Penalization of Barycenters in the Wasserstein Space

Jérémie Bigot, Elsa Cazelles, Nicolas Papadakis

▶ **To cite this version:**

Jérémie Bigot, Elsa Cazelles, Nicolas Papadakis. Penalization of Barycenters in the Wasserstein Space. SIAM Journal on Mathematical Analysis, 2019, 51 (3), pp.2261-2285. hal-01564007

HAL Id: hal-01564007
https://hal.science/hal-01564007v1

Submitted on 18 Jul 2017

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.

Penalized Barycenters in the Wasserstein Space *

Jérémie Bigot, Elsa Cazelles & Nicolas Papadakis

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

July 3, 2017

Abstract

A regularization of Wasserstein barycenters for random measures supported on $\mathbb{R}^d$ is introduced via convex penalization. The existence and uniqueness of such barycenters is proved for a large class of penalization functions. A stability result of regularized barycenters in terms of Bregman distance associated to the penalization term is also given. This allows to compare the case of data made of n probability measures with the more realistic setting where we have only access to a dataset of random variables sampled from unknown distributions. We also analyze the convergence of the regularized empirical barycenter of a set of n iid random probability measures towards its population counterpart, and we discuss its rate of convergence. This approach is shown to be appropriate for the statistical analysis of discrete or absolutely continuous random measures. In this setting, we propose efficient algorithms for the computation of penalized Wasserstein barycenters. This approach is finally illustrated with simulated and real data sets.

1 Introduction

This paper is concerned by the statistical analysis of data sets whose elements may be modeled as random probability measures supported on $\mathbb{R}^d$ that are either discrete or absolutely continuous. In the one dimensional case ($d = 1$), examples can be found in neuroscience [WS11], biodemographic and genomics studies [ZM11], economics [KU01], as well as in biomedical imaging [PM15], while examples for the two dimensional case ($d = 2$) arise in spatial statistics for replicated point processes [Ger16].

In this paper, we focus on first-order statistics methods for the purpose of estimating, from such data, a population mean measure or density function.

The notion of averaging depends on the metric that is chosen to compare elements in a given data set. In this work, we consider the Wasserstein distance W_2 associated to the quadratic cost for the comparison of probability measures. Let Ω be a convex subset of $\mathbb{R}^d$ and $\mathcal{P}_2(\Omega)$ be the set of probability measures supported on Ω with finite order second moment. As introduced in [AC11], an empirical Wasserstein barycenter $\bar{\nu}_n$ of set of n probability measures $\nu_1, \ldots, \nu_n$ (not necessarily random) in $\mathcal{P}_2(\Omega)$ is defined as a minimizer of

$$\mu \mapsto \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i), \text{ over } \mu \in \mathcal{P}_2(\Omega). \tag{1.1}$$

*This work has been carried out with financial support from the French State, managed by the French National Research Agency (ANR) in the frame of the GOTMI project (ANR-16-CE33-0010-01).

However, depending on the data at hand, such a barycenter may be irregular (and not even uniquely defined) which is typically the case when the ν_i's are discrete measures. As an illustrative example, we consider a dataset of the locations of reported incidents of crime (with the exception of murders) in Chicago in 2014 which is publicly available[1] and that has been recently studied in [Ger16]. A sample from this dataset is displayed in Figure 1. For each month $1 \leq i \leq 12$ of 2014, we let $\nu_i = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}$ be the discrete measure whose support is the set of locations of reported crimes for the i-th month. As argued in [Ger16], the locations of crimes $\boldsymbol{X}_{i,j}$ may be considered as a sample from random intensity functions whose values change from one day to another as crime opportunities are not uniformly distributed in time. In this setting, we show that a regularization of the Wasserstein barycenter of the ν_i's (as defined below) is a meaningful way to obtain a mean distribution of crimes locations which is absolutely continuous on the area of the city.

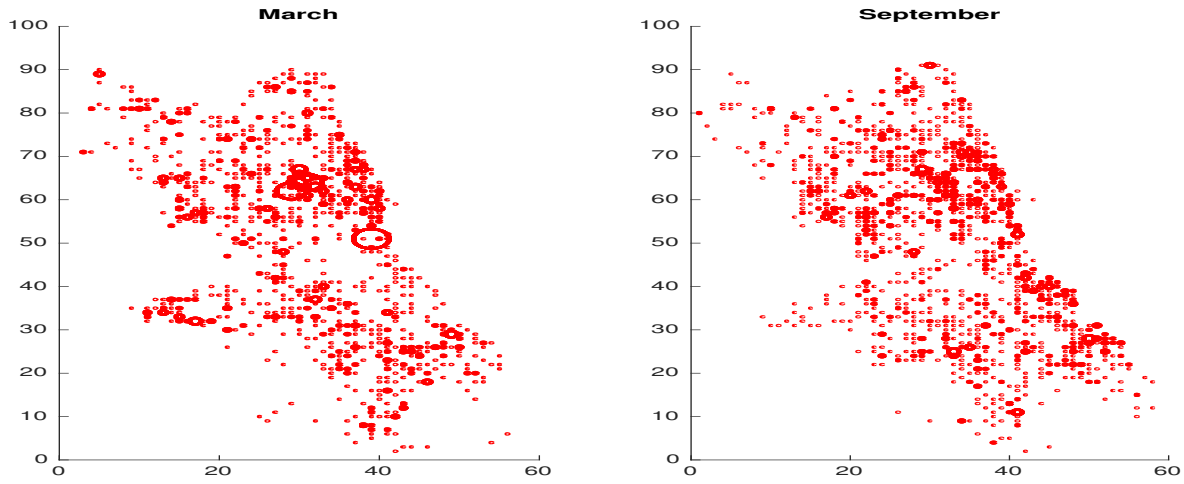


Figure 1: Spatial locations of reported incidents of crime in Chicago for 2 months of 2014. In order to protect the privacy of crime victims, addresses are shown at the block level only and specific locations are not identified.

Definition 1.1. Let $\mathbb{P}_n^\nu = \frac{1}{n} \sum_{i=1}^n \delta_{\nu_i}$ where δ_{ν_i} is the dirac distribution at ν_i. A regularized empirical barycenter $\mu_{\mathbb{P}_n^\nu}^\gamma$ of the discrete measure $\mathbb{P}_n^\nu$ on $\mathcal{P}_2(\Omega)$ is a minimizer of

$$\mu \mapsto \frac{1}{n} \sum_{i=1}^n W_2^2(\mu, \nu_i) + \gamma E(\mu) \text{ over } \mu \in \mathcal{P}_2(\Omega), \tag{1.2}$$

where $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ is a smooth convex penalty function, and $\gamma > 0$ is a regularization parameter.

1.1 Related results in the literature

Statistical inference using optimal transport For $d = 1$, tools from optimal are used in [PZ16] for the registration of multiple point processes which model repeated observations organized in samples from independent subjects or experimental units. In [PZ16], a consistent

[1] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

estimator of the population Wasserstein barycenter of multiple point processes is proposed. It is based on a smoothed empirical Wasserstein barycenter obtained by a preliminary kernel smoothing step of the observed point processes that is followed by quantile averaging. Therefore, the way of constructing a smoothed Wasserstein barycenter in [PZ16] differs from the approach followed in this paper where regularization of the empirical Wasserstein barycenter via a penalty function is considered.

The penalized problem (1.2) is motivated by the nonparametric method introduced in [BFS12] for the classical density estimation problem from discrete samples based on a variational regularization approach to optimal transport with the Wasserstein distance as a data fidelity term. However, the adaptation of this work for the regularization of Wasserstein barycenter has not been considered so far.

Generalized notions of Wasserstein barycenters A detailed characterization of empirical Wasserstein barycenters in terms of existence, uniqueness and regularity is given in [AC11]. There exists also a link between Wasserstein barycenters and the multi-marginal problem in optimal transport as studied in [AC11] and [Pas13]. Recently, the notion of Wasserstein barycenter has been generalized in [LGL16] for random probability measures supported on a locally compact geodesic space. The main contributions in [LGL16] are the proofs of existence, uniqueness and consistency of such barycenters. The case of probability measures supported on a Riemannian manifolds is also studied in [KP14]. Trimmed barycenters in the Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ have been introduced in [ÁEdBCAM15] for the purpose of combining informations from different experimental units in a parallelized or distributed estimation setting.

For applications in image processing, a fast approximation of the Wasserstein distance W_2 between probability measures supported on $\mathbb{R}^d$ with $d \geq 2$ has also been proposed in [BRPP15, PFR12] using a sliced framework based one dimensional Wasserstein distances along radial projections of the input measure.

However, in all these papers, incorporating regularization (through penalization) into the computation of Wasserstein barycenters has not been considered, which is of interest when the data are irregular probability measures.

Regularization of the transport map Alternatively, regularized barycenters may be obtained by adding a convex regularization on optimal transport plans when computing the Wasserstein distance between probability measures. This approach leads to the notion of regularized transportation problems [BCC$^+$15, CP16, DPR16, FPPA14], and it has recently gained popularity in the literature on image processing and machine learning. Recent contributions include the fast approach in [CD14] to compute smooth Wasserstein barycenters of discrete measures via entropic regularization of the transport plan, and the so-called Sinkhorn's algorithm. In these works, such a regularization is motivated by the need to accelerate the computation of the Wasserstein distance between probability measures supported on $\mathbb{R}^d$ with $d \geq 2$.

1.2 Contributions and structure of the paper

The presentation of the main results in this paper is organized as follows.

- In Section 2, we introduce various definitions and notation, and we prove a key result

called subgradient's inequality on which a large part of the developments in the paper lean.

- In Section 3, we analyze the existence, uniqueness and stability of regularized Wasserstein barycenters that are solutions of (1.2) for various of penalty functions E and any regularization parameter $\gamma > 0$.

- In Section 4, for the Bregman distance associated to the penalization term, we derive convergence properties of empirical regularized barycenters toward their population counterpart in the asymptotic setting where n tends to infinity and $\gamma = \gamma_n$ is let going to zero. In this context, we demonstrate that the bias term (as classically referred to in nonparametric statistics) converges to zero when $\gamma \to 0$ in $\mathbb{R}^d$. We also show (with additional regularity assumptions for $d \geq 2$) that the variance term converges to 0 when $\lim_{n\to\infty} \gamma_n^2 n = +\infty$. We mainly focus on penalization functions that enforce the Wasserstein barycenter to be an absolutely continuous (a.c.) measure with a smooth probability density function (pdf). In this case, it is natural to use the Bregman distance to asses the quality of estimation of the pdf of the population barycenter.

- To illustrate the benefits of regularized barycenters for data analysis, we propose to use in Section 5 efficient minimization algorithms for the computation of regularized barycenters as well as a selection strategy for the parameter γ. This approach is finally applied to the statistical analysis of simulated and real data sets in $\mathcal{P}_2(\mathbb{R})$ and $\mathcal{P}_2(\mathbb{R}^2)$.

Finally, a brief overview of the concepts of Bregman divergence and subgradient are gathered in the Appendix A, the proofs in a technical Appendix B and Appendix C contains algorithmic details.

2 Definitions, notation and first results

2.1 Wasserstein distance and Kantorovich's duality

For Ω a convex subset of $\mathbb{R}^d$, we denote by $\mathcal{M}(\Omega)$ the space of bounded Radon measures on Ω. We recall that $\mathcal{P}_2(\Omega)$ is the set of probability measures over $(\Omega, \mathcal{B}(\Omega))$ with finite second order moment, where $\mathcal{B}(\Omega)$ is the σ-algebra of Borel subsets of Ω. In particular, $\mathcal{P}_2(\Omega) \subset \mathcal{M}(\Omega)$.

Definition 2.1. The Wasserstein distance $W_2(\mu, \nu)$ is defined for $\mu, \nu \in \mathcal{P}_2(\Omega)$ by

$$W_2^2(\mu, \nu) = \inf_{\pi} \int_\Omega \int_\Omega |x - y|^2 d\pi(x, y) \tag{2.1}$$

where the infimum is taken over all probability measures π on the product space $\Omega \times \Omega$ with respective marginals μ and ν.

The well known Kantorovich's duality theorem leads to another formulation of the Wasserstein distance.

Theorem 2.2 (Kantorovich's duality)**.** *For any $\mu, \nu \in \mathcal{P}_2(\Omega)$, one has that*

$$W_2^2(\mu, \nu) = \sup_{(\phi,\psi)\in C_W} \int_\Omega \phi(x) d\mu(x) + \int_\Omega \psi(y) d\nu(y) \tag{2.2}$$

where C_W is the set of all measurable functions $(\phi, \psi) \in \mathbb{L}_1(\mu) \times \mathbb{L}_1(\nu)$ satisfying

$$\phi(x) + \psi(y) \leq |x - y|^2, \tag{2.3}$$

for μ-almost every $x \in \Omega$ and ν-almost every $y \in \Omega$. A couple $(\phi, \psi) \in C_W$ that attains the supremum is called an optimal couple for (μ, ν).

For a detailed presentation of the Wasserstein distance and Kantorovich's duality, we refer to [Vil03, Vil08]. For $\mu, \nu \in \mathcal{P}_2(\Omega)$, we denote by $\pi^{\mu,\nu}$ an optimal transport plan, that is a solution of (2.1) satisfying $W_2^2(\mu, \nu) = \iint |x - y|^2 d\pi^{\mu,\nu}(x, y)$. Likewise a pair $(\phi^{\mu,\nu}, \psi^{\mu,\nu}) \in \mathbb{L}_1(d\mu) \times \mathbb{L}_1(d\nu)$ achieving the supremum in (2.2) (under the constraint $\phi^{\mu,\nu}(x) + \psi^{\mu,\nu}(y) \leq |x - y|^2$) stands for the optimal couple in the Kantorovich duality formulation of the Wasserstein distance between μ and ν.

For the sake of completeness, we also introduce the functional space $Y := \{g \in \mathcal{C}(\Omega) : x \mapsto g(x)/(1 + |x|^2) \text{ is bounded}\}$ endowed with the norm $\|g\|_Y = \sup_{x \in \Omega} |g(x)|/(1 + |x|^2)$ where $\mathcal{C}(\Omega)$ is the space of continuous functions from Ω to $\mathbb{R}$. We finally denote as Z the closed subspace of Y given by $Z = \left\{g \in \mathcal{C}(\Omega) : \lim_{|x| \to \infty} g(x)/(1 + |x|^2) = 0\right\}$. The space $\mathcal{M}(\Omega)$ of bounded Radon measures is identified with the dual of $\mathcal{C}_0(\Omega)$ (space of continuous functions that vanish at infinity). Finally, we denote by $\mathbb{L}_1(\mu)$ the set of integrable functions $g : \Omega \to \mathbb{R}$ with respect to the measure μ.

