



HAL
open science

Existence and Consistency of Wasserstein Barycenters

Thibaut Le Gouic, Jean-Michel Loubes

► **To cite this version:**

Thibaut Le Gouic, Jean-Michel Loubes. Existence and Consistency of Wasserstein Barycenters. 2015.
hal-01163262v1

HAL Id: hal-01163262

<https://hal.science/hal-01163262v1>

Preprint submitted on 12 Jun 2015 (v1), last revised 12 Feb 2016 (v2)

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.

Existence and Consistency of Wasserstein Barycenters

Thibaut Le Gouic¹ & Jean-Michel Loubes²

École Centrale de Marseille, Institut de Mathématiques de Marseille¹
Université de Toulouse, Institut de Mathématiques de Toulouse²

Abstract

In this paper, based on the Fréchet mean, we define a notion of *barycenter* corresponding to a usual notion of *statistical mean*. We prove the existence of Wasserstein barycenters of random distributions defined on a geodesic space (E, d) . We also prove the consistency of this barycenter in a general setting, that includes taking barycenters of empirical versions of the distributions or of a growing set of distributions.

Keywords: **Fréchet mean, Wasserstein distance**

Introduction

Giving a sense to the notion of the *mean* of a data sample is one of the major activities of statisticians. When dealing with complex variable data which do not possess an Euclidean structure, the mere issue of defining the mean becomes a difficult task. This problem arises naturally for a wide range of statistical research fields such as functional data analysis for instance in [12], [17], [5] and references therein, image analysis in [19] or [3], shape analysis in [14] or [13] with many applications ranging from biology in [10] to pattern recognition [18] just to name a few.

When dealing with distributions, the issue of finding a central distribution that will convey the information of the distributions is a difficult task. This has been tackled in [1] by considering a notion of barycenter with respect to the Wasserstein distance. This notion coincides with the notion of Fréchet mean. That is, the Fréchet mean of the points $(x_i)_{1 \leq i \leq n}$ of a geodesic space (E, d) given weights $(\lambda_i)_{1 \leq i \leq n}$ is defined as a minimizer of

$$x \mapsto \sum_{i=1}^n \lambda_i d^2(x, x_i).$$

This definition provides a natural extension of the barycenter as it coincides on \mathbb{R}^d with the barycenter $\sum_{i=1}^n \lambda_i x_i$ of the points $(x_i)_{1 \leq i \leq n}$, with weights $(\lambda_i)_{1 \leq i \leq n}$. This function to minimize can be rewritten as

$$x \mapsto \mathbb{E}d^2(X, x)$$

if the distribution of the random variable X is the discrete measure

$$\mu = \sum_{i=1}^n \lambda_i \delta_{x_i},$$

where δ denotes the Dirac measure. We will call any of these minimizer a barycenter of μ , so that the Fréchet mean of $(x_i)_{1 \leq i \leq n}$ with weights $(\lambda_i)_{1 \leq i \leq n}$ is the barycenter of μ . There is then a natural extension of barycenter of measure that is: a point x is said so be a barycenter of a measure μ (not necessarily finitely supported) is it minimizes

$$x \mapsto \mathbb{E}d^2(X, x)$$

when the distribution of the random variable X is μ .

When (E, d) is assumed to be a geodesic locally compact space, the existence of this barycenter is a straightforward consequence of balls of (E, d) being compact. But it is not obvious in more general cases.

In this paper, we consider barycenters in the Wasserstein space of a locally compact geodesic space. Since the Wasserstein space of a locally compact space is, in general, not locally compact, its existence is not as straightforward. We refer to [20] and references therein for definitions on the Wasserstein distance. In this work, the barycenter of the measures $(\mu_i)_{1 \leq i \leq n}$ with weights $(\lambda_i)_{1 \leq i \leq n}$, is thus defined as a minimizer of the following criterion

$$\nu \mapsto \sum_{i=1}^n \lambda_i W_p^p(\nu, \mu_i),$$

which is thus also the barycenter of \mathbb{P} defined by

$$\mathbb{P} = \sum_{i=1}^n \lambda_i \delta_{\mu_i},$$

An important result proved in [1], is the existence and uniqueness of this minimizer under some assumptions. In their paper, the author proves existence and uniqueness of the barycenter when the underlying space (E, d) is the Euclidean space \mathbb{R}^d , and under the assumption that at least one of the μ_i 's *vanishes on small sets*. This *vanishing* property means that the considered measures give probability 0 to sets with Hausdorff dimension less than $d - 1$. Any measure absolutely continuous with respect to the Lebesgue measure *vanishes on small sets*. This work of [1] has been extended in [15] to compact Riemannian

manifolds, with the condition to *vanish on small sets* being replaced by absolute continuity with respect to the volume measure. Since the Wasserstein space of a compact space is also compact, the existence of the barycenter in this setting is straightforward, but their work provides, among other results, an interesting extension of the work of [1], by showing a dual problem called the *multidimensional* problem, for any \mathbb{P} of the form $\sum_{i=1}^n \lambda_i \delta_{\mu_i}$. The same dual problem has been used in a previous work to show existence of barycenter whenever there exists a measurable (not necessarily unique) barycenter application on (E^n, d^n) that associate the barycenter of $\sum_{i=1}^n \lambda_i \delta_{\mu_i}$ to every n -uplets (x_1, \dots, x_n) . It is a first step toward the proof of existence of barycenter for any \mathbb{P} .

