



Consistent estimation of a population barycenter in the Wasserstein space

Jérémie Bigot, Thierry Klein

► To cite this version:

Jérémie Bigot, Thierry Klein. Consistent estimation of a population barycenter in the Wasserstein space. 2013. hal-00763668v4

HAL Id: hal-00763668

<https://hal.science/hal-00763668v4>

Preprint submitted on 9 Mar 2015 (v4), last revised 28 Nov 2017 (v7)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Consistent estimation of a population barycenter in the Wasserstein space

Jérémie Bigot¹ and Thierry Klein²

Institut de Mathématiques de Bordeaux et CNRS (UMR 5251)¹
Université de Bordeaux

Institut de Mathématiques de Toulouse et CNRS (UMR 5219)³
Université de Toulouse

March 9, 2015

Abstract

We define a notion of barycenter for random probability measures in the Wasserstein space. Using a duality argument, we give a precise characterization of the population barycenter for various parametric classes of random probability measures with compact support. In particular, we make a connection between averaging in the Wasserstein space and taking the expectation of optimal transport maps with respect to a fixed reference measure. We also discuss the usefulness of this approach for warping problems in statistics and data analysis. In this setting, the problem of estimating a population barycenter from n independent and identically distributed random probability measures is also considered. In particular, we study the convergence of the empirical barycenter proposed in Agueh and Carlier [2] to its population counterpart as the number of measures n tends to infinity.

Keywords: Wasserstein space; Empirical and population barycenters; Fréchet mean; Convergence of random variables; Optimal transport; Duality; Curve and image warping.

AMS classifications: Primary 62G05; secondary 49J40.

Acknowledgements

The authors acknowledge the support of the French Agence Nationale de la Recherche (ANR) under reference ANR-JCJC-SIMI1 DEMOS. We would like to also thank Jérôme Bertrand for fruitful discussions on Wasserstein spaces and the optimal transport problem.

1 Introduction

In this paper, we consider the problem of defining the barycenter of random probability measures on \mathbb{R}^d . The set of Radon probability measures endowed with the 2-Wasserstein distance is not an Euclidean space. Consequently, to define a notion of barycenter for random probability measures, it is natural to use the notion of Fréchet mean [20] that is an extension of the usual Euclidean barycenter to non-linear spaces endowed with non-Euclidean metrics. If \mathbf{Y} denotes a random

variable with distribution \mathbb{P} taking its value in a metric space $(\mathcal{M}, d_{\mathcal{M}})$, then a Fréchet mean (not necessarily unique) of the distribution \mathbb{P} is a point $m^* \in \mathcal{M}$ that is a global minimum (if any) of the functional

$$J(m) = \frac{1}{2} \int_{\mathcal{M}} d_{\mathcal{M}}^2(m, y) d\mathbb{P}(y) \quad \text{i.e.} \quad m^* \in \arg \min_{m \in \mathcal{M}} J(m).$$

In this paper, a Fréchet mean of a random variable \mathbf{Y} with distribution \mathbb{P} will be also called a barycenter. An empirical Fréchet mean of an independent and identically distributed (iid) sample $\mathbf{Y}_1, \dots, \mathbf{Y}_n$ of distribution \mathbb{P} is

$$\bar{\mathbf{Y}}_n \in \arg \min_{m \in \mathcal{M}} \frac{1}{n} \sum_{j=1}^n \frac{1}{2} d_{\mathcal{M}}^2(m, \mathbf{Y}_j).$$

For random variables belonging to nonlinear metric spaces, a well-known example is the computation of the mean of a set of planar shapes in the Kendall's shape space [29] that leads to the Procrustean means studied in [22]. Many properties of the Fréchet mean in finite dimensional Riemannian manifolds (such as consistency and uniqueness) have been investigated in [1, 5, 7, 8, 27]. For random variables taking their value in metric spaces of nonpositive curvature (NPC), a detailed study of various properties of their barycenter can be found in [34]. However, there is not so much work on Fréchet means in infinite dimensional metric spaces that do not satisfy the global NPC property as defined in [34].

1.1 A parametric class of random probability measures

In this paper, we consider the case where $\mathbf{Y} = \mu$ is a random probability measure belonging to the 2-Wasserstein space on a compact set Ω of \mathbb{R}^d . In this work, we consider the following parametric class of random probability measures. Let $\mathcal{M}_+(\Omega)$ be the set of Radon probability measures on Ω endowed with the 2-Wasserstein distance d_{W_2} between two probability measures. Let $\Phi : (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p)) \rightarrow (\mathcal{M}_+(\Omega), \mathcal{B}(\mathcal{M}_+(\Omega)))$ be a measurable mapping, where $\mathcal{B}(\mathbb{R}^p)$ is the Borel σ -algebra of \mathbb{R}^p and $\mathcal{B}(\mathcal{M}_+(\Omega))$ is the Borel σ -algebra generated by the topology induced by the distance d_{W_2} . For a compact subset Θ of \mathbb{R}^p (with $p \geq 1$), we consider the parametric set of probability measures $\{\mu_\theta = \Phi(\theta), \theta \in \Theta\}$. Furthermore, we assume that, for any $\theta \in \Theta$, the measure $\mu_\theta = \Phi(\theta) \in \mathcal{M}_+(\Omega)$ admits a density with respect to the Lebesgue measure on \mathbb{R}^d . It follows that if $\boldsymbol{\theta} \in \mathbb{R}^p$ is a random vector with distribution \mathbb{P}_Θ admitting a density $g : \Theta \rightarrow \mathbb{R}_+$, then $\mu_{\boldsymbol{\theta}} = \Phi(\boldsymbol{\theta})$ is a random probability measure with distribution \mathbb{P}_g on $(\mathcal{M}_+(\Omega), \mathcal{B}(\mathcal{M}_+(\Omega)))$ that is the push-forward measure defined by

$$\mathbb{P}_g(B) = \mathbb{P}_\Theta(\Phi^{-1}(B)), \text{ for any } B \in \mathcal{B}(\mathcal{M}_+(\Omega)). \quad (1.1)$$

In this paper, we propose to study some properties of the barycenter μ^* of $\mu_{\boldsymbol{\theta}}$ defined as the following Fréchet mean

$$\mu^* = \arg \min_{\nu \in \mathcal{M}_+(\Omega)} J(\nu), \quad (1.2)$$

where

$$J(\nu) = \int_{\mathcal{M}_+(\Omega)} \frac{1}{2} d_{W_2}^2(\nu, \mu) d\mathbb{P}_g(\mu) = \mathbb{E} \left(\frac{1}{2} d_{W_2}^2(\nu, \mu_{\boldsymbol{\theta}}) \right) = \int_{\Theta} \frac{1}{2} d_{W_2}^2(\nu, \mu_{\theta}) g(\theta) d\theta, \quad \nu \in \mathcal{M}_+(\Omega). \quad (1.3)$$

If it exists and is unique, the measure μ^* will be referred to as the population barycenter of the random measure $\mu_{\boldsymbol{\theta}}$ with distribution \mathbb{P}_g .

1.2 Related literature

Throughout the paper, we discuss various statistical models that involve the use of such parametric random measures, and for which the notion of barycenter in the 2-Wasserstein space is relevant. A similar notion (to the one in this paper) of a population barycenter and its connection to optimal transportation with infinitely many marginals have been studied in [31]. In particular, a similar class of parametric random probability measures, where the parameter set Θ is one-dimensional, is also considered in [31] for the purpose of studying the existence and uniqueness of μ^* . Some generalization of the results in [31] for probability measures defined on a Riemannian manifold, equipped with the Wasserstein metric, have been recently proposed in [30]. The empirical counterpart of μ^* is the barycenter $\bar{\mu}_n$ defined as

$$\bar{\mu}_n = \arg \min_{\nu \in \mathcal{M}_+(\Omega)} \frac{1}{n} \sum_{j=1}^n \frac{1}{2} d_{W_2}^2(\nu, \mu_{\boldsymbol{\theta}_j}), \quad (1.4)$$

where $\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n$ are iid random vectors in Θ with density g . A detailed characterization of empirical barycenters, in a broader setting, in terms of existence, uniqueness and regularity, together with its link to the multi-marginal problem in optimal transport can be found in [2].

1.3 Main results of the paper

The main contribution of this paper is to give an explicit characterization of the population barycenter μ^* that can be (informally) stated as follows: let $\mu_0 \in \mathcal{M}_+(\Omega)$ be a fixed reference measure that is absolutely continuous with respect to the Lebesgue measure on \mathbb{R}^d , and let $T_{\boldsymbol{\theta}} : \Omega \rightarrow \Omega$ be the optimal mapping to transport μ_0 onto $\mu_{\boldsymbol{\theta}}$ (i.e. $\mu_{\boldsymbol{\theta}} = T_{\boldsymbol{\theta}} \# \mu_0$). If $\mathbb{E}(T_{\boldsymbol{\theta}}(x)) = x$ for all $x \in \Omega$, then μ_0 is equal to the population barycenter μ^* , and one has that

$$\inf_{\nu \in \mathcal{M}_+(\Omega)} J(\nu) = \frac{1}{2} \int_{\Omega} \mathbb{E}(|T_{\boldsymbol{\theta}}(x) - x|^2) d\mu^*(x) = \frac{1}{2} \int_{\Omega} \int_{\Theta} (|T_{\theta}(x) - x|^2) g(\theta) d\theta d\mu^*(x).$$

This result can also be written in the form $\mu^* = \bar{T} \# \mu_0$, where $\bar{T} = \mathbb{E}(T_{\boldsymbol{\theta}})$ is the mapping defined by $\bar{T}(x) = \mathbb{E}(T_{\boldsymbol{\theta}}(x))$, for all $x \in \Omega$. This suggests that averaging in the Wasserstein space may amount to take the expectation (in the usual sense) of the optimal transport map $T_{\boldsymbol{\theta}}$ with respect to a fixed reference measure μ_0 . However, this result is generally not true if $\mathbb{E}(T_{\boldsymbol{\theta}}) \neq I$, where $I = \Omega \rightarrow \Omega$ denotes the identity mapping. Nevertheless, we propose to consider sufficient conditions (beyond the case $\mathbb{E}(T_{\boldsymbol{\theta}}) = I$) ensuring that

$$\mu^* = \bar{T} \# \mu_0, \quad \text{where } \bar{T} = \mathbb{E}(T_{\boldsymbol{\theta}}). \quad (1.5)$$

To the best of our knowledge, such a result on the population barycenter in the 2-Wasserstein space is new. A similar result has been established in [13] for the empirical barycenter, namely that

$$\bar{\mu}_n = \left(\frac{1}{n} \sum_{j=1}^n T_{\theta_j} \right) \# \mu_0, \quad (1.6)$$

under specific assumptions on the optimal maps T_{θ_j} and the reference measure μ_0 such that $\mu_{\theta_j} = T_{\theta_j} \# \mu_0$ for $j = 1, \dots, n$. The validity of equation (1.6) stated in [13] is restricted to a specific class of optimal transport maps that is assumed to be admissible in the sense that the composition of two maps in this class remains an optimal one. The extension of the results in [13] to the population barycenter has not been considered so far. In this paper, we do not assume that the collection of maps $(T_\theta)_{\theta \in \Theta}$ is an admissible class to prove that $\mu^* = \mu_0$ if $\mathbb{E}(T_{\bar{\theta}}) = I$. Nevertheless, we prove that equation (1.5) may hold under the assumption that $T_\theta \circ \bar{T}^{-1}$ is an optimal map for all $\theta \in \Theta$.

