

Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line

Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo Lopez

▶ **To cite this version:**

Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo Lopez. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. Electronic Journal of Statistics , 2018, 12 (02), pp.2253–2289. 10.1214/18-EJS1400 . hal-01333401v2

HAL Id: hal-01333401
https://hal.science/hal-01333401v2

Submitted on 29 Jan 2018

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.

Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line

Jérémie Bigot[1]*, Raúl Gouet[2], Thierry Klein[3] & Alfredo López[4]

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

Depto. de Ingeniería Matemática and CMM (CNRS, UMI 2807)[2]
Universidad de Chile

ENAC- Ecole nationale de l'aviation civile
et Institut de Mathématiques de Toulouse et CNRS (UMR 5219)[3]
Université de Toulouse

CSIRO Chile International Centre of Excellence[4]

December 8, 2017

Abstract

This paper is focused on the statistical analysis of probability measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ on $\mathbb{R}$ that can be viewed as independent realizations of an underlying stochastic process. We consider the situation of practical importance where the random measures $\boldsymbol{\nu}_i$ are absolutely continuous with densities $\boldsymbol{f}_i$ that are not directly observable. In this case, instead of the densities, we have access to datasets of real random variables $(X_{i,j})_{1\leq i\leq n;\, 1\leq j\leq p_i}$ organized in the form of n experimental units, such that $X_{i,1}, \ldots, X_{i,p_i}$ are iid observations sampled from a random measure $\boldsymbol{\nu}_i$ for each $1 \leq i \leq n$. In this setting, we focus on first-order statistics methods for estimating, from such data, a meaningful structural mean measure. For the purpose of taking into account phase and amplitude variations in the observations, we argue that the notion of Wasserstein barycenter is a relevant tool. The main contribution of this paper is to characterize the rate of convergence of a (possibly smoothed) empirical Wasserstein barycenter towards its population counterpart in the asymptotic setting where both n and $\min_{1\leq i\leq n} p_i$ may go to infinity. The optimality of this procedure is discussed from the minimax point of view with respect to the Wasserstein metric. We also highlight the connection between our approach and the curve registration problem in statistics. Some numerical experiments are used to illustrate the results of the paper on the convergence rate of empirical Wasserstein barycenters.

Keywords: Wasserstein space; Fréchet mean; Barycenter of probability measures; Functional data analysis; Phase and amplitude variability; Smoothing; Minimax optimality.

AMS classifications: Primary 62G08; secondary 62G20.

*J. Bigot is a member of Institut Universitaire de France.

Acknowledgments

We are very much indebted to the referees and the Associate Editor for their constructive criticism, comments and remarks that resulted in a major revision of the original manuscript.

1 Introduction

In this paper, we are concerned with the statistical analysis of a set of absolutely continuous measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ on the real line $\mathbb{R}$, with supports included in (a possibly unbounded) interval $\Omega \subset \mathbb{R}$, that can be viewed as independent copies of an underlying random measure $\boldsymbol{\nu}$. In this setting, it is of interest to define and estimate a mean measure ν_0 of the random probability measure $\boldsymbol{\nu}$. The notion of mean or averaging depends on the metric that is chosen to compare elements in a given data set. In this work, we consider the Wasserstein metric d_W associated to the quadratic cost for the comparison of probability measures and we define ν_0 as the population Wasserstein barycenter of $\boldsymbol{\nu}$, given by

$$\nu_0 = \underset{\mu \in W_2(\Omega)}{\arg\min}\, \mathbb{E}\left[d_W^2(\boldsymbol{\nu}, \mu)\right],$$

where the above expectation is taken with respect to the distribution of $\boldsymbol{\nu}$, and $W_2(\Omega)$ denotes the space of probability measures with support included in Ω and with finite second moment. A Wasserstein barycenter corresponds to the Fréchet mean [Fré48] that is an extension of the usual Euclidean mean to non-linear metric spaces. Throughout the paper, the population mean measure ν_0 is also referred to as the structural mean of $\boldsymbol{\nu}$, which is a terminology borrowed from curve registration (see [ZM11] and references therein). We choose to work with the Wasserstein metric because it has been shown to be successful in the presence of phase variation (we refer to Section 2 for more explanations).

Data sets leading to the analysis of absolutely continuous measures appear in various research fields. Examples can be found in neuroscience [WS11], demographic and genomics studies [Del11, ZM11], economics [KU01], as well as in biomedical imaging [PM16]. Nevertheless, in such applications one does not directly observe raw data in the form of absolutely continuous measures. Indeed, we generally only have access to random observations sampled from different distributions, that represent independent subjects or experimental units.

Thus, we propose to study the estimation of the structural mean measure ν_0 (the population Wasserstein barycenter) from a data set consisting of independent real random variables $(X_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p_i}$ organized in the form of n experimental units, such that (conditionally on $\boldsymbol{\nu}_i$) the random variables $X_{i,1}, \ldots, X_{i,p_i}$ are iid observations sampled from the measure $\boldsymbol{\nu}_i$ with density $\boldsymbol{f}_i$, where p_i denotes the number of observations for the i-th subject or experimental unit. The main purpose of this paper is to propose nonparametric estimators of the structural mean measure ν_0 and to characterize their rates of convergence with respect to the Wasserstein metric in the asymptotic setting, where both n and $\min_{1 \leq i \leq n} p_i$ may go to infinity.

1.1 Main contributions

Two types of nonparametric estimators are considered in this paper. The first one is given by the empirical Wasserstein barycenter of the set of measures $\tilde{\boldsymbol{\nu}}_1, \ldots, \tilde{\boldsymbol{\nu}}_n$, with $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{X_{i,j}}$ for $1 \leq i \leq n$. This estimator will be referred to as the non-smoothed empirical Wasserstein barycenter. Alternatively, since the unknown probability measures $\boldsymbol{\nu}_i$ are supposed to be absolutely continuous, a second estimator is based on a preliminary smoothing step which consists in using standard kernel smoothing to construct estimators $\hat{\boldsymbol{f}}_i$ of the unknown densities $\boldsymbol{f}_i$ for each $1 \leq i \leq n$. Then, an estimator of ν_0 is obtained by taking the empirical Wasserstein barycenter of the measures $\hat{\nu}_i, \ldots, \hat{\nu}_n$, with $\hat{\nu}_i(A) := \int_A \hat{f}_i(x)dx$, $A \subset \mathbb{R}$ measurable. We refer to this class of estimators as smoothed empirical Wasserstein barycenters whose smoothness depend on the choice of the bandwidths in the preliminary kernel smoothing step.

The rates of convergence of both types of estimators are derived for their (squared) Wasserstein risks, defined as their expected (squared) Wassertein distances from ν_0, and their optimality is discussed from the minimax point of view. Finally, some numerical experiments with simulated data are used to illustrate these results.

1.2 Related work in the literature

The notion of barycenter in the Wasserstein space, for a finite set of n probability measures supported on $\mathbb{R}^d$ (for any $d \geq 1$), has been recently introduced in [AC11] where a detailed characterization of such barycenters in terms of existence, uniqueness and regularity is given using arguments from duality and convex analysis. However, the convergence (as $n \to \infty$) of such Wasserstein barycenters is not considered in that work.

In the one dimensional case ($d = 1$), computing the Wasserstein barycenter of a finite set of probability measures simply amounts to averaging (in the usual way) their quantile functions. In statistics, this approach has been referred to as quantile synchronization [ZM11]. In the presence of phase variability in the data, quantile synchronization is known to be an appropriate alternative to the usual Euclidean mean of densities to compute a structural mean density that is more consistent with the data. Various asymptotic properties of quantile synchronization are studied in [ZM11] in a statistical model and asymptotic setting similar to that of this paper with $\min_{1 \leq i \leq n} p_i \geq n$. However, other measures of risk than the one in this paper are considered in [ZM11], but the optimality of the resulting convergence rates of quantile synchronization is not discussed.

The results of this paper are very much connected with those in [PZ16] where a new framework is developed for the registration of multiple point processes on the real line for the purpose of separating amplitude and phase variation in such data. In [PZ16], consistent estimators of the structural mean of multiple point processes are obtained by the use of smoothed Wasserstein barycenters with an appropriate choice of kernel smoothing. Also, rates of convergence of such estimators are derived for the Wasserstein metric. The statistical analysis of multiple point processes is very much connected to the study of repeated observations organized in samples from independent subjects or experimental units. Therefore, some of our results in this paper on smoothed empirical Wasserstein barycenters are built upon the work in [PZ16]. Nevertheless, novel contributions include the derivation of an exact formula to compute the risk

of non-smoothed Wasserstein barycenters in the case of samples of equal size, and new upper bounds on the rate of convergence of the Wasserstein risk of non-smoothed and smoothed empirical Wasserstein barycenters, together with a discussion of their optimality from the minimax point of view.

The construction of consistent estimators of a population Wasserstein barycenter for semi-parametric models of random measures can also be found in [BK17] and [BLGL15], together with a discussion on their connection to the well known curve registration problem in statistics [RL01, WG97].

1.3 Organization of the paper

In Section 2, we first briefly explain why using statistics based on the Wasserstein metric is a relevant approach for the analysis of a set of random measures in the presence of phase and amplitude variations in their densities. Then, we introduce a deformable model for the registration of probability measures that is appropriate to study the statistical properties of empirical Wasserstein barycenters. The two types of nonparametric estimators described above are finally introduced at the end of Section 2. The convergence rates and the optimality of these estimators are studied in Section 3. Some numerical experiments with simulated data are proposed in Section 4 to highlight the finite sample performances of these estimators. Section 5 contains a discussion on the main contributions of this work and their potential extensions. The proofs of the main results are gathered in a technical Appendix. Finally, note that we use bold symbols $\boldsymbol{f}, \boldsymbol{\nu}, \ldots$ to denote random objects (except real random variables).

2 Wasserstein barycenters for the estimation of the structural mean in a deformable model of probability measures

2.1 The need to account for phase and amplitude variations

To estimate a mean measure from the data $(X_{i,j})_{1\leq i\leq n;\, 1\leq j\leq p_i}$, a natural approach is the following one. In a first step, one uses the $X_{i,j}$'s to compute estimators $\hat{\boldsymbol{f}}_1, \ldots, \hat{\boldsymbol{f}}_n$ (e.g. via kernel smoothing) of the unobserved density functions $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ of the measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$. Then, an estimator of a mean density might be defined as the usual Euclidean mean $\bar{\boldsymbol{f}}_n = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{f}}_i$, which is also classically referred to as the cross-sectional mean in curve registration. At the level of measures, it corresponds to computing the arithmetical mean measure $\bar{\boldsymbol{\nu}}_n = \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\nu}}_i$. The Euclidean mean $\bar{\boldsymbol{f}}_n$ is the Fréchet mean of the $\hat{\boldsymbol{f}}_i$'s with respect to the usual squared distance in the Hilbert space $L^2(\Omega)$ of square integrable functions on Ω. Therefore, it only accounts for linear variations in amplitude in the data. However, as remarked in [ZM11], in many applications it is often of interest to also incorporate an analysis of phase variability (i.e. time warping) in such functional objects, since it may lead to a better understanding of the structure of the data. In such settings, the use of the standard squared distance in $L^2(\Omega)$ to compare density functions ignores a possible significant source of phase variability in the data.

To better account for phase variability in the data, it has been proposed in [ZM11] to introduce the so-called method of quantile synchronization as an alternative to the cross sectional

mean $\bar{\boldsymbol{f}}_n$. It amounts to computing the mean measure $\boldsymbol{\nu}_n^{\oplus}$ (and, if it exists, its density $\boldsymbol{f}_n^{\oplus}$) whose quantile function is

$$\bar{\boldsymbol{F}}_n^{-} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{F}_i^{-}, \tag{2.1}$$

where $\boldsymbol{F}_i^{-}$ denotes the quantile function of the measure $\boldsymbol{\nu}_i$ with density $\boldsymbol{f}_i$.

The statistical analysis of quantile synchronization, as studied in [ZM11], complements the quantile normalization method originally proposed in [BIAS03] to align density curves in microarray data analysis. This method is therefore appropriate for the registration of density functions and the estimation of phase and amplitude variations as explained in details in [PZ16].

Let us now assume that $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are random elements taking values in the set of absolutely continuous measures contained in $W_2(\Omega)$. In this setting, it can be checked (see e.g. Proposition 2.1 below) that quantile synchronization corresponds to computing the empirical Wasserstein barycenter of the random measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$, namely

$$\boldsymbol{\nu}_n^{\oplus} = \operatorname*{arg\,min}_{\mu \in W_2(\Omega)} \frac{1}{n}\sum_{i=1}^{n} d_W^2(\boldsymbol{\nu}_i, \mu).$$

Therefore, the notion of averaging by quantile synchronization corresponds to using the Wasserstein distance d_W to compare probability measures, which leads to a notion of measure averaging that may better reflect the structure of the data than the arithmetical mean in the presence of phase and amplitude variability.

To illustrate the differences between using Euclidean and Wasserstein distances to account for phase and amplitude variation, let us assume that the measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ have densities $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ obtained from the following location-scale model: we let f_0 be a density on $\mathbb{R}$ having a finite second moment and, for $(\boldsymbol{a}_i, \boldsymbol{b}_i) \in (0, \infty) \times \mathbb{R}$, $i = 1, \ldots, n$, a given set of iid random vectors, we define

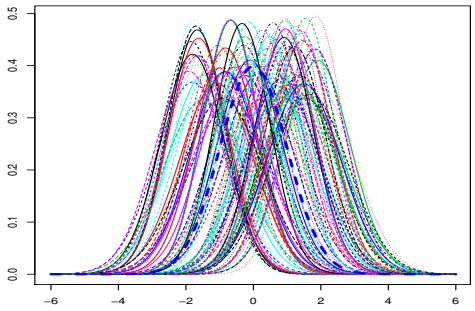
$$\boldsymbol{f}_i(x) := \boldsymbol{a}_i^{-1} f_0\left(\boldsymbol{a}_i^{-1}(x - \boldsymbol{b}_i)\right), \ x \in \mathbb{R}, \ 1 \leq i \leq n. \tag{2.2}$$

The sources of variability of the densities from model (2.2) are the variation in location along the x-axis, and the scaling variation. In Figure 1(a), we plot a sample of $n = 100$ densities from model (2.2) with f_0 being the standard Gaussian density, $\boldsymbol{a}_i \sim \mathcal{U}([0.8, 1.2])$ and $\boldsymbol{b}_i \sim \mathcal{U}([-2, 2])$, where $\mathcal{U}([x, y])$ denotes the uniform distribution on the interval $[x, y]$. In this numerical experiment, there is more variability in phase (i.e. location) than in amplitude (i.e. scaling), which can also be observed at the level of quantile functions as shown by Figure 1(b).