This paper studies, in this setting, the existence of the barycenter and state consistency properties. In a previous work [9] or [16], the authors studied some asymptotic results giving conditions under which a sequence of barycenters of discrete measures converging to a limit measure can be understood as a barycenter of the limit distribution. This result enables to define the barycenter of empirical measures and study its asymptotic behavior. In the following, we propose an improved version of this limit theorem that enables to prove existence of barycenters of distributions in a our more general framework.

This paper falls into the following parts. Section 1 presents general definitions and states a general theorem that ensures the existence of a barycenter of distributions. In Section 2, a consistency result is proven. Section 3 is devoted to some statistical applications. The technical lemmas are presented in Section 5 while the detailed proofs are postponed to Section 4.

1 A definition of a mean for probability in Wasserstein space

Let (E, d) be a locally compact geodesic space and for $p \geq 1$ let $\mathcal{W}_p(E)$ be its Wasserstein space defined as the set of Borel probability measures μ on E with finite p order moments, such that for all $y \in E$,

$$\int_E d^p(x, y) \mu(dx) < +\infty.$$

Given two measures μ, ν in $\mathcal{W}_p(E)$, we denote by $\Gamma(\mu, \nu)$ the set of all probability measures π over the product set $E \times E$ with first, resp. second, marginal μ , resp. ν . The transportation cost with d^p cost function between two measures μ, ν in $\mathcal{W}_p(E)$, is defined as

$$\mathcal{T}_p(\mu, \nu) = \inf_{\pi \in \Gamma(\mu, \nu)} \int d^p(x, y) d\pi.$$

The transportation cost allows to endow the set of probability measures $\mathcal{W}_p(E)$ with a metric by setting

$$W_p(\mu, \nu) = \mathcal{T}_p(\mu, \nu)^{1/p}.$$

This metric is known as the p -Wasserstein distance. With this metric, the set $(\mathcal{W}_p(E), W_p)$ is a geodesic space but not locally compact unless (E, d) is compact.

We consider a random probability $\tilde{\mu}$ in $\mathcal{W}_p(E)$, following a distribution \mathbb{P} . This probability \mathbb{P} is chosen in the space $\mathcal{W}_p(\mathcal{W}_p(E))$ endowed with the metric W_p . Note that we use the same notations for the Wasserstein distances over $\mathcal{W}_p(E)$ and $\mathcal{W}_p(\mathcal{W}_p(E))$. Thus, if $\tilde{\mu} \in \mathcal{W}_p(E)$ is a random measure with distribution \mathbb{P} , then for all $\nu \in \mathcal{W}_p(E)$, we can write

$$W_p^p(\delta_\nu, \mathbb{P}) = \mathbb{E}(W_p^p(\nu, \tilde{\mu})) = \int W_p^p(\nu, \mu) d\mathbb{P}(\mu). \quad (1)$$

For a probability $\mathbb{P} \in \mathcal{W}_p(\mathcal{W}_p(E))$, consider a minimizer over $\nu \in \mathcal{W}_p(E)$ of

$$\nu \mapsto \mathbb{E} [W_p^p(\nu, \tilde{\mu})] = W_p^p(\delta_\nu, \mathbb{P}),$$

where $\tilde{\mu}$ is a random probability of $\mathcal{W}_p(E)$ with distribution \mathbb{P} . If exists, this measure will be called the mean measure of \mathbb{P} or its barycenter. In order to prove its existence, we will need the following definition.

Definition (Measurable barycenter) Let $p \geq 1$ and (E, d) be a geodesic space. The space (E, d) is said to admit a measurable barycenter application if given any $J \in \mathbb{N}^*$ and weights $(\lambda_j)_{1 \leq j \leq J}$, there exists a Borel application $T : E^J \rightarrow E$ that associate $(x_j)_{1 \leq j \leq J}$ to a minimum of $x \mapsto \sum_{j=1}^J \lambda_j d^p(x, x_j)$.

It is not clear whether all locally compact geodesic spaces admits a measurable barycenter application, so it will be supposed in the following theorem which proves existence of Wasserstein barycenters.

Theorem 1 (Existence of a Wasserstein Barycenter) *Let $p \geq 1$. Assume that (E, d) is a geodesic space locally compact that admits a measurable barycenter application. Hence, for $\mathbb{P} \in \mathcal{W}_p(\mathcal{W}_p(E))$, there exists a mean $\bar{\mu}_{\mathbb{P}}$ defined as*

$$\bar{\mu}_{\mathbb{P}} \in \arg \min_{\nu \in \mathcal{W}_p(E)} \mathbb{E} [W_p^p(\nu, \tilde{\mu})], \quad (2)$$

for $\tilde{\mu}$ a random measure with distribution \mathbb{P} .