Equation (1.5) is not difficult to prove in a one-dimensional setting ($d = 1$) i.e. for random measures supported on the real line. The main issue is then to extend this result to higher dimensions $d \geq 2$. To this end, we consider a dual formulation of the optimization problem (1.3) that allows a precise characterization of some properties of the population barycenter that are used to prove equation (1.5). These results are based on an adaptation of the duality arguments developed in [2] for the characterization of an empirical barycenter. Therefore, our approach is very much connected with the theory of optimal mass transport, and with the characterization of the Monge-Kantorovich problem via arguments from convex analysis and duality, see [38] for further details on this topic. Another contribution of this paper is to study the consistency of $\bar{\mu}_n$ to μ^* as the number n of measures tends to infinity.

1.4 Organisation of the paper

The paper is then organized as follows. In Section 2, we introduce some definitions and notation for the framework of the paper, and we discuss the existence and uniqueness of the population barycenter. In Section 3, we characterize the population barycenter in the one-dimensional case, i.e. for random measures supported on $\Omega \subset \mathbb{R}$. In Section 4, we prove the main result of the paper, namely equation (1.5) in dimension $d \geq 2$. As an application of the methodology developed in this paper, we discuss, in Section 5, the usefulness of barycenters in the Wasserstein space for various problems in statistics and data analysis. The consistency of the empirical barycenter is discussed in Section 6. In Section 7, we introduce a dual formulation of the optimization problem (1.3) that is used to prove the main result of this paper. In Section 8, we finally give a conclusion and some perspectives on the extension of this work.

2 Existence and uniqueness of the population barycenter

2.1 Some definitions and notation

We use bold symbols $\mathbf{Y}, \boldsymbol{\mu}, \boldsymbol{\theta}, \dots$ to denote random objects. The notation $|x|$ is used to denote the usual Euclidean norm of a vector $x \in \mathbb{R}^m$, the notation $\langle x, y \rangle$ denotes the usual inner product

for $x, y \in \mathbb{R}^m$, and $\mathcal{M}_+(\Omega)$ is the set of Radon probability measures on Ω . We recall the following assumptions that are made throughout the paper.

Assumption 1. *The sets $\Omega \subset \mathbb{R}^d$ and $\Theta \subset \mathbb{R}^p$ are compact.*

Assumption 2. *The mapping $\Phi : (\Theta \subset \mathbb{R}^p, \mathcal{B}(\mathbb{R}^p)) \rightarrow (\mathcal{M}_+(\Omega), \mathcal{B}(\mathcal{M}_+(\Omega)))$ is measurable. Moreover, it is such that, for any $\theta \in \Theta$, the measure $\mu_\theta = \Phi(\theta) \in \mathcal{M}_+(\Omega)$ admits a density with respect to the Lebesgue measure on \mathbb{R}^d .*

The squared 2-Wasserstein distance between two probability measures $\mu, \nu \in \mathcal{M}_+(\Omega)$ is

$$d_{W_2}^2(\mu, \nu) := \inf_{\gamma \in \Pi(\mu, \nu)} \left\{ \int_{\Omega \times \Omega} |x - y|^2 d\gamma(x, y) \right\},$$

where $\Pi(\mu, \nu)$ is the set of all probability measures on $\Omega \times \Omega$ having μ and ν as marginals, see e.g. [38]. We recall that $\tilde{\gamma} \in \Pi(\mu, \nu)$ is called an optimal transport plan between μ and ν if

$$d_{W_2}^2(\mu, \nu) = \int_{\Omega \times \Omega} |x - y|^2 d\tilde{\gamma}(x, y).$$

Let $T : \Omega \rightarrow \Omega$ be a measurable mapping, and let $\mu \in \mathcal{M}_+(\Omega)$. The push-forward measure $T\#\mu$ of μ through the map T is the measure defined by duality as

$$\int_{\Omega} f(x) d(T\#\mu)(x) = \int_{\Omega} f(T(x)) d\mu(x), \quad \text{for all continuous and bounded functions } f : \Omega \rightarrow \mathbb{R}.$$

We also recall the following well known result in optimal transport (see e.g. [38] or Proposition 3.3 in [2]):

Proposition 2.1. *Let $\mu, \nu \in \mathcal{M}_+(\Omega)$. Then, $\gamma \in \Pi(\mu, \nu)$ is an optimal transport plan between μ and ν if and only if the support of γ is included in the set $\partial\phi$ that is the graph of the subdifferential of a convex and lower semi-continuous function ϕ solution of the problem*

$$\phi = \arg \min_{\psi \in \mathcal{C}} \left\{ \int_{\Omega} \psi(x) d\mu(x) + \int_{\Omega} \psi^*(x) d\nu(x) \right\},$$

where $\psi^*(x) = \sup_{y \in \Omega} \{ \langle x, y \rangle - \psi(y) \}$ is the convex conjugate of ψ , and \mathcal{C} denotes the set of proper convex functions $\psi : \Omega \rightarrow \mathbb{R}$ that are lower semi-continuous

If μ admits a density with respect to the Lebesgue measure on \mathbb{R}^d , then there exists a unique optimal transport plan $\gamma \in \Pi(\mu, \nu)$ that is of the form $\gamma = (\text{id}, T)\#\mu$ where $T = \nabla\phi$ (the gradient of ϕ) is called the optimal mapping between μ and ν . The uniqueness of the transport plan holds in the sense that if $\nabla\phi\#\mu = \nabla\psi\#\mu$, where $\psi : \Omega \rightarrow \mathbb{R}$ is a lower semi-continuous convex function, then $\nabla\phi = \nabla\psi$ μ -almost everywhere. Moreover, one has that

$$d_{W_2}^2(\mu, \nu) = \int_{\Omega} |\nabla\phi(x) - x|^2 d\mu(x).$$

2.2 About the measurability of T_θ

Let μ_0 be a fixed reference measure that is absolutely continuous with respect to the Lebesgue measure on \mathbb{R}^d . In the introduction, for any $\theta \in \Theta$, we have defined $T_\theta : \Omega \rightarrow \Omega$ as the optimal mapping to transport μ_0 onto μ_θ . By Proposition 2.1, such a mapping exists, and it is the μ_0 -almost everywhere unique one that can be written as the gradient of a convex function ϕ_θ . Then, thanks to Assumption 2, it follows, by Theorem 1.1 in [19], that there exists a function $(\theta, x) \mapsto T(\theta, x)$ which is measurable with respect to the σ -algebra $\mathcal{B}(\mathbb{R}^p) \otimes \mathcal{B}(\mathcal{M}_+(\Omega))$, and such that for \mathbb{P}_Θ -almost everywhere $\theta \in \Theta$,

$$T(\theta, x) = T_\theta(x), \text{ for } \mu_0\text{-almost everywhere } x \in \Omega.$$

In particular, the mapping $(x, \theta) \mapsto T_\theta(x)$ is measurable with respect to the completion of $\mathcal{B}(\mathbb{R}^p) \otimes \mathcal{B}(\mathcal{M}_+(\Omega))$ with respect to $g(\theta)d\theta d\mu_0(x)$ (we recall the assumption that $d\mathbb{P}_\Theta(\theta) = g(\theta)d\theta$).

2.3 Existence and uniqueness of μ^*

Let us now prove the existence and uniqueness of the barycenter (i.e. the Fréchet mean) of the parametric random measures μ_θ with distribution \mathbb{P}_g , as defined in equation (1.1). We recall that this amounts to study the solution (if any) of the optimization problem (1.2).

Remark 2.1 (About the existence and finiteness of $J(\nu)$). *Let $\nu \in \mathcal{M}_+(\Omega)$. The integral defining $J(\nu)$ will be defined as soon as $\theta \mapsto d_{W_2}^2(\nu, \mu_\theta)g(\theta)$ is a measurable application. The measurability of this mapping follows from Assumption 2. Moreover, since Ω is compact by Assumption 1, it follows that, for any $\theta \in \Theta$, $d_{W_2}^2(\nu, \mu_\theta) \leq 4\delta^2(\Omega)$, where $\delta^2(\Omega) = \sup_{x \in \Omega} \{|x|^2\} < +\infty$. Therefore, one has that $J(\nu) = \frac{1}{2} \int_\Theta d_{W_2}^2(\nu, \mu_\theta)g(\theta)d\theta \leq 2\delta^2(\Omega) < +\infty$, for any $\nu \in \mathcal{M}_+(\Omega)$.*

As shown below, using standard arguments, the existence and the uniqueness of μ^* are not difficult to prove. To this end, a key property of the functional J defined in (1.3) is the following:

Lemma 2.1. *Suppose that Assumption 2 holds. Then, the functional $J : \mathcal{M}_+(\Omega) \rightarrow \mathbb{R}$ is strictly convex in the sense that*

$$J(\lambda\mu + (1-\lambda)\nu) < \lambda J(\mu) + (1-\lambda)J(\nu), \text{ for any } \lambda \in]0, 1[\text{ and } \mu, \nu \in \mathcal{M}_+(\Omega) \text{ with } \mu \neq \nu. \quad (2.1)$$

Proof. Inequality (2.1) follows immediately from the assumption that, for any $\theta \in \Theta$, the measure μ_θ admits a density with respect to the Lebesgue measure on \mathbb{R}^d , and from the use of Theorem 2.9 in [4] (for similar results to those in [4] on the strict convexity, one can read Lemma 3.2.1 in [31]). \square

Hence, thanks to the strict convexity of J , if a barycenter μ^* exists, then it is necessarily unique. The existence of μ^* is then proved in the next proposition.

Proposition 2.2. *Under Assumption 1 and Assumption 2, the optimization problem (1.2) admits a unique minimizer.*

Proof. Let ν^n be a minimizing sequence of the optimization problem (1.2). Let us first show that the sequence $\int_{\Omega} |x|^2 d\nu^n(x)$ is uniformly bounded. Thanks to Remark 2.1, there exists a constant $C > 0$ such that $\int_{\Theta} \frac{1}{2} d_{W_2}^2(\mu_{\theta}, \nu_n) g(\theta) d\theta \leq C$, for all $n \geq 1$. Since g is a probability density function on Θ with respect to the Lebesgue measure, it follows that $\Theta' := \{\theta, g(\theta) > 0\}$ is a non negligible set (for the Lebesgue measure). Since $\int_{\Theta'} \frac{1}{2} d_{W_2}^2(\mu_{\theta}, \nu_n) g(\theta) d\theta \leq C$, one obtains that there exists $\theta^* \in \Theta'$, such that $g(\theta^*) \neq 0$ and

$$\frac{1}{2} d_{W_2}^2(\mu_{\theta^*}, \nu_n) g(\theta^*) \leq C' \quad (2.2)$$

where $C' = \frac{C}{\int_{\Theta'} d\theta}$. Then, thanks to the Kantorovich duality formula (see e.g. [38], or Lemma 2.1 in [2]), it follows that

$$\frac{1}{2} d_{W_2}^2(\mu_{\theta^*}, \nu_n) g(\theta^*) \geq \int_{\Omega} |x|^2 d\nu_n(x) + \int_{\Omega} C_{g(\theta^*)}(x) d\mu_{\theta^*}(x),$$

where $C_{g(\theta^*)}(x) = \inf_{y \in \Omega} \left\{ \frac{g(\theta^*)}{2} |x - y|^2 - |y|^2 \right\}$. Note that by Assumption 1, Ω is a compact set, and therefore the function $x \mapsto |C_{g(\theta^*)}(x)|$ is bounded on Ω . Hence, combining the above inequality with (2.2), one finally obtains that

$$\int_{\Omega} |x|^2 d\nu_n(x) \leq C' - \int_{\Omega} C_{g(\theta^*)}(x) d\mu_{\theta^*}(x), \text{ for all } n \geq 1,$$

which shows that $\int_{\Omega} |x|^2 d\nu^n(x)$ is a uniformly bounded sequence.