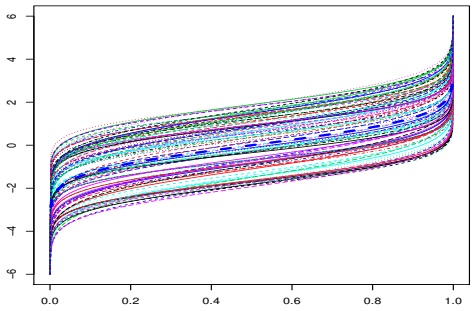
In the location-scale model (2.2), it can be checked, e.g. using the quantile averaging formula (2.1), that the empirical Wasserstein barycenter $\boldsymbol{\nu}_n^{\oplus}$ is the probability measure with density

$$\boldsymbol{f}_n^{\oplus}(x) = \bar{\boldsymbol{a}}_n^{-1} f_0\left(\bar{\boldsymbol{a}}_n^{-1}(x - \bar{\boldsymbol{b}}_n)\right),$$

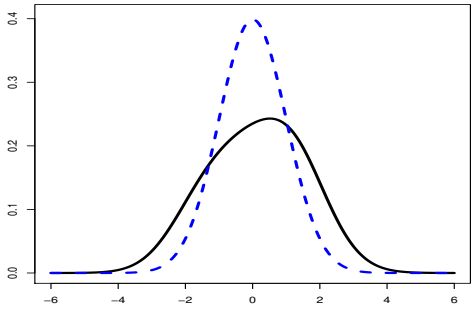
where $\bar{\boldsymbol{a}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{a}_i$ and $\bar{\boldsymbol{b}}_n = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{b}_i$. Hence, if we assume that $\mathbb{E}(\boldsymbol{a}_1) = 1$ and $\mathbb{E}(\boldsymbol{b}_1) = 0$, it follows that $d_W^2(\boldsymbol{\nu}_n^{\oplus}, \nu_0)$ converges almost surely to 0 as $n \to \infty$, meaning that $\boldsymbol{\nu}_n^{\oplus}$ is a consistent estimator of ν_0 as shown by Figure 1(f). On the contrary, the arithmetical mean measure



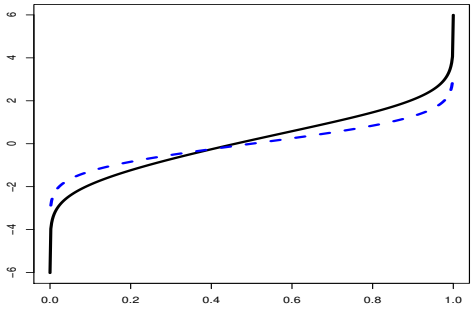
(a) Densities $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ sampled from a location-scale model



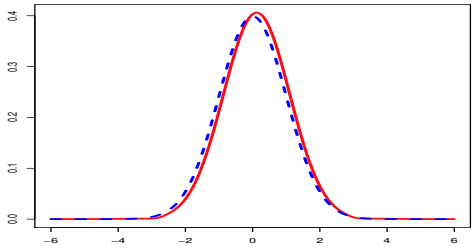
(b) Quantile functions $\boldsymbol{F}_1^-, \ldots, \boldsymbol{F}_n^-$ of $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$



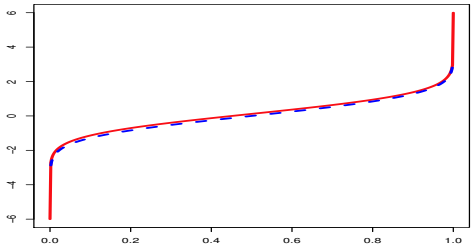
(c) Euclidean mean density $\bar{\boldsymbol{f}}_n$



(d) Quantile function of the arithmetical mean measure $\bar{\boldsymbol{\nu}}_n$ with density $\bar{\boldsymbol{f}}_n$



(e) Density $\boldsymbol{f}_n^{\oplus}$ by quantile synchronization



(f) Quantile function of the Wasserstein barycenter $\boldsymbol{\nu}_n^{\oplus}$ with density $\boldsymbol{f}_n^{\oplus}$

Figure 1: An example of $n = 100$ random densities (a) with quantile functions (b) sampled from the location-scale model (2.2) with f_0 the standard Gaussian density, $\boldsymbol{a}_i \sim \mathcal{U}([0.8, 1.2])$ and $\boldsymbol{b}_i \sim \mathcal{U}([-2, 2])$. (c,d) The solid-black curves are the Euclidean mean $\bar{\boldsymbol{f}}_n$ and its quantile function. (e,f) The solid-red curves are the structural mean $\boldsymbol{f}_n^{\oplus}$ given by quantile synchronization and the quantile function of the empirical Wasserstein barycenter $\boldsymbol{\nu}_n^{\oplus}$. In all the figures, the dashed-blue curves are either the density f_0 or its quantile function in the location-scale model (2.2).

$\bar{\boldsymbol{\nu}}_n$ is clearly not a consistent estimator of ν_0, as it can be observed in Figure 1(d).

Remark 2.1. It is clear that, in the above location-scale model, one may easily prove that $\boldsymbol{f}_n^{\oplus}$ converges almost surely to f_0 as $n \to \infty$ for various distances between density functions as illustrated by Figure 1(e). However, in this paper, we restrict our attention to the problem of how the structural mean measure ν_0 can be estimated from empirical Wasserstein barycenters with respect to the Wasserstein distance between probability measures. Showing that the density (if it exists) of such estimators converges to the density f_0 of ν_0 is not considered in this work.

2.2 Barycenters in the Wasserstein space

Let Ω be an interval of $\mathbb{R}$, possibly unbounded. We let $W_2(\Omega)$ be the set of probability measures over $(\Omega, \mathcal{B}(\Omega))$, with finite second moment, where $\mathcal{B}(\Omega)$ is the σ-algebra of Borel subsets of Ω. We also denote by $W_2^{ac}(\Omega)$ the set of measures $\nu \in W_2(\Omega)$ that are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. The cumulative distribution function (cdf), the quantile function and the density function (if $\nu \in W_2^{ac}(\Omega)$) of ν are denoted respectively by F_ν, F_ν^- and f_ν.

Definition 2.1. The quadratic Wasserstein distance d_W in $W_2(\Omega)$ is defined by

$$d_W^2(\mu,\nu) := \int_0^1 (F_\mu^-(\alpha) - F_\nu^-(\alpha))^2 d\alpha, \text{ for any } \mu,\nu \in W_2(\Omega). \tag{2.3}$$

It can be shown that $W_2(\Omega)$ endowed with d_W is a metric space, usually called Wasserstein space. For a detailed analysis of $W_2(\Omega)$ and its connection with optimal transport theory, we refer to [Vil03].

A $W_2(\Omega)$-valued random probability measure $\boldsymbol{\nu}$ is a measurable function from an abstract probability space to $(W_2(\Omega), \mathcal{B}(W_2(\Omega))$, where $\mathcal{B}(W_2(\Omega))$ is the Borel σ-algebra of $W_2(\Omega)$. We denote by $\mathbb{P}$ the probability on $W_2(\Omega)$, induced by $\boldsymbol{\nu}$.

Definition 2.2 (Square-integrability)**.** The random measure $\boldsymbol{\nu}$ is said to be square-integrable if

$$\mathbb{E}(d_W^2(\mu,\boldsymbol{\nu})) = \int_{W_2(\Omega)} d_W^2(\mu,\nu) d\mathbb{P}(\nu) < +\infty$$

for some (thus for every) $\mu \in W_2(\Omega)$.

Observe that the expectation in the definition above is well defined since $\nu \mapsto d_W(\mu,\nu)$ is continuous and therefore, measurable. In the rest of the paper we assume that random measures $\boldsymbol{\nu}$ are square integrable, in the sense of the previous definition, and so, this property is not explicitly stated in definitions or results involving $\boldsymbol{\nu}$. We use the notations $\boldsymbol{F}$, $\boldsymbol{F}^-$ and $\boldsymbol{f}$ to denote respectively the cumulative distribution function, the quantile function and the density function (if it exists) of $\boldsymbol{\nu}$.

Definition 2.3 (Population and empirical Wasserstein barycenters)**.** A population Wasserstein barycenter of $\boldsymbol{\nu}$ is defined as a minimizer of

$$\mu \mapsto \int_{W_2(\Omega)} d_W^2(\mu,\nu) d\mathbb{P}(\nu) \text{ over } \mu \in W_2(\Omega).$$

An empirical Wasserstein barycenter of $\nu_1, \ldots, \nu_n \in W_2(\Omega)$ is defined as a minimizer of

$$\mu \mapsto \frac{1}{n} \sum_{i=1}^{n} d_W^2(\mu, \nu_i) \text{ over } \mu \in W_2(\Omega).$$

Proposition 2.1.

(i) There exists a unique barycenter of $\boldsymbol{\nu}$, denoted ν_0.

(ii) $F^-_{\nu_0} = \mathbb{E}\left[\boldsymbol{F}^-\right]$.

(iii) $\mathrm{Var}(\boldsymbol{\nu}) := \mathbb{E}\left[d_W^2\left(\boldsymbol{\nu}, \nu_0\right)\right] = \int_0^1 \mathrm{Var}\left(\boldsymbol{F}^-(\alpha)\right) d\alpha < \infty$.

Proof. Observe that $\boldsymbol{F}^-$ is measurable, considered as a random element of $L^2(0,1)$ (the space of real valued functions on $(0,1)$, square-integrable wrt Lebesgue's measure), since the map $\nu \in W_2(\Omega) \mapsto F^-_\nu \in L^2(0,1)$ is continuous, because of (2.3). Assertions (i) and (ii) are contained in Proposition 4.1 in [BGKL17]. Let us prove (iii). From (2.3) and Fubini's theorem, we have

$$\mathrm{Var}\left(\boldsymbol{\nu}\right) = \mathbb{E}\left[\int_0^1 \left(\boldsymbol{F}^-(\alpha) - F^-_{\nu_0}(\alpha)\right)^2 d\alpha\right] = \int_0^1 \mathbb{E}\left[\left(\boldsymbol{F}^-(\alpha) - F^-_{\nu_0}(\alpha)\right)^2\right] d\alpha = \int_0^1 \mathrm{Var}\left(\boldsymbol{F}^-(\alpha)\right) d\alpha.$$

Finiteness of $\mathrm{Var}\left(\boldsymbol{\nu}\right)$ is immediate from the square-integrability of $\boldsymbol{\nu}$.

□

2.3 A deformable model of probability measures

Let $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ be independent copies of the random measure $\boldsymbol{\nu}$, with barycenter ν_0. We consider a deformable model of random probability measures satisfying the following assumptions:

Assumption 2.1. $\boldsymbol{\nu} \in W_2^{ac}(\Omega)$ *a.s.*

Assumption 2.2. $\nu_0 \in W_2^{ac}(\Omega)$.

Assumption 2.3. *Conditionally on $\boldsymbol{\nu}_i$, the observations $X_{i,1}, \ldots, X_{i,p_i}$ are iid random variables sampled from $\boldsymbol{\nu}_i$, where $p_i \geq 1$ is a known integer, for $1 \leq i \leq n$.*

Remark 2.2. Similar assumptions are considered in [PZ16] to characterize a population barycenter in $W_2(\Omega)$ for the purpose of estimating phase and amplitude variations from the observations of multiple point processes. For examples of parametric models satisfying the Assumptions 2.1-2.3, we refer to [BK17] and [BLGL15]. The main restriction of this deformable model is that ν_0 is assumed to be absolutely continuous.

Remark 2.3. Observe that Assumption 2.1 requires $W_2^{ac}(\Omega)$ to be a measurable subset of $W_2(\Omega)$. The proof of this technical result will appear in a forthcoming paper.

2.4 Non-smoothed empirical barycenter

To estimate the structural mean measure ν_0 from the data $(X_{i,j})_{1\leq i\leq n;\, 1\leq j\leq p_i}$, a first approach consists in computing straightaway the barycenter of the empirical measure $\tilde{\boldsymbol{\nu}}_1, \ldots, \tilde{\boldsymbol{\nu}}_n$ where $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i}\sum_{j=1}^{p_i} \delta_{X_{i,j}}$ (δ_x denotes the Dirac mass at point $x \in \Omega$). The non-smoothed empirical barycenter is thus defined as

$$\hat{\boldsymbol{\nu}}_{n,\underline{p}} = \operatorname*{arg\,min}_{\mu\in W_2(\Omega)} \frac{1}{n}\sum_{i=1}^{n} d_W^2(\tilde{\boldsymbol{\nu}}_i, \mu), \tag{2.4}$$

where $\underline{p} = (p_1, \ldots, p_n)$. In the case $p_1 = p_2 = \ldots = p_n = p$, we have the following procedure for computing the non-smoothed empirical barycenter. For each $1 \leq i \leq n$, we denote by $X^*_{i,1} \leq X^*_{i,2} \leq \ldots \leq X^*_{i,p}$ the order statistics corresponding to the i-th sample of observations $(X_{i,j})_{1\leq j\leq p}$, and define

$$\bar{X}^*_j = \frac{1}{n}\sum_{i=1}^{n} X^*_{i,j}, \text{ for all } 1 \leq j \leq p.$$

Thanks to Proposition 2.1, the quantile function of the empirical Wasserstein barycenter is the average of the quantile functions of $\tilde{\boldsymbol{\nu}}_1, \ldots, \tilde{\boldsymbol{\nu}}_n$, and thus we obtain the formula

$$\hat{\boldsymbol{\nu}}_{n,p} = \frac{1}{p}\sum_{j=1}^{p} \delta_{\bar{X}^*_j}. \tag{2.5}$$

Note that we use $\hat{\boldsymbol{\nu}}_{n,p}$ instead of $\hat{\boldsymbol{\nu}}_{n,\underline{p}}$ to denote the non-smoothed empirical barycenter in the case $p_1 = p_2 = \ldots = p_n = p$.

2.5 Smoothed empirical barycenter

An alternative approach is to use a smoothing step to obtain estimated densities and then compute the barycenter.

1. In a first step we use kernel smoothing to obtain estimators $\hat{\boldsymbol{f}}_1^{h_1}, \ldots, \hat{\boldsymbol{f}}_n^{h_n}$ of $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$, where $h_1, \ldots, h_n$ are positive bandwidth parameters, that may be different for each subject or experimental unit. In this paper we investigate a non-standard choice for the kernel function, proposed in [PZ16], to analyze the convergence of smoothed empirical barycenter in $W_2(\Omega)$. In Section 3 we give a precise definition of the resulting estimators. However, at this point it is not necessary to go into such details.