Using the expression (1), we can see that Theorem 1 implies the existence of the metric projection of \mathbb{P} onto the subset of $\mathcal{W}_p(\mathcal{W}_p(E))$ made of Dirac measures.

Proof The proof of Theorem 1 relies on the existence of barycenters of finitely supported measures in $\mathcal{W}_p(E)$. It is split in three steps.

- First, consider a set of probabilities $(\mu_j), j = 1, \dots, J$ of $\mathcal{W}_p(E)$ and assume that \mathbb{P} is a discrete measure defined, for positive weights $\lambda_1, \dots, \lambda_J$ such that $\sum_{j=1}^J \lambda_j = 1$, as

$$\mathbb{P} = \sum_{j=1}^J \lambda_j \delta_{\mu_j}.$$

In this case

$$W_p^p(\delta_\nu, \mathbb{P}) = \mathbb{E}W_p^p(\nu, \tilde{\mu}) = \sum_{j=1}^J \lambda_j W_p^p(\nu, \mu_j).$$

Within this framework, Theorem 1 reduces to an already solved problem, in the case $p = 2$ in [1] or [9] and for general p in [16]. It is recalled in this paper as Theorem 6.

- To prove the theorem in the general case, we will show that if there is a sequence of probabilities $(\mathbb{P}_j)_{j \geq 1}$ converging to a limit distribution \mathbb{P} and if for each \mathbb{P}_j , its mean $\bar{\mu}_{\mathbb{P}_j}$ is well defined (i.e. exists), then the limit distribution has a mean that can be seen as the limit of a subsequence of the means $(\bar{\mu}_{\mathbb{P}_j})_{j \geq 1}$. This result is stated as Theorem 2 in the following section.
- Finally Proposition 11 concludes the proof enabling to approximate any probability on $\mathcal{W}_p(\mathcal{W}_p(E))$ by a discrete probability.

2 Consistency of the barycenter of a sequence of measures

The following theorem deals with a continuity issue of the barycenters of measures. Consider a sequence of measures $(\mathbb{P}_j)_{j \geq 1} \subset \mathcal{W}_p(E)$ converging in $\mathcal{W}_p(\mathcal{W}_p(E))$ to \mathbb{P} . If these measures all admit a barycenter, it is natural to ask whether the sequence of barycenters also converges to a barycenter of \mathbb{P} . Theorem 2 provides a positive answer.

Theorem 2 *Let $(\mathbb{P}_j)_{j \geq 1} \subset \mathcal{W}_p(\mathcal{W}_p(E))$ a sequence of probability measures on $\mathcal{W}_p(E)$ and set μ_j a barycenter of \mathbb{P}_j , for all $j \in \mathbb{N}$. Suppose that for some $\mathbb{P} \in \mathcal{W}_p(\mathcal{W}_p(E))$, we have that $W_p(\mathbb{P}, \mathbb{P}_j) \xrightarrow{j \rightarrow +\infty} 0$. Then, the sequence $(\mu_j)_{j \geq 1}$ is precompact in $\mathcal{W}_p(E)$ and any limit is a barycenter of \mathbb{P} .*

Sketch of proof The proof of Theorem 2 can be split into three steps.

- The first step shows that the sequence of barycenters $(\mu_j)_{j \geq 1}$ is tight. It is a consequence of the fact that balls on (E, d) are compact together with Markov's inequality applied to these balls.

- The second step uses Skorokhod representation theorem and lower semicontinuity of $\nu \mapsto W_p(\mu, \nu)$ for any μ , to show that any weak limit of the sequence $(\mu_j)_{j \geq 1}$ is a barycenter of \mathbb{P} .
- The final step proves that the convergence of the $(\mu_j)_{j \geq 1}$ actually holds in $\mathcal{W}_p(E)$.

Applying this result to a constant sequence gives the following corollary.

Corollary 3 *The set of all barycenters of a given measure $\mathbb{P} \in \mathcal{W}_p(\mathcal{W}_p(E))$ is compact.*

An interesting and immediate corollary follows from the assumption that \mathbb{P} has a unique barycenter.

Corollary 4 *Suppose $\mathbb{P} \in \mathcal{W}_p(\mathcal{W}_p(E))$ has a unique barycenter. Then for any sequence $(\mathbb{P}_j)_{j \geq 1} \subset \mathcal{W}_p(\mathcal{W}_p(E))$ converging to \mathbb{P} , any sequence $(\mu_j)_{j \geq 1}$ of their barycenters converges to the barycenter of \mathbb{P} .*

On $E = \mathbb{R}^d$ and $p = 2$, there exists a simple condition under which the barycenter is unique.

Proposition 5 *Let $\mathbb{P} \in \mathcal{P}_2(\mathcal{P}_2(\mathbb{R}^d))$ such that there exists a set $A \subset \mathcal{P}_2(\mathbb{R}^d)$ of measures such that for all $\mu \in A$,*

$$B \in \mathcal{B}(\mathbb{R}^d), \dim(B) \leq d - 1 \implies \mu(B) = 0, \quad (3)$$

and $\mathbb{P}(A) > 0$, then, \mathbb{P} admits a unique barycenter.