Hence, by Chebyshev's inequality, the sequence ν_n is tight and by Prokhorov's Theorem there exists a (non relabeled) subsequence that weakly converges to some $\mu^* \in \mathcal{M}_+(\Omega)$. Therefore, $\frac{1}{2} d_{W_2}^2(\mu_{\theta}, \mu^*) \leq \liminf_{n \rightarrow +\infty} \frac{1}{2} d_{W_2}^2(\mu_{\theta}, \nu^n)$, and thus, by Fatou's Lemma

$$\int_{\Theta} \frac{1}{2} d_{W_2}^2(\mu_{\theta}, \mu^*) g(\theta) d\theta \leq \int_{\Theta} \liminf_{n \rightarrow +\infty} \frac{1}{2} d_{W_2}^2(\mu_{\theta}, \nu^n) g(\theta) d\theta \leq \liminf_{n \rightarrow +\infty} \int_{\Theta} \frac{1}{2} d_{W_2}^2(\mu_{\theta}, \nu_n) g(\theta) d\theta.$$

Therefore, $J(\mu^*) = \inf_{\nu \in \mathcal{M}_+(\Omega)} \frac{1}{2} \int_{\Theta} d_{W_2}^2(\nu, \mu_{\theta}) g(\theta) d\theta$, which proves that the optimization problem (1.2) admits a minimizer.

By the strict convexity of the functional J (as stated in Lemma 2.1), it follows that the barycenter of μ_{θ} is necessarily unique. \square

3 Barycenter for measures supported on the real line

In this section, we study some properties of the population barycenter μ^* of random measures supported on the real line i.e. we consider the case $d = 1$ where Ω is a compact subset of \mathbb{R} . In this setting, our characterization of μ^* will follow from the well known fact that if μ and ν are measures belonging to $\mathcal{M}_+(\Omega)$ then

$$d_{W_2}^2(\nu, \mu) = \int_0^1 |F_{\nu}^{-1}(x) - F_{\mu}^{-1}(x)|^2 dx, \quad (3.1)$$

where F_ν^{-1} (resp. F_μ^{-1}) is the quantile function of ν (resp. μ). This explicit expression for the Wasserstein distance allows a simple characterization of the barycenter of random measures. In particular, we prove that equation (1.5) always holds for $d = 1$. This result means that, in dimension 1, computing a barycenter in the Wasserstein space amounts to take the expectation (in the usual sense) of the optimal mapping to transport a fixed (non-random) reference measure μ_0 onto μ_θ .

Theorem 3.1. *Let μ_0 be any fixed measure in $\mathcal{M}_+(\Omega)$ that is absolutely continuous with respect to the Lebesgue measure. Suppose that Assumption 1 and Assumption 2 hold. Let $\mu_\theta = \Phi(\theta)$ be a random measure where θ is a random vector in Θ with density g . Let T_θ be the random optimal mapping between μ_0 and μ_θ given by $T_\theta(x) = F_{\mu_\theta}^{-1}(F_{\mu_0}(x))$, $x \in \Omega$, where $F_{\mu_\theta}^{-1}$ is the quantile function of μ_θ , and F_{μ_0} is the cumulative distribution function of μ_0 .*

Then, the barycenter of μ_θ exists, it is unique, and it satisfies the equation

$$\mu^* = \bar{T} \# \mu_0, \quad (3.2)$$

where $\bar{T} : \Omega \rightarrow \Omega$ denotes the optimal mapping between μ_0 and μ^* that is defined by

$$\bar{T}(x) = \mathbb{E}(T_\theta(x)) = \int_{\Theta} T_\theta(x) g(\theta) d\theta, \text{ for all } x \in \Omega.$$

Furthermore, the quantile function of μ^* is

$$F_{\mu^*}^{-1}(y) = \mathbb{E}\left(F_{\mu_\theta}^{-1}(y)\right) = \int_{\Theta} F_{\mu_\theta}^{-1}(y) g(\theta) d\theta,$$

for $y \in [0, 1]$. Thus, the barycenter μ^* does not depend on the choice of μ_0 , and one has that

$$\inf_{\nu \in \mathcal{M}_+(\Omega)} J(\nu) = \frac{1}{2} \int_{\Omega} \mathbb{E}(|\bar{T}_\theta(x) - x|^2) d\mu^*(x) = \frac{1}{2} \int_{\Omega} \int_{\Theta} |\bar{T}_\theta(x) - x|^2 g(\theta) d\theta d\mu^*(x), \quad (3.3)$$

where $\bar{T}_\theta = T_\theta \circ \bar{T}^{-1}$.

Proof. Let $\nu \in \mathcal{M}_+(\Omega)$ then

$$J(\nu) = \int_{\Theta} \frac{1}{2} d_{W_2}^2(\nu, \mu_\theta) g(\theta) d\theta = \frac{1}{2} \int_{\Theta} \int_0^1 |F_\nu^{-1}(y) - F_{\mu_\theta}^{-1}(y)|^2 dy g(\theta) d\theta.$$

In the proof, we will repeatedly use Fubini's Theorem to interchange the order of integration in the right-hand side of the above equation. To be valid, Fubini's Theorem requires that the application $(y, \theta) \mapsto |F_\nu^{-1}(y) - F_{\mu_\theta}^{-1}(y)|^2 g(\theta)$ is measurable. This application will be measurable as soon as $(y, \theta) \mapsto F_{\mu_\theta}^{-1}(y)$ is measurable. Since, by definition, $F_{\mu_\theta}^{-1}(y) = T_\theta(F_{\mu_0}^{-1}(y))$ for any $y \in [0, 1]$, the measurability of $(y, \theta) \mapsto F_{\mu_\theta}^{-1}(y)$ is ensured by the continuity of $y \mapsto F_{\mu_0}^{-1}(y)$ and the measurability of the mapping $(x, \theta) \mapsto T_\theta(x)$ which has been stated in Section 2.2.

Let us now consider the function $\mathbb{E}\left(F_{\mu_\theta}^{-1}\right) : [0, 1] \rightarrow \Omega$ defined by

$$\mathbb{E}\left(F_{\mu_\theta}^{-1}\right)(y) := \mathbb{E}\left(F_{\mu_\theta}^{-1}(y)\right) = \int_{\Theta} F_{\mu_\theta}^{-1}(y) g(\theta) d\theta < +\infty$$

for any $y \in [0, 1]$. We will now prove that $\mathbb{E} \left(F_{\mu_{\theta}}^{-1} \right)$ is a quantile function. To this end, let $\delta^2(\Omega) = \sup_{x \in \Omega} \{|x|^2\}$ that is finite by Assumption 1 (compactness of Ω). Then, by Assumption 2, the measure μ_{θ} admits a density with respect to the Lebesgue measure, and thus the function $y \mapsto F_{\mu_{\theta}}^{-1}(y)$ is continuous on $[0, 1]$, for any $\theta \in \Theta$. Hence, since $|F_{\mu_{\theta}}^{-1}(y)| \leq \delta(\Omega)$ for any $(\theta, y) \in \Theta \times [0, 1]$ it follows, by Lebesgue's dominated convergence theorem, that $y \mapsto \mathbb{E} \left(F_{\mu_{\theta}}^{-1} \right)(y)$ is a continuous function on $[0, 1]$. Since $y \mapsto F_{\mu_{\theta}}^{-1}(y)$ is increasing for any $\theta \in \Theta$, it is also clear that $y \mapsto \mathbb{E} \left(F_{\mu_{\theta}}^{-1} \right)(y)$ is an increasing function. Hence, thanks to the property that any continuous and increasing function from $[0, 1]$ to Ω is the quantile of some probability measure belonging to $\mathcal{M}_+(\Omega)$ (see e.g. [18]), one obtains that $\mathbb{E} \left(F_{\mu_{\theta}}^{-1} \right)$ is a quantile function.

Now, it is clear that $\mathbb{E} \left(|F_{\mu_{\theta}}^{-1}(y)|^2 \right) \leq \delta^2(\Omega) < +\infty$ for any $y \in [0, 1]$. Hence, applying Fubini's Theorem, and the fact that $\mathbb{E} |a - X|^2 \geq \mathbb{E} |\mathbb{E}(X) - X|^2$ for any squared integrable real random variable X and real number a , we obtain that

$$\begin{aligned}
\int_{\Theta} d_{W_2}^2(\nu, \mu_{\theta}) g(\theta) d\theta &= \int_0^1 \int_{\Theta} |F_{\nu}^{-1}(y) - F_{\mu_{\theta}}^{-1}(y)|^2 g(\theta) d\theta dy = \int_0^1 \mathbb{E} \left| F_{\nu}^{-1}(y) - F_{\mu_{\theta}}^{-1}(y) \right|^2 dy \\
&\geq \int_0^1 \mathbb{E} \left| \mathbb{E} \left(F_{\mu_{\theta}}^{-1}(y) \right) - F_{\mu_{\theta}}^{-1}(y) \right|^2 dy \\
&= \int_0^1 \int_{\Theta} \left| \mathbb{E} \left(F_{\mu_{\theta}}^{-1}(y) \right) - F_{\mu_{\theta}}^{-1}(y) \right|^2 g(\theta) d\theta dy \\
&= \int_{\Theta} d_{W_2}^2(\mu^*, \mu_{\theta}) g(\theta) d\theta,
\end{aligned} \tag{3.4}$$

where μ^* is the measure in $\mathcal{M}_+(\Omega)$ with quantile function given by $F_{\mu^*}^{-1} = \mathbb{E} \left(F_{\mu_{\theta}}^{-1} \right)$. The above inequality shows that $J(\nu) \geq J(\mu^*)$ for any $\nu \in \mathcal{M}_+(\Omega)$. Therefore, μ^* is a barycenter of the random measure μ_{θ} , and the unicity of μ^* follows from the strict convexity of the functional J as stated in Lemma 2.1. Finally, let μ_0 be any fixed measure in $\mathcal{M}_+(\Omega)$ that is absolutely continuous with respect to the Lebesgue measure. Hence, one has that $F_{\mu_0} \circ F_{\mu_0}^{-1}(y) = y$ for any $y \in [0, 1]$. Therefore, equation (3.2) follows from the equalities

$$F_{\mu^*}^{-1} = \mathbb{E} \left(F_{\mu_{\theta}}^{-1} \right) = \mathbb{E} \left(F_{\mu_{\theta}}^{-1} \circ F_{\mu_0} \right) \circ F_{\mu_0}^{-1} = \mathbb{E} (T_{\theta}) \circ F_{\mu_0}^{-1} = \bar{T} \circ F_{\mu_0}^{-1},$$

where $\bar{T} = \mathbb{E} (T_{\theta}) = F_{\mu^*}^{-1} \circ F_{\mu_0}$. Note that it is clear that $\bar{T} : \Omega \rightarrow \Omega$ is an increasing and continuous function, and thus it is the optimal mapping between μ_0 and μ^* by Proposition 2.1.