2.2 Regularized barycenters of a random measure

A probability measure $\boldsymbol{\nu}$ in $\mathcal{P}_2(\Omega)$ is said to be random if it is sampled from a distribution $\mathbb{P}$ on $(\mathcal{P}_2(\Omega), \mathcal{B}(\mathcal{P}_2(\Omega)))$, where $\mathcal{B}(\mathcal{P}_2(\Omega))$ is the Borel σ-algebra generated by the topology induced by the distance W_2. Throughout the paper, we use bold symbols $\boldsymbol{\nu}, \boldsymbol{X}, \boldsymbol{f}, \ldots$ to denote random objects. Then, we introduce a Wasserstein distance between distributions of random measures (see [LGL16] and [ÁEdBCAM15] for similar concepts), and the notion of Wasserstein barycenter of a random probability measure $\boldsymbol{\nu}$.

Definition 2.3. Let $W_2(\mathcal{P}_2(\Omega))$ be the space of distributions $\mathbb{P}$ on $\mathcal{P}_2(\Omega)$ (endowed with the Wasserstein distance W_2) such that for some (thus for every) $\mu \in \mathcal{P}_2(\Omega)$

$$\mathcal{W}_2^2(\delta_\mu, \mathbb{P}) := \mathbb{E}_{\mathbb{P}}(W_2^2(\mu, \boldsymbol{\nu})) = \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) < +\infty$$

where $\boldsymbol{\nu} \in \mathcal{P}_2(\Omega)$ is a random measure with distribution $\mathbb{P}$. The Wasserstein barycenter of a random probability measure with distribution $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ is defined as a minimizer of

$$\mu \in \mathcal{P}_2(\Omega) \mapsto \mathcal{W}_2^2(\delta_\mu, \mathbb{P}) = \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) \text{ over } \mu \in \mathcal{P}_2(\Omega) \tag{2.4}$$

where δ_μ denotes the Dirac measure at the point μ.

Thanks to the results in [LGL16], there exists a minimizer of (2.4), and thus the notion of Wasserstein barycenter of a random probability measure is well defined. Throughout the paper, the following assumptions are made on the regularizing function E.

Assumption 2.1. *Let $E : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ be a proper, lower semicontinuous and differentiable function that is strictly convex on its domain*

$$\mathcal{D}(E) = \{\mu \in \mathcal{P}_2(\Omega) \text{ such that } E(\mu) < +\infty\}. \tag{2.5}$$

Regularized barycenters of a random measure are then defined as follows.

Definition 2.4. For a distribution $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ and a regularization parameter $\gamma \geq 0$, the functional $J^\gamma_\mathbb{P} : \mathcal{P}_2(\Omega) \to \mathbb{R}_+$ is defined as

$$J^\gamma_\mathbb{P}(\mu) = \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) d\mathbb{P}(\nu) + \gamma E(\mu), \ \mu \in \mathcal{P}_2(\Omega). \tag{2.6}$$

If it exists, a minimizer $\mu^\gamma_\mathbb{P}$ of $J^\gamma_\mathbb{P}$ is called a regularized Wasserstein barycenter of the random measure $\boldsymbol{\nu}$ with distribution $\mathbb{P}$. In particular, if $\mathbb{P}$ is the discrete (or empirical) measure defined by $\mathbb{P} = \mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\nu_i}$ where $\nu_i \in \mathcal{P}_2(\Omega)$ for $i = 1, \dots, n$, then $J^\gamma_\mathbb{P}$ becomes

$$J^\gamma_{\mathbb{P}_n}(\mu) = \frac{1}{n}\sum_{i=1}^n W_2^2(\mu, \nu_i) + \gamma E(\mu). \tag{2.7}$$

Note that $J^\gamma_\mathbb{P}$ is strictly convex on $\mathcal{D}(E)$ by Assumption 2.1.

A typical example of a regularizing function satisfying Assumption 2.1 is the negative entropy [BFS12] (see e.g. Lemma 1.4.3 in [DE97]) defined as

$$E(\mu) = \begin{cases} \int_{\mathbb{R}^d} f(x)\log(f(x))dx, & \text{if } \mu \text{ admits a density } f \text{ with respect to} \\ & \text{the Lebesgue measure on } \Omega, \\ +\infty & \text{otherwise.} \end{cases}$$

which enforces the barycenter to be a.c. with respect to the Lebesgue measure on $\mathbb{R}^d$.

2.3 Subgradient's inequality

In order to analyze the stability of the minimizers of $J^\gamma_\mathbb{P}$ with respect to the distribution $\mathbb{P}$, the notion of Bregman divergence together with the concept of subgradient will be needed. An overview of these tools is presented in an C.

In our case, since E is supposed differentiable, for $\mu \in \mathcal{M}(\Omega)$ we have that $\partial E(\mu) = \nabla E(\mu)$. Then for $\mu, \nu \in \mathcal{M}(\Omega)$, the (symmetric) Bregman distance is defined by

$$d_E(\mu, \nu) = \langle \nabla E(\mu) - \nabla E(\nu), \mu - \nu \rangle \tag{2.8}$$

where the linear form is understood as $\langle f, \mu \rangle = \int_\Omega f(x) d\mu(x)$, for $\mu \in \mathcal{M}(\Omega)$ and $f \in \mathcal{C}_b(\Omega)$ the space of continuous bounded function from Ω to $\mathbb{R}$.

Theorem 2.5 (Subgradient's inequality)**.** *Let ν be a probability measure in $\mathcal{P}_2(\Omega)$, and define the functional*

$$J : \mu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu, \nu) + \gamma E(\mu)$$

where $E : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is a proper, differentiable and convex function, and $\gamma \geq 0$. If $\mu \in \mathcal{P}_2(\Omega)$ minimizes J, then there exists $\phi^{\mu,\nu} \in \mathbb{L}_1(\mu)$ and $\psi \in \mathbb{L}_1(\nu)$ verifying $\phi^{\mu,\nu}(x) + \psi(y) \leq |x - y|^2$ for all x, y in Ω such that $(\phi^{\mu,\nu}, \psi)$ is an optimal couple of the Kantorovich's dual problem associated to μ, ν (Theorem 2.2). Moreover, for all $\eta \in \mathcal{P}_2(\Omega)$,

$$\gamma \ \langle \nabla E(\mu), \mu - \eta \rangle \leq -\int \phi^{\mu,\nu} d(\mu - \eta). \tag{2.9}$$

The proof of Theorem 2.5 is based on the two following lemmas whose proof can be found in the Appendix B.1.

Lemma 2.6. *The two following assertions are equivalent:*

1. $\mu \in \mathcal{P}_2(\Omega)$ *minimizes* J *over* $\mathcal{P}_2(\Omega)$,
2. *there exists a subgradient* $\phi \in \partial J(\mu)$ *such that* $\langle \phi, \eta - \mu \rangle \geq 0$ *for all* $\eta \in \mathcal{P}_2(\Omega)$.

Lemma 2.7. *Let* $\mu \in \mathcal{P}_2(\Omega)$ *and* $\phi \in \mathbb{L}_1(\mu)$, *then*

$$\phi \in \partial_1 W_2^2(\mu, \nu) \Leftrightarrow \exists\ \psi \in \mathbb{L}_1(\nu)\ /\ \phi(x) + \psi(y) \leq |x - y|^2$$

and $W_2^2(\mu, \nu) = \int \phi d\mu + \int \psi d\nu$ *where* $\partial_1 W_2^2(\mu, \nu)$ *denote the subdifferential of the function* $W_2^2(\cdot, \nu)$ *at* μ.

Proof of Theorem 2.5. Let $\mu \in \mathcal{P}_2(\Omega)$ be a minimizer of J. From Lemma 2.6, we know that there exists ϕ a subgradient of J in μ such that $\langle \phi, \eta - \mu \rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$. Since $\zeta \mapsto E(\zeta)$ is convex differentiable, $\zeta \mapsto W_2^2(\zeta, \nu)$ is a continuous convex function and μ minimizes J, we have by the subdifferential of the sum (Theorem 4.10 in [Cla13]) that $\partial J(\mu) = \partial_1 W_2^2(\mu, \nu) + \gamma\ \nabla E(\mu)$. This implies that all $\phi \in \partial J(\mu)$ is written $\phi = \phi_1 + \phi_2$ with $\phi_1 = \phi^{\mu,\nu}$ optimal for the couple (μ, ν) (by Lemma 2.7) and $\phi_2 = \gamma \nabla E(\mu)$. Finally, we have that $\langle \phi^{\mu,\nu} + \gamma \nabla E(\mu), \eta - \mu \rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$ that is $\gamma\ \langle \nabla E(\mu), \mu - \eta \rangle \leq -\int \phi^{\mu,\nu} d(\mu - \eta), \forall \eta \in \mathcal{P}_2(\Omega)$. □

3 Existence, uniqueness and stability of regularized barycenters

In this section, we present some properties of the minimizers of the functional $J_{\mathbb{P}}^{\gamma}$ (see Definition 2.4) in terms of existence and stability.

3.1 Existence and uniqueness

In a first part, we state that the minimization problem (2.6) admits a unique minimum in the particular setting where $\mathbb{P}$ is a discrete distribution on $\mathcal{P}_2(\Omega)$ that is we study the problem

$$\min_{\mu \in \mathcal{P}_2(\Omega)} J_{\mathbb{P}_n}^{\gamma}(\mu) = \int W_2^2(\mu, \nu) d\mathbb{P}_n(\nu) + \gamma E(\mu) = \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i) + \gamma E(\mu) \tag{3.1}$$

where $\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{\nu_i} \in W_2(\mathcal{P}_2(\Omega))$ with $\nu_1, \ldots, \nu_n$ measures in $\mathcal{P}_2(\Omega)$. In a second part, we prove the existence and uniqueness of (2.6) in a general case.

Theorem 3.1. *Suppose that Assumption 2.1 holds and that* $\gamma > 0$. *Then, the functional* $J_{\mathbb{P}_n}^{\gamma}$ *defined by* (3.1) *admits a unique minimizer on* $\mathcal{P}_2(\Omega)$ *which belongs to the domain* $\mathcal{D}(E)$ *of the regularizing function* E, *as defined in* (2.5).

The proof of Theorem 3.1 is given in the Appendix B.2. Thanks to this result, one may impose the regularized Wasserstein barycenter $\mu_{\mathbb{P}_n}^{\gamma}$ to be a.c. with respect to Lebesgue measure on Ω by choosing the negative entropy $E(\mu) = \int_{\Omega} \log(d\mu(x))\, d\mu(x)$ for the regularization function E. For this choice, (3.1) becomes a problem of minimization over a set of pdf with

entropy regularization. Examples of the use of the negative entropy as a regularization term are given in Section 5 on numerical experiments.

From Theorem 3.1, it is possible to prove the general case (whose proof is also given in Appendix B.2).

Theorem 3.2. *Suppose that Assumption 2.1 holds and that $\gamma > 0$. Then, the functional $J^{\gamma}_{\mathbb{P}}$ defined by (2.6) admits a unique minimizer.*

3.2 Stability

We now study the stability of the solution with respect to the symmetric Bregman distance d_E (2.8) associated to the differentiable function E. Let $\nu_1, \ldots, \nu_n \in \mathcal{P}_2(\Omega)$ and $\eta_1, \ldots, \eta_n \in \mathcal{P}_2(\Omega)$. We denote by $\mathbb{P}^{\nu}_n$ (resp. $\mathbb{P}^{\eta}_n$) the discrete measure $\frac{1}{n}\sum_{i=1}^n \delta_{\nu_i}$ (resp. $\frac{1}{n}\sum_{i=1}^n \delta_{\eta_i}$) in $W_2(\mathcal{P}_2(\Omega))$.

Theorem 3.3. *Suppose that Ω is bounded. Let $\mu_\nu, \mu_\eta \in \mathcal{P}_2(\Omega)$ with μ_ν minimizing $J^{\gamma}_{\mathbb{P}^{\nu}_n}$ and μ_η minimizing $J^{\gamma}_{\mathbb{P}^{\eta}_n}$ defined by (3.1). Then, the symmetric Bregman distance associated to E can be upper bounded as follows*

$$d_E(\mu_\nu, \mu_\eta) \leq \frac{2}{\gamma n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n W_2(\nu_i, \eta_{\sigma(i)}), \tag{3.2}$$

where $\mathcal{S}_n$ is the permutation group of the set $\{1, \ldots, n\}$.

The proof of Theorem 3.3 is given in Appendix B.3. To better interpret the upper bound (3.2), we need the notion of Kantorovich transport distance $\mathcal{T}_{W_2}$ on the metric space $(\mathcal{P}_2(\Omega), W_2)$, see [Vil03]. For $\mathbb{P}, \mathbb{Q} \in W_2(\mathcal{P}_2(\Omega))$ endowed with the Wasserstein distance W_2, we have that

$$\mathcal{T}_{W_2}(\mathbb{P}, \mathbb{Q}) := \inf_{\Pi} \int_{\mathcal{P}_2(\Omega)\times\mathcal{P}_2(\Omega)} W_2(\mu, \nu) d\Pi(\mu, \nu),$$

where the minimum is taken over all probability measures Π on the product space $\mathcal{P}_2(\Omega) \times \mathcal{P}_2(\Omega)$ with marginals $\mathbb{P}$ and $\mathbb{Q}$. Since $\mathbb{P}^{\nu}_n$ and $\mathbb{P}^{\eta}_n$ are discrete probability measures supported on $\mathcal{P}_2(\Omega)$, it follows that the upper bound (3.2) in Theorem 3.3 can also be written as (by Birkhoff's theorem for bi-stochastic matrices, see e.g. [Vil03])

$$d_E(\mu_\nu, \mu_\eta) \leq \frac{2}{\gamma} \mathcal{T}_{W_2}(\mathbb{P}^{\nu}_n, \mathbb{P}^{\eta}_n).$$

The above upper bound means that the Bregman distance between the regularized Wasserstein barycenters μ_ν and μ_η is controlled by the Kantorovich transport distance between the distributions $\mathbb{P}^{\nu}_n$ and $\mathbb{P}^{\eta}_n$.