Proof It is a consequence of the fact that if ν satisfies (3), then $\mu \mapsto W_2(\mu, \nu)$ is strictly convex and thus, so is $\mu \mapsto \mathbb{E}W_2^2(\mu, \tilde{\mu})$.

3 Statistical applications

The existence of a mean of a probability measure (or its barycenter) plays an important role in two statistical problems concerning two different empirical points of view. Two asymptotic behaviors can be taken into account, whether the number of distributions goes to infinity or whether the distributions are not observed directly but through empirical samples. We describe here how Theorem 2 handles both of these settings.

The first point of view concerns the case where the distribution $\mathbb{P} \in \mathcal{W}_p(\mathcal{W}_p(E))$ is approximated by a growing discrete distribution \mathbb{P}_J supported on J elements, with J

growing to infinity. Consider a collection of measures $\mu_j \in \mathcal{W}_p(E)$ for $j \geq 1$, and weights $\lambda_j^J \geq 0$, and define the sequence of measures $\mathbb{P}_J, J \geq 1$ as follows

$$\mathbb{P}_J = \sum_{i=1}^J \lambda_i^J \delta_{\mu_j}.$$

Assume that \mathbb{P}_J converges to some measure \mathbb{P} with respect to Wasserstein distance. Hence Theorem 2 states that the barycenter (or any barycenter if not unique) of \mathbb{P}_J converges to the barycenter of \mathbb{P} (provided \mathbb{P} has a unique barycenter).

In a previous work [9], a similar result has been proved under a more restrictive assumption on the μ_j 's: this result was proven in the case where the $(\mu_j)_{j \geq 1}$ are *admissible deformations* in the sense that they can be written as the gradient of a convex functions. Thus, this result enables to tackle the framework of the estimation of a template distribution μ observed in a deformation setting. This setting has also been considered in [2]. More precisely, the distributions μ_j are warped from the template by a random center deformation operator T with realizations T_j , such that

$$\mu_j = T_{j\#}\mu = \mu \circ T_j^{-1}.$$

Then the barycenter of the μ_j 's is a proper estimate for the unobserved template.

In [7] this problem is also tackled in the particular case where the $(\mu_j)_{j \geq 1}$ have compact support, are absolutely continuous with respect to the Lebesgue measure and are indexed on a compact set Θ of \mathbb{R}^d . They state more precisely that given a probability measure on Θ , one can induce a probability measure \mathbb{P} on $\mathcal{W}_p(\mathbb{R}^d)$, and if the $(\mu_j)_{j \geq 1}$ are chosen randomly under $\mathbb{P}^{\otimes \infty}$, the (unique) barycenter of $\frac{1}{J} \sum_{j=1}^J \delta_{\mu_j}$ converges to the barycenter of \mathbb{P} , \mathbb{P} -almost surely.

The second asymptotic point of view deals with the case where the measures μ_j are unknown but approximated by a sequence of measures μ_j^n converging with respect to the Wasserstein distance to measures μ_j when n grows to infinity. Compared to the first framework, the number of measures here is fixed but only an estimation of the measures is known. This covers the interesting case where we observe i.i.d sample $X_{i,j}$ with $i = 1, \dots, n$ with distribution $\mu_j \in \mathcal{W}_p(E)$. Here $\mu_j^n = \frac{1}{n} \sum_{i=1}^n \delta_{X_{i,j}}$ is the empirical measure. Given positive weights $(\lambda_i)_{1 \leq i \leq J}$ (or a sequence of weights converging to them) the issue is whether the barycenter of the observed measure $\sum_{j=1}^J \lambda_j \delta_{\mu_j^n}$ converges to the barycenter of the limit $\sum_{j=1}^J \lambda_j \delta_{\mu_j}$ in the case where this barycenter is unique. This problem has been answered positively in [9], up to extracting a subsequence, since the barycenter is not unique. Within this framework, set

$$\mathbb{P}_n = \sum_{j=1}^J \lambda_j \delta_{\mu_j^n}$$

with positive weights λ_j and measures $(\mu_j^n)_{1 \leq j \leq J, n \geq 1} \subset \mathcal{W}_p(E)^J$ converging to some limit measures $(\mu_j)_{1 \leq j \leq J} \in \mathcal{W}_p(E)^J$. Then Theorem 2 states that the barycenter (or any if not unique) converges to the barycenter of $\sum_{j=1}^J \lambda_j \delta_{\mu_j^n}$ (if unique).

The existence of Wasserstein barycenters has several implications in statistics. In deformation models, as pointed out in [2] and [9], the barycenter provides a good estimate for the template to be estimated. In biology gene expressions suffer from a huge variability and the first task preliminary to any analysis is a normalization procedure to extract a mean feature. This procedure requires the existence of a Wasserstein barycenter as proved in [10] or [11]. In information fusion, the goal is to find a mean measure that aggregates the information provided by different input measures. Hence the Wasserstein barycenter is a natural way to aggregate this information as pointed out in [8]. In multi-target tracking, the main issue is the estimation of both the number and locations of multiple moving targets such as airplanes based on sensor measurements. In [4] the Wasserstein barycenter provides an alternative to the MOSPA (Mean Optimal Sub-Pattern Alignment) distance.