Finally, by equation (3.4) and Fubini's Theorem, one has that

$$\begin{aligned}
\int_{\Theta} d_{W_2}^2(\mu^*, \mu_{\theta}) g(\theta) d\theta &= \int_0^1 \mathbb{E} \left| F_{\mu^*}^{-1}(y) - F_{\mu_{\theta}}^{-1}(y) \right|^2 dy = \mathbb{E} \left(\int_0^1 \left| F_{\mu^*}^{-1}(y) - F_{\mu_{\theta}}^{-1}(y) \right|^2 dy \right) \\
&= \mathbb{E} \left(\int_{\Omega} \left| F_{\mu_{\theta}}^{-1} \circ F_{\mu^*}(x) - x \right|^2 d\mu^*(x) \right) \\
&= \mathbb{E} \left(\int_{\Omega} \left| F_{\mu_{\theta}}^{-1} \circ F_{\mu_0} \circ F_{\mu_0}^{-1} \circ F_{\mu^*}(x) - x \right|^2 d\mu^*(x) \right) \\
&= \mathbb{E} \left(\int_{\Omega} \left| T_{\theta} \circ \bar{T}^{-1}(x) - x \right|^2 d\mu^*(x) \right)
\end{aligned}$$

where the last equality follows from the fact that $T_{\theta} = F_{\mu_{\theta}}^{-1} \circ F_{\mu_0}$ and $\bar{T}^{-1} = F_{\mu_0}^{-1} \circ F_{\mu^*}$. Hence, this proves equation (3.3), and it completes the proof. \square

To illustrate Theorem 3.1, we consider a simple construction of random probability measures in the case where Ω is a compact interval of \mathbb{R} . Let $\bar{\mu} \in \mathcal{M}_+^2(\Omega)$ admitting the density \bar{f} with respect to the Lebesgue measure on \mathbb{R} , and cumulative distribution function (cdf) \bar{F} . We assume that the density \bar{f} is continuous on Ω , and that it is compactly supported in a sub-interval Ω_0 of Ω . Let $\theta = (\mathbf{a}, \mathbf{b}) \in]0, +\infty[\times \mathbb{R}$ be a two dimensional random vector with density g , such that $\mathbf{a}x + \mathbf{b} \in \Omega$ for any $x \in \Omega_0$. We denote by μ_{θ} the random probability measure admitting the density

$$f_{\theta}(x) = \frac{1}{\mathbf{a}} \bar{f} \left(\frac{x - \mathbf{b}}{\mathbf{a}} \right), \quad x \in \Omega,$$

where we have extended \bar{f} outside Ω by letting $\bar{f}(u) = 0$ for $u \notin \Omega$. Thanks to our assumptions, the density f_{θ} is compactly supported in a sub-interval of Ω . Moreover, the cdf and quantile function of μ_{θ} are given by

$$F_{\mu_{\theta}}(x) = \bar{F} \left(\frac{x - \mathbf{b}}{\mathbf{a}} \right), \quad x \in \Omega, \quad \text{and} \quad F_{\mu_{\theta}}^{-1}(y) = \mathbf{a} \bar{F}^{-1}(y) + \mathbf{b}, \quad y \in [0, 1].$$

By Theorem 3.1, it follows that the barycenter of μ_{θ} is the probability measure μ^* whose quantile function is given by

$$F_{\mu^*}^{-1}(y) = \mathbb{E}(\mathbf{a}) \bar{F}^{-1}(y) + \mathbb{E}(\mathbf{b}), \quad y \in [0, 1].$$

Therefore, μ^* admits the density

$$f^*(x) = \frac{1}{\mathbb{E}(\mathbf{a})} \bar{f} \left(\frac{x - \mathbb{E}(\mathbf{b})}{\mathbb{E}(\mathbf{a})} \right), \quad x \in \Omega,$$

with respect to the Lebesgue measure on \mathbb{R} . Moreover, if μ_0 is any fixed measure in $\mathcal{M}_+(\Omega)$, that is absolutely continuous with respect to the Lebesgue measure, then

$$\mu^* = T_0 \# \mu_0, \quad \text{where} \quad T_0(x) = \mathbb{E}(\mathbf{a}) \bar{F}^{-1}(F_{\mu_0}(x)) + \mathbb{E}(\mathbf{b}), \quad x \in \Omega.$$

Now, we remark that extending Theorem 3.1 to dimension $d \geq 2$ is not straightforward. Indeed, the two key ingredients in the proof of Theorem 3.1 are the use of the well-known

characterization (3.1) of the Wasserstein distance in dimension $d = 1$ via the quantile functions, and the fact that, in dimension $d = 1$, the composition of two optimal maps (which are increasing functions) remains an optimal one (thanks to Proposition 2.1). However, the property (3.1) which explicitly relates the Wasserstein distance $d_{W_2}(\nu, \mu_\theta)$ to the marginal distributions of ν and μ_θ is not valid in higher dimensions. Moreover, the composition of two optimal maps is not necessarily an optimal one in dimension $d \geq 2$. Nevertheless, we show in the next section that analogs of Theorem 3.1 can still be obtained in dimension $d \geq 2$.

4 An explicit characterization of the population barycenter

In this section, we extend the results of Theorem 3.1 to dimension $d \geq 2$. Let $\mu_\theta \in \mathcal{M}_+(\Omega)$ denote the parametric random measure with distribution \mathbb{P}_g , as defined in equation (1.1). Let μ_0 be a fixed (reference) measure in $\mathcal{M}_+(\Omega)$ admitting a density with respect to the Lebesgue measure on \mathbb{R}^d . Then, by Proposition 2.1, there exists, for any $\theta \in \Theta$, a unique optimal mapping $T_\theta : \Omega \rightarrow \Omega$ such that $\mu_\theta = T_\theta \# \mu_0$, where $T_\theta = \nabla \phi_\theta$ μ_0 -almost everywhere, and $\phi_\theta : \Omega \rightarrow \mathbb{R}$ is a lower semi-continuous convex (l.s.c) function.

In Theorem 4.1 below, we give sufficient conditions on the expectation of T_θ which imply that the barycenter of μ_θ is given by $\mu^* = \mathbb{E}(T_\theta) \# \mu_0$. This result means that computing a barycenter in the Wasserstein space amounts to take the expectation (in the usual sense) of the optimal mapping T_θ to transport μ_0 onto μ_θ . The proof of this result relies on the use of a dual formulation of the optimization problem (1.2) that is stated in equation (7.2), and we refer to Section 7 for further details.

Before stating the main result of this section, we introduce, for any $\theta \in \Theta$, the notation

$$\bar{T}_\theta = T_\theta \circ \bar{T}^{-1}, \quad (4.1)$$

where $\bar{T} : \Omega \rightarrow \Omega$ is the mapping defined by $\bar{T}(x) = \mathbb{E}(T_\theta(x)) = \int_\Theta T_\theta(x)g(\theta)d\theta$, $x \in \Omega$. Note that equation (4.1) is only valid if \bar{T}^{-1} is well defined. Sufficient conditions to ensure the existence of \bar{T} as a one-to-one mapping from Ω to Ω are given in the following theorem which contains the main results of this paper.

Theorem 4.1. *Let $\theta \in \mathbb{R}^p$ be a random vector with a density $g : \Theta \rightarrow \mathbb{R}$ such that $g(\theta) > 0$ for all $\theta \in \Theta$. Let $\mu_\theta = \Phi(\theta)$ be a parametric random measure with distribution \mathbb{P}_g . Let μ_0 be a fixed measure in $\mathcal{M}_+(\Omega)$ admitting a density with respect to the Lebesgue measure on \mathbb{R}^d . Suppose that Assumption 1 and Assumption 2 hold, and that the following condition are satisfied: for any $\theta \in \Theta$*

$$(i) \quad T_\theta \text{ is a } C^1 \text{ diffeomorphism of } \Omega, \quad (4.2)$$

$$(ii) \quad \bar{T}_\theta = \nabla \bar{\phi}_\theta \text{ where } \bar{\phi}_\theta : \Omega \rightarrow \mathbb{R} \text{ is a l.s.c. convex function}, \quad (4.3)$$

such that for any $x \in \Omega$, $\bar{\phi}_\theta$ is integrable and satisfies the normalization condition

$$\int_\Theta \bar{\phi}_\theta(x)g(\theta)d\theta = \frac{1}{2}|x|^2 \text{ for all } x \in \Omega. \quad (4.4)$$

Then, the population barycenter is the measure $\mu^* \in \mathcal{M}_+(\Omega)$ given by

$$\mu^* = \bar{T} \# \mu_0, \quad (4.5)$$

and the optimization problem (1.2) satisfies

$$\inf_{\nu \in \mathcal{M}_+(\Omega)} J(\nu) = \frac{1}{2} \int_{\Theta} d_{W_2}^2(\mu^*, \mu_\theta) g(\theta) d\theta = \frac{1}{2} \int_{\Omega} \mathbb{E} (|\bar{T}_\theta(x) - x|^2) d\mu^*(x). \quad (4.6)$$

Theorem 4.1 shows that computing the population barycenter in the Wasserstein space of the parametric random measure μ_θ may amount to transport the reference measure μ_0 by the expected amount of deformation measured by \bar{T} . In the case where $\bar{T} = I$ is the identity mapping, it can be checked that Conditions (4.3) and (4.4) are necessarily satisfied. Hence, if $\bar{T} = I$, then the population barycenter μ^* is equal to the reference measure μ_0 . In the case where $\bar{T} \neq I$, Condition (4.3) is not necessarily satisfied since the composition $T_\theta \circ \bar{T}^{-1}$ between the optimal mapping T_θ and the map \bar{T}^{-1} is not always an optimal one. In the following section, we describe statistical models to illustrate the usefulness of the barycenter in the Wasserstein space for data analysis, and we discuss various assumptions on \bar{T} to ensure that Conditions (4.3) and (4.4) are satisfied.

Condition (4.2) is generally more delicate to verify. In the next section, we provide some statistical models for which such a condition holds. For a general setting, we refer to the review in [16] for a discussion on sufficient assumptions on the smoothness of Ω , μ_θ and μ_0 which ensure that the optimal mapping T_θ is a C^1 diffeomorphism.

Proof. Under the assumptions of Theorem 4.1, it follows from Proposition 2.2 that the barycenter μ^* exists and is unique.

Let us first prove that Condition (4.2) implies that $\bar{T}_\theta = T_\theta \circ \bar{T}^{-1}$ is a C^1 diffeomorphism. By our remarks in Section 2.2, one has that $(x, \theta) \mapsto T_\theta(x)$ is a measurable mapping with respect to the completion of $\mathcal{B}(\mathbb{R}^p) \otimes \mathcal{B}(\mathcal{M}_+(\Omega))$ with respect to $g(\theta) d\theta d\mu_0(x)$. Therefore, $x \mapsto \bar{T}(x)$ is also a measurable mapping. Using Assumption 1 (compactness of Ω), Condition (4.2), and Lebesgue's dominated convergence theorem, it can be checked that \bar{T} is a C^1 diffeomorphism of Ω , which implies that $\bar{T}_\theta = T_\theta \circ \bar{T}^{-1}$ is also a C^1 diffeomorphism of Ω for any $\theta \in \Theta$.