3.3 Discussion

Theorem 3.3 is of particular interest in the setting where the ν_i's and η_i's are discrete probability measures on $\mathbb{R}^d$. If we assume that $\nu_i = \frac{1}{p}\sum_{j=1}^p \delta_{\boldsymbol{X}_{i,j}}$ and $\eta_i = \frac{1}{p}\sum_{j=1}^p \delta_{\boldsymbol{Y}_{i,j}}$ where $(\boldsymbol{X}_{i,j})_{1\leq i\leq n; 1\leq j\leq p}$ and $(\boldsymbol{Y}_{i,j})_{1\leq i\leq n; 1\leq j\leq p}$ are (possibly random) vectors in $\mathbb{R}^d$, then by (3.2),

$$d_E(\mu_\nu, \mu_\eta) \leq \frac{2}{\gamma n} \inf_{\sigma \in \mathcal{S}_n} \sum_{i=1}^n \left(\inf_{\lambda \in \mathcal{S}_p} \left\{ \frac{1}{p} \sum_{j=1}^p |\boldsymbol{X}_{i,j} - \boldsymbol{Y}_{\sigma(i),\lambda(j)}|^2 \right\} \right)^{1/2}.$$

Theorem 3.3 is also useful to compare the computation of regularized Wasserstein barycenter between the case of data made of n a.c. probability measures $\nu_1, \dots, \nu_n$, with the more realistic setting where we have only access to random variables $\boldsymbol{X} = (\boldsymbol{X}_{i,j})_{1 \leq i \leq n;\, 1 \leq j \leq p_i}$ organized in the form of n experimental units, such that $\boldsymbol{X}_{i,1}, \dots, \boldsymbol{X}_{i,p_i}$ are iid observations in $\mathbb{R}^d$ sampled from the measure ν_i for each $1 \leq i \leq n$. If we denote by $\boldsymbol{\nu}_{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}$ the usual empirical measure associated to ν_i, it follows from inequality (3.2) that

$$\mathbb{E}\left(d_E^2\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \boldsymbol{\mu}^\gamma_{\boldsymbol{X}}\right)\right) \leq \frac{4}{\gamma^2 n} \sum_{i=1}^n \mathbb{E}\left(W_2^2(\nu_i, \boldsymbol{\nu}_{p_i})\right),$$

where $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}}$ is the random density satisfying

$$\boldsymbol{\mu}^\gamma_{\boldsymbol{X}} = \operatorname*{argmin}_{\mu \in \mathcal{P}_2(\Omega)} \frac{1}{n} \sum_{i=1}^n W_2^2\left(\mu, \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{i,j}}\right) + \gamma E(\mu).$$

This result allows to discuss the rate of convergence (for the symmetric squared Bregman distance) of $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}}$ to $\mu^\gamma_{\mathbb{P}^\nu_n}$ as a function of the rate of convergence (for the squared Wasserstein distance) of the empirical measure $\boldsymbol{\nu}_{p_i}$ to ν_i for each $1 \leq i \leq n$ (in the asymptotic setting where $p = \min_{1 \leq i \leq n} p_i$ is let going to infinity).

As an illustrative example, in the one-dimensional case (that is $d = 1$), one may use Theorem 5.1 in [BL14], to obtain that

$$\mathbb{E}\left(W_2^2(\nu_i, \boldsymbol{\nu}_{p_i})\right) \leq \frac{2}{p_i + 1} J_2(\nu_i), \text{ with } J_2(\nu_i) = \int_\Omega \frac{F_i(x)(1 - F_i(x))}{f_i(x)} dx,$$

where f_i is the pdf of ν_i, and F_i denotes its cumulative distribution function. Therefore, provided that $J_2(\nu_i)$ is finite for each $1 \leq i \leq n$, one obtains the following rate of convergence of $\boldsymbol{\mu}^\gamma_{\boldsymbol{X}}$ to $\mu^\gamma_{\mathbb{P}^\nu_n}$ (for $d = 1$)

$$\mathbb{E}\left(d_E^2\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \boldsymbol{\mu}^\gamma_{\boldsymbol{X}}\right)\right) \leq \frac{8}{\gamma^2 n} \sum_{i=1}^n \frac{J_2(\nu_i)}{p_i + 1} \leq \frac{8}{\gamma^2} \left(\frac{1}{n} \sum_{i=1}^n J_2(\nu_i)\right) p^{-1}. \tag{3.3}$$

When the measures $\nu_1, \dots, \nu_n$ are supported on $\mathbb{R}^d$ with $d \geq 2$, we refer to [FG15] for further results on the rate of convergence of an empirical measure in Wasserstein distance that may be used to derive rates of convergence for $d_E\left(\mu^\gamma_{\mathbb{P}^\nu_n}, \boldsymbol{\mu}^\gamma_{\boldsymbol{X}}\right)$.

4 Convergence properties of regularized empirical barycenters

In this section, when Ω is a compact of $\mathbb{R}^d$, we study the convergence of the regularized Wasserstein barycenter of a set $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n$ of independent random measures sampled from a distribution $\mathbb{P}$ towards a minimizer of $J^0_{\mathbb{P}}$, that is a population Wasserstein barycenter of the probability distribution $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$. To this end, we first introduce and recall some notation.

Definition 4.1. For $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n$ iid random measures in $\mathcal{P}_2(\Omega)$ sampled from a distribution

$\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$, we let $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{\nu}_i}$. Moreover, we use the notation

$$\boldsymbol{\mu}^\gamma_{\mathbb{P}_n} \in \operatorname*{argmin}_{\mu\in\mathcal{P}_2(\Omega)} J^\gamma_{\mathbb{P}_n}(\mu) = \int W_2^2(\mu,\nu)d\mathbb{P}_n(\nu) + \gamma E(\mu) \tag{4.1}$$

$$\mu^\gamma_{\mathbb{P}} \in \operatorname*{argmin}_{\mu\in\mathcal{P}_2(\Omega)} J^\gamma_{\mathbb{P}}(\mu) = \int W_2^2(\mu,\nu)d\mathbb{P}(\nu) + \gamma E(\mu) \tag{4.2}$$

$$\mu^0_{\mathbb{P}} \in \operatorname*{argmin}_{\mu\in\mathcal{P}_2(\Omega)} J^0_{\mathbb{P}}(\mu) = \int W_2^2(\mu,\nu)d\mathbb{P}(\nu), \tag{4.3}$$

that will be respectively referred as to the empirical Wasserstein barycenter (4.1), the regularized population Wasserstein barycenter (4.2) and the population Wasserstein barycenter (4.3).

In what follows, we obtain a rate of convergence in expected squared Bregman distance between $\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}$ and $\mu^\gamma_{\mathbb{P}}$ which depends on n and γ. A general result is first stated in Section 4.1. Complementary results are given in Section 4.2 for $d = 1$ and in Section 4.3 for $d \geq 2$. Then, in Section 4.4, we prove the convergence in Bregman divergence (as $\gamma \to 0$) of the regularized population Wasserstein barycenter $\mu^\gamma_{\mathbb{P}}$ towards $\mu^0_{\mathbb{P}}$.

4.1 Rate of convergence of $\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}$ towards $\mu^\gamma_{\mathbb{P}}$ in symmetrized Bregman distance

To compute a rate of convergence between $\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}$ and $\mu^\gamma_{\mathbb{P}}$, we will need results from the empirical process theory. Thus, we first introduce some notions borrowed from [VDVW96].

Definition 4.2. Let $\mathcal{F} = \{f : U \mapsto \mathbb{R}\}$ be a class of real-valued functions defined on a given set U, endowed with a norm $\|\cdot\|$. An envelope function of $\mathcal{F}$ is any function $u \mapsto F(u)$ such that $|f(u)| \leq F(u)$ for every $u \in U$ and $f \in \mathcal{F}$. The minimal envelope function is $u \mapsto \sup_f |f(u)|$. The covering number $N(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of balls $\{\|g - f\| < \epsilon\}$ of radius ϵ and center g needed to cover the set $\mathcal{F}$. The metric entropy is the logarithm of the covering number. Finally, we define

$$I(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\epsilon\|F\|_{\mathbb{L}_2(Q)}, \mathcal{F}, \|\cdot\|_{\mathbb{L}_2(Q)})}d\epsilon \tag{4.4}$$

where the supremum is taken over all discrete probability measures Q supported on U with $\|F\|_{\mathbb{L}_2(Q)} = \left(\int |F(u)|^2 dQ(u)\right)^{1/2} > 0$.

Theorem 4.3. *If Ω is a compact of $\mathbb{R}^d$, then one has that*

$$\mathbb{E}(d_E^2(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}})) \leq \frac{CI(1,\mathcal{H})\|H\|_{\mathbb{L}_2(\mathbb{P})}}{\gamma^2 n} \tag{4.5}$$

where C is a positive constant, and

$$\mathcal{H} = \{h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu) \in \mathbb{R}; \mu \in \mathcal{P}_2(\Omega)\} \tag{4.6}$$

is a class of functions defined on $\mathcal{P}_2(\Omega)$ with envelope H.

The proof of Theorem 4.3 is given in the Appendix B.4. To complete this result, one needs to prove that $I(1, \mathcal{H}) < \infty$, which depends on the rate of of convergence of the metric entropy towards infinity as ϵ tends to zero.

4.2 The one-dimensional case

For probability measures $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ supported in the real line, we can prove that the right-hand side of (4.5) is finite by using existing results on the notion of bracketing number defined below.

Definition 4.4. Given two real-valued functions l and r, the bracket $[l, r]$ is the set of all functions f with $l \leq f \leq r$. An ϵ-bracket is a bracket $[l, r]$ with $\|l - r\| < \epsilon$. The bracketing number $N_{[]}(\epsilon, \mathcal{F}, \|\cdot\|)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{F}$.

Theorem 4.5. *If Ω is a compact of $\mathbb{R}$, then there exists a finite constant $C > 0$ such that $\mathbb{E}(d_E^2\left(\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}, \mu_{\mathbb{P}}^{\gamma}\right)) \leq \frac{C}{\gamma^2 n}$.*

Proof. In what follows, C denotes a universal constant whose value may change from line to line. We define the envelope function $H : \nu \in \mathcal{P}_2(\Omega) \mapsto \sup_{\mu \in \mathcal{P}_2(\Omega)} \{W_2(\mu, \nu); W_2^2(\mu, \nu)\}$. Since for $h_\mu \in \mathcal{H}$ we have

$$|h_\mu(\nu)| \leq 2 \int |x|^2 d\mu(x) + 2 \int |y|^2 d\nu(y) \leq 4\delta(\Omega) \quad \text{for all } \nu \in \mathcal{P}_2(\Omega)$$

where $\delta(\Omega) = \sup_{x \in \Omega} |x|^2$, then for all $Q \in W_2(\mathcal{P}_2(\Omega))$,

$$\|H\|_{\mathbb{L}_2(Q)} = \left(\int H(\nu)^2 dQ(\nu) \right)^{1/2} \leq \left(16\delta(\Omega)^2 \int dQ(\nu) \right)^{1/2} \leq 4\delta(\Omega).$$

Now, it remains to control the term $I(1, \mathcal{H})$ in the upper bound (4.5). By the triangle reverse inequality, we have

$$\begin{aligned} |h_\mu(\nu) - h_{\mu'}(\nu)| &= |W_2(\nu, \mu) - W_2(\nu, \mu')| \, (W_2(\nu, \mu) + W_2(\nu, \mu')) \\ &\leq W_2(\mu, \mu') \, 2H(\nu). \end{aligned}$$

Then, from Theorem 2.7.11 in [VDVW96], and since Theorem 4 in [KT59] allows us to bound the metric entropy by the bracket entropy, we get

$$\begin{aligned} \log N(\epsilon \|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)}) &\leq \log N_{[]}(\epsilon \|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)}) \\ &\leq \log N(\epsilon, \mathcal{P}_2(\Omega), W_2) \leq \log N_{[]}(\epsilon, \mathcal{P}_2(\Omega), W_2). \end{aligned} \tag{4.7}$$

Also, for $d = 1$, we have

$$W_2(\mu, \mu') = \left(\int_0^1 |F_\mu^-(t) - F_{\mu'}^-(t)|^2 dt \right)^{1/2} = \|F_\mu^- - F_{\mu'}^-\|_{\mathbb{L}_2([0,1])} \tag{4.8}$$

where F_μ^- is the quantile function of the cumulative distribution function F_μ of μ. We denote by $\mathcal{G} = \{F_\mu^-, \mu \in \mathcal{P}_2(\Omega)\}$ the class of quantile functions of probability measures μ in $\mathcal{P}_2(\Omega)$, which are monotonic functions. Moreover, we can observe that $F_\mu^- : [0, 1] \to [F_\mu^-(0), F_\mu^-(1)] \subseteq \Omega$, where Ω is a compact included in $\mathbb{R}$. Hence, $\mathcal{G}$ is uniformly bounded, say by a constant $M > 0$. Finally, by Theorem 2.7.5. of [VDVW96] concerning the bracket entropy of the class of monotonic functions, we obtain that $\log N_{[]}(\epsilon, \mathcal{G}, \mathbb{L}_2[0, 1]) \leq \frac{CM}{\epsilon}$, for some constant $C > 0$. Finally, from relations (4.7) and (4.8), we can deduce that

$$I(1, \mathcal{H}) = \sup_Q \int_0^1 \sqrt{1 + \log N(\epsilon \|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \mathbb{L}_2(Q))} d\epsilon \leq \int_0^1 \sqrt{1 + \frac{CM}{\epsilon}} d\epsilon < \infty.$$

□

Therefore, when $\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_n$ are iid random measures with support included in a compact interval Ω, it follows from Theorem 4.5 that if $\gamma = \gamma_n$ is such that $\lim_{n\to\infty} \gamma_n^2 n = +\infty$ then $\lim_{n\to\infty} \mathbb{E}(d_E^2\left(\boldsymbol{\mu}_{\mathbb{P}_n^\nu}^\gamma, \mu_{\mathbb{P}}^0\right)) = 0$.

4.3 The $\mathbb{R}^d$ case with additional regularization

In the case $d \geq 2$, the class of functions $\mathcal{H}$ defined in (4.6) is too large to control the metric entropy so that $I(1, \mathcal{H})$ is finite. To solve this issue, we impose more smoothness on the regularized Wasserstein barycenter as follows.