2. In a second step, an estimator of ν_0 is given by $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$, defined as the measure whose quantile function is given by

$$\hat{\boldsymbol{F}}^-_h(\alpha) = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{F}}^-_i(\alpha), \ \alpha \in [0,1], \tag{2.6}$$

where $\hat{\boldsymbol{F}}^-_i$ denotes the quantile function of the density $\hat{\boldsymbol{f}}_i^{h_i}$, $1 \leq i \leq n$. If we denote by $\hat{\boldsymbol{\nu}}_i^{h_i}$ the measure with density $\hat{\boldsymbol{f}}_i^{h_i}$ then, by Proposition 2.1, one has that $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$ is also defined as

the following smoothed empirical Wasserstein barycenter

$$\hat{\boldsymbol{\nu}}_{n,\underline{p}}^{h} = \underset{\mu \in W_2(\Omega)}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} d_W^2(\hat{\boldsymbol{\nu}}_i^{h_i}, \mu). \tag{2.7}$$

3 Convergence rate for estimators of the population Wasserstein barycenter

In this section we discuss the rates of convergence of the estimators $\hat{\boldsymbol{\nu}}_{n,\underline{p}}$ and $\hat{\boldsymbol{\nu}}_{n,\underline{p}}^{h}$, that are respectively characterized by equations (2.4) and (2.7). Some of the results presented below are using the work in [BL17], on a detailed study of the variety of rates of convergence of an empirical measure on the real line toward its population counterpart in the Wasserstein metric. Then, we discuss the optimality of these estimators from the minimax point of view following the guidelines in nonparametric statistics to derive optimal rates of convergence (see e.g. [Tsy09] for an introduction to this topic).

3.1 Non-smoothed empirical barycenter in the case of samples of equal size

Let us first characterize the rate of convergence of $\hat{\boldsymbol{\nu}}_{n,p}$, in the specific case where samples of observations per unit are of equal size, namely when $p_1 = p_2 = \ldots = p_n = p$. In what follows, we let $Y_1, \ldots, Y_p$ be iid random variables sampled from the population mean measure ν_0 (independently of the data $(X_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p}$), and we denote by $\boldsymbol{\mu}_p = \frac{1}{p} \sum_{k=1}^{p} \delta_{Y_k}$ the corresponding empirical measure.

Theorem 3.1. *If Assumptions 2.1, 2.2 and 2.3 are satisfied and if $p_1 = p_2 = \ldots = p_n = p$, then the estimator $\hat{\boldsymbol{\nu}}_{n,p}$ satisfies*

$$\begin{aligned} \mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] &= \frac{1}{n}\operatorname{Var}(\boldsymbol{\nu}) + \frac{1}{pn}\sum_{j=1}^{p}\operatorname{Var}\left(Y_j^*\right) + \sum_{j=1}^{p}\int_{(j-1)/p}^{j/p}\left(\mathbb{E}\left[Y_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha, \\ &= \frac{1}{n}\operatorname{Var}(\boldsymbol{\nu}) + \frac{1-n}{pn}\sum_{j=1}^{p}\operatorname{Var}\left(Y_j^*\right) + \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right], \end{aligned} \tag{3.1}$$

where $Y_1^* \leq Y_2^* \leq \ldots \leq Y_p^*$ denote the order statistics of the sample $Y_1, \ldots, Y_p$.

Theorem 3.1 provides exact formulas to compute the rate of convergence (for the expected squared Wasserstein distance) of $\hat{\boldsymbol{\nu}}_{n,p}$. Formula (3.1) relies on the computation of the variances of the order statistics of iid variables $Y_1, \ldots, Y_p$ sampled from the population mean measure ν_0, and on the computation of the rate of convergence of $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]$. However, deriving a sharp rate of convergence for $\hat{\boldsymbol{\nu}}_{n,p}$ using inequality (3.1) requires computing the variances of the order statistics of iid random variables. To the best of our knowledge, obtaining a sharp estimate for $\operatorname{Var}\left(Y_j^*\right)$ for any $1 \leq j \leq p$ remains a difficult task except for specific distributions. For example, if ν_0 is assumed to be a log-concave measure, then it is possible to use the results in Section 6 in [BL17] which provide sharp bounds on the variances of order statistics for such

probability measures. We discuss below some examples where equality (3.1) may be used to derive a sharp rate of convergence for $\hat{\boldsymbol{\nu}}_{n,p}$.

The case where ν_0 is the uniform distribution on $[0,1]$. In this setting, it is known (see e.g. Section 4.2 in [BL17]) that

$$\mathrm{Var}\left(Y_j^*\right) = \frac{j(p-j+1)}{(p+1)^2(p+2)} \text{ and thus } \sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right) = \frac{p}{6(p+1)}.$$

Moreover, from Theorem 4.7 in [BL17], it follows that $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] = \frac{1}{6p}$. Therefore, thanks to (3.1) we obtain

$$\begin{aligned} \mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] &= \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1-n}{6n(p+1)} + \frac{1}{6p} \\ &= \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1}{6}\left(\frac{1}{n(p+1)} + \frac{1}{p(p+1)}\right). \end{aligned} \tag{3.2}$$

Equality (3.2) thus shows that, when ν_0 is the uniform distribution on $[0,1]$, the rate of convergence of $\hat{\boldsymbol{\nu}}_{n,p}$ is given by

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \asymp \frac{1}{n} + \frac{1}{np} + \frac{1}{p^2}, \tag{3.3}$$

and this rate is sharp.

Beyond the specific case where ν_0 is a uniform distribution, it is in general difficult to compute $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]$. Nevertheless, thanks to Theorem 4.3 in [BL17], we have the following bounds

$$\frac{1}{2p}\sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right) \leq \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] \leq \frac{2}{p}\sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right), \tag{3.4}$$

for any distribution $\nu_0 \in W_2(\Omega)$.

The case where ν_0 is the one-sided exponential distribution. Combining inequality (3.4) with (3.1), it follows that

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \leq \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1+n}{pn}\sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right). \tag{3.5}$$

Now (using e.g. Remark 6.13 in [BL17]) one has that if ν_0 is the one-sided exponential distribution (with density e^{-x}, for $x \geq 0$) then

$$\sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right) = \sum_{j=1}^{p} \frac{1}{j} \sim \log p, \text{ as } p \to \infty.$$

Therefore, there exist a constant $c > 0$ such that

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \leq \frac{1}{n}\operatorname{Var}(\boldsymbol{\nu}) + c\left(1 + \frac{1}{n}\right)\frac{\log p}{p}, \tag{3.6}$$

for all sufficiently large p. Hence, when ν_0 is the exponential distribution the above inequalities show that the rate of convergence of $\hat{\boldsymbol{\nu}}_{n,p}$ is $\mathcal{O}\left(\frac{1}{n} + \left(\frac{1}{np} + \frac{1}{p}\right)\log p\right)$.

The case where ν_0 is a Gaussian distribution. By Theorem 4.3 and Corollary 6.14 in [BL17] there exist constants $c_1, c_2 > 0$ such that

$$c_1\frac{\log\log p}{p} \leq \frac{1}{p}\sum_{j=1}^{p}\operatorname{Var}\left(Y_j^*\right) \leq c_2\frac{\log\log p}{p}. \tag{3.7}$$

Therefore, combining the above upper bound with (3.5), one finally has that

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \leq \frac{1}{n}\operatorname{Var}(\boldsymbol{\nu}) + c_2\left(\frac{1}{n} + 1\right)\frac{\log\log p}{p}. \tag{3.8}$$

when ν_0 is the standard Gaussian. In this setting, the rate of convergence is thus $\mathcal{O}\left(\frac{1}{n} + \left(\frac{1}{np} + \frac{1}{p}\right)\log\log p\right)$.

Upper bounds in more general cases. If one is interested in deriving an upper bound on $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right]$ for a larger class of measures $\nu_0 \in W_2(\Omega)$ (e.g. beyond the log-concave case), another approach is as follows. Noting that the term $\frac{1-n}{pn}\sum_{j=1}^{p}\operatorname{Var}(Y_j^*)$ in equality (3.1) is negative, a straightforward consequence of Theorem 3.1 is the following upper bound

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \leq \frac{1}{n}\operatorname{Var}(\boldsymbol{\nu}) + \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]. \tag{3.9}$$

Then, thanks to inequality (3.9), to derive the rate of convergence of $\hat{\boldsymbol{\nu}}_{n,p}$, it remains to control the rate of convergence of the empirical measure $\boldsymbol{\mu}_p$ to ν_0 for the expected squared Wasserstein distance. This issue is discussed in detail in [BL17]. In particular, the work in [BL17] describes a variety of rates for the expected distance $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]$, from the standard one $\mathcal{O}\left(\frac{1}{p}\right)$ to slower rates. For example, by Theorem 5.1 in [BL17], the following upper bound holds

$$\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] \leq \frac{2}{p+1}J_2(\nu_0), \tag{3.10}$$

where, the so-called J_2-functional is defined as $J_2 : W_2^{ac}(\Omega) \to \mathbb{R}_+ \cup \{\infty\}$, with

$$J_2(\nu) = \int_\Omega \frac{F_\nu(x)(1 - F_\nu(x))}{f_\nu(x)}dx, \quad \nu \in W_2^{ac}(\Omega). \tag{3.11}$$

The J_2 functional is shown to be measurable in Proposition A.1 of the Appendix.

Therefore, provided that $J_2(\nu_0)$ is finite, the empirical measure $\boldsymbol{\mu}_p$ converges to ν_0 at the rate $\mathcal{O}\left(\frac{1}{p}\right)$. Hence, using inequality (3.10), we have:

Corollary 3.1. *Suppose that Assumptions 2.1, 2.2 and 2.3 are satisfied. Then, the estimator* $\hat{\boldsymbol{\nu}}_{n,p}$ *satisfies*

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \leq \frac{1}{n}\operatorname{Var}(\boldsymbol{\nu}) + \frac{2}{p+1}J_2(\nu_0). \tag{3.12}$$

By Corollary 3.1, if $J_2(\nu_0) < \infty$, then it follows that $\hat{\boldsymbol{\nu}}_{n,p}$ converges to ν_0 at the rate $\mathcal{O}\left(\frac{1}{n} + \frac{1}{p}\right)$. Hence, in the setting where $p \geq n$ and $J_2(\nu_0) < \infty$, $\hat{\boldsymbol{\nu}}_{n,p}$ converges at the classical parametric rate $\mathcal{O}\left(\frac{1}{n}\right)$. The case $p \geq n$, usually refereed to as the dense case in the literature on functional data analysis (see e.g. [LH10] and references therein), corresponds to the situation where the number of observations per unit/subject is larger than the sample size n of functional objects. In the sparse case (when $p < n$), the non-smoothed Wasserstein barycenter converges at the rate $\mathcal{O}\left(\frac{1}{p}\right)$, if $J_2(\nu_0) < \infty$.

Remark 3.1. When ν_0 is the uniform distribution on $[0,1]$ one has that $J_2(\nu_0) < \infty$, but we have shown that $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \asymp \frac{1}{n} + \frac{1}{np} + \frac{1}{p^2}$. Hence, in this setting, $\hat{\boldsymbol{\nu}}_{n,p}$ converges at the parametric rate $\mathcal{O}\left(\frac{1}{n}\right)$ provided that $p \geq \sqrt{n}$, which is a dense regime condition weaker than $p \geq n$.

To conclude this discussion on the rate of convergence of the non-smoothed Wasserstein barycenter in the case of samples of equal size, we study in more detail the control of the rate of convergence of the term $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]$ in inequality (3.9). As pointed out in many works (see for example [dBGU05, BL17] and the references therein) the finiteness of $J_2(\nu_0)$ is the key point to control the convergence of the empirical measure $\boldsymbol{\mu}_p$ to the population measure ν_0 in the Wasserstein space. Some known facts concerning this issue are the following.

1. If $J_2(\nu_0) < \infty$ then ν_0 is supported on an interval of $\mathbb{R}$ and its density is a.e. strictly positive on this interval.

2. If ν_0 is compactly supported with a density bounded away from zero or with a log-concave density then $J_2(\nu_0) < \infty$.

3. If the density of ν_0 is of the form $C_\alpha e^{-|x|^\alpha}$ then $J_2(\nu_0)$ is finite if and only if $\alpha > 2$. In particular, $J_2(\nu_0) = \infty$ for the Gaussian distribution.

Some further comments can be made in the case where ν_0 is Gaussian. In this setting, one has that $J_2(\nu_0) = \infty$ and the rate of convergence of $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]$ to zero is slower than $\mathcal{O}\left(\frac{1}{p}\right)$. Indeed, from Corollary 6.14 in [BL17], if ν_0 is the standard Gaussian, then the rate of convergence of $\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right]$ is $\mathcal{O}\left(\frac{\log\log p}{p}\right)$, which leads to the upper bound (3.8) for the non-smoothed Wasserstein barycenter $\hat{\boldsymbol{\nu}}_{n,p}$ in the Gaussian case. Hence, thanks to (3.8), one has that if p is sufficiently large with respect to n (namely when $p \geq n \log\log p$), then $\hat{\boldsymbol{\nu}}_{n,p}$ also converges at the classical parametric rate $\mathcal{O}\left(\frac{1}{n}\right)$ when ν_0 is the standard Gaussian distribution.

Remark 3.2. Following the work in [BL17], if ν_0 has a log-concave distribution, then one may obtain rates of convergence for $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right]$ that are slower than the standard $\mathcal{O}\left(\frac{1}{p}\right)$ rate

(e.g. for beta or exponential distributions). Moreover, it is also possible to considerer for any $q \geq 1$ and for any absolutely continuous probability ν on Ω, the functional

$$J_q(\nu) = \int_\Omega \frac{(F_\nu(x)(1-F_\nu(x)))^{q/2}}{f_\nu(x)^{q-1}} dx,$$

in order to control the rate of convergence of the empirical measure to ν for the q-Wasserstein distance.

3.2 Non-smoothed empirical barycenter in the general case

Let us now consider the general situation where the p_i's are possibly different. The result below gives an upper bound on the rate of convergence of $\hat{\boldsymbol{\nu}}_{n,\underline{p}}$ where $\underline{p} = (p_1, \ldots, p_n)$.

Theorem 3.2. *Suppose that Assumptions 2.1, 2.2 and 2.3 are satisfied. Then*

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \nu_0)\right] \leq n^{-1/2}\sqrt{\mathrm{Var}(\boldsymbol{\nu})} + \frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}\left[d_W^2(\tilde{\boldsymbol{\nu}}_i, \boldsymbol{\nu}_i)\right]},$$

where $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i}\sum_{j=1}^{p_i} \delta_{X_{i,j}}$ *for each* $1 \leq i \leq n$.