The rest of the proof relies on the dual formulation (7.2) of the optimization problem (1.2), and we refer to Section 7 for further details and notation. To prove the results stated in Theorem 4.1, we will use the dual characterization of the barycenter μ^* that is stated in Proposition 7.1. For this purpose, we need to find a maximizer $f^\Theta = (f_\theta)_{\theta \in \Theta} \in L^1(\Theta, X)$ of the dual problem (\mathcal{P}^*) , see equation (7.2). In the proof, we repeatedly use the fact that $\mu_\theta = \bar{T}_\theta \# \bar{\mu}$ where, by definition, $\bar{\mu} = \bar{T}\mu_0$, and the property that $\bar{T}_\theta = T_\theta \circ \bar{T}^{-1}$ is a C^1 diffeomorphism.

a) Let us first compute an upper bound of $J_{\mathcal{P}^*}$. Let $f^\Theta \in L^1(\Theta, X)$ be such that $\int_{\Theta} f_\theta(x) d\theta = 0$ for all $x \in \Omega$. By definition of $S_{g(\theta)} f_\theta(x)$ one has that

$$S_{g(\theta)} f_\theta(x) \leq \frac{g(\theta)}{2} |x - y|^2 - f_\theta(y) \quad (4.7)$$

for any $y \in \Omega$. By using inequality (4.7) with $y = \bar{T}_\theta^{-1}(x)$ for $x \in \Omega$, one obtains that

$$\begin{aligned} \int_{\Theta} \int_{\Omega} S_{g(\theta)} f_{\theta}(x) d\mu_{\theta}(x) d\theta &\leq \int_{\Theta} \int_{\Omega} \left(\frac{g(\theta)}{2} |x - \bar{T}_\theta^{-1}(x)|^2 - f_{\theta}(\bar{T}_\theta^{-1}(x)) \right) d\mu_{\theta}(x) d\theta \\ &\leq \int_{\Theta} \int_{\Omega} \left(\frac{g(\theta)}{2} |\bar{T}_\theta(u) - u|^2 - f_{\theta}(u) \right) d\bar{\mu}(u) d\theta \\ &\leq \int_{\Theta} \int_{\Omega} \left(\frac{g(\theta)}{2} |\bar{T}_\theta(u) - u|^2 \right) d\bar{\mu}(u) d\theta \end{aligned}$$

Note that to obtain the second inequality above, we have used the change of variable $u = \bar{T}_\theta^{-1}(x)$, while the third inequality has been obtained using the fact that $\int_{\Theta} f_{\theta}(u) d\theta = 0$ for any $u \in \Omega$ combined with Fubini's theorem. Thanks to Condition 4.3, \bar{T}_θ is the gradient of a convex function, and it is such that $\mu_{\theta} = \bar{T}_\theta \# \bar{\mu}$. Hence, by Proposition 2.1, one obtains that $\int_{\Omega} |\bar{T}_\theta(u) - u|^2 d\bar{\mu}(u) = d_{W_2}^2(\bar{\mu}, \mu_{\theta})$ which implies that

$$\begin{aligned} J(\bar{\mu}) &= \int_{\Theta} \frac{1}{2} d_{W_2}^2(\bar{\mu}, \mu_{\theta}) g(\theta) d\theta, \\ &= \int_{\Theta} \int_{\Omega} \left(\frac{g(\theta)}{2} |\bar{T}_\theta(u) - u|^2 \right) d\bar{\mu}(u) d\theta = \frac{1}{2} \int_{\Omega} \mathbb{E}(|\bar{T}_{\boldsymbol{\theta}}(u) - u|^2) d\bar{\mu}(u), \end{aligned}$$

where the last equality above is a consequence of Fubini's Theorem which follows from the measurability of the mapping $(u, \theta) \mapsto \bar{T}_\theta(u)$ that has been stated in Section 2.2.

Therefore, we have shown that

$$\int_{\Theta} \int_{\Omega} S_{g(\theta)} f_{\theta}(x) d\mu_{\theta}(x) d\theta \leq J(\bar{\mu}),$$

for any $f^{\Theta} \in L^1(\Theta, X)$ such that $\int_{\Theta} f_{\theta}(x) d\theta = 0$ for all $x \in \Omega$. This inequality implies that

$$J_{\mathcal{P}^*} \leq J(\bar{\mu}) = \frac{1}{2} \int_{\Omega} \mathbb{E}(|\bar{T}_{\boldsymbol{\theta}}(u) - u|^2) d\bar{\mu}(u), \quad (4.8)$$

where $J_{\mathcal{P}^*}$ denotes the maximal value of the dual problem (7.2).

b) Let us recall that we have assumed that $g(\theta) > 0$ for any $\theta \in \Theta$. Now, for any $\theta \in \Theta$, we define the function

$$f_{\theta}(x) = -g(\theta) \left(\bar{\phi}_{\theta}(x) - \frac{1}{2}|x|^2 \right), \quad x \in \Omega, \quad (4.9)$$

where $\bar{\phi}_{\theta}$ is the l.s.c. convex function introduced in Condition (4.4). First, Assumption ii) ensures that $f^{\Theta} = (f_{\theta})_{\theta \in \Theta}$ belongs to $L^1(\Theta, X)$. Moreover, by Condition (4.4), one has that $\int_{\Theta} f_{\theta}(x) d\theta = 0$ for all $x \in \Omega$. For any $(x, \theta) \in \Omega \times \Theta$, let us now consider the function $F : \Omega \rightarrow \mathbb{R}$ defined by

$$F(y) = \frac{g(\theta)}{2} |x - y|^2 - f_{\theta}(y) = \frac{g(\theta)}{2} |x - y|^2 + g(\theta) \left(\bar{\phi}_{\theta}(y) - \frac{1}{2}|y|^2 \right), \quad y \in \Omega.$$

Note that $F(y) = \frac{g(\theta)}{2}|x|^2 - g(\theta)\langle x, y \rangle + g(\theta)\bar{\phi}_\theta(y)$. Hence, F is a convex function, since, by assumption, $\bar{\phi}_\theta$ is convex. Searching for some $y \in \Omega$, where the gradient of F vanishes, leads to the equation

$$0 = -g(\theta)(x - y) + g(\theta)(\nabla \bar{\phi}_\theta(y) - y) = -g(\theta)(x - y) + g(\theta)(\bar{T}_\theta(y) - y).$$

Since $g(\theta) > 0$, it follows that the convex function $y \mapsto F(y)$ has a minimum at $y = \bar{T}_\theta^{-1}(x) \in \Omega$. Therefore,

$$S_{g(\theta)}f_\theta(x) = \frac{g(\theta)}{2}|x - \bar{T}_\theta^{-1}(x)|^2 + g(\theta)\left(\bar{\phi}_\theta(\bar{T}_\theta^{-1}(x)) - \frac{1}{2}|\bar{T}_\theta^{-1}(x)|^2\right) \quad (4.10)$$

Let us introduce the notation $J^*(f^\Theta) = \int_\Theta \int_\Omega S_{g(\theta)}f_\theta(x)d\mu_\theta(x)d\theta$. By equation (4.10) and using the change of variable $u = \bar{T}_\theta^{-1}(x)$, one obtains that

$$\begin{aligned} J^*(f^\Theta) &= \int_\Theta \int_\Omega \left(\frac{g(\theta)}{2}|x - \bar{T}_\theta^{-1}(x)|^2 + g(\theta)\left(\bar{\phi}_\theta(\bar{T}_\theta^{-1}(x)) - \frac{1}{2}|\bar{T}_\theta^{-1}(x)|^2\right) \right) d\mu_\theta(x)d\theta \\ &= \int_\Theta \int_\Omega \left(\frac{g(\theta)}{2}|\bar{T}_\theta(u) - u|^2 + g(\theta)\left(\bar{\phi}_\theta(u) - \frac{1}{2}|u|^2\right) \right) d\bar{\mu}(u)d\theta \\ &= \int_\Theta \int_\Omega \frac{g(\theta)}{2}|\bar{T}_\theta(u) - u|^2 d\bar{\mu}(u) = \frac{1}{2} \int_\Omega \mathbb{E}(|\bar{T}_\theta(u) - u|^2) d\bar{\mu}(u). \end{aligned}$$

where the last equality follows as an application of Fubini's theorem combined with Condition (4.4). Hence, thanks to the upper bound (4.8), we finally have that

$$J^*(f^\Theta) = J_{\mathcal{P}^*} = \frac{1}{2} \int_\Omega \mathbb{E}(|\bar{T}_\theta(u) - u|^2) d\bar{\mu}(u) = J(\bar{\mu}),$$

which proves that f^Θ is a maximizer of the dual problem (\mathcal{P}^*). Moreover, by Proposition 7.1, one also obtains that $J(\bar{\mu}) \leq J(\nu)$ for any $\nu \in \mathcal{M}_+(\Omega)$. Hence, $\bar{\mu} = \mu^*$ is the barycenter of the random measure μ_θ . This completes the proof of Theorem 4.1. \square

5 Applications

5.1 A connection with statistical models for curve and image warping

In this section, we propose to discuss some applications of Theorem 4.1. To define probability models where averaging in the Wasserstein space amounts to take the expectation of an optimal transport mapping, we first introduce some statistical models for which the notion of population and empirical barycenters in the 2-Wasserstein space is relevant.

In many applications observations are in the form of a set of n gray-level curves or images $\mathbf{X}_1, \dots, \mathbf{X}_n$ (e.g. in geophysics, biomedical imaging or in signal processing for neurosciences), which can be considered as iid random variables belonging to the set $L^2(\Omega)$ of square-integrable and real-valued functions on a compact domain of $\Omega \subset \mathbb{R}^d$. In many situations the observed curves or images share the same structure. This may lead to the assumption that these observations are random elements which vary around the same but unknown mean pattern (also called

reference template). Estimating such a mean pattern and characterizing the modes of individual variations around this template is of fundamental interest.

Due to additive noise and geometric variability in the data, this mean pattern is typically unknown, and it has to be estimated. In this setting, a widely used approach is Grenander's pattern theory [23, 24, 36, 37] that models geometric variability by the action of a Lie group on an infinite dimensional space of curves (or images). Following the ideas of Grenander's pattern theory, a simple assumption is to consider that the data $\mathbf{X}_1, \dots, \mathbf{X}_n$ are obtained through the deformation of the same reference template $h \in L^2(\Omega)$ via the so-called deformable model

$$\mathbf{X}_i = h \circ \varphi_i^{-1}, \quad i = 1, \dots, n, \quad (5.1)$$

where $\varphi_1, \dots, \varphi_n$ are iid random variables belonging to the set of smooth diffeomorphisms of Ω . In signal and image processing, there has been recently a growing interest on the statistical analysis of deformable models (5.1) using either rigid or non-rigid random diffeomorphisms φ_i , see e.g. [3, 9, 11, 12, 10, 21, 39] and references therein. In a data set of curves or images, one generally observes not only a source of variability in geometry, but also a source of photometric variability (the intensity of a pixel changes from one image to another) that cannot be only captured by a deformation of the domain Ω via a diffeomorphism as in model (5.1).

It is always possible to transform the data $\mathbf{X}_1, \dots, \mathbf{X}_n$ into a set of n iid random probability densities by computing the random variables

$$\mathbf{Y}_i(x) = \frac{\widetilde{\mathbf{X}}_i(x)}{\int_{\Omega} \widetilde{\mathbf{X}}_i(u)} du, \quad x \in \Omega, \quad \text{where } \widetilde{\mathbf{X}}_i(x) = \mathbf{X}_i(x) - \min_{u \in \Omega} \{\mathbf{X}_i(u)\}, \quad i = 1, \dots, n.$$

Let $q_0 \in L^2(\Omega)$ be a probability density function. We consider the following deformable model of densities:

$$\mathbf{Y}_i(x) = |\det(D\varphi_i^{-1})(x)| q_0(\varphi_i^{-1}(x)), \quad x \in \Omega, \quad i = 1, \dots, n, \quad (5.2)$$

where $\det(D\varphi_i^{-1})(x)$ denotes the determinant of the Jacobian matrix of the random diffeomorphism φ_i^{-1} at point x . If we denote by $\mu_1, \dots, \mu_n \in \mathcal{M}_+(\Omega)$ the random probability measures with densities $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, and by μ_0 the measure with density q_0 , then (5.2) can also be written as the following deformable model of measures

$$\mu_i = \varphi_i \# \mu_0, \quad i = 1, \dots, n. \quad (5.3)$$

In model (5.3), computing the empirical barycenter in the Wasserstein space of the random measures μ_1, \dots, μ_n may lead to consistent and meaningful estimators of the reference measure μ_0 and thus of the mean pattern q_0 . In the rest of this section, we discuss some examples of model (5.3). In particular, we show how the results of Section 4 can be used to characterise the population barycenter of random measures satisfying the deformable model (5.3).