4 Proofs

Theorem 6 (Barycenter and multi-marginal problem) *Let (E, d) be a complete separable geodesic space, $p \geq 1$ and $J \in \mathbb{N}^*$. Given $(\mu_i)_{1 \leq i \leq J} \in \mathcal{P}_p(E)^J$ and weights $(\lambda_i)_{1 \leq i \leq J}$, there exists a measure $\gamma \in \Gamma(\mu_1, \dots, \mu_J)$ minimizing*

$$\hat{\gamma} \mapsto \int \inf_{x \in E} \sum_{1 \leq i \leq J} \lambda_i d(x_i, x)^p d\hat{\gamma}(x_1, \dots, x_J).$$

Moreover, if (E, d) admits a measurable barycenter application $T : E^J \rightarrow E$ then the measure $\nu = T_{\#}\gamma$ is a barycenter of $(\mu_i)_{1 \leq i \leq J}$ and if this application is unique, any barycenter ν is of the form $\nu = T_{\#}\gamma$.

Proof of Theorem 6 This proof is adapted from proposition 4.2 of [1].

Existence of the solution of the multi-marginal problem is a direct consequence of lemma 10.

Denote by γ a solution of the multi-marginal problem and set $\nu = T_{\#}\gamma$. Then, by definition of the Wasserstein distance,

$$W_p^p(\mu_i, \nu) \leq \int d^p(x_i, T(x)) d\gamma(x).$$

and consequently,

$$\sum_{1 \leq i \leq J} \lambda_i W_p^p(\mu_i, \nu) \leq \int \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, T(x)) d\gamma(x). \quad (4)$$

Also, for $\hat{\nu} \in \mathcal{P}_p(E)$, denote $\pi_i \in \Gamma(\mu_i, \hat{\nu})$ the optimal transport plan between $\hat{\nu}$ and μ_i . Using disintegration theorem, for any $1 \leq i \leq J$, there exists a (conditional) measure μ_i^y defined for $\hat{\nu}$ -almost any y , which satisfies $\pi_i(x, y) = \mu_i^y(x) \otimes \hat{\nu}(y)$. Set then,

$$\theta(x, y) = \mu_1^y(x_1) \otimes \dots \otimes \mu_J^y(x_J) \otimes \hat{\nu}(y),$$

and denote $\hat{\gamma}$ the law of the J first marginals of θ . Then, by construction of θ ,

$$\begin{aligned} \sum_{1 \leq i \leq J} \lambda_i W_p^p(\mu_i, \hat{\nu}) &= \sum_{1 \leq i \leq J} \lambda_i \int d^p(x_i, y) d\theta(x, y) \\ &= \int \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, y) d\theta(x, y) \\ &\geq \int \inf_{z \in E} \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, z) d\theta(x, y) \\ &= \int \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, T(x)) d\theta(x, y) \\ &= \int \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, T(x)) d\hat{\gamma}(x) \\ &\geq \int \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, T(x)) d\gamma(x) \\ &\geq \sum_{1 \leq i \leq J} \lambda_i W_p^p(\mu_i, \nu), \end{aligned} \tag{5}$$

where the last inequality is an application of (4).

Since $\hat{\nu}$ is arbitrary, we have just shown that $T_{\#}\gamma$ is a barycenter.

Also, taking $\hat{\nu}$ a barycenter, (5) becomes an equality, so that for θ -almost any $(x, y) \in E^J \times E$,

$$\sum_{1 \leq i \leq J} \lambda_i d^p(x_i, y) = \inf_{z \in E} \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, z) = \sum_{1 \leq i \leq J} \lambda_i d^p(x_i, T(x)),$$

and thus, if the barycenter application T is unique, $T(x) = y$, θ -almost surely and so $T_{\#}\hat{\gamma} = \hat{\nu}$. Also, optimality of $\hat{\nu}$, and (4) show that $\hat{\gamma}$ is a solution of the multi-marginal problem.

Proof of Theorem 2 Denotes μ_j a barycenter of \mathbb{P}_j . The proof is in three steps.

1. Proving the tightness of the sequence of the barycenters $(\mu_j)_{j \geq 1}$.
2. Proving that any limit μ of $(\mu_j)_{j \geq 1}$ (in the sense of the weak convergence of measures) is a barycenter.

3. Proving that there exists $\nu \in \mathcal{W}_p(E)$ such that $W_p(\nu, \mu_j) \rightarrow W_p(\nu, \mu)$. The conclusion of the proof will be derived from Lemma 12.

Let $\tilde{\mu}$ and $\tilde{\mu}_j$ random measures with distribution respectively \mathbb{P} and \mathbb{P}_j .

1. First prove that the moments of order p of the random measures considered as random variables μ_j can be bounded from above by a constant $M < \infty$.