For the random measure $\boldsymbol{\nu}$, we consider the extended random variable $J_2(\boldsymbol{\nu})$ (see Proposition A.1). Since the $\boldsymbol{\nu}_i$'s are independent copies of $\boldsymbol{\nu}$, by applying inequality (3.10) it follows that

$$\sqrt{\mathbb{E}\left[d_W^2(\tilde{\boldsymbol{\nu}}_i, \boldsymbol{\nu}_i)\right]} \leq \sqrt{2\mathbb{E}\left[J_2(\boldsymbol{\nu})\right]} p_i^{-1/2}.$$

Hence, from Theorem 3.2, we finally obtain the following upper bound on the rate of convergence for the non-smoothed empirical barycenter

Corollary 3.2. *Suppose that Assumptions 2.1, 2.2 and 2.3 are satisfied. If $J_2(\boldsymbol{\nu})$ has a finite expectation, then*

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \nu_0)\right] \leq n^{-1/2}\sqrt{\mathrm{Var}(\boldsymbol{\nu})} + \sqrt{2\mathbb{E}\left[J_2(\boldsymbol{\nu})\right]}\left(\frac{1}{n}\sum_{i=1}^n p_i^{-1/2}\right).$$

From Corollary 3.2, one has that if $\min_{1\leq i\leq n} p_i \geq n$ (dense case), then $\frac{1}{n}\sum_{i=1}^n p_i^{-1/2} \leq n^{-1/2}$, and thus, the non-smoothed empirical barycenter converges at the parametric rate $n^{-1/2}$ (provided that $\mathbb{E}\left[J_2(\boldsymbol{\nu})\right] < \infty$), namely

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \nu_0)\right] \leq \left(\sqrt{\mathrm{Var}(\boldsymbol{\nu})} + \sqrt{2\mathbb{E}\left[J_2(\boldsymbol{\nu})\right]}\right) n^{-1/2}. \tag{3.13}$$

Remark 3.3. Knowing if $J_2(\boldsymbol{\nu})$ has a finite expectation is in general a difficult task. But, if we assume that the density $\boldsymbol{f}$ of $\boldsymbol{\nu}$ is bounded below by a non-random positive constant then (obviously) $\mathbb{E}\left[J_2(\boldsymbol{\nu})\right] < \infty$.

3.3 The case of smoothed empirical barycenters

In this section, we assume that $\Omega = [0,1]$ and we discuss the rate of convergence of smoothed empirical barycenters $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$ (note that the following results hold if Ω is any compact interval).

To choose an appropriate kernel function to study the convergence rate of the estimator $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$, we follow the proposal made in [PZ16]. We let ψ be a positive, smooth and symmetric density on the real line, such that $\int_{\mathbb{R}} x^2\psi(x)dx = 1$. We also denote by Ψ the cdf of the density ψ and, for a bandwidth parameter $h > 0$, we let $\psi_h(x) = \frac{1}{h}\psi\left(\frac{x}{h}\right)$. Then, for any $y \in [0,1]$ and $h > 0$, we denote by μ_h^y the measure supported on $[0,1]$ whose density $f_{\mu_h^y}$ is defined as

$$f_{\mu_h^y}(x) = \psi_h(x-y) + 2b_2\psi_h(x-y)\mathbb{1}_{\{x-y>0\}} + 2b_1\psi_h(x-y)\mathbb{1}_{\{x-y<0\}} + 4b_1b_2, \quad x \in [0,1], \tag{3.14}$$

where $b_1 = 1 - \Psi\left((1-y)/h\right)$ and $b_2 = \Psi\left(-y/h\right)$. Then, for each $1 \leq i \leq n$, we construct a kernel density estimator of $\boldsymbol{f}_i$ by defining $\hat{\boldsymbol{f}}_i^{h_i}$ as the density associated to the measure

$$\hat{\boldsymbol{\nu}}_i^{h_i} = \frac{1}{p_i}\sum_{j=1}^{p_i} \mu_{h_i}^{X_{i,j}}, \tag{3.15}$$

where $h_i > 0$ is a bandwidth parameter depending on i. For a discussion on the intuition for this choice of kernel smoothing, we refer to [PZ16]. A key property to analyze the convergence rate of $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$ is the following lemma which relates the Wasserstein distance between $\hat{\boldsymbol{\nu}}_i^{h_i}$ and the empirical measure $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i}\sum_{j=1}^{p_i}\delta_{X_{i,j}}$.

Remark 3.4. The results in [PZ16] strongly depend on the assumption that Ω is compact. To go beyond this assumption, one should be able to extend (in an appropriate way) the density $f_{\mu_h^y}(x)$ to a non-compact setting, which we believe to be a difficult task. Nevertheless, it should be remarked that our results on non-smoothed empirical Wasserstein barycenter hold in the general case where $\Omega = \mathbb{R}$.

Lemma 3.1. *Let $1 \leq i \leq n$. Suppose that $0 < h_i \leq 1/4$, then one has the following upper bound*

$$d_W^2(\hat{\boldsymbol{\nu}}_i^{h_i}, \tilde{\boldsymbol{\nu}}_i) \leq 3h_i^2 + 4\Psi(-1/\sqrt{h_i}), \qquad 1 \leq i \leq n. \tag{3.16}$$

Furthermore, if there exist constants $C > 0$ and $\alpha \geq 5$ satisfying

$$\psi(x) \leq Cx^{-\alpha}, \text{ for all sufficiently large } x, \tag{3.17}$$

then

$$d_W^2(\hat{\boldsymbol{\nu}}_i^{h_i}, \tilde{\boldsymbol{\nu}}_i) \leq C_\psi h_i^2,$$

for h_i small enough and some constant $C_\psi > 0$ depending only on ψ.

Proof. The upper bound (3.16) follows immediately from Lemma 1 in [PZ16] and the symmetry of ψ. Then, by applying inequality (3.17) and since ψ is symmetric, it follows that for h small enough

$$\Psi(-1/\sqrt{h}) = \int_{-\infty}^{-1/\sqrt{h}} \psi(x)dx = \int_{1/\sqrt{h}}^{+\infty} \psi(x)dx \leq C\int_{1/\sqrt{h}}^{+\infty} x^{-\alpha}dx = \frac{C}{\alpha-1}h^{(\alpha-1)/2}.$$

Hence, the second part of Lemma 3.1 is a consequence of the above inequality, the fact that $\alpha \geq 5$, and the upper bound (3.16), which completes the proof. □

The result below gives a rate of convergence for the estimator $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$.

Theorem 3.3. *Suppose that Assumptions 2.1, 2.2 and 2.3 are satisfied, and that the density ψ, used to define kernel smoothing in (3.15), satisfies inequality (3.17). If $J_2(\boldsymbol{\nu})$ has a finite expectation, and the bandwidth parameters h_i are small enough, then we have*

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \nu_0)\right] \leq n^{-1/2}\sqrt{\mathrm{Var}(\boldsymbol{\nu})} + C_\psi^{1/2}\left(\frac{1}{n}\sum_{i=1}^n h_i\right) + \sqrt{2\mathbb{E}\left[J_2(\boldsymbol{\nu})\right]}\left(\frac{1}{n}\sum_{i=1}^n p_i^{-1/2}\right). \quad (3.18)$$

Theorem 3.3 can then be used to discuss choices of bandwidth parameters that may lead to a parametric rate of convergence. For example, if $0 < h_i \leq n^{-1/2}$ for all $1 \leq i \leq n$ and $\min_{1\leq i\leq n} p_i \geq n$ (dense case), then Theorem 3.3 implies that (for sufficiently large n to ensure that $\max_{1\leq i\leq n}\{h_i\}$ is small enough)

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \nu_0)\right] \leq \left(\sqrt{\mathrm{Var}(\boldsymbol{\nu})} + C_\psi^{1/2} + \sqrt{2\mathbb{E}\left[J_2(\boldsymbol{\nu})\right)}\right) n^{-1/2}. \quad (3.19)$$

Remark 3.5. In the dense case (namely $\min_{1\leq i\leq n} p_i \geq n$) it can be seen, by comparing the upper bounds (3.13) and (3.19), that a preliminary smoothing step of the data (namely kernel smoothing the empirical measures $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i}\sum_{j=1}^{p_i}\delta_{X_{i,j}}$) does not improve the parametric rate of convergence $n^{-1/2}$. Moreover, the bandwidth values have to be small to ensure the rate of convergence $n^{-1/2}$ for $\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \nu_0)\right]$. This result comes from the fact that we evaluate the risk of empirical barycenters at the level of measures in $W_2(\Omega)$, and that we do not aim to control an estimation of the density f_0 of the population mean measure ν_0.

Remark 3.6. Theorem 3.3 shares similarities with the results from Theorem 2 in [PZ16], which gives the rate of convergence for smoothed Wasserstein barycenters, computed from the realizations of multiple Poisson processes in a deformable model of measures similar to that of this paper. The main difference in [PZ16] is that the number $\boldsymbol{p}_i$ of observations for each experimental unit are independent Poisson random variables with expectation $\mathbb{E}(\boldsymbol{p}_i) = \tau_n$ for each $1 \leq i \leq n$ (they are not deterministic integers). From such observations and under similar assumptions, it is proved in [PZ16] that the following upper bound holds (in probability)

$$d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \nu_0) \leq \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^n h_i\right) + \mathcal{O}_{\mathbb{P}}\left(\frac{1}{\sqrt[4]{\tau_n}}\right). \quad (3.20)$$

The upper bound (3.20) is very similar to (3.18), proposed in this paper. Both approaches essentially split the distance $d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \nu_0)$ into three terms:

1. a parametric term of the order $\sqrt{n}$ coming from the observation of a sample of size n of the random measure $\boldsymbol{\nu}$,

2. a term involving kernel smoothing and

3. a term involving the Wasserstein distance between ν_0 and its empirical counterpart $\boldsymbol{\mu}_p = \frac{1}{p}\sum_{k=1}^{p}\delta_{Y_k}$, with $Y_1, \ldots, Y_p$ iid ν_0-distributed random variables.

Under the conditions that $\tau_n \geq \mathcal{O}(n^2)$ and $\max_{1\leq i\leq n} h_i \leq \mathcal{O}_{\mathbb{P}}\left(n^{-1/2}\right)$, it follows from Theorem 2 in [PZ16] that $\hat{\boldsymbol{\nu}}^h_{n,p}$ converges at the parametric rate $\mathcal{O}\left(n^{-1/2}\right)$, for the Wasserstein distance. The quantity τ_n represents the averaged number of points observed for each Poisson process. As remarked in [PZ16] the condition $\tau_n \geq \mathcal{O}(n^2)$ corresponds to a dense sampling regime where the number n of observed Poisson processes should not grow too fast with respect to the expected number of points observed for each process. Comparing the upper bounds (3.18) and (3.20), the main difference in the control of the risk of $\hat{\boldsymbol{\nu}}^h_{n,p}$ between our approach and the one in [PZ16] is that we use the condition $\mathbb{E}\left[J_2(\boldsymbol{\nu})\right] < \infty$. Under such an assumption, the smoothed Wasserstein barycenter (for the model considered in this paper) may be shown to converge at the rate $\mathcal{O}\left(n^{-1/2}\right)$, for the expected Wasserstein distance, under the dense case setting $p := \min\{p_i,\ 1 \leq i \leq n\} \geq n$ which is somehow a weaker condition than $\mathbb{E}(\boldsymbol{p}_i) \geq n^2$ for all $1 \leq i \leq n$, as in [PZ16]. Therefore, it might be argued that there is a sort of "optimality gap" in [PZ16], and that the results in this paper are a first step towards closing this gap. But, on the other hand, the result in [PZ16] is more general because nothing is assumed about the finiteness of the functional J_2. In particular, the upper bound (3.20) given in [PZ16] also holds when J_2 is infinite, which is not the case for the upper bound (3.18) in this paper.

3.4 A lower bound on the minimax risk

In the rest of this section, we show that, in the dense case and for the expected squared Wasserstein distance, the rate of convergence $\mathcal{O}\left(n^{-1}\right)$ for non-smoothed empirical Wasserstein barycenters is optimal from the minimax point of view over a large class of random measures $\boldsymbol{\nu}$ satisfying the deformable model defined in Section 2.3 through Assumptions 2.1, 2.2 and 2.3.

Definition 3.1. For $\nu_0 \in W_2^{ac}(\Omega)$ and $\sigma > 0$, we define $\mathcal{D}(\Omega, \nu_0, \sigma^2)$ as the class of $W_2(\Omega)$-valued random measures $\boldsymbol{\nu}$ that satisfy the deformable model defined in Section 2.3 with $\mathrm{Var}(\boldsymbol{\nu}) < \sigma^2$.

Definition 3.2. Let $A > 0$. We denote by $\mathcal{F}(\mathbb{R}, A) \subseteq W_2^{ac}(\mathbb{R})$ a given set of measures with variance bounded by A, which contains at least all Gaussian distributions with variance bounded by A.

Then, by inequality (3.9), we obtain the following corollary giving a uniform rate of convergence for the non-smoothed empirical barycenter in the case of samples of equal size.

Corollary 3.3. *Let $A > 0$ and $\sigma > 0$. Suppose that $p_1 = p_2 = \ldots = p_n = p$. Then, if there exists a constant $c_0 > 0$ such that*

$$\sup_{\nu_0 \in \mathcal{F}(\mathbb{R},A)} \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] \leq \frac{c_0}{n}, \tag{3.21}$$

it follows that

$$\sup_{\nu_0 \in \mathcal{F}(\mathbb{R},A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R},\nu_0,\sigma^2)} \mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] \leq \frac{\sigma^2 + c_0}{n}. \tag{3.22}$$

The condition (3.21) may be interpreted as the generalization of the dense case setting that has been discussed in the previous sections as it is valid only if p is sufficiently large with respect to n. As an example, let $A \geq 0$ and suppose that the set $\mathcal{F}(\mathbb{R}, A)$ can be partitioned as

$$\mathcal{F}(\mathbb{R}, A) = \mathcal{F}_0(\mathbb{R}, A) \cup \mathcal{G}(\mathbb{R}, A),$$

where $\mathcal{F}_0(\mathbb{R}, A)$ denotes a set of measures $\nu_0 \in W_2^{ac}(\mathbb{R})$ with variance bounded by A satisfying

$$A_0 := \sup_{\nu_0 \in \mathcal{F}_0(\mathbb{R}, A)} J_2(\nu_0) < +\infty,$$

while $\mathcal{G}(\mathbb{R}, A)$ denotes the set of Gaussian distributions with variance bounded by A. For this example, it follows from inequalities (3.4), (3.7) and (3.10) in Section 3.1 (with samples of equal size) that

$$\sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] \leq \max\left(A_0 \frac{2}{p+1}, c_2 A \frac{\log\log p}{p}\right) \leq \max(2A_0, c_2 A) \frac{\log\log p}{p},$$

provided that $\log\log p \geq 1$, where c_2 is a constant from inequality (3.7). Hence, if p is such that $p \geq n \log\log p$ then condition (3.21) is satisfied with

$$c_0 = \max(2A_0, c_2 A) = \max\left(2 \sup_{\nu_0 \in \mathcal{F}_0(\mathbb{R}, A)} J_2(\nu_0), c_2 A\right).$$

The following theorem shows that the upper bound (3.22) in Corollary 3.3 is optimal (in term of rate of convergence) from the minimax point of view in nonparametric statistics.