5.2 A parametric class of diffeomorphisms

Let μ_0 be a measure on \mathbb{R}^d having a density q_0 (with respect to the Lebesgue measure dx on \mathbb{R}^d) whose support is contained in compact set $\Omega_{q_0} \subset \mathbb{R}^d$. We propose to characterise the population

barycenter of a random measure μ satisfying the deformable model

$$\mu = \varphi \# \mu_0, \quad (5.4)$$

for a specific class of random diffeomorphisms $\varphi : \mathbb{R}^d \rightarrow \mathbb{R}^d$. Let $\mathbb{S}_d^+(\mathbb{R})$ be the set of positive definite $d \times d$ symmetric matrices with real entries. Let

$$\phi : (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p)) \rightarrow \left(\mathbb{S}_d^+(\mathbb{R}) \times \mathbb{R}^d, \mathcal{B}(\mathbb{S}_d^+(\mathbb{R}) \times \mathbb{R}^d) \right)$$

be a measurable mapping, where $\mathcal{B}(\mathbb{S}_d^+(\mathbb{R}) \times \mathbb{R}^d)$ is the Borel σ -algebra of $\mathbb{S}_d^+(\mathbb{R}) \times \mathbb{R}^d$. For $\theta \in \mathbb{R}^p$, we will use the notations

$$\phi(\theta) = (A_\theta, b_\theta), \text{ with } A_\theta \in \mathbb{S}_d^+(\mathbb{R}), b_\theta \in \mathbb{R}^d,$$

and

$$T_\theta(x) = A_\theta x + b_\theta, \quad x \in \mathbb{R}^d.$$

Since the matrix A_θ belongs to $\mathbb{S}_d^+(\mathbb{R})$, it is clear that T_θ is the gradient of the convex function

$$\phi_\theta(x) = \frac{1}{2} x' A_\theta x + b_\theta' x, \text{ for } x \in \mathbb{R}^d.$$

Moreover, for any $\theta \in \mathbb{R}^d$, it is clear that $T_\theta : \mathbb{R}^d \rightarrow \mathbb{R}^d$ is a C^1 diffeomorphism with inverse given by

$$T_\theta^{-1}(x) = A_\theta^{-1}(x - b_\theta), \quad x \in \mathbb{R}^d.$$

Let $\Theta \subset \mathbb{R}^p$ be a compact set. One can then define a parametric class of diffeomorphisms of \mathbb{R}^d as follows

$$D_\phi(\Theta) = \{T_\theta, \theta \in \Theta\}. \quad (5.5)$$

Finally, let $\theta \in \mathbb{R}^p$ be a random vector with density g (with respect to the Lebesgue measure $d\theta$ on \mathbb{R}^p) having a support included in the compact set Θ . We propose to study the population barycenter in the 2-Wasserstein space of the random measure μ_θ satisfying the following deformable model:

$$\mu_\theta = T_\theta \# \mu_0. \quad (5.6)$$

The above equation may also be interpreted as a semi-parametric model of random densities, where T_θ is the optimal mapping between μ_0 and μ_θ . For any $\theta \in \Theta$ (not necessarily a random vector), we define $\mu_\theta = T_\theta \# \mu_0$. Since T_θ is a smooth diffeomorphism and μ_0 is a measure with density q_0 whose support is included in the compact set Ω_{q_0} , it follows that μ_θ admits a density q_θ on \mathbb{R}^d given by

$$q_\theta(x) = \begin{cases} \det(A_\theta^{-1}) q_0(A_\theta^{-1}(x - b_\theta)) & \text{if } x \in \mathcal{R}(T_\theta), \\ 0 & \text{if } x \notin \mathcal{R}(T_\theta). \end{cases} \quad (5.7)$$

where $\mathcal{R}(T_\theta) = \{T_\theta(y), y \in \Omega_{q_0}\} = \{A_\theta y + b_\theta, y \in \Omega_{q_0}\}$. If we further assume that the mapping $\phi : \Theta \rightarrow \mathbb{S}_d^+(\mathbb{R}) \times \mathbb{R}^d$ is continuous, it follows that there exists a compact set $\Omega \subset \mathbb{R}^d$ such that

$\mathcal{R}(T_\theta) \subset \Omega$ for all $\theta \in \Theta$. Thus, under this assumption, the random measure μ_θ takes its values in $\mathcal{M}_+(\Omega)$.

If we assume that $\mathbb{E}(A_\theta) = I$ and $\mathbb{E}(b_\theta) = 0$, then, by applying Theorem 4.1, it follows that $\mu_0 = \mu^*$ is the population barycenter in the Wasserstein space of the random measure μ_θ satisfying the deformable model (5.6). In the case where $\mathbb{E}(A_\theta) \neq I$, we define, for any $\theta \in \Theta$, the following quantities

$$\bar{A}_\theta = A_\theta \bar{A}^{-1} \quad \text{and} \quad \bar{b}_\theta = b_\theta - A_\theta \bar{A}^{-1} \bar{b}, \quad (5.8)$$

where $\bar{A} = \mathbb{E}(A_\theta)$ and $\bar{b} = \mathbb{E}(b_\theta)$. For any $\theta \in \Theta$, we also define the mapping

$$\bar{T}_\theta(x) = T_\theta(\bar{T}^{-1}(x)) = \bar{A}_\theta x + \bar{b}_\theta, \quad x \in \mathbb{R}^d,$$

where $\bar{T}(x) = \mathbb{E}(A_\theta)x + \mathbb{E}(b_\theta)$. To apply Theorem 4.1, it is necessary to assume that \bar{T}_θ is the gradient of a convex function, which means assuming that $\bar{A}_\theta = A_\theta \bar{A}^{-1}$ is a positive definite matrix. A sufficient condition to have this property is that A_θ is a diagonal matrix for any θ . Another condition ensuring this property is to assume that $A_\theta A_{\theta'} = A_{\theta'} A_\theta$ for any θ and θ' in Θ , namely that $(A_\theta)_{\theta \in \Theta}$ is a class of commutative matrices. Under such assumptions, it is then possible to apply Theorem 4.1 to obtain that the population barycenter in the Wasserstein space of the random measure μ_θ satisfying (5.6) is given by

$$\mu^* = \bar{T} \# \mu_0.$$

5.3 Related results in the literature on signal and image processing

In the literature, there exists various applications of the notion of an empirical barycenter in the Wasserstein space for signal and image processing. For example, it has been successfully used for texture analysis in image processing [14, 32]. There also exists a growing interest on the development of fast algorithms for the computation of empirical barycenters with various applications in image processing [6, 15]. The theory of optimal transport for image warping has also been shown to be useful in various applications, see e.g. [25, 26] and references therein. Some properties of the empirical barycenter in the 2-Wasserstein space of random measures satisfying a deformable model similar to (5.3) have also been studied in [13].

Nevertheless, the results in this paper are novel in various aspects. First, we have also shown the benefits of considering the dual formulation (7.2) of the (primal) problem (1.2) to characterize the population barycenter in the 2-Wasserstein space for a large class of deformable models of measures. To the best of our knowledge, the characterization of a population barycenter in deformable models throughout such duality arguments is novel.

6 Convergence of the empirical barycenter

Let us now prove the convergence of the empirical barycenter for the set of measures considered in this paper under Assumption 1 (i.e. that they are compactly supported). We recall that this assumption implies that the Wasserstein space $(\mathcal{M}_+(\Omega), d_{W_2})$ is compact. This is the key property that we use to study the convergence of the empirical barycenter.

Let $\theta_1, \dots, \theta_n$ be iid random variables in \mathbb{R}^p with distribution \mathbb{P}_Θ . Then, let us define the functional

$$J_n(\nu) = \frac{1}{n} \sum_{j=1}^n \frac{1}{2} d_{W_2}^2(\nu, \mu_{\theta_j}), \quad \nu \in \mathcal{M}_+(\Omega), \quad (6.1)$$

and consider the optimization problem: find an empirical barycenter

$$\bar{\mu}_n \in \arg \min_{\nu \in \mathcal{M}_+(\Omega)} J_n(\nu), \quad (6.2)$$

Thanks to the results in [2], the following lemma holds:

Lemma 6.1. *Suppose that Assumption 2 holds. Then, for any $n \geq 1$, there exists a unique minimizer $\bar{\mu}_n$ of $J_n(\cdot)$ over $\mathcal{M}_+(\Omega)$.*

Let us now give our main result on the convergence of the empirical barycenter $\bar{\mu}_n$.

Theorem 6.1. *Suppose that Assumption 1 and Assumption 2 hold. Let μ^* be the population barycenter defined by (1.2), and $\bar{\mu}_n$ be the empirical barycenter defined by (6.2). Then,*

$$\lim_{n \rightarrow +\infty} d_{W_2}(\bar{\mu}_n, \mu^*) = 0 \text{ almost surely (a.s.)}$$

Proof. For $\nu \in \mathcal{M}_+(\Omega)$, let us define

$$\Delta_n(\nu) = J_n(\nu) - J(\nu).$$

The proof is divided in two steps. It follows standard arguments to obtain a strong law of large number for random variables belonging to compact metric spaces that can be found in [35]. Theorem 6.1 can be seen as a particular case of the general results in [35]. But in order to be self contained we include a detailed proof that is inspired by the arguments in [35]. First, we prove the uniform convergence to zero of Δ_n over $\mathcal{M}_+(\Omega)$. Then, we show that any converging subsequence of $\bar{\mu}_n$ converges a.s. to μ^* for the 2-Wasserstein distance.