We assume that Ω is a smooth and uniformly convex set, and we choose

$$E(\mu) = \begin{cases} \int_{\mathbb{R}^d} f(x)\log(f(x))dx + \|f\|^2_{H^k(\Omega)}, & \text{if } f = \frac{d\mu}{dx} \text{ and } f > \alpha, \\ +\infty & \text{otherwise.} \end{cases} \tag{4.9}$$

where $\|\cdot\|_{H^k(\Omega)}$ denotes the Sobolev norm associated to the $\mathbb{L}^2(\Omega)$ space and $\alpha > 0$ is arbitrarily small. Then, the following result holds.

Theorem 4.6. *Suppose that Ω is a compact and uniformly convex set with a C^1 boundary. Assume that the penalty function E is given by (4.9) for some $\alpha > 0$ and $k > d-1$. Then, there exists a finite contant $C > 0$ such that $\mathbb{E}(d_E^2\left(\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma, \mu_{\mathbb{P}}^\gamma\right)) \leq \frac{C}{\gamma^2 n}$.*

Proof. Supposing that Ω has a C^1 boundary, we have by the Sobolev embedding theorem that $H^k(\Omega)$ is included in the Hölder space $C^{m,\beta}(\bar{\Omega})$ for any integer m and $\beta \in]0,1]$ satisfying $m + \beta = k - d/2$. Hence, the densities of $\boldsymbol{\mu}_{\mathbb{P}_n}^\gamma$ and $\mu_{\mathbb{P}}^\gamma$ given by (4.1) and (4.2) belong to $C^{m,\beta}(\bar{\Omega})$.

Arguing as in the proof of Theorem 4.5, we have $|h_\mu(\nu) - h_{\mu'}(\nu)| \leq W_2(\mu,\mu')\ 2H(\nu)$ and $\|H\|_{\mathbb{L}_2(Q)} < \infty$, where $H(\nu) = \sup_{\mu\in\mathcal{D}(E)} \{W_2(\mu,\nu); W_2^2(\mu,\nu)\}$ where $\mathcal{D}(E)$ is defined by (2.5). Thus, instead of controlling the metric entropy $N(\epsilon\|H\|_{\mathbb{L}_2(Q)}, \mathcal{H}, \|\cdot\|_{\mathbb{L}_2(Q)})$, it is enough to bound the metric entropy $N(\epsilon, \mathcal{D}(E), W_2)$ thanks to Theorem 2.7.11 in [VDVW96].

To this end, since $\mu, \mu' \in \mathcal{D}(E)$ are a.c. measures, one has that

$$W_2(\mu,\mu') \leq \left(\int_\Omega |T(x) - T'(x)|^2 dx\right)^{1/2} \text{ where } T\#\lambda^d = \mu \text{ and } T'\#\lambda^d = \mu',$$

with λ^d denoting the Lebesgue measure on Ω. Thanks to Theorem 3.3 in [DPF14] on the regularity of optimal maps (results initally due to Caffarelli, [Caf92] and [Caf96]), the coordinates of T and T' are $C^{m+1,\beta}(\bar{\Omega})$ functions $\lambda^d - a.e.$. Thus, we can bound $N(\epsilon, \mathcal{D}(E), W_2)$ by the bracket entropy $N_{[]}(\epsilon, C^{m+1,\beta}(\bar{\Omega}), \mathbb{L}_2(\Omega))$ since $|T(x) - T'(x)|^2 = \sum_{j=1}^d |T_j(x_j) - T_j'(x_j)|^2$ where $T_j, T_j' : \Omega \to \mathbb{R}$. Now, by Corollary 2.7.4 in [VDVW96],

$$\log N_{[]}(\epsilon, C^{m+1,\beta}(\bar{\Omega}), \mathbb{L}_2(\Omega)) \leq K\left(\frac{1}{\epsilon}\right)^V$$

for any $V \geq d/(m+1)$. Hence, as soon as $V/2 < 1$ (for which the condition $k > d-1$ is sufficient if $V = d/(m+1)$), the upper bound in (4.5) is finite for $\mathcal{H} = \{h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu) \in \mathbb{R}; \mu \in \mathcal{D}(E)\}$, which yields the result of Theorem 4.6 by finally following the arguments in the proof of Theorem 4.5. □

4.4 Convergence of $\mu_{\mathbb{P}}^{\gamma}$ towards $\mu_{\mathbb{P}}^{0}$ in Bregman divergence

Theorem 4.7. *If Ω is a compact of $\mathbb{R}^d$ and $\nabla E(\mu_{\mathbb{P}}^0)$ is a bounded function on Ω then*

$$\lim_{\gamma \to 0} D_E(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^{0}) = 0,$$

where D_E denotes the Bregman divergence between two measures μ and ν defined as $D_E(\mu, \nu) = E(\mu) - E(\nu) - \langle \nabla E(\nu), \mu - \nu \rangle$.

Proof. By definition (4.2) of $\mu_{\mathbb{P}}^{\gamma}$, we get that

$$\int W_2^2(\mu_{\mathbb{P}}^{\gamma}, \nu) d\mathbb{P}(\nu) - \int W_2^2(\mu_{\mathbb{P}}^{0}, \nu) d\mathbb{P}(\nu) + \gamma(E(\mu_{\mathbb{P}}^{\gamma}) - E(\mu_{\mathbb{P}}^{0})) \leq 0. \tag{4.10}$$

By definition (4.3) of $\mu_{\mathbb{P}}^{0}$, one has $\int W_2^2(\mu_{\mathbb{P}}^{\gamma}, \nu) d\mathbb{P}(\nu) - \int W_2^2(\mu_{\mathbb{P}}^{0}, \nu) d\mathbb{P}(\nu) \geq 0$. Therefore, by definition of the Bregman divergence, inequality (4.10) gives

$$D_E(\mu_{\mathbb{P}}^{\gamma}, \mu_{\mathbb{P}}^{0}) \leq \langle \nabla E(\mu_{\mathbb{P}}^{0}), \mu_{\mathbb{P}}^{0} - \mu_{\mathbb{P}}^{\gamma} \rangle \leq C \sup_{\|\phi\|_{BL} \leq 1} \langle \phi, \mu_{\mathbb{P}}^{0} - \mu_{\mathbb{P}}^{\gamma} \rangle \leq C d_{BL^*}(\mu_{\mathbb{P}}^{0}, \mu_{\mathbb{P}}^{\gamma}),$$

where d_{BL^*} is the bounded Lipschitz distance and $\|\phi\|_{BL} := \|\phi\|_{\infty} + \|\phi\|_{Lip}$. We denote by $\|\cdot\|_{Lip}$ the norm define on the space of all Lipschitz functions on (Ω, d) with $d(x, y) = \mathbb{1}_{x \neq y}$. By hypothesis, $\|\nabla E(\mu_{\mathbb{P}}^{0})\|_{BL}$ is finite. For sequence of probability measures, convergence in distance d_{BL^*} is equivalent to weak convergence (e.g. Section 1.2.1 of [Vil03]). Hence, by Theorem 2.1.(d) in [Bra06], $J_{\mathbb{P}}^{\gamma}$ Γ-converges to $J_{\mathbb{P}}^{0}$. Indeed for every sequence $(\mu_\gamma)_\gamma \subset \mathcal{P}_2(\Omega)$ converging to $\mu \in \mathcal{P}_2(\Omega)$ in bounded Lipchitz distance,

$$J_{\mathbb{P}}^{0}(\mu) \leq \liminf_{\gamma \to 0} J_{\mathbb{P}}^{\gamma}(\mu_\gamma)$$

by lower semicontinuity of $J_{\mathbb{P}}^{\gamma}$ in W_2. Moreover, there exists a sequence $(\mu_\gamma)_\gamma$ converging to μ (for instance take $(\mu_\gamma)_\gamma$ constant equals to μ) such that $\lim_{\gamma \to 0} J_{\mathbb{P}}^{\gamma}(\mu_\gamma) = \lim_{\gamma \to 0} J_{\mathbb{P}}^{\gamma}(\mu) = J_{\mathbb{P}}^{0}(\mu)$. One can also notice that $J_{\mathbb{P}}^{\gamma} : \mathcal{P}_2(\Omega) \to \mathbb{R}$ is equi-coercive: for all $t \in \mathbb{R}$, the set $\{\nu \in \mathcal{P}_2(\Omega) \text{ such that } J_{\mathbb{P}}^{\gamma}(\nu) \leq t\}$ is included in a compact K_t since it is closed in the compact set $\mathcal{P}_2(\Omega)$ (by compactness of Ω). Therefore, we can apply the fundamental theorem of Γ-convergence (Theorem 2.10 in [Bra06]) to the bounded Lipschitz metric to obtain that $d_{BL^*}(\mu_{\mathbb{P}}^{0}, \mu_{\mathbb{P}}^{\gamma}) \underset{\gamma \to 0}{\longrightarrow} 0$. □

5 Algorithmic approach and numerical experiments

In this section, we first present a method to automatically choose the parameter γ. Then, we present numerical experiments on simulated and real data sets in $\mathbb{R}$ and $\mathbb{R}^2$. A discretization of the minimization problem (1.2) is used to compute a numerical approximation of a regularized Wasserstein barycenter $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$. It consists of using a fixed grid $\{x^k\}_{k=1}^N$ of equally spaced points $x^k \in \mathbb{R}^d$, and to approximate $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$ by the discrete measure $\sum_{k=1}^N f^k \delta_{x^k}$ where the f^k are positive weights summing to one which minimize a discrete version of the optimisation problem (1.2). Further algorithmic details are given in an C. In these numerical experiments, we focus on the case where $E(\mu) = +\infty$ if μ is not a.c. to enforce the regularized Wasserstein barycenter to have a smooth pdf (we write $E(f) = E(\mu_f)$ if μ has a density f). In this setting, if the grid of points is of sufficiently large size, then the weights f^k yield a good approximation of this pdf.

5.1 Choice of the parameter γ

By analogy with the work in [BM07] based on the Lepskii balancing principle, we use an automatic selection of the regularization parameter γ. The method proposed in [BM07] requires the knowledge of an upper bound on the decay to zero of the variance term $\mathbb{E}(d_E^2(\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}, \mu_{\mathbb{P}}^{\gamma}))$ as $\gamma \to 0$ which is given by (4.5). To match their notation, we set $\lambda = 1/\gamma$. Hence $\lambda \to +\infty$ corresponds to $\gamma \to 0$.

Without the knowledge of $\mu_{\mathbb{P}}$, the Lepskii balancing principle described in [BM07] to select an appropriate λ works as follows:

- For $\sigma > 1$ and a threshold $\Lambda > 0$, the Look-Ahead function is defined as $l_{\Lambda,\sigma}(\lambda) = \min\{\min\{\kappa | \rho(\lambda) > \sigma\rho(\kappa)\}, \Lambda\}$ where the choice $\rho : \lambda \mapsto \frac{1}{\lambda}$ comes from the upper bound (4.5).

- For $\delta > 0$, the balancing functional reads

$$b_{\Lambda,\sigma}(\lambda) = \max_{\lambda < \kappa \leq l_{\Lambda,\sigma}(\lambda)} \{\frac{1}{4\delta} d_E(\boldsymbol{\mu}_{\mathbb{P}_n}^{1/\lambda}, \boldsymbol{\mu}_{\mathbb{P}_n}^{1/\kappa})\rho(\kappa)\}$$

- The data-driven choice is given by $\lambda_{\Lambda,\sigma,\epsilon} = \min\{\lambda \leq \Lambda \ ; \ B_{\Lambda,\sigma}(\lambda) \leq \epsilon\}$ where $B_{\Lambda,\sigma}(\lambda) = \max_{\lambda < \kappa \leq \Lambda} \{b_{\Lambda,\sigma}(\kappa)\}$ is the smooth balancing functional, and $\epsilon > 0$ is a parameter to control the stability of regularized barycenters for successive choices of $\lambda = 1/\gamma$.

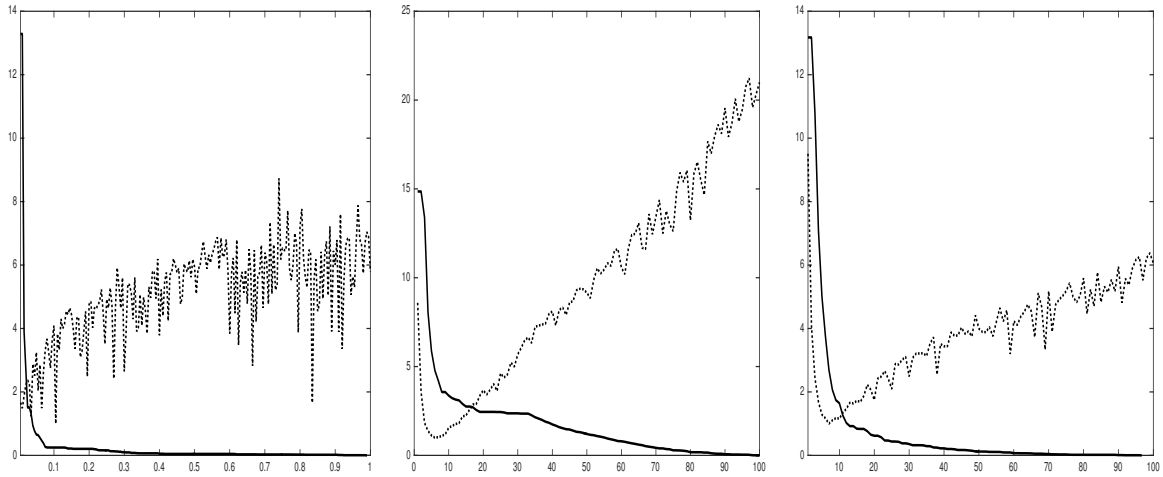
We display in Figure 2 an example of the smooth balancing functional $B_{\Lambda,\sigma}(\lambda)$ associated to the simulated Gaussian data from Section 5.2 for three different choices for the penalty function E.

In this paper, instead of choosing a value for ϵ, we use the fact that $B_{\Lambda,\sigma}$ is a decreasing function, and that it has the shape of a L-curve. Hence, a data-driven value $\hat{\lambda}_{\Lambda,\sigma}$ is chosen by determining the location where the curve $B_{\Lambda,\sigma}$ has an "elbow" (change of curvature). In Figure 2, this strategy leads to a data-driven value for λ that is close to the ideal choice given by the minimizer of $\lambda \mapsto d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{1/\lambda}) / \min_\lambda d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{1/\lambda})$.