Theorem 3.4. *Let $A > 0$ and $\sigma > 0$. Then the following lower bound holds*

$$\inf_{\hat{\boldsymbol{\nu}}} \sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}, \nu_0)\right] \geq \frac{e^{-2} \min(A^{1/2}, \sigma)}{4} n^{-1/2}, \tag{3.23}$$

where $\hat{\boldsymbol{\nu}} = \hat{\nu}\left((X_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p_i}\right)$ denotes any estimator taking values in $(W_2(\mathbb{R}), \mathcal{B}(W_2(\mathbb{R})))$ with $\hat{\nu}$ denoting a measurable function of the data $(X_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p_i}$ sampled from the deformable model defined in Section 2.3.

Now, by using inequalities (3.13) and (3.19) and Definitions 3.1 and 3.2 introduced above, we also obtain the following corollary giving uniform rates of convergence for the non-smoothed Wasserstein barycenter in the general situation where the p_i's are possibly different.

Corollary 3.4. *Let $A > 0$ and $\sigma > 0$. Suppose that the assumptions of Corollary 3.2 are satisfied, and that $p_i \geq n$, for all $1 \leq i \leq n$. Then, the following upper bound holds*

$$\sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \nu_0)\right] \leq n^{-1/2} \left(\sigma + \sqrt{2} \sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \sqrt{\mathbb{E}\left[J_2(\boldsymbol{\nu})\right]}\right).$$

Hence, under the assumptions made in Corollary 3.4, the estimator $\hat{\boldsymbol{\nu}}_{n,\underline{p}}$ converges at the optimal rate of convergence $n^{-1/2}$ provided that

$$\sup_{\nu_0 \in \mathcal{F}(\mathbb{R},A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R},\nu_0,\sigma^2)} \mathbb{E}\left[J_2(\boldsymbol{\nu})\right] < +\infty.$$

We conclude this discussion by a few remarks on the rate of convergence that may be obtained in the sparse case.

Remark 3.7. In the case of samples of equal size, the results above show that the rate of convergence n^{-1} is optimal in the dense case (for the risk $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right]$), namely when the number $p = p_1 = \ldots = p_n$ of observations per units is sufficiently large with respect to n. We believe that deriving a lower bound on the minimax risk depending on p in the sparse case (e.g. when $p < n$) is more involved. Indeed, from the discussion in Section 3.1 on the rate of convergence of the non-smoothed empirical barycenter, it appears that the exact decay of $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right]$ as a function of p is difficult to establish as it depends on ν_0. Indeed, from Section 3.1, one has that

- if ν_0 is the uniform distribution on $[0,1]$, then $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right] \asymp \frac{1}{n} + \frac{1}{np} + \frac{1}{p^2}$,
- if ν_0 is the one-sided exponential distribution, then $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right] = \mathcal{O}\left(\frac{1}{n} + \log p\left(\frac{1}{np} + \frac{1}{p}\right)\right)$,
- if ν_0 is the standard Gaussian distribution, then $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right] = \mathcal{O}\left(\frac{1}{n} + \log\log p\left(\frac{1}{np} + \frac{1}{p}\right)\right)$,
- if ν_0 is such that $J_2(\nu_0) < +\infty$, then $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right] = \mathcal{O}\left(\frac{1}{n} + \frac{1}{p}\right)$.

From Theorem 3.1, one has that the risk of the non-smoothed empirical barycenter may be bounded from below as follows

$$\sup_{\nu_0 \in \mathcal{F}(\mathbb{R},A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R},\nu_0,\sigma^2)} \mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right] \geq \sup_{\nu_0 \in \mathcal{F}(\mathbb{R},A)} \sum_{j=1}^{p} \int_{(j-1)/p}^{j/p} \left(\mathbb{E}\left[Y_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha. \quad (3.24)$$

The quantity $\sum_{j=1}^{p} \int_{(j-1)/p}^{j/p} \left(\mathbb{E}\left[Y_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha$ may be interpreted as a bias term when estimating the unknown measure by the nonparametric estimator $\boldsymbol{\mu}_p = \frac{1}{p}\sum_{j=1}^{p} \delta_{Y_j}$. Therefore, for samples of equal size and in the sparse case (when $p < n$), the lower bound (3.24) may be used to control (as a function of p) the best rate of convergence for $\hat{\boldsymbol{\nu}}_{n,p}$ that may be obtained over the class of measures $\nu_0 \in \mathcal{F}(\mathbb{R}, A)$.

Remark 3.8. Finally, we remark that better rates of convergence may be obtained if one assumes a parametric model for the random measure $\boldsymbol{\nu}$. Indeed, suppose that $\mu_0 \in W_2^{ac}(\Omega)$ denotes a *known probability measure* with expectation m_0 and variance σ_0^2 and consider that the data $(X_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p_i}$ are sampled from iid random measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ satisfying the location model

$$F_{\boldsymbol{\nu}_i}^-(\alpha) = F_{\mu_0}^-(\alpha) + \boldsymbol{a}_i,\ \alpha \in [0,1],\ 1 \leq i \leq n, \quad (3.25)$$

where $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are iid random variables with unknown expectation $\bar{a}$ and variance γ^2. In this model, the population Wasserstein barycenter is the measure ν_0 with quantile function

$F^-_{\nu_0}(\cdot) = F^-_{\mu_0}(\cdot) + \bar{a}$. Since, the measure μ_0 is assumed to be known, a natural estimator for ν_0 is to take the measure $\hat{\boldsymbol{\nu}}_0$ with quantile function $F^-_{\hat{\boldsymbol{\nu}}_0}(\cdot) = F^-_{\mu_0}(\cdot) + \hat{\boldsymbol{a}}$, with

$$\hat{\boldsymbol{a}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\sum_{j=1}^{p_i} X_{ij} - m_0.$$

Then, it is clear that

$$\begin{aligned}\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_0, \nu_0)\right] &= \int_0^1 \mathbb{E}\left(F^-_{\hat{\boldsymbol{\nu}}_0}(\alpha) - F^-_{\nu_0}(\alpha)\right)^2 d\alpha = \mathbb{E}\left(\hat{\boldsymbol{a}} - \bar{a}\right)^2 \\ &= \frac{\sigma_0^2 + \gamma^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i}\right) + \frac{\gamma^2}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{p_i - 1}{p_i}\right).\end{aligned}$$

In the case where all the p_i's are equal to p, then the above equality simplifies to

$$\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_0, \nu_0)\right] = \frac{\sigma_0^2 + \gamma^2}{np} + \frac{\gamma^2}{n}\frac{p-1}{p},$$

and thus the parametric estimator $\hat{\boldsymbol{\nu}}_0$ converges at the rate $\mathcal{O}\left(\frac{1}{n} + \frac{1}{np}\right)$. Therefore, either in the dense ($p \geq n$) or sparse case ($p < n$), the parametric estimator $\hat{\boldsymbol{\nu}}_0$ converges at the rate $\mathcal{O}\left(\frac{1}{n}\right)$ in the location model (3.25) when the "reference measure" μ_0 is known. Moreover, in the sparse case ($p < n$) the parametric estimator $\hat{\boldsymbol{\nu}}_0$ converges faster than the non-smoothed empirical Wasserstein barycenter $\hat{\boldsymbol{\nu}}_{n,p}$, thanks to the results in Section 3.1.

4 Numerical experiments

In this simulation study we perform Monte Carlo experiments to compare the decay of the squared Wassertein risks $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}^h_{n,p}, \nu_0)\right]$ and $\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right]$, of the smoothed and non-smoothed empirical Wasserstein barycenters $\hat{\boldsymbol{\nu}}^h_{n,p}$ and $\hat{\boldsymbol{\nu}}_{n,p}$, as a function of the number n of units and the sample size p. The theoretical results in this paper indicate that, in the dense case, both estimators converge at the optimal parametric rate $\mathcal{O}\left(\frac{1}{n}\right)$. However, in the sparse case it remains unclear if a preliminary smoothing step may improve the quality of estimation of the population Wasserstein barycenter. The purpose of these numerical experiments is thus to compare the behavior of smoothed and non-smoothed empirical Wasserstein barycenters, in these two settings, and analyze the influence of the number n of measures and the sample size p.

We analyze the case of random samples $(X_{i,j})_{1\leq i\leq n;\, 1\leq j\leq p}$, with $10 \leq n \leq 200$ and $10 \leq p \leq 200$. Data are generated from densities supported on a compact interval Ω that are sampled from the following model, accounting for vertical and horizontal variations

$$\boldsymbol{f}_i(x) = \boldsymbol{a}_i^{-1} f\left(\boldsymbol{a}_i^{-1}(x - \boldsymbol{b}_i)\right),\ x \in \Omega,\ 1 \leq i \leq n, \tag{4.1}$$

where f is either the density of the standard Gaussian law (truncated to the interval $[-3, 3]$) or the uniform density on the interval $[0, 1]$, $\boldsymbol{a}_i \sim \mathcal{U}([0.8, 1.2])$, $\boldsymbol{b}_i \sim \mathcal{U}([-2, 2])$. This setting

corresponds to the the simulation study conducted in [PM16]. For each choice of f (either the Gaussian or Uniform case), the interval Ω is taken such that each random function $\boldsymbol{f}_i$ has a compact support included in Ω. Therefore, the population Wasserstein barycenter in model (4.1) is the measure with density f thanks to the fact that $\mathbb{E}(\boldsymbol{b}_i) = 0$ and $\mathbb{E}(\boldsymbol{a}_i) = 1$. The Gaussian case (resp. Uniform case) corresponds to the estimation of a Wasserstein barycenter having smooth (resp. non-differentiable) density f.

For given values of n and p, we evaluate the Wasserstein risk of $\hat{\boldsymbol{\nu}}^h_{n,p}$ by repeating $M = 100$ times the following experiment. First, data are simulated from model (4.1). Then, for each $1 \leq i \leq n$, we use kernel smoothing to compute the density $\hat{\boldsymbol{f}}_i^{h_i}$ and its associated measure $\hat{\boldsymbol{\nu}}_i^{h_i}$. We slightly deviate from the analysis carried out in Section 3, as we use a Gaussian kernel to smooth the data $(X_{i,j})_{1\leq j\leq p}$, with bandwidth h_i chosen by cross validation, instead of the specific kernel defined in (3.14), that has been proposed for the convergence analysis of $\hat{\boldsymbol{\nu}}^h_{n,p}$. We found that this modification has no substantial effect on the finite sample performance of the procedure, and a similar choice has been made in the numerical experiments in [PZ16]. In Figure 2(a) (resp. Figure 3(a)) we display an example of densities estimated from realizations of the model (4.1) with $n = p = 100$ and f the truncated Gaussian density (resp. f the Uniform density). After computing the quantile function $F^-_{\hat{\boldsymbol{\nu}}^h_{n,p}}$ of the empirical smoothed Wasserstein barycenter $\hat{\boldsymbol{\nu}}^h_{n,p}$, we approximate $d^2_W(\hat{\boldsymbol{\nu}}^h_{n,p}, \nu_0) = \int_0^1 (F^-_{\hat{\boldsymbol{\nu}}^h_{n,p}}(\alpha) - F^-_{\nu_0}(\alpha))^2 d\alpha$ by discretizing the integral over a fine grid of values for $\alpha \in]0, 1[$. This approximated value of $d^2_W(\hat{\boldsymbol{\nu}}^h_{n,p}, \nu_0)$ is then averaged over the $M = 100$ repeated experiments to approximate $\mathbb{E}\left[d^2_W(\hat{\boldsymbol{\nu}}^h_{n,p}, \nu_0)\right]$.

Thanks to the explicit expression (2.5) of the non-smoothed empirical Wasserstein barycenter $\hat{\boldsymbol{\nu}}_{n,p}$, its quantile function $F^-_{\hat{\boldsymbol{\nu}}_{n,p}}$ is straightforward to compute on a grid of values for α, and the Wasserstein risk $\mathbb{E}\left(d^2_W(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right)$ is then approximated in the same way by using Monte Carlo repetitions.

For values of n and p ranging from 10 to 200, we display in Figure 2(c) and 2(d) (resp. Figure 3(c) and 3(d)) these approximations of $\mathbb{E}\left[d^2_W(\hat{\boldsymbol{\nu}}^h_{n,p}, \nu_0)\right]$ and $\mathbb{E}\left[d^2_W(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right]$ (in logarithmic scale) for f the truncated Gaussian density (resp. f the Uniform density). For both estimators, it appears that the Wasserstein risk is clearly a decreasing function of the number n of units. To the contrary, increasing p does not lead to a significant decay of this risk. This suggest that $\frac{1}{n}\operatorname{Var}(\boldsymbol{\nu})$ is the most significant term in the upper bound (3.12) of the Wasserstein risk of $\hat{\boldsymbol{\nu}}_{n,p}$.

In Figure 2(b) and Figure 3(b), we also display the logarithm of the ratio

$$\mathbb{E}\left[d^2_W(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] / \mathbb{E}\left[d^2_W(\hat{\boldsymbol{\nu}}^h_{n,p}, \nu_0)\right].$$

It can be observed that:

- When the population Wasserstein barycenter has a smooth density f (Gaussian case) then, for values of p larger than 100, both estimators (smoothed and non-smoothed empirical Wasserstein barycenters) appear to have squared Wasserstein risks of approximately the same magnitude. This tends to confirm the results on convergence rates obtained in Section 3, in the dense case (when p is sufficiently large with respect to n), which show that a preliminary smoothing is not necessary in this setting. For smaller values of p (between

10 and 50), the smoothed empirical Wasserstein barycenter has a smaller Wasserstein risk. This suggests that introducing a smoothing step through kernel smoothing of the data, in each experimental unit, may improve the quality of the estimation of ν_0, when the sample size p is small (which corresponds to the sparse case), in the setting where the population Wasserstein barycenter is a distribution with a smooth density. In this example, Gaussian kernel smoothing is particularly well suited, which may explain the better performances obtained with a smoothed empirical Wasserstein barycenter in the sparse case.