Step 1. For $\nu \in \mathcal{M}_+(\Omega)$, let us denote by $f_\nu : \mathcal{M}_+(\Omega) \rightarrow \mathbb{R}$ the real-valued function defined by

$$f_\nu(\mu) = \frac{1}{2} d_{W_2}^2(\nu, \mu).$$

Then, let us define the following class of functions

$$\mathcal{F} = \{f_\nu, \nu \in \mathcal{M}_+(\Omega)\}.$$

Since Ω is compact with diameter $\delta(\Omega)$, \mathcal{F} is a class of functions uniformly bounded by $2\delta^2(\Omega)$ (for the supremum norm). Now, let $\nu, \mu, \mu' \in \mathcal{M}_+(\Omega)$. By the triangle reverse inequality

$$\begin{aligned} |f_\nu(\mu) - f_\nu(\mu')| &= \frac{1}{2} |d_{W_2}^2(\nu, \mu) - d_{W_2}^2(\nu, \mu')| \leq \sqrt{2}\delta(\Omega) |d_{W_2}(\nu, \mu) - d_{W_2}(\nu, \mu')| \\ &\leq \sqrt{2}\delta(\Omega) d_{W_2}(\mu, \mu'). \end{aligned}$$

The above inequality proves that \mathcal{F} is an equicontinuous family of functions. Now, let $\theta_1, \dots, \theta_n$ be iid random vectors in \mathbb{R}^p with density g , and let us define the random empirical measure on $(\mathcal{M}_+(\Omega), \mathcal{B}(\mathcal{M}_+(\Omega)))$

$$\mathbb{P}_g^n = \frac{1}{n} \sum_{i=1}^n \delta_{\mu_{\theta_i}} = \frac{1}{n} \sum_{i=1}^n \delta_{\phi(\theta_i)},$$

where δ_μ denotes the Dirac measure. It is clear that

$$\Delta_n(\nu) = \int_{\mathcal{M}_+(\Omega)} f_\nu(\mu) d\mathbb{P}_g^n(\mu) - \int_{\mathcal{M}_+(\Omega)} f_\nu(\mu) d\mathbb{P}_g(\mu).$$

Let $f : \mathcal{M}_+(\Omega) \rightarrow \mathbb{R}$ be a real-valued function that is continuous (for the topology induced by d_{W_2}) and bounded. Thanks to the mesurability of the mapping ϕ , one has that the real random variable $\int_{\mathcal{M}_+(\Omega)} f(\mu) d\mathbb{P}_g^n(\mu)$ converges a.s. to $\int_{\mathcal{M}_+(\Omega)} f(\mu) d\mathbb{P}_g(\mu)$ as $n \rightarrow +\infty$, meaning that the random measure \mathbb{P}_g^n a.s. converges to \mathbb{P}_g in the weak sense. Therefore, since \mathcal{F} is a uniformly bounded and equicontinuous family of functions, one can use Theorem 6.2 in [33] to obtain that

$$\sup_{\nu \in \mathcal{M}_+(\Omega)} |\Delta_n(\nu)| = \sup_{f \in \mathcal{F}} \left| \int_{\mathcal{M}_+(\Omega)} f(\mu) d\mathbb{P}_g^n(\mu) - \int_{\mathcal{M}_+(\Omega)} f(\mu) d\mathbb{P}_g(\mu) \right| \rightarrow 0 \text{ as } n \rightarrow +\infty, \text{ a.s.} \quad (6.3)$$

which proves the uniform convergence of Δ_n to zero over $\mathcal{M}_+(\Omega)$.

Step 2. By Lemma 6.1, there exists a unique sequence $(\bar{\mu}_n)_{n \geq 1}$ of empirical barycenters defined by (6.2). Thanks to the compactness of the Wasserstein space $(\mathcal{M}_+(\Omega), d_{W_2})$, one can extract a converging sub-sequence of empirical barycenters $(\bar{\mu}_{n_k})_{k \geq 1}$ such that $\lim_{k \rightarrow +\infty} d_{W_2}(\bar{\mu}_{n_k}, \bar{\mu}) = 0$ for some measure $\bar{\mu} \in \mathcal{M}_+(\Omega)$.

Let us now prove that $\bar{\mu} = \mu^*$. To this end, let us first note that by the definition of $\bar{\mu}_{n_k}$ and μ^* as the unique minimizer of $J_{n_k}(\cdot)$ and $J(\cdot)$ respectively, it follows that

$$\begin{aligned} |J(\bar{\mu}_{n_k}) - J(\mu^*)| &= J(\bar{\mu}_{n_k}) - J_{n_k}(\bar{\mu}_{n_k}) + J_{n_k}(\bar{\mu}_{n_k}) - J_{n_k}(\mu^*) + J_{n_k}(\mu^*) - J(\mu^*) \\ &\leq 2 \sup_{\nu \in \mathcal{M}_+(\Omega)} |\Delta_{n_k}(\nu)|, \end{aligned}$$

where we have used the fact that $J_{n_k}(\bar{\mu}_{n_k}) - J_{n_k}(\mu^*) \leq 0$. Therefore, thanks to the uniform convergence (6.3) of Δ_n to zero over $\mathcal{M}_+(\Omega)$, one obtains that

$$\lim_{k \rightarrow +\infty} J(\bar{\mu}_{n_k}) = J(\mu^*). \quad (6.4)$$

Therefore, using that

$$\begin{aligned} |J_{n_k}(\bar{\mu}_{n_k}) - J(\mu^*)| &\leq |J_{n_k}(\bar{\mu}_{n_k}) - J(\bar{\mu}_{n_k})| + |J(\bar{\mu}_{n_k}) - J(\mu^*)| \\ &\leq \sup_{\nu \in \mathcal{M}_+(\Omega)} |\Delta_{n_k}(\nu)| + |J(\bar{\mu}_{n_k}) - J(\mu^*)|, \end{aligned}$$

one finally obtains by (6.3) and (6.4) that

$$\lim_{k \rightarrow +\infty} J_{n_k}(\bar{\mu}_{n_k}) = J(\mu^*). \quad (6.5)$$

Since $|J_{n_k}(\bar{\mu}) - J(\bar{\mu})| \leq \sup_{\nu \in \mathcal{M}_+(\Omega)} |\Delta_{n_k}(\nu)|$, it follows by equation (6.3) that

$$\lim_{k \rightarrow +\infty} J_{n_k}(\bar{\mu}) = J(\bar{\mu}) \text{ a.s.} \quad (6.6)$$

Moreover, for any $\epsilon > 0$, there exists $k_\epsilon \in \mathbb{N}$ such that $d_{W_2}(\bar{\mu}_{n_k}, \bar{\mu}) \leq \epsilon$ for all $k \geq k_\epsilon$. Therefore, using the triangle inequality, it follows that for all $k \geq k_\epsilon$

$$\begin{aligned} (J_{n_k}(\bar{\mu}))^{1/2} &= \left(\frac{1}{n_k} \sum_{j=1}^{n_k} \frac{1}{2} d_{W_2}^2(\bar{\mu}, \mu_{\theta_j}) \right)^{1/2} \\ &\leq \left(\frac{1}{n_k} \sum_{j=1}^{n_k} \frac{1}{2} d_{W_2}^2(\bar{\mu}_{n_k}, \mu_{\theta_j}) \right)^{1/2} + \left(\frac{1}{n_k} \sum_{j=1}^{n_k} \frac{1}{2} d_{W_2}^2(\bar{\mu}, \bar{\mu}_{n_k}) \right)^{1/2} \\ &\leq \left(\frac{1}{n_k} \sum_{j=1}^{n_k} \frac{1}{2} d_{W_2}^2(\bar{\mu}_{n_k}, \mu_{\theta_j}) \right)^{1/2} + \frac{\epsilon}{\sqrt{2}}, \end{aligned}$$

and thus by equations (6.5) and (6.6), we obtain that

$$J(\bar{\mu}) \leq \lim_{k \rightarrow +\infty} J_{n_k}(\bar{\mu}_{n_k}) = J(\mu^*) \text{ a.s.} \quad (6.7)$$

which finally proves that $\bar{\mu} = \mu^*$ a.s. since μ^* is the unique minimizer of $J(\nu)$ over $\nu \in \mathcal{M}_+(\Omega)$.

Hence, any converging subsequence of empirical barycenters converges a.s. to μ^* for the 2-Wasserstein distance. Since $(\mathcal{M}_+(\Omega), d_{W_2})$ is compact, this finally shows that $(\bar{\mu}_n)_{n \geq 1}$ is a converging sequence such that $\lim_{n \rightarrow +\infty} d_{W_2}(\bar{\mu}_n, \mu^*) = 0$ a.s. which completes the proof of Theorem 6.1. \square

7 Dual formulation

In this section, we introduce a dual formulation of problem (1.2) that is inspired by the one proposed in [2] to study the properties of empirical barycenters. Let us recall the optimization problem (1.2) as

$$(\mathcal{P}) \quad J_{\mathcal{P}} := \inf_{\nu \in \mathcal{M}_+(\Omega)} J(\nu), \text{ where } J(\nu) = \frac{1}{2} \int_{\Theta} d_{W_2}^2(\nu, \mu_{\theta}) g(\theta) d\theta. \quad (7.1)$$

Then, let us introduce some definitions. Let $X = C(\Omega, \mathbb{R})$ be the space of continuous functions $f : \Omega \rightarrow \mathbb{R}$ equipped with the supremum norm

$$\|f\|_X = \sup_{x \in \Omega} \{|f(x)|\}.$$

We also denote by $X' = \mathcal{M}(\Omega)$ the topological dual of X . The notation $f^{\Theta} = (f_{\theta})_{\theta \in \Theta} \in L^1(\Theta, X)$ will denote any mapping

$$\begin{cases} f^{\Theta} : \Theta & \rightarrow X \\ \theta & \mapsto f_{\theta} \end{cases}$$

such that for any $x \in \Omega$

$$\int_{\Theta} |f_{\theta}(x)| d\theta < +\infty.$$

Then, following the terminology in [2], we introduce the dual optimization problem

$$(\mathcal{P}^*) \quad J_{\mathcal{P}^*} := \sup \left\{ \int_{\Theta} \int_{\Omega} S_{g(\theta)} f_{\theta}(x) d\mu_{\theta}(x) d\theta; f^{\Theta} \in L^1(\Theta, X) \text{ such that } \int_{\Theta} f_{\theta}(x) d\theta = 0, \forall x \in \Omega \right\}, \quad (7.2)$$

where

$$S_{g(\theta)} f(x) := \inf_{y \in \Omega} \left\{ \frac{g(\theta)}{2} |x - y|^2 - f(y) \right\}, \forall x \in \Omega \text{ and } f \in X.$$

Let us also define

$$H_{g(\theta)}(f) := - \int_{\Omega} S_{g(\theta)} f(x) d\mu_{\theta}(x),$$

and the Legendre-Fenchel transform of $H_{g(\theta)}$ for $\nu \in X'$ as

$$H_{g(\theta)}^*(\nu) := \sup_{f \in X} \left\{ \int_{\Omega} f(x) d\nu(x) - H_{g(\theta)}(f) \right\} = \sup_{f \in X} \left\{ \int_{\Omega} f(x) d\nu(x) + \int_{\Omega} S_{g(\theta)} f(x) d\mu_{\theta}(x) \right\}.$$

In what follows, we show that the problems (\mathcal{P}) and (\mathcal{P}^*) are dual to each other in the sense that the minimal value $J_{\mathcal{P}}$ in problem (\mathcal{P}) is equal to the supremum $J_{\mathcal{P}^*}$ in problem (\mathcal{P}^*) . This duality is the key tool for the proof of Theorem 4.1.

Proposition 7.1. *Suppose that Assumption 1 and Assumption 2 are satisfied. Then,*

$$J_{\mathcal{P}} = J_{\mathcal{P}^*}.$$

Proof. 1. Let us first prove that $J_{\mathcal{P}} \geq J_{\mathcal{P}^*}$.

By definition for any $f^{\Theta} \in L^1(\Theta, X)$ such that $\forall x \in \Omega, \int_{\Theta} f_{\theta}(x) d\theta = 0$, and for all $y \in \Omega$ we have

$$S_{g(\theta)} f_{\theta}(x) + f_{\theta}(y) \leq \frac{g(\theta)}{2} |x - y|^2.$$

Let $\nu \in \mathcal{M}_+(\Omega)$ and $\gamma_{\theta} \in \Pi(\mu_{\theta}, \nu)$ be an optimal transport plan between μ_{θ} and ν . By integrating the above inequality with respect to γ_{θ} we obtain

$$\int_{\Omega} S_{g(\theta)} f_{\theta}(x) d\mu_{\theta}(x) + \int_{\Omega} f_{\theta}(y) d\nu(y) \leq \int_{\Omega \times \Omega} \frac{g(\theta)}{2} |x - y|^2 d\gamma_{\theta}(x, y) = \frac{g(\theta)}{2} d_{W_2}^2(\mu_{\theta}, \nu).$$

Integrating now with respect to $d\theta$ and using Fubini's Theorem we get

$$\int_{\Theta} \int_{\Omega} S_{g(\theta)} f_{\theta}(x) d\mu_{\theta}(x) d\theta \leq \int_{\Theta} \frac{g(\theta)}{2} d_{W_2}^2(\mu_{\theta}, \nu) d\theta.$$

Therefore we deduce that $J_{\mathcal{P}} \geq J_{\mathcal{P}^*}$.

