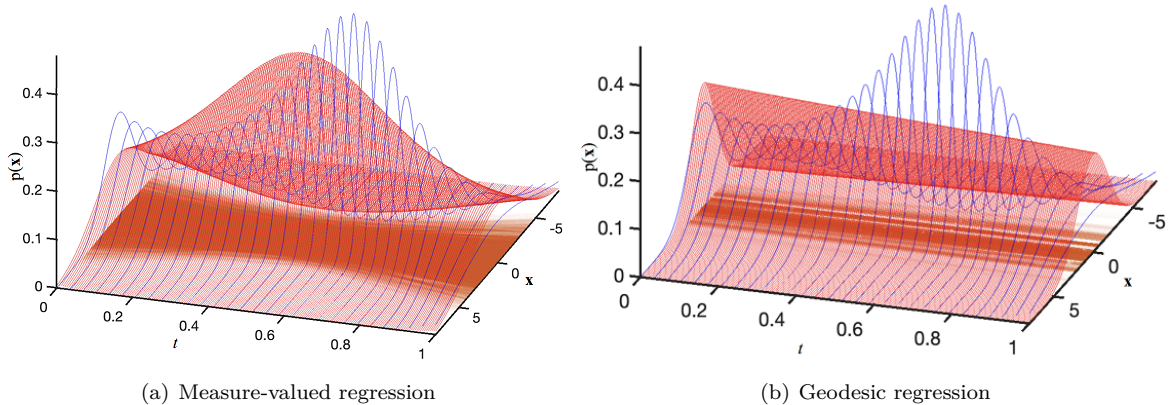|(a) Measure-valued regression|(b) Geodesic regression|

Fig. 2. Interpolation of one-dimensional Gaussian marginals. Blue curves are the given distributions and red ones are the interpolation. The intensity of color in lines is proportional to the likelihood of each path.

We conclude with numerical examples of density curves that interpolate a set of given Gaussian marginals. To do so, first we consider 30 different marginals to be one-dimensional Gaussian distributions with zero mean (blue curves in Fig. 2), where the focus is on interpolation of the respective variances. Then, the SDP in (8) is solved to obtain the optimal joint distributions γ and $\pi(\mathbf{x}_0, \mathbf{x}_1) = (\mathbf{x}_0, \mathbf{x}_1)_\# \gamma$. Figure 2(a) depicts the density curve for different values of t along with the dataset. Furthermore, to compare the result to geodesic regression, in Fig. 2(b) the optimal geodesic in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$, which interpolates the dataset, is represented. One can notice that the measure-valued regression captures the variation in the dataset better than geodesic one. This ensues from the fact that in geodesic regression an interpolating curve in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ with highest correlated end-points is sought. However, in the framework of this paper, this constraint is relaxed which can also moderate underfitting.

VI. CONCLUSIONS AND FUTURE WORKS

In this paper we introduce an approach for statistical learning in the space of distributions, metrized by the Wasserstein metric, which can be regarded as a generalization of regression analysis and PCA to the space of probability measures. To alleviate the issues with non-linearity of this formulation, we capitalize on a link between this problem and multi-marginal transport. In addition to computational benefits and consistency with Euclidean regression, our model can readily incorporate correlations between distributions (i.e., available higher-dimensionality marginals) in the dataset. The potential of the theory is envisaged in aggregate data inference with application to metapopulation dynamics [27] where the identity of individuals is not available, particle & power spectra tracking [4], system identification [2], and measurement noise and uncertainty treatment in dynamical systems. A natural future research direction is towards utilizing higher-order curves in $(\mathcal{P}_2(\mathbf{R}^d), W_2)$ along a dynamic formulation analogous to the proposed method.

REFERENCES

[1] R. E. Kalman, "Identification from real data," in *Current developments in the interface: Economics, Econometrics, Mathematics*. Springer, 1982, pp. 161–196.

[2] L. Ljung, "Perspectives on system identification," *Annual Reviews in Control*, vol. 34, no. 1, pp. 1–12, 2010.

[3] M. Niethammer, Y. Huang, and F.-X. Vialard, "Geodesic regression for image time-series," in *International conference on medical image computing and computer-assisted intervention*. Springer, 2011, pp. 655–662.