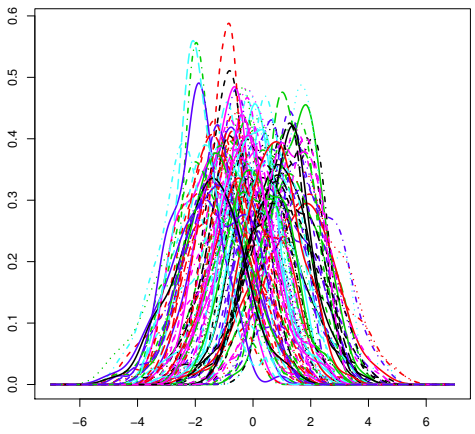
- When the population Wasserstein barycenter has a non-smooth density f (Uniform case) then, except for very small values of $p \leq 10$, the non-smoothed empirical Wasserstein barycenter has always a lower squared Wasserstein risk, both in the sparse and dense cases. A preliminary smoothing with a Gaussian kernel does not improve the estimation of a Wasserstein barycenter having piecewise constant density and, in this setting, such a step is thus not necessary.

5 Conclusion and perspectives

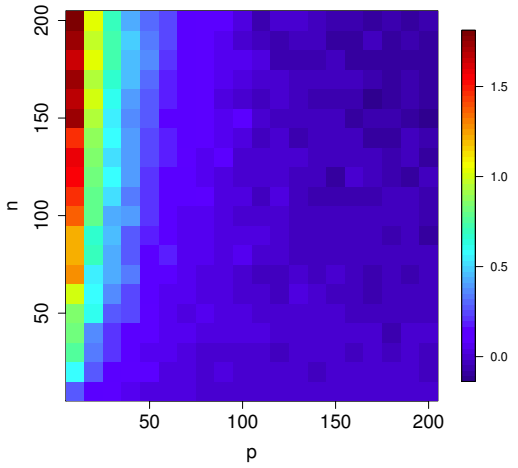
In this paper we have studied the rate of convergence for the (squared) Wasserstein distance of (possibly smoothed) empirical barycenters in a deformable model of measures. The main contributions of this work can be summarized as follows. In the case of samples of equal size, we have derived a closed-form expression for the risk of non-smooth empirical barycenter, as a function of n and p, which allows to derive sharp rates of convergence whose decay in p depends on the population mean measure ν_0. A second conclusion of the paper is that, in the dense case (when the minimal number $\min_{1\leq i\leq n} p_i \geq n$ of observations per unit is sufficiently large with respect to the number n of observed measures), the non-smoothed empirical barycenter converges at the parametric rate of convergence n^{-1}. Moreover, this rate is shown to be a lower bound on the decay of a novel notion of minimax risk, in the deformable model of measures introduced in this paper. In the dense case, the numerical experiments that have been carried out are in agreement with the theoretical results which show that, in this setting, one may only consider the non-smoothed empirical Wasserstein barycenter and that a preliminary smoothing step is not necessary to obtain an optimal estimator.

A first perspective would be to find a lower bound on the minimax risk depending on p in the sparse case. However, to this end, we believe that one has to first obtain sharper rates of convergence as a function of p, for the non-smoothed empirical barycenter.

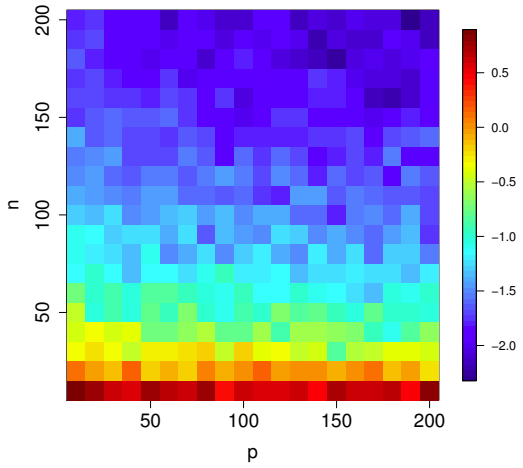
Finally, a natural perspective is to ask how these results can be extended to higher dimensional settings for measures supported on $\mathbb{R}^d$, with $d > 1$. However, we believe that this is far from being obvious as the results in this paper rely heavily on the closed-form formula of Wasserstein barycenters, in the one-dimensional setting though quantile averaging. Such results do not hold in higher-dimension, for data sets consisting of iid random vectors, sampled from unknown random measures supported on $\mathbb{R}^2$ or $\mathbb{R}^3$, for example.



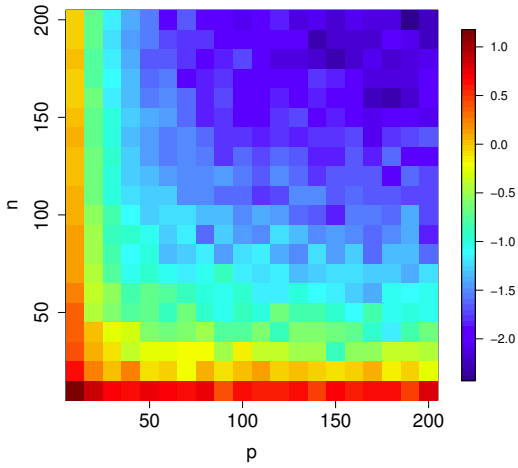
(a) An example of estimated densities.



(b) $\log\left(\mathbb{E}\left(d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right)/\mathbb{E}\left(d_W^2(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}},\nu_0)\right)\right)$.

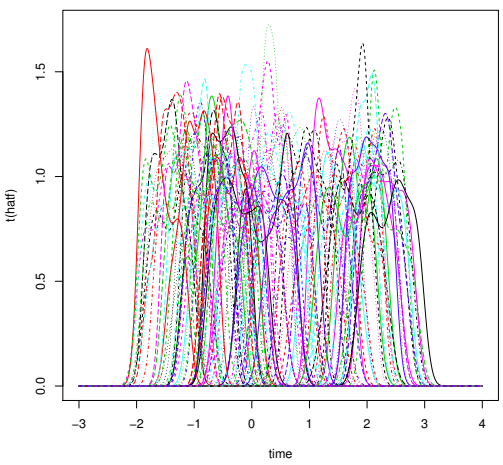


(c) Log-Wasserstein risk of $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$.

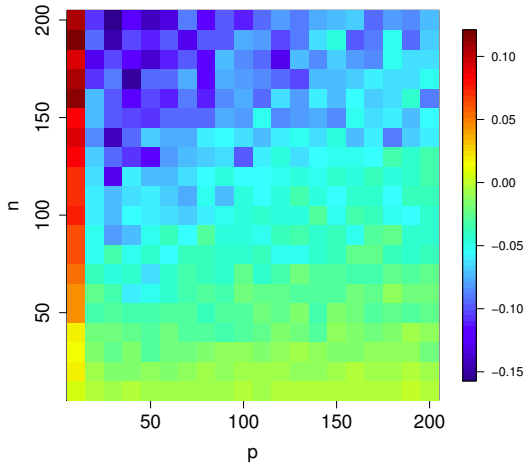


(d) Log-Wasserstein risk of $\hat{\boldsymbol{\nu}}_{n,p}$.

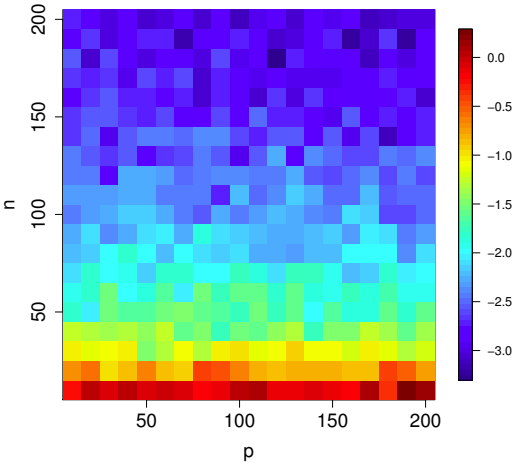
Figure 2: Gaussian case - (a) An example of $n = 100$ densities estimated from data sampled from model (4.1) with the choice of a standard Gaussian density f truncated to the interval $[-3, 3]$ and $n = p = 100$, (b) Logarithm of the ratio $\mathbb{E}(d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0))/\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}},\nu_0)\right]$, (c) Wasserstein risk of the smoothed empirical barycenter $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$ with kernel bandwidths chosen by cross-validation, (c) Wasserstein risk of the non-smoothed empirical barycenter $\hat{\boldsymbol{\nu}}_{n,p}$. The values of n and p vary from 10 to 200 by an increment of 10.



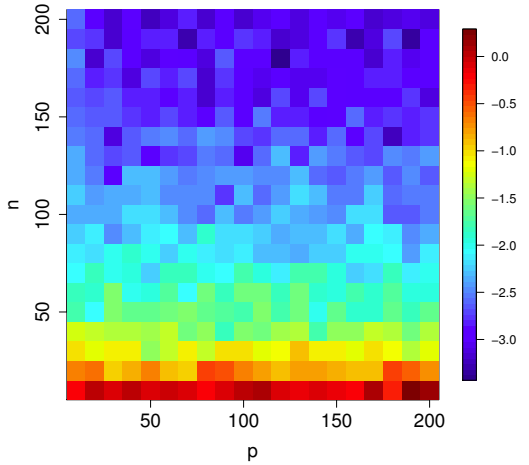
(a) An example of estimated densities.



(b) $\log\left(\mathbb{E}\left(d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0)\right)/\mathbb{E}\left(d_W^2(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}},\nu_0)\right)\right)$.



(c) Log-Wasserstein risk of $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$.



(d) Log-Wasserstein risk of $\hat{\boldsymbol{\nu}}_{n,p}$.

Figure 3: Uniform case - (a) An example of $n = 100$ densities estimated from data sampled from model (4.1) with the choice of a uniform density f on the interval $[0, 1]$ and $n = p = 100$, (b) Logarithm of the ratio $\mathbb{E}(d_W^2(\hat{\boldsymbol{\nu}}_{n,p},\nu_0))/\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}},\nu_0)\right]$, (c) Wasserstein risk of the smoothed empirical barycenter $\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}$ with kernel bandwidths chosen by cross-validation, (c) Wasserstein risk of the non-smoothed empirical barycenter $\hat{\boldsymbol{\nu}}_{n,p}$. The values of n and p vary from 10 to 200 by an increment of 10.

A Appendix

A.1 Auxiliary results

We recall that $Y_1, \dots, Y_p$ denote iid random variables sampled from the measure ν_0 (independently of the data), and that the associated empirical measure is $\boldsymbol{\mu}_p = \frac{1}{p}\sum_{j=1}^p \delta_{Y_j}$. By Corollary 4.5 in [BL17], it follows that

$$\mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] = \frac{1}{p}\sum_{j=1}^p \mathrm{Var}\left(Y_j^*\right) + \sum_{j=1}^p \int_{(j-1)/p}^{j/p} \left(\mathbb{E}\left[Y_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha, \tag{A.1}$$

where $Y_1^* \le Y_2^* \le \ldots \le Y_p^*$ denote the order statistics of the sample $Y_1, \dots, Y_p$.

It is well known that the j-th order statistic Y_j^* admits the density (see e.g. [BL17])

$$f_{Y_j^*}(y) = \frac{p!}{(j-1)!(p-j)!} f_0(y)[F_0(y)]^{j-1}[1-F_0(y)]^{p-j},\ y \in \Omega. \tag{A.2}$$

Moreover, under Assumption 2.2, one has, conditionally on $\boldsymbol{F}_i$, that the j-th order statistic $X_{i,j}^*$ admits the density

$$f_{X_{i,j}^*}(x) = \frac{p!}{(j-1)!(p-j)!} \boldsymbol{f}_i(x)[\boldsymbol{F}_i(x)]^{j-1}[1-\boldsymbol{F}_i(x)]^{p-j},\ x \in \Omega. \tag{A.3}$$

Let us recall the notation $\bar{X}_j^* = \frac{1}{n}\sum_{i=1}^n X_{i,j}^*$. Then the following result holds.

Lemma A.1. *If Assumptions 2.1, 2.2 and 2.3 are satisfied, then, for each $1 \le j \le p$, one has*

$$\mathbb{E}\left[\bar{X}_j^*\right] = \mathbb{E}\left[Y_j^*\right].$$

Moreover,

$$\frac{1}{p}\sum_{j=1}^p \mathrm{Var}\left(\bar{X}_j^*\right) - \mathrm{Var}\left(Y_j^*\right) = \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1-n}{pn}\sum_{j=1}^p \mathrm{Var}\left(Y_j^*\right).$$

Proof. Let $1 \le j \le p$ and $1 \le i \le n$. Thanks to the expression (A.3) for the density of $X_{i,j}^*$, one has that

$$\begin{aligned}
\mathbb{E}\left[X_{i,j}^* | \boldsymbol{F}_i\right] &= \int_\Omega x \frac{p!}{(j-1)!(p-j)!} \boldsymbol{f}_i(x)[\boldsymbol{F}_i(x)]^{j-1}[1-\boldsymbol{F}_i(x)]^{p-j} dx \\
&= \int_0^1 \boldsymbol{F}_i^-(\alpha) \frac{p!}{(j-1)!(p-j)!} \alpha^{j-1}(1-\alpha)^{p-j} d\alpha,
\end{aligned}$$

where we used the change of variable $\alpha = \boldsymbol{F}_i(x)$ to obtain the last equality. By Proposition 2.1,

$\mathbb{E}\left[\boldsymbol{F}_i^-\right] = F_0^-(\alpha)$ for each $1 \le i \le n$. Therefore, using Fubini's theorem, it follows that

$$\begin{aligned}
\mathbb{E}\left[X_{i,j}^*\right] &= \mathbb{E}\left[\mathbb{E}\left[X_{i,j}^*|\boldsymbol{F}_i\right]\right] \\
&= \mathbb{E}\left[\int_0^1 \boldsymbol{F}_i^-(\alpha)\frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha\right] \\
&= \int_0^1 \mathbb{E}\left[\boldsymbol{F}_i^-(\alpha)\right]\frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha \\
&= \int_0^1 F_0^-(\alpha)\frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha \\
&= \int_\Omega y\frac{p!}{(j-1)!(p-j)!}f_0(y)[F_0(y)]^{j-1}[1-F_0(y)]^{p-j}dy = \mathbb{E}\left[Y_j^*\right], \qquad \text{(A.4)}
\end{aligned}$$

where we used the change of variable $y = F_0^-(\alpha)$ and (A.2) to obtain the last equality above. Given that $\mathbb{E}\left[\bar{X}_j^*\right] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}\left[X_{i,j}^*\right]$, the first statement of Lemma A.1 follows from equality (A.4).