5.2 Numerical experiments for $d = 1$

Simulated data We consider a simulated example where the measures $\boldsymbol{\nu}_i$ are discrete and supported on a small number p_i of data points. To this end, for each $i = 1, \ldots, n$, we simulate a sequence $(\boldsymbol{X}_{ij})_{1 \leq j \leq p_i}$ of iid random variables sampled from a Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i^2)$ where the p_i's are ranging from 5 to 10, and the $\boldsymbol{\mu}_i$'s (resp. $\boldsymbol{\sigma}_i$) are iid random variables such that $-2 \leq \boldsymbol{\mu}_i \leq 2$ and $0 \leq \boldsymbol{\sigma}_i \leq 1$ with $\mathbb{E}(\boldsymbol{\mu}_i) = 0$ and $\mathbb{E}(\boldsymbol{\sigma}_i) = 1/2$. The target measure that we wish to estimate in these simulations is the population (or true) Wasserstein barycenter of the random distribution $\mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\sigma}_1^2)$ which is $\mathcal{N}(0, 1/4)$ thanks to the assumptions $\mathbb{E}(\boldsymbol{\mu}_1) = 0$ and $\mathbb{E}(\boldsymbol{\sigma}_1) = 1/2$. Then, we define the random discrete measure $\boldsymbol{\nu}_i = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\boldsymbol{X}_{ij}}$.

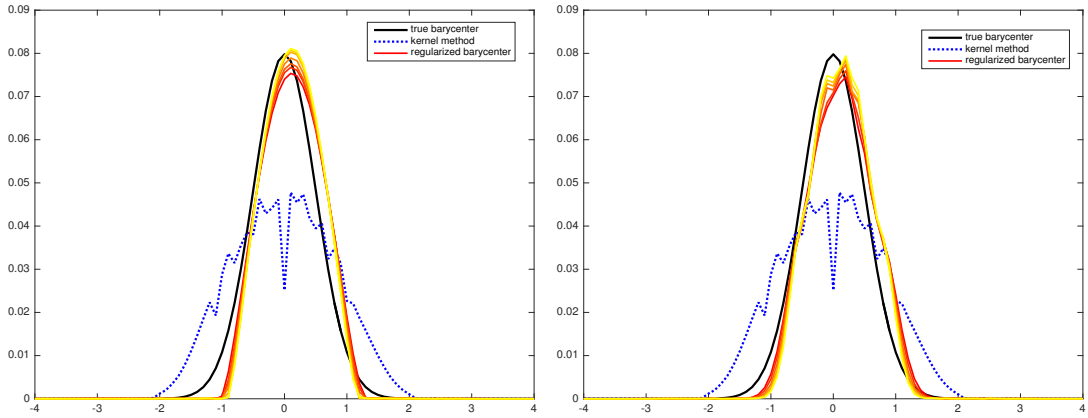
To illustrate the benefits of regularizing the Wasserstein barycenter of the $\boldsymbol{\nu}_i$'s, we compare our estimator with the one obtained by the following procedure which we refer to as the kernel method. In a preliminary step, each measure $\boldsymbol{\nu}_i$ is smoothed using a standard kernel density estimator whose bandwidth h_i is chosen by cross-validation. An alternative estimator is then defined as the Wasserstein barycenter of these smoothed measures which can be easily computed thanks to the quantile averaging formula for measures supported on $\mathbb{R}$ (see e.g.



(a) $E(f) = ||f'||^2$ Dirichlet (b) $E(f) = \int f \log(f)$ Entropy (c) Dirichlet + Entropy

Figure 2: Simulated Gaussian data from Section 5.2. Smooth balancing functional $B_{\Lambda,\sigma}(\lambda)$ (times 10^{-6}) in solid lines for three different regularizations (with $\sigma = 3$). For these simulated data, we have access to $\mu_{\mathbb{P}}$, and the dotted lines represent $d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{1/\lambda}) / \min_\lambda d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{1/\lambda})$ as functions of λ.

Section 6.1 in [AC11]). This estimator corresponds to the notion of smoothed Wasserstein barycenter of multiple point processes as considered in [PZ16]. The density of this smoothed Wasserstein barycenter is displayed in Figure 3. For this example, it appears a preliminary smoothing of the $\boldsymbol{\nu}_i$ followed by quantile averaging is not sufficient to recover a satisfactory Gaussian shape when the number p_i of observations per unit is small.



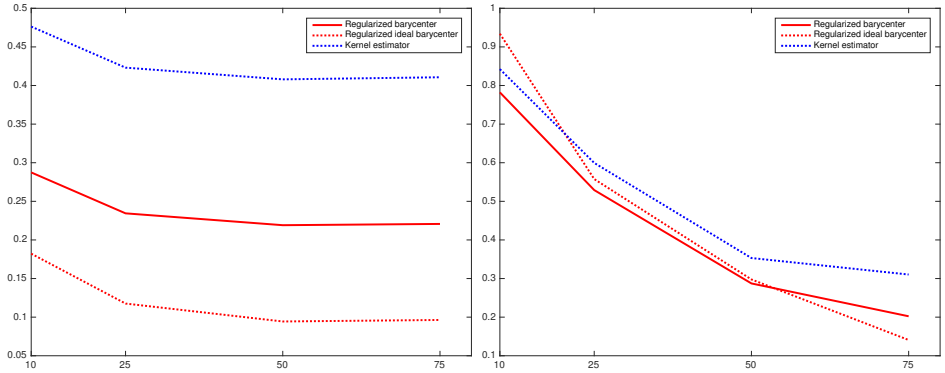
(a) Dirichlet regularization (b) Entropy regularization

Figure 3: Simulated data from Gaussian distributions with random means and variances. In all the figures, the black curve is the density of the true Wasserstein barycenter. The blue and dotted curve represents the pdf of the smoothed Wasserstein barycenter obtained by a preliminary kernel smoothing step. Pdf of the regularized Wasserstein barycenter $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$ (a) for $20 \leq \gamma \leq 50$ with $E(f) = ||f'||^2$ (Dirichlet), and (b) for $0.08 \leq \gamma \leq 14$ with $E(f) = \int f \log(f)$ (negative entropy).

Alternatively, we have applied the algorithm described Section 3.1 of the C directly on

the (non-smoothed) discrete measures $\boldsymbol{\nu}_i$ to obtain a numerical approximation of regularized barycenter $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$ with two different choices for the penalty function E. The results are displayed in Figure 3 for different values of γ around the data-driven choice $1/\hat{\lambda}_{\Lambda,\sigma}$ from Section 5.1. For both penalty functions and despite a small number of observations per experimental units, the shape of these densities better reflects the fact that the population Wasserstein barycenter is a Gaussian distribution.

Finally, we provide Monte-Carlo simulations to illustrate the influence of the number n of observed measures on the convergence of these estimators. For a given $10 \leq n_0 \leq n$, we randomly draw n_0 measures $\boldsymbol{\nu}_i$ from the whole sample, and we compute the following estimators: a smoothed barycenter via the kernel method, a regularized barycenter using a data-driven choice for γ, and an *ideal regularized barycenter* obtained by the Lepskii method if we had access to the true population barycenter by setting γ as the minimizer of $\gamma \mapsto d_E(\mu_{\mathbb{P}}, \mu_{\mathbb{P}_n}^{\gamma})$. For given value of n_0, this procedure is repeated 200 times, which allows to obtain an approximation of the expected error $\mathbb{E}\left(d(\hat{\mu}, \mu_{\mathbb{P}})\right)$ of each estimator $\hat{\mu}$, where d is either the Bregman or the Wasserstein distance. The penalty used is a linear combinaison of Dirichlet and negative entropy functions. The results are displayed in Figure 4. It can be observed that our approach yields better results than the kernel method for both types of error (using either the Bregman or Wasserstein distance).



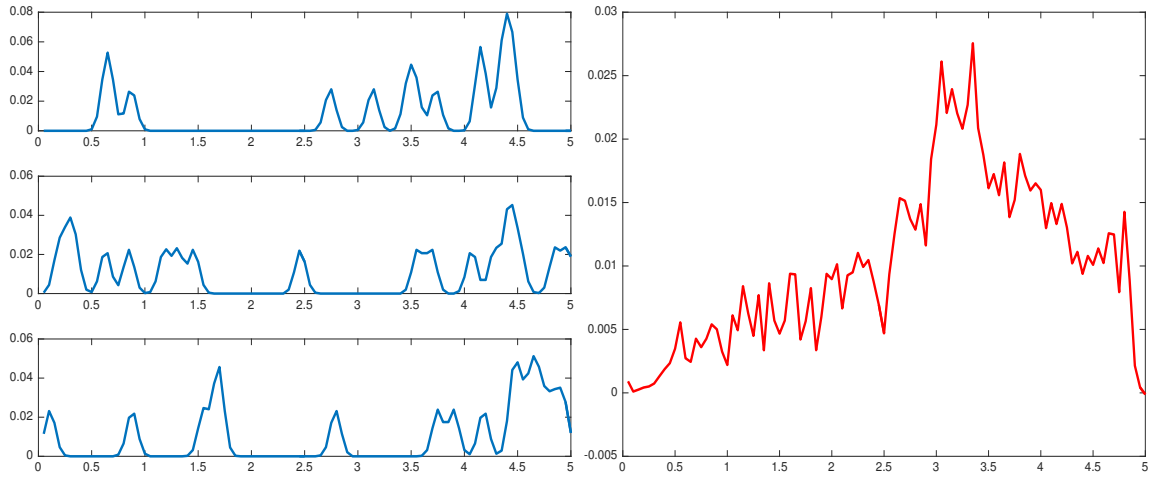
(a) Error with Bregman distance (b) Error with Wasserstein distance

Figure 4: Errors in terms of expected Bregman and Wasserstein distances between the population barycenter and the estimated barycenters (kernel method in dashed blue, regularized barycenter in full red and ideal regularized barycenter in dashed red) for a sample of size $n_0 = 10, 25, 50$ and 75 (the whole sample is made of $n = 100$ discrete measures).

A real data set We consider now a real data set of neural spike trains which is publicly available from the MBI website[2]. During a squared-path task, the spiking activity of a movement-encoded neuron of a monkey has been recorded during 5 seconds over $n = 60$ repeated trials. Each spike train is then smoothed using a Gaussian kernel (further details on the data collection can be found in [WS11]). For each trial $1 \leq i \leq n$, we let ν_i be the measure with pdf proportional to the sum of these Gaussian kernels centered at the times of spikes, see Figure 5(a). This is an example of a dataset made of a.c. measures (histograms).

[2] http://mbi.osu.edu/2012/stwdescription.html

The pdf of the Wasserstein barycenter $\bar{\nu}_n$ of these measures is displayed in Figure 5(b). This approach leads to a an irregular mean density of spiking activity.



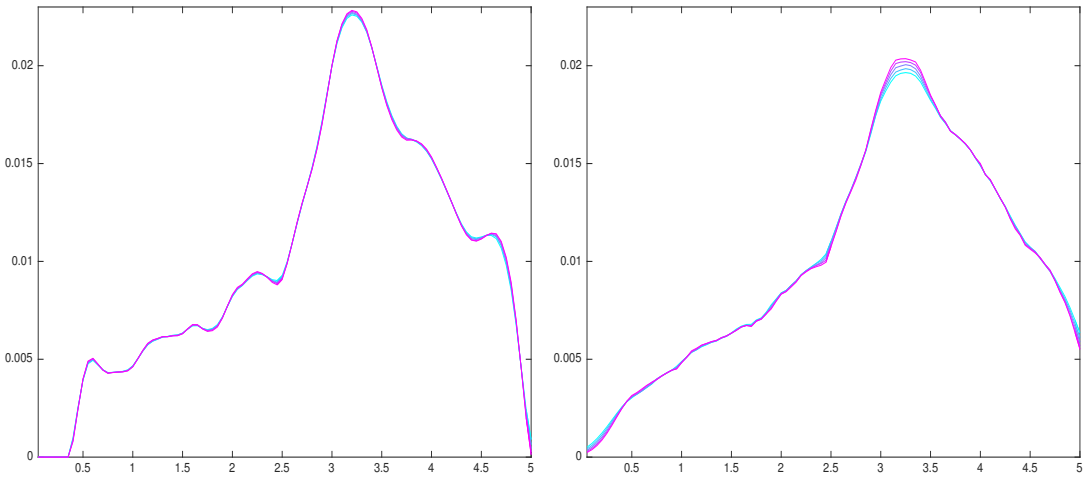
(a) (b)

Figure 5: Neural spike trains data. (a) A subset of 3 smoothed spikes out of $n = 60$ of the neural activity of a monkey during 5 seconds. Each row represents one trial and the pdf obtained by smoothing each spike train with a Gaussian kernel of width 50 milliseconds. (b) Pdf of the empirical Wasserstein barycenter $\bar{\nu}_n$ for this data set.

To compute a regularized Wasserstein barycenter from this dataset, we apply the algorithm described in Section 3.1 of the C. In Figure 6(a), we display the densities of the regularized Wasserstein barycenters with a Dirichlet regularization obtained for $6 \leq \gamma \leq 10$, where this range of values is chosen by the adaptative Lepskii balancing principle described in Section 5.1. In Figure 6(b), we display the results obtained with a negative entropy regularization for the data-driven range of values $0.08 \leq \gamma \leq 0.12$. Comparing Figures 5 and 6, this approach allows to clearly smooth the result obtained by quantile averaging of the ν_i's (which corresponds to $\gamma = 0$).