[4] X. Jiang, Z.-Q. Luo, and T. T. Georgiou, "Geometric methods for spectral analysis," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1064–1074, 2011.

[5] Y. Hong, R. Kwitt, N. Singh, B. Davis, N. Vasconcelos, and M. Niethammer, "Geodesic regression on the grassmannian," in *European Conference on Computer Vision*. Springer, 2014, pp. 632–646.

[6] M. Agueh and G. Carlier, "Barycenters in the Wasserstein space," *SIAM J. Math. Anal.*, vol. 43, no. 2, pp. 904–924, 2011. [Online]. Available: https://doi.org/10.1137/100805741

[7] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis, "Geodesic pca versus log-pca of histograms in the wasserstein space," *SIAM Journal on Scientific Computing*, vol. 40, no. 2, pp. B429–B456, 2018.

[8] B. Pass, "Multi-marginal optimal transport: theory and applications," *ESAIM: Mathematical Modelling and Numerical Analysis*, vol. 49, no. 6, pp. 1771–1790, 2015.

[9] J. Bigot, R. Gouet, T. Klein, A. López *et al.*, "Geodesic pca in the wasserstein space by convex pca," in *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, vol. 53, no. 1. Institut Henri Poincaré, 2017, pp. 1–26.

[10] L. Ambrosio, N. Gigli, and G. Savaré, "Gradient flows with metric and differentiable structures, and applications to the wasserstein space," *Atti Accad. Naz. Lincei Cl. Sci. Fis. Mat. Natur. Rend. Lincei (9) Mat. Appl*, vol. 15, no. 3-4, pp. 327–343, 2004.

[11] W. Wang, D. Slepčev, S. Basu, J. A. Ozolek, and G. K. Rohde, "A linear optimal transportation framework for quantifying and visualizing variations in sets of images," *International journal of computer vision*, vol. 101, no. 2, pp. 254–269, 2013.

[12] V. Seguy and M. Cuturi, "Principal geodesic analysis for probability measures under the optimal transport metric," in *Advances in Neural Information Processing Systems*, 2015, pp. 3312–3320.

[13] L. Ambrosio, N. Gigli, and G. Savaré, *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[14] P. T. Fletcher, "Geodesic regression and the theory of least squares on riemannian manifolds," *International journal of computer vision*, vol. 105, no. 2, pp. 171–185, 2013.

[15] Y. Chen, G. Conforti, and T. T. Georgiou, "Measure-valued spline curves: An optimal transport viewpoint," *SIAM Journal on Mathematical Analysis*, vol. 50, no. 6, pp. 5947–5968, 2018.

[16] J.-D. Benamou, T. Gallouët, and F.-X. Vialard, "Second order models for optimal transport and cubic splines on the wasserstein space," *arXiv preprint arXiv:1801.04144*, 2018.

[17] A. Karimi, L. Ripani, and T. Georgiou, "Statistical learning in W_2," *In preparation*, 2020.

[18] C. Villani, *Topics in optimal transportation*. American Mathematical Soc., 2003, no. 58.

[19] L. Ambrosio and N. Gigli, "A user's guide to optimal transport," in *Modelling and optimisation of flows on networks*. Springer, 2013, pp. 1–155.

[20] C. Villani, *Optimal transport: old and new*. Springer Science & Business Media, 2008, vol. 338.

[21] L. Malagò, L. Montrucchio, and G. Pistone, "Wasserstein riemannian geometry of gaussian densities," *Information Geometry*, vol. 1, no. 2, pp. 137–179, 2018.

[22] L. Nenna, "Numerical methods for multi-marginal optimal transportation," Ph.D. dissertation, Université Paris-Dauphine, 2016.

[23] J.-D. Benamou, G. Carlier, and L. Nenna, "Generalized incompressible flows, multi-marginal transport and sinkhorn algorithm," *Numerische Mathematik*, vol. 142, no. 1, pp. 33–54, 2019.

[24] R. J. McCann, "A convexity principle for interacting gases," *Advances in mathematics*, vol. 128, no. 1, pp. 153–179, 1997.

[25] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, pp. A1111–A1138, 2015.

[26] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[27] J. M. Nichols, J. A. Spendelow, and J. D. Nichols, "Using optimal transport theory to estimate transition probabilities in metapopulation dynamics," *Ecological Modelling*, vol. 359, pp. 311–319, 2017.