Let $\tilde{\mu}_j$ be a random measure drawn according to a distribution \mathbb{P}_j . Then, for any $x \in E$

$$\begin{aligned}
W_p(\mu_j, \delta_x) &= W_p(\delta_{\mu_j}, \delta_{\delta_x}) \\
&\leq W_p(\delta_{\mu_j}, \mathbb{P}_j) + W_p(\mathbb{P}_j, \delta_{\delta_x}) \\
&= (EW_p^p(\mu_j, \tilde{\mu}_j))^{1/p} + (EW_p^p(\tilde{\mu}_j, \delta_x))^{1/p} \\
&\leq 2 (EW_p^p(\tilde{\mu}_j, \delta_x))^{1/p} \text{ since } \mu_j \text{ is a minimizer of } \nu \mapsto EW_p^p(\nu, \tilde{\mu}_j) \\
&= 2W_p(\mathbb{P}_j, \delta_{\delta_x}) \\
&\leq 2(W_p(\mathbb{P}_j, \mathbb{P}) + W_p(\mathbb{P}, \delta_{\delta_x})) \leq M < \infty \text{ since } W_p(\mathbb{P}, \mathbb{P}_j) \rightarrow 0.
\end{aligned}$$

Denote $B(x, r)$ the ball of E centered in x with radius r . Then Markov's inequality entails that

$$\mu_j(B(x, r)^c) \leq \frac{E_{\mu_j} d^p(X, x)}{r^p} = \frac{W_p^p(\mu_j, \delta_x)}{r^p} \leq \frac{M^p}{r^p}.$$

The compactness of the balls of E entails that the sequence $(\mu_j)_{j \geq 1}$ is tight. So it can be extracted a sequence which converges towards a distribution that will be denoted μ . For ease of notations, the subsequence will be denoted as the initial sequence.

2. Let $\nu \in \mathcal{W}_p(E)$ and $\tilde{\mu}$ a random measure with distribution \mathbb{P} . We get

$$\begin{aligned}
EW_p^p(\nu, \tilde{\mu}) &= W_p^p(\delta_\nu, \mathbb{P}) \\
&= \lim_{j \rightarrow \infty} W_p^p(\delta_\nu, \mathbb{P}_j) \text{ since } W_p(\mathbb{P}_j, \mathbb{P}) \rightarrow 0 \\
&= \lim_{j \rightarrow \infty} EW_p^p(\nu, \tilde{\mu}_j) \\
&\geq \lim_{j \rightarrow \infty} EW_p^p(\mu_j, \tilde{\mu}_j) \text{ since } \mu_j \text{ is a barycenter} \tag{6} \\
&\geq E \liminf_{j \rightarrow \infty} W_p^p(\mu_j, \tilde{\mu}_j) \text{ using Fatou's Lemma for any coupling of the } \tilde{\mu}_j \text{'s} \\
&\geq EW_p^p(\mu, \tilde{\mu}) \text{ since } W_p \text{ is lower semi-continuous.}
\end{aligned}$$

For the last inequality, we used that since $\mathbb{P}_j \rightarrow \mathbb{P}$, Skorokhod's representation theorem enables to build $\tilde{\mu}_j \rightarrow \tilde{\mu}$ a.s.. This proves that μ is a barycenter of \mathbb{P} .

3. For $\nu = \mu$, the inequality (6) is in fact an equality which implies that

$$W_p(\delta_{\mu_j}, \mathbb{P}_j) \rightarrow W_p(\delta_\mu, \mathbb{P}).$$

Hence

$$W_p(\delta_{\mu_j}, \mathbb{P}) - W_p(\delta_\mu, \mathbb{P}) \leq W_p(\delta_{\mu_j}, \mathbb{P}_j) + W_p(\mathbb{P}_j, \mathbb{P}) - W_p(\delta_\mu, \mathbb{P}) \rightarrow 0.$$

This implies that

$$\begin{aligned} EW_p^p(\mu, \tilde{\mu}) &= W_p(\delta_\mu, \mathbb{P}) \\ &= \lim_{j \rightarrow \infty} W_p(\delta_{\mu_j}, \mathbb{P}) \\ &= \lim_{j \rightarrow \infty} EW_p^p(\mu_j, \tilde{\mu}) \\ &\geq E \liminf_{j \rightarrow \infty} W_p^p(\mu_j, \tilde{\mu}) \text{ using Fatou's Lemma} \\ &\geq EW_p^p(\mu, \tilde{\mu}) \text{ using again semi-lower continuity of } W_p \text{ for weak convergence.} \end{aligned}$$

So \mathbb{P} -a.s. (since $\liminf W_p(\mu_j, \tilde{\mu}) \geq W_p(\mu, \tilde{\mu})$)

$$\liminf W_p(\mu_j, \tilde{\mu}) = W_p(\mu, \tilde{\mu}).$$

So all along a subsequence and for a $\nu \in \mathcal{W}(E)$, $W_p(\mu_j, \nu) \rightarrow W_p(\mu, \nu)$. So using Lemma 12, we get that

$$W_p(\mu_j, \mu) \rightarrow 0,$$

which concludes the proof.

5 Technical Lemmas

The following five results are well known. They are recalled here for the purpose of clarity of the proofs.