5.3 Numerical experiments for $d = 2$

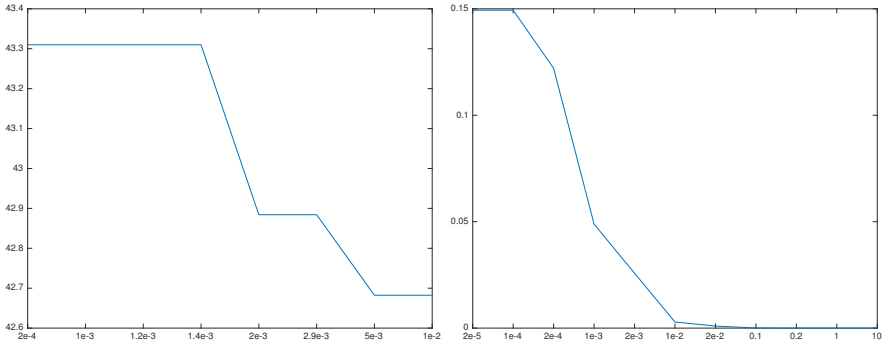
We consider the real dataset described previously on the locations of reported incidents of crime in Chicago. The city of Chicago is represented as an image of size 92×59, and a crime is considered as a Dirac located at some pixel (see Figure 1). To compute a regularized Wasserstein barycenter from this dataset, we apply the algorithm described in Section 3.2 of the C for $d = 2$. For a Dirichlet regularization, the adaptative Lepskii's strategy from Section 5.1 leads us to choose $\lambda = 2.10^{-3}$ as it can be observed from Figure 7(a). We compare our approach with the one in [CP16] which consists in using a regularized barycenter associated to an entropically regularized transportation cost. The computation of such regularized barycenters is obtained via the so-called Sinkhorn's algorithm, see [CP16] for further details. The parameter which controls the amount of transportation plan regularization is fixed to $\epsilon = 1$ (following the notation in [CP16]). The parameter γ controlling the amount of regularization of such Wasserstein barycenter is again chosen in a data-driven way using the Lepskii's strategy. Using transportation plan regularization, it follows than the best choice is



(a) $E(f) = ||f'||^2$ Dirichlet (b) $E(f) = \int f \log(f)$ Entropy

Figure 6: Neural spike trains data. (a) Pdf of the regularized empirical Wasserstein barycenter with a Dirichlet regularization for $6 \leq \gamma \leq 10$. (b) Pdf of the regularized empirical Wasserstein barycenter with negative entropy regularization and $0.08 \leq \gamma \leq 0.12$

$\lambda = 10^{-2}$, as it can be seen from Figure 7(b).



(a) Our approach: regularized Wasserstein barycenter with unregularized transport cost.

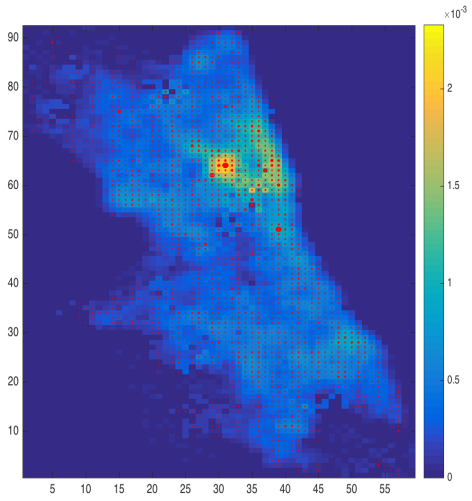
(b) The approach in [CP16]: regularized Wasserstein barycenter with regularized transport cost and Sinkhorn's algorithm.

Figure 7: Real dataset of reported incidents of crime in Chicago. Smooth balancing functional $B_{\Lambda,\sigma}(\lambda)$ associated to regularized Wasserstein barycenters for different values of λ.

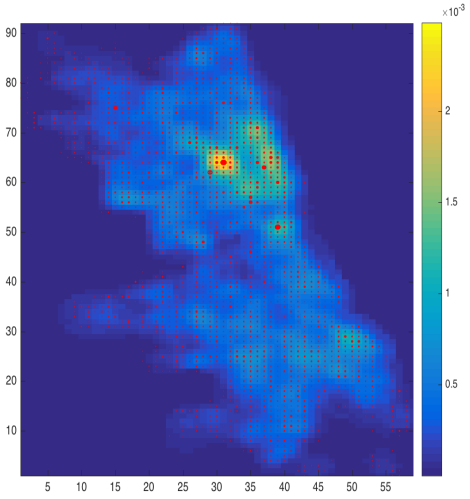
These two choices for γ lead to the regularized Wasserstein barycenters whose pdf are displayed in Figure (8)(a) and Figure (8)(b). The main differences between the estimator is the spreading of the mass in Figure (8)(b) due to the use of an entropically regularized transportation cost. Finally, we present in Figure (8)(c) a comparison with a standard kernel density estimator using the whole dataset (over all the year 2014) obtained from an implementation in Matlab[3] of a bivariate Gaussian kernel density estimator which uses a data-driven choice for the bandwidth suggested in [BA97]. The method of kernel density yields a much

[3] https://www.mathworks.com/examples/matlab/community/20312-bivariate-kernel-density-estimation

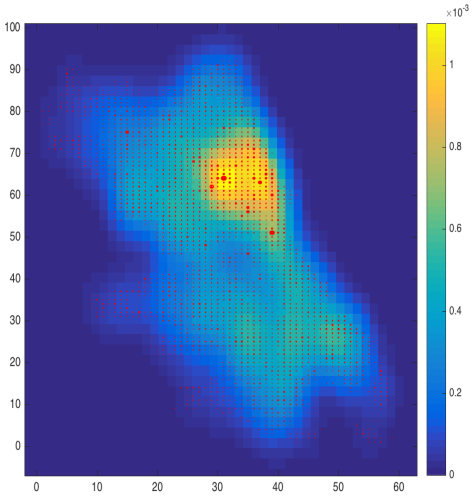
smoothed estimator than those obtained with regularized Wasserstein barycenters.



(a) Our approach: pdf of regularized Wasserstein barycenter with unregularized transport cost.



(b) The approach in [CP16]: regularized Wasserstein barycenter with regularized transport cost and Sinkhorn's algorithm.



(c) Standard Kernel density estimator of the whole dataset with a data-driven bandwidth.

Figure 8: (a) Barycenter and (b) kernel density estimator associated to the location of crimes in the city of Chicago during the year 2014 are represented in color scale. The red points represent locations of all crimes during this month.

A Bregman divergence and symmetric distance

Definition A.1 (Subdifferential)**.** Let X be a vector space and let X' its dual. For a convex functional $G : X \to \mathbb{R}$, we say that $q \in X'$ is a subgradient of G at u if it satisfies the inequality

$$G(v) \geq G(u) + \langle q, v - u \rangle \text{ for every } v \in X \tag{A.1}$$

where $\langle \cdot, \cdot \rangle$ is the duality pairing (or linear form) of X and X'. The set of subgradients at u is the subdifferential of G, denoted $\partial G(u)$.

Definition A.2 (Bregman divergence)**.** The (generalized) Bregman divergence related to a convex functional $G : X \to \mathbb{R}$ is defined by

$$D_G^q(u, v) = G(u) - G(v) - \langle q, u - v \rangle \text{ for } u, v \in X,$$

where $q \in \partial G(v)$. The symmetric Bregman distance is defined by

$$d_G^{p,q}(u, v) = D_G^q(u, v) + D_G^p(v, u) = \langle p - q, u - v \rangle$$

where $p \in \partial G(u)$ and $q \in \partial G(v)$.

B Proofs of main Theorems

B.1 Proof of the subgradient's inequality

Proof of Lemma 2.6. 2⇒1. Let $\phi \in \partial J(\mu)$ such that $\langle \phi, \eta - \mu \rangle \geq 0$ for all $\eta \in \mathcal{P}_2(\Omega)$. By definition of the subgradient, $\forall\ \eta \in \mathcal{P}_2(\Omega)$, we have $J(\eta) \geq J(\mu) + \langle \phi, \eta - \mu \rangle$ which is greater than $J(\mu)$ by assertion. Hence μ minimizes J.
1⇒2. Take $\mu \in \text{int}(\text{dom } J)$ (that is $J(\mu) < +\infty$) such that μ is a minimum of J over $\mathcal{P}_2(\Omega)$. Then the directional derivative of J at the point μ along $(\eta - \mu)$ exists (Proposition 2.22 in [Cla13]) and satisfies

$$J'(\mu; \eta - \mu) := \lim_{\substack{t \to 0 \\ t > 0}} \frac{J(\mu + t(\eta - \mu)) - J(\mu)}{t} \geq 0. \tag{B.1}$$

Remark that $\mathcal{P}_2(\Omega)$ is a convex set. By Proposition 4.3 of [Cla13], since J is a proper convex function and $\mu \in \text{dom}(J)$, we obtain the equivalence

$$\phi \in \partial J(\mu) \ \Leftrightarrow\ \langle \phi, \Delta \rangle \leq J'(\mu; \Delta) \text{ for all } \Delta \in \mathcal{P}_2(\Omega).$$

Moreover, since J is proper convex and lower semi-continuous, so is $J'(f; \cdot)$. Given that $\mathcal{P}_2(\Omega)$ is a Hausdorff convex space, we get by Theorem 7.6 of [AB06], that for all $(\eta - \mu) \in \mathcal{P}_2(\Omega)$, $J'(\mu; \eta - \mu) = \sup\{\langle \phi, \eta - \mu \rangle \text{ where } \phi \text{ is such that } \langle \phi, \Delta \rangle \leq J'(\mu; \Delta), \forall \Delta \text{ in } \mathcal{P}_2(\Omega)\}$. Hence by (B.1) we get $\sup_{\phi \in \partial J(\mu)} \langle \phi, \eta - \mu \rangle \geq 0$. We then define the ball $B_\epsilon = \{\eta + \mu \in \mathcal{M}(\Omega)$ such that $\|\eta\|_{TV} \leq \epsilon\}$, where $\| \cdot \|_{TV}$ is the norm of total variation. We still have

$$\inf_{\eta \in B_\epsilon \cap \mathcal{P}_2(\Omega)} \sup_{\phi \in \partial J(\mu)} \langle \phi, \eta - \mu \rangle \geq 0.$$

Note that $\partial J(\mu)$ in a convex set. Moreover $B_\epsilon \cap \mathcal{P}_2(\Omega)$ is compact, and $(\phi, \eta) \mapsto \langle \phi, \eta - \mu \rangle$ is bilinear. Thus we can switch the infimum and the supremum by the Ky Fan's theorem

(4.36 in [Cla13]). In that way, there exists $\phi \in \partial J(f)$ such that $\inf_{\eta \in B_\epsilon \cap \mathcal{P}_2(\Omega)} \langle \phi, \eta - \mu \rangle \geq 0$. By convexity of $\mathcal{P}_2(\Omega)$, any $\zeta \in \mathcal{P}_2(\Omega)$ can be written as $t(\eta - \mu) + \mu$ for some $t \geq 0$ and $\eta \in B_\epsilon \cap \mathcal{P}_2(\Omega)$. This concludes the proof of the lemma. □

Proof of Lemma 2.7. ($\Leftarrow$). We first assume that for $\phi^{\mu,\nu} \in \mathbb{L}_1(\mu)$, there exists $\psi^{\mu,\nu} \in \mathbb{L}_1(\nu)$ such that $W_2^2(\mu,\nu) = \int \phi^{\mu,\nu} d\mu + \int \psi^{\mu,\nu} d\nu$ and $\phi^{\mu,\nu}(x) + \psi^{\mu,\nu}(y) \leq |x-y|^2$. Then for all $\eta \in \mathcal{P}_2(\Omega)$, denoting $(\phi^{\eta,\nu}, \psi^{\eta,\nu})$ an optimal couple for η and ν, we get

$$\begin{aligned} W_2^2(\eta,\nu) &= \sup_{\phi(x)+\psi(y) \leq |x-y|^2} \int \phi d\eta + \int \psi d\nu = \int \phi^{\eta,\nu} d\eta + \int \psi^{\eta,\nu} d\nu \\ &\geq W_2^2(\mu,\nu) + \int \phi^{\mu,\nu} d(\eta - \mu). \end{aligned}$$

Hence, from the definition of a subgradient, $\phi^{\mu,\nu} \in \partial_1 W_2^2(\mu,\nu)$.
($\Rightarrow$). We denote by F the function $\mu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu)$. Let $\phi^* \in \partial F(\mu)$, then by the Legendre-Fenchel theory, we have that $F^*(\phi^*) + F(\mu) = \int \phi^* d\mu$, where F^* denote the Fenchel conjugate of F. Recall we want to show that there exists $\psi \in \mathbb{L}_1(\nu)$ verifying $\phi^*(x) + \psi(y) \leq |x-y|^2$ such that

$$\int \phi^* d\mu - W_2^2(\mu,\nu) = -\int \psi d\nu,$$

which is equivalent to $F^*(\phi^*) = -\int \psi d\nu$. In that purpose, we first define $\psi^\phi(\cdot) := \inf_{y \in \Omega} \{ |\cdot - y|^2 - \phi(y) \}$ and $H(\phi) := -\int \psi^\phi d\nu$.

By definition, $H^*(\mu) = \sup_{\phi \in Y} \{ \int \phi d\mu - H(\phi) \}$. Observing that H is convex, l.s.c. on Y and proper as :

$$\begin{aligned} H(\phi) &= -\int \psi^\phi d\nu = \int \sup_{y \in \Omega} \{ \phi(y) - |x-y|^2 \} \, d\nu(x) \\ &\geq \int (\phi(y_0) - 2|y_0| - 2|x|^2) d\nu(x) > -\infty \text{ by definition of } \nu, \end{aligned}$$

where $y_0 \in \Omega$ is such that $\phi(y_0)$ is finite. We get $H^{**}(\phi) = H(\phi)$ by Theorem 2.3.3. in [Zal02]. Moreover, for $\mu \in \mathcal{P}_2(\Omega)$, we have by the duality formulation of Kantorovich (e.g Lemma 2.1. of [AC11]) that

$$\begin{aligned} W_2^2(\mu,\nu) &= \sup \left\{ \int_\Omega \phi d\mu + \int_\Omega \psi d\nu; \ \phi, \psi \in \mathcal{C}_b, \phi(x) + \psi(y) \leq |x-y|^2 \right\} \\ &= \sup \left\{ \int_\Omega \phi d\mu + \int_\Omega \psi d\nu; \ \phi, \psi \in \mathcal{C}_b, \psi(y) \leq \inf_x \{ |x-y|^2 - \phi(x) \} \right\} \\ &= \sup_\phi \left\{ \int_\Omega \phi d\mu + \int_\Omega \psi^\phi \right\} = H^*(\mu). \end{aligned}$$

We deduce that $H^{**}(\phi) = \sup_{f \in \mathcal{P}_2(\Omega)} \{ \int \phi d\mu - W_2^2(\mu,\nu) \} = F^*(\phi)$, which implies $F^*(\phi^*) = H(\phi^*)$. Thus we end up with the equality $F(\mu) = \int \phi^* d\mu - F^*(\phi^*) = \int \phi^* d\mu + \int \psi^{\phi^*} d\nu$. This exactly means that for $\phi^* \in \partial_1 W_2^2(\mu,\nu)$, there exists ψ^{ϕ^*} such that $\phi^*(x) + \psi^{\phi^*}(y) \leq |x-y|^2$ and $W_2^2(\mu,\nu) = \int \phi^* d\mu + \int \psi^{\phi^*} d\nu$, which concludes the proof. □