Now, let us prove the second statement of Lemma A.1. Thanks to formula (A.3) for the density of $X_{i,j}^*$, one has that, for each $1 \le j \le p$ and $1 \le i \le n$,

$$\begin{aligned}
\mathbb{E}\left[|X_{i,j}^*|^2\right] &= \mathbb{E}\left[\mathbb{E}\left[|X_{i,j}^*|^2|\boldsymbol{F}_i\right]\right] \\
&= \mathbb{E}\left[\int_\Omega x^2\frac{p!}{(j-1)!(p-j)!}\boldsymbol{f}_i(x)[\boldsymbol{F}_i(x)]^{j-1}[1-\boldsymbol{F}_i(x)]^{p-j}dx\right] \\
&= \int_0^1 \mathbb{E}\left[\left|\boldsymbol{F}_i^-(\alpha)\right|^2\right]\frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha, \qquad \text{(A.5)}
\end{aligned}$$

where, again, we use the change of variable $\alpha = \boldsymbol{F}_i(x)$, and Fubini's theorem to obtain the last equality. Similarly, from (A.2) it follows that, for each $1 \le j \le p$,

$$\begin{aligned}
\mathbb{E}\left[|Y_j^*|^2\right] &= \int_\Omega y^2\frac{p!}{(j-1)!(p-j)!}f_0(y)[F_0(y)]^{j-1}[1-F_0(y)]^{p-j}dy \\
&= \int_0^1 \left|F_0^-(\alpha)\right|^2\frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha. \qquad \text{(A.6)}
\end{aligned}$$

Since $\bar{X}_j^* = \frac{1}{n}\sum_{i=1}^n X_{i,j}^*$, we obtain by independence that

$$\mathrm{Var}\left(\bar{X}_j^*\right) = \frac{1}{n^2}\sum_{i=1}^n \mathrm{Var}\left(X_{i,j}^*\right) = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}\left[|X_{i,j}^*|^2\right] - \left|\mathbb{E}\left[X_{ij}^*\right]\right|^2.$$

Hence, using equalities (A.4), (A.5) and (A.6), and the fact that $\mathbb{E}\left[\left|\boldsymbol{F}_i^-(\alpha)\right|^2\right] = \mathbb{E}\left[\left|\boldsymbol{F}^-(\alpha)\right|^2\right]$

for each $1 \leq i \leq n$, we obtain

$$\begin{aligned}
\mathrm{Var}\left(\bar{X}_j^*\right) &= \frac{1}{n}\left(\int_0^1 \mathbb{E}\left[\left|\boldsymbol{F}^-(\alpha)\right|^2\right] \frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha - \left|\mathbb{E}\left[Y_j^*\right]\right|^2\right) \\
&= \frac{1}{n}\left(\int_0^1 \mathbb{E}\left[\left|\boldsymbol{F}^-(\alpha)\right|^2\right] \frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha + \mathrm{Var}\left(Y_j^*\right) - \mathbb{E}\left[|Y_j^*|^2\right]\right) \\
&= \frac{1}{n}\left(\int_0^1 \left(\mathbb{E}\left[\left|\boldsymbol{F}^-(\alpha)\right|^2\right] - \left|F_0^-(\alpha)\right|^2\right) \frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha + \mathrm{Var}\left(Y_j^*\right)\right) \\
&= \frac{1}{n}\left(\int_0^1 \mathrm{Var}\left(\boldsymbol{F}^-(\alpha)\right) \frac{p!}{(j-1)!(p-j)!}\alpha^{j-1}(1-\alpha)^{p-j}d\alpha + \mathrm{Var}\left(Y_j^*\right)\right),
\end{aligned}$$

where, for the last inequalities, we used that $\mathbb{E}\left[\boldsymbol{F}^-\right] = F_0^-$ by Proposition 2.1. Therefore, from the above equality, one finally obtains

$$\frac{1}{p}\sum_{j=1}^p \mathrm{Var}\left(\bar{X}_j^*\right) - \mathrm{Var}\left(Y_j^*\right) = \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1-n}{pn}\sum_{j=1}^p \mathrm{Var}\left(Y_j^*\right),$$

which completes the proof of Lemma A.1. □

A.2 Proof of Theorem 3.1

By Definition 2.1 of the Wasserstein distance, and since $\hat{\boldsymbol{\nu}}_{n,p} = \frac{1}{p}\sum_{j=1}^p \delta_{\bar{X}_j^*}$, it follows by using Fubini's theorem that

$$\begin{aligned}
\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] &= \mathbb{E}\left[\int_0^1 \left(F^-_{\hat{\boldsymbol{\nu}}_{n,p}}(\alpha) - F_0^-(\alpha)\right)^2 d\alpha\right] \\
&= \mathbb{E}\left[\sum_{j=1}^p \int_{(j-1)/p}^{j/p} \left(\bar{X}_j^* - F_0^-(\alpha)\right)^2 d\alpha\right] \\
&= \sum_{j=1}^p \int_{(j-1)/p}^{j/p} \mathbb{E}\left[\bar{X}_j^* - F_0^-(\alpha)\right]^2 d\alpha \\
&= \sum_{j=1}^p \int_{(j-1)/p}^{j/p} \mathbb{E}\left[\bar{X}_j^* - \mathbb{E}\left[\bar{X}_j^*\right]\right]^2 + \left(\mathbb{E}\left[\bar{X}_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha \\
&= \frac{1}{p}\sum_{j=1}^p \mathrm{Var}\left(\bar{X}_j^*\right) + \sum_{j=1}^p \int_{(j-1)/p}^{j/p} \left(\mathbb{E}\left[\bar{X}_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha. \qquad \text{(A.7)}
\end{aligned}$$

From Lemma A.1, one has that $\mathbb{E}\left[\bar{X}_j^*\right] = \mathbb{E}\left[Y_j^*\right]$. Therefore, by combining (A.7) with (A.1), we obtain

$$\begin{aligned}
\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_{n,p}, \nu_0)\right] &= \frac{1}{p}\sum_{j=1}^{p} \mathrm{Var}\left(\bar{X}_j^*\right) + \sum_{j=1}^{p} \int_{(j-1)/p}^{j/p} \left(\mathbb{E}\left[Y_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha \\
&= \frac{1}{p}\sum_{j=1}^{p} \left(\mathrm{Var}\left(\bar{X}_j^*\right) - \mathrm{Var}\left(Y_j^*\right)\right) + \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] \\
&= \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1-n}{pn}\sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right) + \mathbb{E}\left[d_W^2(\boldsymbol{\mu}_p, \nu_0)\right] \qquad \text{(A.8)} \\
&= \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}) + \frac{1}{pn}\sum_{j=1}^{p} \mathrm{Var}\left(Y_j^*\right) + \sum_{j=1}^{p} \int_{(j-1)/p}^{j/p} \left(\mathbb{E}\left[Y_j^*\right] - F_0^-(\alpha)\right)^2 d\alpha,
\end{aligned}$$

where the last equalities also follow from Lemma A.1 and (A.1), which completes the proof of Theorem 3.1.

A.3 Proof of Theorem 3.2

We recall that $\boldsymbol{\nu}_n^{\oplus}$ denotes the measure with quantile function given by equation (2.1). By the triangle inequality, we have that

$$d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \nu_0) \leq d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus}) + d_W(\boldsymbol{\nu}_n^{\oplus}, \nu_0). \qquad \text{(A.9)}$$

Thanks to Definition 2.1 of the Wasserstein distance, it follows by Fubini's theorem that

$$\mathbb{E}\left[d_W^2(\boldsymbol{\nu}_n^{\oplus}, \nu_0)\right] = \int_0^1 \mathbb{E}\left[\bar{\boldsymbol{F}}_n^-(\alpha) - F_0^-(\alpha)\right]^2 d\alpha = \int_0^1 \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \boldsymbol{F}_i^-(\alpha) - F_0^-(\alpha)\right]^2 d\alpha.$$

By Assumption 2.3, one has that $\mathbb{E}\left[\boldsymbol{F}_i^-(\alpha)\right] = F_0^-(\alpha)$ for any $1 \leq i \leq n$, and thus, by independence of the random variables $\boldsymbol{F}_i^-(\alpha)$, one obtains

$$\mathbb{E}\left[d_W^2(\boldsymbol{\nu}_n^{\oplus}, \nu_0)\right] = \frac{1}{n}\mathrm{Var}(\boldsymbol{\nu}). \qquad \text{(A.10)}$$

Hence, by (A.10) and the inequality $\mathbb{E}\left[d_W(\boldsymbol{\nu}_n^{\oplus}, \nu_0)\right] \leq \sqrt{\mathbb{E}\left[d_W^2(\boldsymbol{\nu}_n^{\oplus}, \nu_0)\right]}$, one obtains

$$\mathbb{E}\left[d_W(\boldsymbol{\nu}_n^{\oplus}, \nu_0)\right] \leq n^{-1/2}\sqrt{\mathrm{Var}(\boldsymbol{\nu})}. \qquad \text{(A.11)}$$

Now, let us remark that

$$d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus}) = \left\| \frac{1}{n}\sum_{i=1}^{n} F_{\tilde{\boldsymbol{\nu}}_i}^- - \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{F}_i^- \right\| \leq \frac{1}{n}\sum_{i=1}^{n} \left\| F_{\tilde{\boldsymbol{\nu}}_i}^- - \boldsymbol{F}_i^- \right\|,$$

where $F^-_{\tilde{\boldsymbol{\nu}}_i}$ denotes the quantile function of the measure $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i}\sum_{j=1}^{p_i}\delta_{X_{i,j}}$ for each $1 \leq i \leq n$, and $\|\cdot\|$ denotes the usual norm in $L^2([0,1],dx)$. Hence, the above inequality leads to the following upper bound

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus})\right] \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}\left[d_W^2(\tilde{\boldsymbol{\nu}}_i, \boldsymbol{\nu}_i)\right]}. \tag{A.12}$$

Therefore, Theorem 3.2 follows from inequality (A.9) combined with (A.11) and (A.12), which completes its proof.

A.4 Proof of Theorem 3.3

The proof follows the same lines as the proof of Theorem 3.2. By the triangle inequality, we have that

$$d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \nu_0) \leq d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus}) + d_W(\boldsymbol{\nu}_n^{\oplus}, \nu_0), \tag{A.13}$$

where $\boldsymbol{\nu}_n^{\oplus}$ is the measure with quantile function given by equation (2.1). The expectation of the second term in the right-hand size of inequality (A.13) is controlled by inequality (A.11). Then, to control the first term, it suffices to remark that

$$d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus}) = \left\| \frac{1}{n}\sum_{i=1}^{n} F^-_{\hat{\boldsymbol{\nu}}_i^{h_i}} - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{F}_i^- \right\|,$$

where $F^-_{\hat{\boldsymbol{\nu}}_i^{h_i}}$ denotes the quantile function of the measure $\hat{\boldsymbol{\nu}}_i^{h_i}$ defined in (3.15), and $\|\cdot\|$ denotes the usual norm in $L^2([0,1],dx)$. Therefore, one has that

$$\begin{aligned} d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus}) &\leq \frac{1}{n}\sum_{i=1}^{n}\left\| F^-_{\hat{\boldsymbol{\nu}}_i^{h_i}} - \boldsymbol{F}_i^- \right\| \\ &= \frac{1}{n}\sum_{i=1}^{n} d_W(\hat{\boldsymbol{\nu}}_i^{h_i}, \boldsymbol{\nu}_i) \leq \frac{1}{n}\sum_{i=1}^{n} d_W(\hat{\boldsymbol{\nu}}_i^{h_i}, \tilde{\boldsymbol{\nu}}_i) + \frac{1}{n}\sum_{i=1}^{n} d_W(\tilde{\boldsymbol{\nu}}_i, \boldsymbol{\nu}_i), \end{aligned}$$

where $\tilde{\boldsymbol{\nu}}_i = \frac{1}{p_i}\sum_{j=1}^{p_i}\delta_{X_{i,j}}$ for each $1 \leq i \leq n$. Hence, the above inequalities lead to the following upper bound

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus})\right] \leq \frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}\left[d_W^2(\hat{\boldsymbol{\nu}}_i^{h_i}, \tilde{\boldsymbol{\nu}}_i)\right]} + \frac{1}{n}\sum_{i=1}^{n}\sqrt{\mathbb{E}\left[d_W^2(\tilde{\boldsymbol{\nu}}_i, \boldsymbol{\nu}_i)\right]}.$$

Finally, by applying Lemma 3.1 and inequality (3.10), we obtain that

$$\mathbb{E}\left[d_W(\hat{\boldsymbol{\nu}}^h_{n,\underline{p}}, \boldsymbol{\nu}_n^{\oplus})\right] \leq C_\psi^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} h_i\right) + \sqrt{2\mathbb{E}\left[J_2(\boldsymbol{\nu})\right]}\left(\frac{1}{n}\sum_{i=1}^{n} p_i^{-1/2}\right). \tag{A.14}$$

Therefore, Theorem 3.3 follows from inequality (A.13) combined with (A.11) and (A.14), which completes the proof.

A.5 Proof of Theorem 3.4

Let $A > 0$ and $\sigma > 0$. To derive Theorem 3.4, we follow the classical scheme in nonparametric statistics to obtain optimal rates of convergence (see Chapter 2 in [Tsy09]). To this end we introduce appropriate random measures in $W_2(\Omega)$, satisfying the deformable model defined in Section 2.3, that will serve as the basic hypotheses to obtain a lower bound.