Lemma 7 (Consistency in L^1) *Let $(X_n)_{n \geq 1}$ be a sequence of real valued random variables such that*

$$\begin{aligned} X_n &\rightarrow X \text{ a.s.} \\ \mathbb{E}|X_n| &\rightarrow \mathbb{E}|X|. \end{aligned}$$

Then, $X_n \rightarrow X$ in L^1 .

Lemma 8 (Uniform integrability) *A family of real valued random variables \mathcal{H} is uniformly integrable (in the sense that $\sup_{X \in \mathcal{H}} \int_{\{|X| > a\}} |X| d\mathbb{P} \rightarrow 0$ as $a \rightarrow +\infty$) if and only if the two following conditions hold*

- i) $\sup_{X \in \mathcal{H}} E|X| < \infty$ (bounded in L^1)*
- ii) $\forall \varepsilon > 0, \exists \alpha > 0$ such that $\forall A \in \mathcal{A}, (\mathbb{P}(A) < \alpha \implies \sup_{X \in \mathcal{H}} \int_A |X| d\mathbb{P} < \varepsilon)$ (equicontinuity).*

Lemma 9 (Consistency in L^1 and uniform integrability) *Let $X_n \rightarrow X$ in probability, then the sequence $(X_n)_{n \geq 1}$ is uniformly integrable if and only if $X_n \rightarrow X$ in L^1 .*

Lemma 10 (Tightness of fixed marginals set of measures) *Let (E, d) be a Polish space (i.e. a complete separable metric space). Let C_1, \dots, C_J be compact sets of $\mathcal{W}_p(E)$. Then, the set $\Gamma(C_1, \dots, C_J)$ defined as the set of probability measures on E^J with marginals respectively in C_1, \dots, C_J , is compact.*

Proposition 11 (Approximation by finitely supported measures) *For all \mathbb{P} there is a sequence of finitely supported distributions \mathbb{P}_j such that*

$$W_p(\mathbb{P}_j, \mathbb{P}) \longrightarrow 0.$$

Here is a lemma used for the proof of Theorem 2.

Lemma 12 *Let $(\mu_n)_{n \geq 1}$ be a sequence of measures on a Polish space (E, d) which converges weakly towards μ . If there exists a measure ν such that*

$$W_p(\mu_n, \nu) \rightarrow W_p(\mu, \nu),$$

then

$$W_p(\mu_n, \mu) \rightarrow 0. \tag{7}$$

Proof Note first that if $\nu = \delta_x$ for a given $x \in E$, then (7) is true, due to the fact that Wasserstein convergence is equivalent to the weak convergence plus convergence of the order p moments (see [20]).

First using the Gluing Lemma (see for instance in [20] or [6]), build three sequences $(X_n)_{n \geq 1}, (Y_n)_{n \geq 1}, (Z_n)_{n \geq 1}$ with distribution respectively μ_n, ν and μ such that

$$(X_n, Y_n) \sim \pi_n^1, (Y_n, Z_n) \sim \pi_n^2,$$

where π_n^1 and π_n^2 are the optimal transport maps between respectively μ_n and ν and between ν and μ . Let Π_n be the distribution of (X_n, Y_n, Z_n) . Since the three marginals weakly converge, the sequence $(\Pi_n)_{n \geq 1}$ is tight. Thus, we can extract a subsequence such that

$$\Pi_n \rightarrow \Pi \text{ weakly,}$$

where Π has marginal distributions μ, ν and μ .

Then Skorokhod's representation Theorem enables to construct a space $(\Omega, \mathcal{A}, \mathbb{P})$ on which there exist X, Y, Z with joint distribution Π and copies of (X_n, Y_n, Z_n) with law Π_n such that

$$d(X_n, X) + d(Y_n, Y) + d(Z_n, Z) \rightarrow 0 \text{ } \mathbb{P}\text{-a.s.}$$

If we show that $(d^p(X_n, X))_{n \geq 1}$ is uniformly integrable then using Lemma (9), we get

$$\mathbb{E}d^p(X_n, X) \rightarrow 0,$$

which implies the result since $W_p(\mu_n, \mu) \leq \mathbb{E}d^p(X_n, X)$.

Uniform integrability remains to be proven. Note that Lemma (8) entails that it is equivalent to prove the two following assumptions

- i) $\sup_{n \geq 1} \mathbb{E}d^p(X_n, X) < \infty$ (bounded in L^1)
- ii) $\forall \varepsilon > 0, \exists \alpha > 0$ such that $\forall A \in \mathcal{A}, (\mathbb{P}(A) < \alpha \implies \int_A d^p(X_n, X) d\mathbb{P} < \varepsilon)$ (equicontinuity).