B.2 Proof of existence, uniqueness and stability of regularized barycenters

Proof of Theorem 3.1. Let $(\mu^k)_k \subset \mathcal{P}_2(\Omega)$ a minimizing sequence of probability measures of $J^\gamma_{\mathbb{P}_n}$. Hence, there exists a constant $M \geq 0$ such that $\forall k,\ J^\gamma_{\mathbb{P}_n}(\mu^k) \leq M$. It follows that for all k, $\frac{1}{n}\sum_{i=1}^n W_2^2(\mu^k, \nu_i) \leq M$. By Lemma 2.1 of [AC11] we thus have

$$\frac{1}{n}\sum_{i=1}^n W_2^2(\nu^i, \mu^k) = 2\sum_{i=1}^n \sup_{f\in Z}\ \left\{\int_{\mathbb{R}^d} f d\mu^k + \int_{\mathbb{R}^d} Sf(x) d\nu^i(x)\right\} \leq M,$$

where $Sf(x) = \inf_{y\in\Omega}\{\frac{1}{2n}|x-y|^2 - f(y)\}$. Since the function $x \mapsto |x|^\alpha$ (with $1 < \alpha < 2$) belongs to Z, we have that $\int_{\mathbb{R}^d}|x|^\alpha d\mu^k(x)$ is bounded by a constant $L \geq 0$ for all k. We deduce that $(\mu^k)_k$ is tight (for instance, take the compact $K^c = \{x \in \Omega \text{ such that } |x|^\alpha > \frac{L}{\epsilon}\}$). Since $(\mu^k)_k$ is tight, by Prokhorov's theorem, there exists a subsequence of $(\mu^k)_k$ (still denoted (μ^k)) which weakly converges to a probability measure μ. Moreover, one can prove that $\mu \in \mathcal{P}_2(\Omega)$. Indeed for all lower semicontinuous functions bounded from below by f, we have that $\liminf_{k\to\infty}\int_\Omega f(x)d\mu^k(x) \geq \int_{\mathbb{R}^d} f(x)d\mu(x)$ by weak convergence. Hence for $f : x \mapsto |x|^2$, we get $\int_\Omega |x|^2 d\mu(x) \leq \liminf_{k\to\infty}\int_\Omega |x|^2 d\mu^k(x) < +\infty$, and thus $\mu \in \mathcal{P}_2(\Omega)$.

Let $(\pi_i^k)_{1\leq i\leq n, 1\leq k}$ be a sequence of optimal transport plans where π_i^k is an optimal transport plan between μ^k and ν_i. Since $\sup_k W_2^2(\mu^k, \nu_i) = \sup_k \iint_{\Omega\times\Omega}|x-y|^2 d\pi_i^k(x,y) < +\infty$, we may apply Proposition 7.1.3 of [AGS08]: $(\pi_i^k)_k$ is weakly relatively compact on the probability space over $\Omega\times\Omega$ and every weak limit π_i is an optimal transport plan between μ and ν_i with, for all $1 \leq i \leq n$, $W_2^2(\mu,\nu_i) \leq \liminf_{k\to\infty}\int_{\mathbb{R}^d\times\mathbb{R}^d}|x-y|^2 d\pi_i^k(x,y) < +\infty$. Since E is lower semicontinuous, we get that

$$\begin{aligned}\liminf_{k\to\infty} J^\gamma_{\mathbb{P}_n}(\mu^k) &= \liminf_{k\to\infty}\ \frac{1}{n}\sum_{i=1}^n W_2^2(\mu^k,\nu_i) + \gamma E(\mu^k)\\ &\geq \frac{1}{n}\sum_{i=1}^n W_2^2(\mu,\nu_i) + \gamma E(\mu) = J^\gamma_{\mathbb{P}_n}(\mu).\end{aligned}$$

Hence $J^\gamma_{\mathbb{P}_n}$ admits at least $\mu \in \mathcal{P}_2(\Omega)$ as a minimizer. Finally, by the strict convexity of $J^\gamma_{\mathbb{P}_n}$ on its domain, the minimizer is unique and it belongs to $\mathcal{D}(E)$ as defined in (2.5), which completes the proof. □

Proof of Theorem 3.2. First, let us prove the existence of a minimizer. For that purpose, we decide to follow the sketch of the proof of the existence of a Wasserstein barycenter given by Theorem 1 in [LGL16]. We suppose that $(\mathbb{P}_n)_{n\geq 0} \subseteq W_2(\mathcal{P}_2(\Omega))$ is a sequence of measures, such that $\mu^n \in \mathcal{P}_2(\Omega)$ is a probability measure minimizing $J^\gamma_{\mathbb{P}_n}$, for all n. Furthermore, we suppose that there exists $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ such that $\mathcal{W}_2(\mathbb{P},\mathbb{P}_n) \underset{n\to+\infty}{\longrightarrow} 0$. We then have to prove that $(\mu^n)_{n\geq 1}$ is precompact and that all limits minimize $J^\gamma_{\mathbb{P}}$. We denote $\tilde{\boldsymbol{\mu}}$ a random measure with distribution $\mathbb{P}$ and $\tilde{\boldsymbol{\mu}}^n$ a random measure with distribution $\mathbb{P}_n$. Hence

$$\begin{aligned}W_2(\mu^n,\delta_x) &= \mathcal{W}_2(\delta_{\mu^n},\delta_{\delta_x}) \leq \mathcal{W}_2(\delta_{\mu^n},\mathbb{P}_n) + \mathcal{W}_2(\mathbb{P}_n,\delta_{\delta_x})\\ &= \mathbb{E}(W_2^2(\mu^n,\tilde{\boldsymbol{\mu}}^n))^{1/2} + \mathbb{E}(W_2^2(\tilde{\boldsymbol{\mu}}^n,\delta_x))^{1/2}.\end{aligned}$$

Moreover, $\mathbb{E}(W_2^2(\mu^n,\tilde{\boldsymbol{\mu}}^n))^{1/2} \leq M$ for a constant $M \geq 0$ since μ_n minimizes $J^\gamma_{\mathbb{P}_n}$ and $\tilde{\boldsymbol{\mu}}^n$ is of law $\mathbb{P}_n$. Then

$$W_2(\mu^n,\delta_x) \leq M + \mathcal{W}_2(\mathbb{P}_n,\delta_{\delta_x}) \leq M + \mathcal{W}_2(\mathbb{P}_n,\mathbb{P}) + \mathcal{W}_2(\mathbb{P},\delta_{\delta_x}) \leq L$$

since $\mathcal{W}_2(\mathbb{P}_n, \mathbb{P}) \underset{n\to+\infty}{\longrightarrow} 0$ and $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$ by hypothesis. By Markov inequality, we have for $r > 0$

$$\mu^n(B(x,r)^c) = \mathbb{P}_{\mu^n}(|X-x|^2 \leq r^2) \leq \frac{\mathbb{E}_{\mu^n}(|X-x|^2)}{r^2} = \frac{W_2^2(\mu^n, \delta_x)}{r^2}$$

and $\mu^n(B(x,r)^c) \leq \frac{L^2}{r^2}$. Hence $(\mu^n)_n$ is tight: it is possible to extract a subsequence (still denoted (μ^n)) which converges weakly to a measure μ by Prokhorov's theorem. Let us show that μ minimizes $J_{\mathbb{P}}^{\gamma}$. Let $\eta \in \mathcal{P}_2(\Omega)$ and $\nu \in \mathcal{P}_2(\Omega)$ with distribution $\mathbb{P}$.

$$\begin{aligned} J_{\mathbb{P}}^{\gamma}(\eta) &= \mathbb{E}_{\mathbb{P}}(W_2^2(\eta,\nu)) + \gamma E(\eta) = \mathcal{W}_2^2(\delta_\eta, \mathbb{P}) + \gamma E(\eta) \\ &= \lim_{n\to+\infty} \mathcal{W}_2^2(\delta_\eta, \mathbb{P}_n) + \gamma E(\eta) \ \text{ since by hypothesis } \mathcal{W}_2(\mathbb{P}_n, \mathbb{P}) \to 0 \\ &\geq \liminf_{n\to+\infty} \mathcal{W}_2^2(\delta_{\mu^n}, \mathbb{P}_n) + \gamma E(\mu^n) \text{ since } \mu^n \text{ minimizes } J_{\mathbb{P}_n}^{\gamma} \end{aligned} \tag{B.2}$$

Moreover, we have by the inverse triangle inequality that

$$\liminf_{n\to+\infty} \mathcal{W}_2(\delta_{\mu^n}, \mathbb{P}_n) \geq \liminf_{n\to+\infty} \left(\mathcal{W}_2(\delta_\mu, \mathbb{P}_n) - \mathcal{W}_2(\delta_\mu, \delta_{\mu^n})\right) = \mathcal{W}_2(\delta_\mu, \mathbb{P}).$$

The last inequality comes from the two convergences $\mathcal{W}_2(\mathbb{P}_n, \mathbb{P}) \to 0$ and $\mathcal{W}_2(\delta_\mu, \delta_{\mu^n}) \to 0$. From (B.2) and by lower semicontinuity of E, we get $J_{\mathbb{P}}^{\gamma}(\eta) \geq \mathcal{W}_2^2(\delta_\mu, \mathbb{P}) + \gamma E(\mu) = J_{\mathbb{P}}^{\gamma}(\mu)$. Hence μ minimizes $J_{\mathbb{P}}^{\gamma}$. To finish the proof of the existence of a minimizer, we need the following result whose proof can be found in [LGL16].

Theorem B.1. *For all $\mathbb{P} \in W_2(\mathcal{P}_2(\Omega))$, there is a sequence of finitely supported distributions $\mathbb{P}_n$ (that is $\mathbb{P}_n = \sum_{k=1}^K \lambda_k \delta_{\kappa_k}$ where $\sum_{k=1}^K \lambda_k = 1$) such that $\mathcal{W}_2^2(\mathbb{P}_n, \mathbb{P}) \underset{n\to+\infty}{\longrightarrow} 0$.*

Now, by Theorem B.1 it follows that for a given distribution $\mathbb{P}$, one can find a sequence of finitely supported distributions $\mathbb{P}_n$ such that for all n there exists a unique measure $\mu^n \in \mathcal{P}_2(\Omega)$ minimizing $J_{\mathbb{P}_n}^{\gamma}$ using Theorem 3.1 and such that $W_2^2(\mathbb{P}_n, \mathbb{P}) \underset{n\to+\infty}{\longrightarrow} 0$ thanks to Theorem B.1. Therefore there is a probability measure μ which minimizes $J_{\mathbb{P}}^{\gamma}$. Let us make sure that μ is indeed in the space $\mathcal{P}_2(\Omega)$. From Theorem 3.1, we also have that $\mu^n \in \mathcal{P}_2(\Omega)$ for all n. Thus by weak convergence, $\int_\Omega |x|^2 d\mu(x) \leq \liminf_{n\to+\infty} \int_\Omega |x|^2 d\mu^n(x) < +\infty$. Finally, the uniqueness of the minimum is obtained by the strict convexity of the functional $\mu \mapsto \mathbb{E}_{\mathbb{P}}(W_2^2(\mu,\nu)) + \gamma E(\mu)$ on the domain $\mathcal{D}(E)$, which completes the proof. □

B.3 Proof of the stability's Theorem 3.3

Proof. We denote by $\mu, \zeta \in \mathcal{P}_2(\Omega)$ the probability measures such that μ minimizes $J_{\mathbb{P}_n^\nu}^{\gamma}$ and ζ minimizes $J_{\mathbb{P}_n^\eta}^{\gamma}$. For each $1 \leq i \leq n$, one has that $\theta \mapsto \frac{1}{n} W_2^2(\theta, \nu_i)$ is a convex, proper and continuous function. Therefore, Theorem 4.10 in [Cla13], we have that $\partial J_{\mathbb{P}_n^\nu}(\mu) = \frac{1}{n}\sum_{i=1}^n \partial_1 W_2^2(\mu, \nu_i) + \gamma \nabla E(\mu)$. Hence by Lemma 2.7, any $\phi \in \partial J_{\mathbb{P}_n^\nu}(\mu)$ is of the form $\phi = \frac{1}{n}\sum_{i=1}^n \phi_i + \gamma \ \nabla E(\mu)$ where for all $i = 1, \dots, n$, $\phi_i = \phi^{\mu,\nu_i}$ is optimal in the sense that $(\phi^{\mu,\nu_i}, \psi^{\mu,\nu_i})$ is an optimal couple associated to (μ, ν_i) in the Kantorovich formulation of the Wasserstein distance (see Theorem 2.2). Therefore by Lemma 2.6, there exists $\phi = \frac{1}{n}\sum_{i=1}^n \phi^{\mu,\nu_i} + \gamma \nabla E(\mu)$ such that $\langle \phi, \theta - \mu \rangle \geq 0$ for all $\theta \in \mathcal{P}_2(\Omega)$. Likewise, there exists $\breve{\phi} = \frac{1}{n}\sum_{i=1}^n \phi^{\zeta,\eta_i} + \gamma \nabla E(\zeta)$ such that $\langle \breve{\phi}, \theta - \zeta \rangle \geq 0$ for all $\theta \in \mathcal{P}_2(\Omega)$. Finally, we obtain that

$$\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle \leq -\int_\Omega \left(\frac{1}{n}\sum_{i=1}^n (\phi^{\mu,\nu_i} - \phi^{\zeta,\eta_i})\right) d(\mu-\zeta).$$