Let $m^{(1)}$ and $m^{(2)}$ be two real numbers such that

$$|m^{(1)} - m^{(2)}| = 2Cn^{-1/2}, \tag{A.15}$$

where C is a positive constant to be specified later on. For $k = 1, 2$, we let $\boldsymbol{a}^{(k)}$ be independent Gaussian random variables with $\mathbb{E}\left[\boldsymbol{a}^{(k)}\right] = m^{(k)}$ and $\mathrm{Var}(\boldsymbol{a}^{(k)}) = \gamma^2$ with $\gamma = \min(A^{1/2}, \sigma)$. We also let $\mathcal{H}^{(k)}$ denote the hypothesis that the data are sampled according the following deformable model:

$$X_{i,j}^{(k)} = \boldsymbol{a}_i^{(k)} + Z_{i,j}^{(k)}, \quad 1 \le i \le n,\ 1 \le j \le p_i, \tag{A.16}$$

where $\boldsymbol{a}_1^{(k)}, \dots, \boldsymbol{a}_n^{(k)}$ are independent copies of $\boldsymbol{a}^{(k)}$, and the $Z_{i,j}^{(k)}$'s are iid random variables sampled from the Gaussian distribution with zero mean and variance γ^2, that are independent of the $\boldsymbol{a}_i^{(k)}$'s. If we let $\boldsymbol{X}_i^{(k)}$ be the random vector in $\mathbb{R}^{p_i}$ whose component are the random variables $(X_{i,j}^{(k)})_{1 \le j \le p_i}$, then the deformable model (A.16) corresponds to the assumption that $\boldsymbol{X}_1^{(k)}, \dots, \boldsymbol{X}_n^{(k)}$ are independent random vectors, such that $\boldsymbol{X}_i^{(k)}$ is a Gaussian vector with

$$\mathbb{E}\left[\boldsymbol{X}_i^{(k)}\right] = m^{(k)}\boldsymbol{e}_i \quad \text{and} \quad \mathrm{Var}\left(\boldsymbol{X}_i^{(k)}\right) = \gamma^2\left(\boldsymbol{e}_i\boldsymbol{e}_i^t + \boldsymbol{I}_i\right), \tag{A.17}$$

where $\boldsymbol{e}_i$ is the vector in $\mathbb{R}^{p_i}$ with all entries equal to one, the notation $\mathrm{Var}\,(\boldsymbol{X})$ denotes the covariance matrix of a random vector $\boldsymbol{X}$, and $\boldsymbol{I}_i$ is the identity $p_i \times p_i$ matrix. For each $k = 1, 2$, if we denote by $\boldsymbol{\nu}_i^{(k)}$ the measure from which $(X_{i,j}^{(k)})_{1 \le j \le p_i}$ are sampled, it follows, from model (A.16), that $\boldsymbol{\nu}_1^{(k)}, \dots, \boldsymbol{\nu}_n^{(k)}$ are independent copies of the random measure $\boldsymbol{\nu}^{(k)}$ with density $\frac{1}{\gamma}\phi_0\left(\frac{x - \boldsymbol{a}^{(k)}}{\gamma}\right), x \in \mathbb{R}$, where ϕ_0 is the density of the standard Gaussian distribution. It can be easily checked that the barycenter $\nu_0^{(k)}$ in $W_2(\mathbb{R})$ of the random measure $\boldsymbol{\nu}^{(k)}$ is the Gaussian distribution with mean $m^{(k)}$ and variance γ^2, and that

$$d_W(\nu_0^{(1)}, \nu_0^{(2)}) = |m^{(1)} - m^{(2)}| = 2Cn^{-1/2}. \tag{A.18}$$

Hence, $\nu_0^{(k)}$ belongs to the class of distributions $\mathcal{F}(\mathbb{R}, A)$ introduced in Definition 3.2, for $k = 1, 2$. Moreover, since $F^-_{\boldsymbol{\nu}^{(k)}}(\alpha) = \Phi_0^-(\alpha) + \boldsymbol{a}^{(k)}$, $t \in [0, 1]$, where Φ_0^- is the quantile function of the standard Gaussian distribution, it follows that

$$\mathrm{Var}(\boldsymbol{\nu}^{(k)}) = \int_0^1 \mathrm{Var}\left(\boldsymbol{a}^{(k)}\right) d\alpha = \gamma^2 \le \sigma^2.$$

Therefore, the random measure $\boldsymbol{\nu}^{(k)}$ belongs to the class of distributions $\mathcal{D}(\mathbb{R}, \nu_0^{(k)}, \sigma^2)$ introduced in Definition 3.1, for $k = 1, 2$.

Then, for $k = 1, 2$, we let $\mathbb{P}^{(k)}$ be the probability measure of the data in model (A.16) under the hypothesis $\mathcal{H}^{(k)}$. From our remark above, one has that $\mathbb{P}^{(k)}$ is the product of n Gaussian measures $\mathbb{P}_i^{(k)}$ on $\mathbb{R}^{p_i}$ with mean and covariance given by (A.17) for $1 \leq i \leq n$. Hence, the Kullback divergence $K\left(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}\right)$ between $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$ can be decomposed as follows

$$\begin{aligned} K\left(\mathbb{P}^{(1)}, \mathbb{P}^{(2)}\right) &= \sum_{i=1}^{n} K\left(\mathbb{P}_i^{(1)}, \mathbb{P}_i^{(2)}\right) \\ &= \frac{1}{2\gamma^2}|m^{(1)} - m^{(2)}|^2 \sum_{i=1}^{n} \boldsymbol{e}_i^t \left(\boldsymbol{e}_i \boldsymbol{e}_i^t + \boldsymbol{I}_i\right)^{-1} \boldsymbol{e}_i \\ &= \frac{1}{2\gamma^2}|m^{(1)} - m^{(2)}|^2 \sum_{i=1}^{n} \frac{p_i}{p_i + 1} \leq \frac{n}{2\gamma^2}|m^{(1)} - m^{(2)}|^2 \\ &\leq 2C^2 \max(A^{-1}, \sigma^{-2}), \end{aligned} \tag{A.19}$$

where the last inequality follows from (A.15) and the fact that $\gamma^2 = \min(A, \sigma^2)$.

To conclude the proof, we finally follow the arguments from Section 2.2 in [Tsy09] on a reduction scheme to a finite number M of hypotheses (here $M = 2$). First, thanks to Markov's inequality, one has that

$$\inf_{\hat{\boldsymbol{\nu}}} \sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \mathbb{E}\left[n^{1/2} d_W(\hat{\boldsymbol{\nu}}, \nu_0)\right] \geq C \inf_{\hat{\boldsymbol{\nu}}} \sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \mathbb{P}\left(d_W(\hat{\boldsymbol{\nu}}, \nu_0) \geq C n^{-1/2}\right),$$

and thus, the following lower bound holds

$$\inf_{\hat{\boldsymbol{\nu}}} \sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \mathbb{E}\left[n^{1/2} d_W(\hat{\boldsymbol{\nu}}, \nu_0)\right] \geq C \inf_{\hat{\boldsymbol{\nu}}} \max_{k \in \{1,2\}} \mathbb{P}^{(k)}\left(d_W(\hat{\boldsymbol{\nu}}, \nu_0^{(k)}) \geq C n^{-1/2}\right), \tag{A.20}$$

where $\mathbb{P}^{(k)}$ denotes the probability measure of the data in model (A.16) under the hypothesis $\mathcal{H}^{(k)}$ for $k = 1, 2$. Now, thanks to equality (A.18), the two hypotheses $\mathcal{H}^{(1)}$ and $\mathcal{H}^{(2)}$ are $2s$-separated in the sense of condition (2.7) in [Tsy09] (with $s = Cn^{-1/2}$). Hence, by inequality (2.9) in [Tsy09], one has that

$$\inf_{\hat{\boldsymbol{\nu}}} \max_{k \in \{1,2\}} \mathbb{P}^{(k)}\left(d_W(\hat{\boldsymbol{\nu}}, \nu_0^{(k)}) \geq C n^{-1/2}\right) \geq p_{e,1}, \tag{A.21}$$

where $p_{e,1}$ is defined by equation (2.10) in [Tsy09]. Then, by the upper bound (A.19) on the Kullback divergence between $\mathbb{P}^{(1)}$ and $\mathbb{P}^{(2)}$, we can combine the Kullback version of Theorem 2.2 in [Tsy09] with inequalities (A.20) and (A.21) to obtain that

$$\inf_{\hat{\boldsymbol{\nu}}} \sup_{\nu_0 \in \mathcal{F}(\mathbb{R}, A)} \sup_{\boldsymbol{\nu} \in \mathcal{D}(\mathbb{R}, \nu_0, \sigma^2)} \mathbb{E}\left[n^{1/2} d_W(\hat{\boldsymbol{\nu}}, \nu_0)\right] \geq C p_{e,1} \geq C \max\left(\frac{1}{4}\exp(-\alpha), \frac{1 - \sqrt{\alpha/2}}{2}\right),$$

with $\alpha = 2C^2 \max(A^{-1}, \sigma^{-2})$. Therefore, taking $C = \min(A^{1/2}, \sigma)$ completes the proof of Theorem 3.4.

A.6 Measurability of J_2

Let μ be an absolutely continuous measure on $\mathbb{R}$ and let $I \subset \mathbb{R}$ be a (possibly unbounded) interval. We denote by $L^2_\mu(I)$ the space of μ-square-integrable functions from I to $\mathbb{R}$ and by $H_\mu(I)$ the Sobolev space of functions u in $L^2_\mu(I)$ having μ-square-integrable weak derivative Du. Finally, for $u \in L^2_\mu(I)$, we define

$$D^h u(y) := \frac{1}{h}(u(y+h) - u(y)), \tag{A.22}$$

for $y \in I, h \in \mathbb{R}$, such that $y + h \in I$.

Lemma A.2. *$H_\mu((0,1))$ is a Borel subset of $L^2_\mu((0,1))$.*

Proof. For $k \in \mathbb{N}$, $\epsilon \in (0, \frac{1}{2})$ and $|h| < \epsilon$, let us define

$$B_{k,\epsilon,h} := \{u \in L^2_\mu((0,1)) : \|D^h u(y))\|_{L^2_\mu([\epsilon,1-\epsilon])} \leq k\}.$$

It is easy to see that $B_{k,\epsilon,h}$ is closed in $L^2_\mu((0,1))$. Moreover, from Lemmas 9.48 and 9.49 in [RR93],

$$H_\mu((0,1)) = \bigcup_{k\in\mathbb{N}} \bigcap_{\epsilon<\frac{1}{2}} \bigcap_{|h|<\epsilon} B_{k,\epsilon,h}. \tag{A.23}$$

Therefore $H_\mu((0,1))$ is a countable union of closed sets in $L^2_\mu((0,1))$, which concludes the proof. □

Lemma A.3. *Let $u \in H_\mu(\mathbb{R})$. Then $\lim_{h\to 0} D^h u = Du$ in the $L^2_\mu(\mathbb{R})$ convergence.*

Proof. See the first exercise of Chapter 7 in [RR93]. □

Lemma A.4. *The operator $K_2 : H_\mu(I) \to \mathbb{R}_+$, defined by $K_2(u) := \|Du\|_{L^2_\mu(I)}$, is measurable with respect to the Borel σ-algebra in $L^2_\mu(I)$.*

Proof. Let us first show the result for $I = \mathbb{R}$. To that end we define $K_2^h : H_\mu(\mathbb{R}) \to \mathbb{R}_+$ as $K_2^h(u) = \|D^h u\|_{L^2_\mu(\mathbb{R})}$. It is clear that K_2^h is continuous and, moreover, from Lemma A.3,

$$\lim_{h\to 0} K_2^h(u) = K_2(u), \tag{A.24}$$

for all $u \in H_\mu(\mathbb{R})$, and so K_2 is measurable.

Now, for proving the general case note that $K_2(u) = \|Du\|_{L^2_\mu(I)} = \|DTu\|_{L^2_\mu(\mathbb{R})}$, where $T : L^2_\mu(I) \to L^2_\mu(\mathbb{R})$ maps u to its extension by 0 outside I. As T is an isometry (hence measurable) we obtain the result. □

Proposition A.1. *$J_2 : W_2^{ac}(\Omega) \to \mathbb{R}_+ \cup \{\infty\}$ is measurable.*

Proof. Let $\nu \in W_2^{ac}(\Omega)$. Observe that (3.11) is equivalent to

$$J_2(\nu) = \int_{\{x\,:\,0<F_\nu(x)<1\}} \frac{F_\nu(x)(1-F_\nu(x))}{f_\nu(x)} dx. \tag{A.25}$$

From Corollary A.22 in [BL17], if F_ν^- is absolutely continuous then

$$J_2(\nu) = \int_0^1 y(1-y)(DF_\nu^-(y))^2 dy. \tag{A.26}$$

Also, from Proposition A.17 in [BL17], F_ν^- is absolutely continuous if and only if f_ν is almost everywhere positive on the interval $\{x\,:\,0<F_\nu(x)<1\}$. Therefore, from (A.25) and (A.26),

$$J_2(\nu) = \begin{cases} \|DF_\nu^-\|_{L^2_\mu((0,1))} = K_2(F_\nu^-), & \text{if } F_\nu^- \in H_\mu((0,1)) \\ \infty, & \text{if not.} \end{cases} \tag{A.27}$$

where K_2 is defined in Lemma A.4 and μ is the measure having density $f_\mu(y) = y(1-y)$, $y \in (0,1)$. Finally, recall from the proof of Proposition 2.1, that the map $\nu \in W_2(\Omega) \to F_\nu^- \in L^2(0,1)$ is measurable. This measurability can be easily extended from $L^2(0,1)$ to $L^2_\mu(0,1)$, since f_μ is bounded. Then, the result follows from Lemma A.2, Lemma A.4 and (A.27). □

References

[AC11] M. Agueh and G. Carlier. Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[BGKL17] J. Bigot, R. Gouet, T. Klein, and A. Lopez. Geodesic PCA in the Wasserstein space by Convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.*, 53(1):1–26, 2017.

[BIAS03] B. M. Bolstad, R. A. Irizarry, M. Astrand, 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.

[BK17] J. Bigot and T. Klein. Characterization of barycenters in the Wasserstein space by averaging optimal transport maps. *ESAIM: Probability & Statistics*, To be published, 2017.

[BL17] S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics and Kantorovich transport distances.* Memoirs of the American Mathematical Society, 2017. Available at https://perso.math.univ-toulouse.fr/ledoux/files/2016/12/MEMO.pdf.

[BLGL15] E. Boissard, T. Le Gouic, and J.-M. Loubes. Distribution's template estimate with Wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[dBGU05] E. del Barrio, E. Giné, and F. Utzet. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189, 2005.

[Del11] P. Delicado. Dimensionality reduction when data are density functions. *Comput. Statist. Data Anal.*, 55(1):401–420, 2011.

[Fré48] 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.

[KU01] A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, 96(454):519–542, 2001.

[LH10] Y. Li and T. Hsing. Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data. *Annals of Statistics*, 38(6):3321–3351, 2010.

[PM16] K. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a Hilbert space. *Annals of Statistics*, 44(1):183–218, 2016.

[PZ16] V.M. Panaretos and Y. Zemel. Amplitude and phase variation of point processes. *Annals of Statistics*, 44(2):771–812, 2016.

[RL01] J.O. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society (B)*, 63:243–259, 2001.

[RR93] M. Renardy and R. C. Rogers. *An Introduction to Partial Differential Equations*, volume 13 of *Texts in Applied Mathematics*. Springer Verlag, 1993.

[Tsy09] A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[Vil03] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[WG97] K. Wang and T. Gasser. Alignment of curves by dynamic time warping. *Annals of Statistics*, 25(3):1251–1276, 1997.

[WS11] W. Wu and A. Srivastava. An information-geometric framework for statistical inferences in the neural spike train space. *Journal of Computational Neuroscience*, 31(3):725–748, 2011.

[ZM11] Z. Zhang and H.-G. Müller. Functional density synchronization. *Computational Statistics & Data Analysis*, 55(7):2234–2249, 2011.