Assertion i) is a consequence of - since $\mathbb{E}d^p(X_n, X) \leq \mathbb{E}d^p(X_n, Z_n)$,

$$\begin{aligned} \mathbb{E}d^p(X_n, X) &\leq C_p [\mathbb{E}d^p(X_n, Y_n) + \mathbb{E}d^p(Y_n, x) + \mathbb{E}d^p(x, Z_n)] \\ &= C_p (W_p^p(\mu_n, \nu) + W_p^p(\nu, \delta_x) + W_p^p(\delta_x, \mu)) \\ &\leq M < \infty \text{ since we assumed that } W_p^p(\mu_n, \nu) \rightarrow W_p^p(\mu, \nu). \end{aligned}$$

To prove Assertion ii), set $A \in \mathcal{A}$. We have that

$$\mathbb{E}d^p(X_n, X)\mathbf{1}_A \leq C_p [\mathbb{E}d^p(X_n, Y_n)\mathbf{1}_A + \mathbb{E}d^p(Y_n, x)\mathbf{1}_A + \mathbb{E}d^p(x, Z_n)\mathbf{1}_A]. \quad (8)$$

Note that $d^p(X_n, Y_n)$, $d^p(Y_n, x)$ and $d^p(x, Z_n)$ converge towards respectively $d^p(X, Y)$, $d^p(Y, x)$ and $d^p(x, Z)$ a.s. Their L^1 norm converge also, for the first term by assumption and since Y_n and Z_n are identically distributed, for all $n \geq 1$. Hence using Lemma 7 they converge in L^1 and thus are equicontinuous sequences. Hence this implies that for all $\varepsilon > 0$, there exists $\alpha > 0$ such that the three terms

$$\mathbb{E}d^p(X_n, Y_n)\mathbf{1}_A + \mathbb{E}d^p(Y_n, x)\mathbf{1}_A + \mathbb{E}d^p(x, Z_n)\mathbf{1}_A < 3\varepsilon$$

for any A such that $\mathbb{P}(A) < \alpha$.

Hence inequality (8) implies that $(d^p(X_n, X))_{n \geq 1}$ is equicontinuous. Since it is also bounded in L^1 , this sequence is uniformly integrable, which proves the result.

References

- [1] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.
- [2] Marina Agulló-Antolín, J. A. Cuesta-Albertos, H el ene Lescornel, and Jean-Michel Loubes. A parametric registration model for warped distributions with Wasserstein’s distance. *J. Multivariate Anal.*, 135:117–130, 2015.
- [3] Y. Amit, U. Grenander, and M. Piccioni. Structural Image Restoration through deformable template. *Journal of the American Statistical Association*, 86:376–387, 1991.
- [4] Marcus Baum, Peter Willett, and Uwe D Hanebeck. On wasserstein barycenters and mmospa estimation. *Signal Processing Letters, IEEE*, 22(10):1511–1515, 2015.
- [5] Bernard Bercu and Philippe Fraysse. A Robbins-Monro procedure for estimation in semiparametric regression models. *Ann. Statist.*, 40(2):666–693, 2012.
- [6] Patrizia Berti, Luca Pratelli, and Pietro Rigo. Gluing lemmas and skorohod representations. *ArXiv e-prints*, 2014.
- [7] J. Bigot and T. Klein. Consistent estimation of a population barycenter in the Wasserstein space. *ArXiv e-prints*, December 2012.
- [8] Adrian N Bishop. Information fusion via the wasserstein barycenter in the space of probability measures: Direct fusion of empirical measures and gaussian fusion with unknown correlation. In *Information Fusion (FUSION), 2014 17th International Conference on*, pages 1–7. IEEE, 2014.
- [9] Emmanuel Boissard, Thibaut Le Gouic, and Jean-Michel Loubes. Distribution’s template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.
- [10] B. M. Bolstad, R. A. Irizarry, M.  strand, and T. P. Speed. A Comparison of Normalization Methods for High Density Oligonucleotide Array Data Based on Variance and Bias. *Bioinformatics*, 19(2):185–193, 2003.
- [11] Santiago Gallon, Jean-Michel Loubes, and Elie Maza. Statistical properties of the quantile normalization method for density curve alignment. *Mathematical Biosciences*, 242(2):129–142, 2013.
- [12] Fabrice Gamboa, Jean-Michel Loubes, and Elie Maza. Semi-parametric estimation of shifts. *Electron. J. Stat.*, 1:616–640, 2007.

- [13] U. Grenander. General Pattern Theory—A Mathematical Study Of Regular Structures, Oxford University Press. New, York:1994.
- [14] D. G. Kendall, D. Barden, T. K. Carne, and H. Le. *Shape and shape theory*. Wiley Series in Probability and Statistics. John Wiley & Sons Ltd., Chichester, 1999.
- [15] Y.-H. Kim and B. Pass. Wasserstein Barycenters over Riemannian manifolds. *ArXiv e-prints*, December 2014.
- [16] Thibaut Le Gouic. *Localisation de masse et espaces de Wasserstein*. PhD thesis, Université de Toulouse, Université de Toulouse III-Paul Sabatier, 2013.
- [17] J. O. Ramsay and B. W. Silverman. *Functional data analysis*. Springer Series in Statistics. Springer, New York, second edition, 2005.
- [18] H. Sakoe and S. Chiba. Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 26(1):43–49, 1978.
- [19] Alain Trounev and Laurent Younes. Metamorphoses Through Lie Group Action. *Foundations of Computational Mathematics*, 5(2):173–198, 2005.
- [20] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.