2. Let us now prove the converse inequalities $J_P \leq J_{P^*}$.

Thanks to the Kantorovich duality formula (see e.g. [38], or Lemma 2.1 in [2]) we have that $H_{g(\theta)}^*(\nu) = \frac{1}{2}d_{W_2}^2(\mu_\theta, \nu)g(\theta)$ for any $\nu \in \mathcal{M}_+(\Omega)$. Therefore, it follows that

$$J_P = \inf \left\{ \int_{\Theta} H_{g(\theta)}^*(\nu) d\theta, \nu \in X' \right\} = - \left(\int_{\Theta} H_{g(\theta)}^* d\theta \right)^* (0). \quad (7.3)$$

Define the inf-convolution of $(H_{g(\theta)})_{\theta \in \Theta}$ by

$$H(f) := \inf \left\{ \int_{\Theta} H_{g(\theta)}(f_\theta) d\theta; f^\Theta \in L^1(\Theta, X), \int_{\Theta} f_\theta(x) d\theta = f(x), \forall x \in \Omega \right\}, \quad \forall f \in X.$$

We have in the other hand that

$$J_{P^*} = -H(0).$$

Using Theorem 1.6 in [28], one has that for any $\nu \in \mathcal{M}_+(\Omega)$

$$H^*(\nu) = \int_{\Theta} H_{g(\theta)}^*(\nu) d\theta.$$

Then, thanks to (7.3), it follows that

$$J_P = -H^{**}(0) \geq -H(0) = J_{P^*}.$$

Let us now prove that $H^{**}(0) = H(0)$. Since H is convex it is sufficient to show that H is continuous at 0 for the supremum norm of the space X (see e.g. Proposition 4.1 in [17]). For this purpose, let $f^\Theta \in L^1(\Theta, X)$ and remark that it follows from the definition of $H_{g(\theta)}$ that

$$\begin{aligned} H_{g(\theta)}(f_\theta) &= \int_{\Omega} \sup_{y \in \Omega} \left\{ f_\theta(y) - \frac{g(\theta)}{2} |x - y|^2 \right\} d\mu_\theta(x) \\ &\geq f_\theta(0) - \frac{g(\theta)}{2} \int_{\Omega} |x|^2 d\mu_\theta(x), \end{aligned}$$

which implies that

$$H(f) \geq f(0) - \int_{\Theta} \frac{g(\theta)}{2} \int_{\Omega} |x|^2 d\mu_\theta(x) d\theta > -\infty, \quad \forall f \in X.$$

Let $f \in X$ such that $\|f\|_X \leq 1/4$ and choose $f^\Theta \in L^1(\Theta, X)$ defined by $f_\theta(x) = f(x)g(\theta)$ for all $\theta \in \Theta$ and $x \in \Omega$. It follows that

$$\begin{aligned} H(f) &\leq \int_{\Theta} H_{g(\theta)}(f(\cdot)g(\theta)) d\theta \leq \int_{\Theta} \int_{\Omega} \sup_{y \in \Omega} \left\{ \frac{g(\theta)}{4} - \frac{g(\theta)}{2} |x - y|^2 \right\} d\mu_\theta(x) d\theta \\ &\leq \int_{\Theta} \int_{\Omega} \frac{g(\theta)}{4} d\mu_\theta(x) d\theta = \frac{1}{4}. \end{aligned}$$

Hence, the convex function H never takes the value $-\infty$ and is bounded from above in a neighborhood of 0 in X . Therefore, by standard results in convex analysis (see e.g. Lemma 2.1 in [17]), H is continuous at 0, and therefore $H^{**}(0) = H(0)$ which completes the proof. \square

8 Beyond the compactly supported case

To conclude the paper, we briefly discuss the case of a random measure μ with distribution \mathbb{P} whose support is not included in a compact set Ω of \mathbb{R}^d . In the one-dimensional case i.e. $d = 1$, let us denote by F_μ its cumulative distribution function, and by F_μ^{-1} its generalized inverse (quantile function). Provided that one can define the measure μ^* with quantile function $F_{\mu^*}^{-1}(y) = \mathbb{E} \left(F_\mu^{-1}(y) \right)$ for all $y \in [0, 1]$, it is expected that arguments similar to those used in the proof of Theorem 3.1 can be used to prove that μ^* is the unique population barycenter of the random measure μ with distribution \mathbb{P} .

The multi-dimensional case (i.e. $d \geq 2$) is more involved. Indeed, the arguments that we used to prove the main results of the paper for $d \geq 2$ strongly depend on the compactness assumptions on the support of the random measure μ_θ and on the set Θ to which the random vector θ belong. Adapting these arguments for the extension of this paper to non-compactly supported measures is an interesting topic for future investigations.

References

- [1] B. Afsari. Riemannian l^p center of mass: existence, uniqueness, and convexity. *Proceedings of the American Mathematical Society*, 139(2):655–673, 2011.
- [2] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, 43(2):904–924, 2011.
- [3] S. Allasonnière, Y. Amit, and A. Trouvé. Towards a coherent statistical framework for dense deformable template estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(1):3–29, 2007.
- [4] P. C. Álvarez-Esteban, E. del Barrio, J. A. Cuesta-Albertos, and C. Matrán. Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré Probab. Stat.*, 47(2):358–375, 2011.
- [5] M. Arnaudon, C. Dombry, A. Phan, and L. Yang. Stochastic algorithms for computing means of probability measures. *Stochastic Process. Appl.*, 122(4):1437–1455, 2012.
- [6] J.-D. Benamou, G. Carlier, Cuturi M., Nenna L., and G. Peyré. Iterative bregman projections for regularized transportation problems. Technical report, Preprint Arxiv:1412.5154, 2014.
- [7] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds (i). *Annals of statistics*, 31(1):1–29, 2003.
- [8] R. Bhattacharya and V. Patrangenaru. Large sample theory of intrinsic and extrinsic sample means on manifolds (ii). *Annals of statistics*, 33:1225–1259, 2005.
- [9] J. Bigot and B. Charlier. On the consistency of Fréchet means in deformable models for curve and image analysis. *Electronic Journal of Statistics*, 5:1054–1089, 2011.

- [10] J. Bigot and S. Gadat. A deconvolution approach to estimation of a common shape in a shifted curves model. *Ann. Statist.*, 38(4):2422–2464, 2010.
- [11] J. Bigot, S. Gadat, and J.M. Loubes. Statistical M-estimation and consistency in large deformable models for image warping. *Journal of Mathematical Imaging and Vision*, 34(3):270–290, 2009.
- [12] J. Bigot, J.M. Loubes, and M. Vimond. Semiparametric estimation of shifts on compact Lie groups for image registration. *Probability Theory and Related Fields*, 152:425–473, 2010.
- [13] E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution’s template estimate with Wasserstein metrics. *Bernoulli*, To be published, 2014.
- [14] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon wasserstein barycenters of measures. Technical report, Preprint Hal-00881872, 2013.
- [15] M. Cuturi and A. Doucet. Fast computation of wasserstein barycenters. In *In Tony Jebara and Eric P. Xing, editors, Proceedings of the 31st International Conference on Machine Learning (ICML-14)*, pages 685–693. JMLR Workshop and Conference Proceedings, 2014.
- [16] G. De Philippis and A. Figalli. The monge-ampère equation and its link to optimal transportation. *Bull. Amer. Math. Soc.*, 51:527–580, 2014.
- [17] I. Ekeland and R. Témam. *Convex analysis and variational problems*, volume 28 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, english edition, 1999. Translated from the French.
- [18] P. Embrechts and M. Hofert. A note on generalized inverses. *Mathematical Methods of Operations Research*, 77:423–432, 2013.
- [19] J. Fontbona, H. Guérin, and S. Méléard. Measurability of optimal transportation and strong coupling of martingale measures. *Electron. Commun. Probab.*, 15:124–133, 2010.
- [20] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H.Poincaré, Sect. B, Prob. et Stat.*, 10:235–310, 1948.
- [21] F. Gamboa, J.-M. Loubes, and E. Maza. Semi-parametric estimation of shifts. *Electron. J. Stat.*, 1:616–640, 2007.
- [22] C. Goodall. Procrustes methods in the statistical analysis of shape. *J. Roy. Statist. Soc. Ser. B*, 53(2):285–339, 1991.
- [23] U. Grenander. *General pattern theory - A mathematical study of regular structures*. Clarendon Press, Oxford, 1993.
- [24] U. Grenander and M. Miller. *Pattern Theory: From Representation to Inference*. Oxford Univ. Press, Oxford, 2007.

- [25] S. Haker and A. Tannenbaum. On the Monge-Kantorovich problem and image warping. In *Mathematical methods in computer vision*, volume 133 of *IMA Vol. Math. Appl.*, pages 65–85. Springer, New York, 2003.
- [26] S. Haker, L. Zhu, A. Tannenbaum, and S. Angenent. Optimal mass transport for registration and warping. *International Journal on Computer Vision*, 60:225–240, 2004.
- [27] S. F. Huckemann. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *Ann. Statist.*, 39(2):1098–1124, 2011.
- [28] A. D. Ioffe and V. M. Tihomirov. Duality of convex functions, and extremal problems. *Uspehi Mat. Nauk*, 23(6 (144)):51–116, 1968.
- [29] D.G. Kendall. Shape manifolds, procrustean metrics, and complex projective spaces. *Bull. London Math Soc.*, 16:81–121, 1984.
- [30] Y. H. Kim and B. Pass. Wasserstein barycenters over riemannian manifolds. *Preprint, arXiv:1412.7726*, 2014.
- [31] B. Pass. Optimal transportation with infinitely many marginals. *J. Funct. Anal.*, 264(4):947–963, 2013.
- [32] J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its applications to texture mixing. In *LNCS, Proc. SSVM’11*, volume 6667, pages 435–446. Springer, 2011.
- [33] R. Ranga Rao. Relations between weak and uniform convergence of measures with applications. *Ann. Math. Statist.*, 33:659–680, 1962.
- [34] K.-T. Sturm. Probability measures on metric spaces of nonpositive curvature. In *Heat kernels and analysis on manifolds, graphs, and metric spaces (Paris, 2002)*, volume 338 of *Contemp. Math.*, pages 357–390. Amer. Math. Soc., Providence, RI, 2003.
- [35] H. Sverdrup-Thygeson. Strong law of large numbers for measures of central tendency and dispersion of random variables in compact metric spaces. *Ann. Statist.*, 9(1):141–145, 1981.
- [36] A. Trounev and L. Younes. Local geometry of deformable templates. *SIAM J. Math. Anal.*, 37(1):17–59 (electronic), 2005.
- [37] A. Trounev and L. Younes. Shape spaces. In *Handbook of Mathematical Methods in Imaging*. Springer, 2011.
- [38] C. Villani. *Topics in Optimal Transportation*. American Mathematical Society, 2003.
- [39] M. Vimond. Efficient estimation for a subclass of shape invariant models. *Ann. Statist.*, 38(3):1885–1912, 2010.