Following the proof of Kantorovich duality's theorem in [Vil03], we can restrict the supremum over $(\phi,\psi) \in C_W$ in Kantorovich's duality Theorem 2.2 to the admissible pairs (ϕ^{cc},ϕ^c) where $\phi^c(y) = \inf_x\{|x-y|^2 - \phi(x)\}$ and $\phi^{cc}(x) = \inf_y\{|x-y|^2 - \phi^c(y)\}$. Then, we replace ϕ^{μ,ν_i} by $(\phi^{\mu,\nu_i})^{cc}$ (resp. ϕ^{ζ,η_i} by $(\phi^{\zeta,\eta_i})^{cc}$) and ψ^{μ,ν_i} by $(\phi^{\mu,\nu_i})^c$ (resp. ψ^{ζ,η_i} by $(\phi^{\zeta,\eta_i})^c$) and obtain that

$$\begin{aligned}\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle &\leq -\frac{1}{n}\sum_{i=1}^n \int_\Omega \left[(\phi^{\mu,\nu_i})^{cc}(x) - (\phi^{\zeta,\eta_i})^{cc}(x)\right] d(\mu-\zeta)(x)\\ &= -\frac{1}{n}\sum_{i=1}^n \iint_{\Omega\times\Omega} \left[(\phi^{\mu,\nu_i})^{cc}(x) - (\phi^{\zeta,\eta_i})^{cc}(x)\right] d(\pi^{\mu,\nu_i} - \pi^{\zeta,\eta_i})(x,y),\end{aligned}$$

where π^{μ,ν_i} is an optimal transport plan on $\Omega\times\Omega$ with marginals μ and ν_i for $i \in \{1,\dots,n\}$ (and π^{ζ,η_i} optimal with marginals ζ and η_i). Developing the right-hand size expression in the above inequality, we get

$$\begin{aligned}&\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle\\ &\leq \frac{1}{n}\sum_{i=1}^n \left[-\iint (\phi^{\mu,\nu_i})^{cc}(x) d\pi^{\mu,\nu_i}(x,y) - \iint (\phi^{\zeta,\eta_i})^{cc}(x) d\pi^{\zeta,\eta_i}(x,y)\right]\\ &\quad + \frac{1}{n}\sum_{i=1}^n \left[\iint (\phi^{\mu,\nu_i})^{cc}(x) d\pi^{\zeta,\eta_i}(x,y) + \iint (\phi^{\zeta,\eta_i})^{cc}(x) d\pi^{\mu,\nu_i}(x,y)\right].\end{aligned}$$

From the condition (2.3) in the Kantorovich's dual problem, we have that $(\phi^{\mu,\nu_i})^{cc}(x) \leq |x-y|^2 - (\phi^{\mu,\nu_i})^c(y)$ and $(\phi^{\zeta,\eta_i})^{cc}(x) \leq |x-y|^2 - (\phi^{\zeta,\eta_i})^c(y)$ for all $i \in \{1,\dots,n\}$. Moreover, we have that $(\phi^{\mu,\nu_i})^{cc}(x) d\pi^{\mu,\nu_i}(x,y) = \left[-(\phi^{\mu,\nu_i})^c(y) + |x-y|^2\right] d\pi^{\mu,\nu_i}(x,y)$ and likewise $(\phi^{\zeta,\eta_i})^{cc}(x) d\pi^{\zeta,\eta_i}(x,y) = \left[-(\phi^{\zeta,\eta_i})^c(y) + |x-y|^2\right] d\pi^{\zeta,\eta_i}(x,y)$. We therefore deduce that

$$\begin{aligned}\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle \leq &-\frac{1}{n}\sum_{i=1}^n \iint \left[-(\phi^{\mu,\nu_i})^c(y) + |x-y|^2\right] d\pi^{\mu,\nu_i}(x,y)\\ &-\frac{1}{n}\sum_{i=1}^n \iint \left[-(\phi^{\zeta,\eta_i})^c(y) + |x-y|^2\right] d\pi^{\zeta,\eta_i}(x,y)\\ &+\frac{1}{n}\sum_{i=1}^n \iint \left[-(\phi^{\mu,\nu_i})^c(y) + |x-y|^2\right] d\pi^{\zeta,\eta_i}(x,y)\\ &+\frac{1}{n}\sum_{i=1}^n \iint \left[-(\phi^{\zeta,\eta_i})^c(y) + |x-y|^2\right] d\pi^{\mu,\nu_i}(x,y)\\ =&\frac{1}{n}\sum_{i=1}^n \int_\Omega \left[(\phi^{\mu,\nu_i})^c(y) - (\phi^{\zeta,\eta_i})^c(y)\right] d(\nu_i - \eta_i)(y).\end{aligned}$$

For all $1 \leq i \leq n$, and $y, y' \in \Omega$, we have

$$\begin{aligned}(\phi^{\mu,\nu_i})^c(y) - (\phi^{\mu,\nu_i})^c(y') =&\sup_x\{\phi^{\mu,\nu_i}(x) - |x-y'|^2\} + \inf_x\{|x-y|^2 - \phi^{\mu,\nu_i}(x)\}\\ \leq&\sup_x\{|x-y|^2 - |x-y'|^2\} \leq 4c_\Omega|y-y'|\end{aligned}$$

where $c_\Omega = \sup_x |x|$. As a consequence, $\frac{1}{4c_\Omega}(\phi^{\mu,\nu_i})^c$ is 1-Lipschitz and so is $\frac{1}{4c_\Omega}(\phi^{\zeta,\eta_i})^c$, which implies that $\frac{1}{8c_\Omega}\left[(\phi^{\mu,\nu_i})^c - (\phi^{\zeta,\eta_i})^c\right]$ is 1-Lipschitz for all $1 \leq i \leq n$.

We then conclude

$$\begin{aligned}\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle \leq & \frac{8c_\Omega}{n}\sum_{i=1}^n \sup\left\{\int \phi\, d(\nu_i - \eta_i);\ \phi \in \cap \mathbb{L}^1(|\nu_i - \eta_i|), \|\phi\|_{Lip} \leq 1\right\} \\ = & \frac{8c_\Omega}{n}\sum_{i=1}^n W_1(\nu_i, \eta_i) \leq \frac{8c_\Omega}{n}\sum_{i=1}^n W_2(\nu_i, \eta_i),\end{aligned}$$

by the Kantorovich-Rubinstein theorem presented in [Vil03], while the last inequality above comes from Hölder inequality between the distance W_2 and the distance W_1 defined for θ_1, θ_2 (probability measures on Ω with moment of order 1) as

$$W_1(\theta_1, \theta_2) = \inf_\pi \int_\Omega \int_\Omega |x - y| d\pi(x, y)$$

where π is a probability measures on $\Omega \times \Omega$ with respective marginals θ_1 and θ_2. Since μ and ζ are independent, we can assign to ν_i any $\eta_{\sigma(i)}$ for $\sigma \in \mathcal{S}_n$ the permutation group of $\{1, \ldots, n\}$ and hence we obtain $\gamma\langle \nabla E(\mu) - \nabla E(\zeta), \mu - \zeta\rangle \leq \frac{2}{n}\inf_{\sigma\in\mathcal{S}_n}\sum_{i=1}^n W_2(\nu_i, \eta_{\sigma(i)})$, which completes the proof. □

B.4 Proof of convergence properties

Proof of Theorem 4.3. We denote by C a universal constant whose value may change from line to line. From the subgradient's inequality (2.9) and following the same process used in the proof of Theorem 3.3, we have that, for each $\boldsymbol{\nu}_i$, $i = 1, \ldots, n$, there exists $\phi^{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \boldsymbol{\nu}_i}$ integrable with respect to $\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}(x)dx$ such that for all $\eta \in \mathcal{P}_2(\Omega)$:

$$\left\langle \frac{1}{n}\sum_{i=1}^n \phi^{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \boldsymbol{\nu}_i} + \gamma \nabla E(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}), \eta - \boldsymbol{\mu}^\gamma_{\mathbb{P}_n}\right\rangle \geq 0. \tag{B.3}$$

By applying once again the subgradient's inequality, we get

$$\mu^\gamma_{\mathbb{P}} \text{ minimizes } J^\gamma_{\mathbb{P}} \ \Leftrightarrow \exists \phi \in \partial J^\gamma_{\mathbb{P}}(\mu^\gamma_{\mathbb{P}}) \text{ s. t. } \langle \phi, \eta - \mu^\gamma_{\mathbb{P}}\rangle \geq 0 \text{ for all } \eta \in \mathcal{P}_2(\Omega).$$

Let us explicit the form of a subgradient $\phi \in \partial J^\gamma_{\mathbb{P}}(\mu^\gamma_{\mathbb{P}})$ using again the Theorem of the subdifferential of a sum. We have that $\mu \mapsto W_2^2(\mu, \nu)$ is continuous for all $\nu \in \mathcal{P}_2(\Omega)$. Moreover by symmetry, $\nu \mapsto W_2^2(\mu, \nu)$ is measurable for all $\mu \in \mathcal{P}_2(\Omega)$ and $W_2^2(\mu, \nu) \leq \iint |x - y|^2 d\mu(x) d\nu(y) \leq 2\int |x|^2 d\mu(x) + 2\int |y|^2 d\nu(y) \leq C$ is integrable with respect to $d\mathbb{P}(\nu)$ (by compactness of Ω). Hence, by Theorem of continuity under integral sign, we deduce that $\mu \mapsto \mathbb{E}[W_2^2(\mu, \boldsymbol{\nu})]$ is continuous. Thus we can manage the subdifferential of the following sum and one has that $\partial J^\gamma_{\mathbb{P}}(\mu^\gamma_{\mathbb{P}}) = \partial_1[\mathbb{E}(W_2^2(\mu^\gamma_{\mathbb{P}}, \boldsymbol{\nu}))] + \gamma \nabla E(\mu^\gamma_{\mathbb{P}})$, where $\boldsymbol{\nu}$ is still a random measure with distribution $\mathbb{P}$. Also the Theorem 23 in [Roc74] implies $\partial_1 \mathbb{E}[W_2^2(\mu^\gamma_{\mathbb{P}}, \boldsymbol{\nu})] = \mathbb{E}[\partial_1 W_2^2(\mu^\gamma_{\mathbb{P}}, \boldsymbol{\nu})]$. Hence, we can sum up

$$\mu^\gamma_{\mathbb{P}} \text{ minimizes } J^\gamma_{\mathbb{P}} \Leftrightarrow \left\langle \int \phi^{\mu^\gamma_{\mathbb{P}}, \nu} d\mathbb{P}(\nu) + \gamma \nabla E(\mu^\gamma_{\mathbb{P}}), \eta - \mu^\gamma_{\mathbb{P}}\right\rangle \geq 0,\ \forall\ \eta \in \mathcal{P}_2(\Omega). \tag{B.4}$$

In the sequel, to simplify the notation, we use $\boldsymbol{\mu} := \boldsymbol{\mu}^{\gamma}_{\mathbb{P}_n}$ and $\eta := \mu^{\gamma}_{\mathbb{P}}$. Therefore thanks to (B.3) and (B.4)

$$\begin{aligned} d_E(\boldsymbol{\mu}, \eta) &= \langle \nabla E(\boldsymbol{\mu}) - \nabla E(\eta), \boldsymbol{\mu} - \eta \rangle \\ &\leq -\frac{1}{\gamma} \left\langle \frac{1}{n} \sum_{i=1}^{n} \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i} - \int \phi^{\eta,\nu} d\mathbb{P}(\nu), \boldsymbol{\mu} - \eta \right\rangle \\ &= \frac{1}{\gamma} \left(-\frac{1}{n} \sum_{i=1}^{n} \int \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) d\boldsymbol{\mu}(x) + \frac{1}{n} \sum_{i=1}^{n} \int \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) d\eta(x) \right. \\ &\quad \left. + \iint \phi^{\eta,\nu} d\mathbb{P}(\nu) d\boldsymbol{\mu}(x) - \iint \phi^{\eta,\nu} d\mathbb{P}(\nu) d\eta(x) \right). \end{aligned} \tag{B.5}$$

We would like to switch integrals of the two last terms. In that purpose, we use that

$$\int W_2^2(\eta, \nu) d\mathbb{P}(\nu) \leq C + 2 \sup_{\nu \in \mathcal{P}_2(\Omega)} \left[\int_{\Omega} |y|^2 d\nu(y) \right] < +\infty.$$

As $0 \leq \int W_2^2(\eta,\nu) d\mathbb{P}(\nu) = \int \left(\int \phi^{\eta,\nu}(x) d\eta(x) + \int \psi^{\eta,\nu}(x) d\nu(y) \right) d\mathbb{P}(\nu)$, we also have that $\iint \phi^{\eta,\nu}(x) d\eta(x) d\mathbb{P}(\nu) < +\infty$. Since $x \mapsto \phi^{\eta,\nu}(x)$ and $\nu \mapsto \phi^{\eta,\nu}(x)$ are measurables, we obtain by Fubini's theorem $\int_{\Omega} \int_{\mathcal{P}_2(\Omega)} \phi^{\eta,\nu} d\mathbb{P}(\nu) d\eta(x) = \int_{\mathcal{P}_2(\Omega)} \int_{\Omega} \phi^{\eta,\nu} d\eta(x) d\mathbb{P}(\nu)$. By the same tools, since

$$\begin{aligned} \int W_2^2(\boldsymbol{\mu}, \nu) d\mathbb{P}(\nu) &= \int \left(\int \phi^{\boldsymbol{\mu},\nu}(x) d\boldsymbol{\mu}(x) + \int \psi^{\boldsymbol{\mu},\nu}(x) d\nu(y) \right) d\mathbb{P}(\nu) \\ &\geq \int \left(\int \phi^{\eta,\nu}(x) d\boldsymbol{\mu}(x) + \int \psi^{\eta,\nu}(x) d\nu(y) \right) d\mathbb{P}(\nu), \end{aligned}$$

we get that $\int \left(\int \phi^{\eta,\nu}(x) d\mu(x) \right) d\mathbb{P}(\nu) < +\infty$, so $\int_{\Omega} \int_{\mathcal{P}_2(\Omega)} \phi^{\eta,\nu} d\mathbb{P}(\nu) d\boldsymbol{\mu}(x) = \int_{\mathcal{P}_2(\Omega)} \int_{\Omega} \phi^{\eta,\nu} d\boldsymbol{\mu}(x) d\mathbb{P}(\nu)$. Therefore, by the dual formulation of Kantorovich, we have that

$$-\int \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i} d\boldsymbol{\mu}(x) = \int \psi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(y) d\boldsymbol{\nu}_i(y) - \iint |x-y|^2 d\pi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x,y) \tag{B.6}$$

$$-\int \phi^{\eta,\nu} d\eta(x) = \int \psi^{\eta,\nu}(y) d\nu(y) - \iint |x-y|^2 d\pi^{\eta,\nu}(x,y) \tag{B.7}$$

where $\pi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}$ and $\pi^{\eta,\nu}$ are optimal transport plans for the Wasserstein distance. Also, $\phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}$ and $\phi^{\eta,\nu}$ verify the Kantorovich condition, that is

$$\phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) \leq -\psi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(y) + |x-y|^2 \tag{B.8}$$

$$\phi^{\eta,\nu}(x) \leq -\psi^{\eta,\nu}(y) + |x-y|^2. \tag{B.9}$$

Next, the trick is to write $\int \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) d\eta(x) = \iint \phi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x) d\pi^{\eta,\boldsymbol{\nu}_i}(x,y)$ and $\int \phi^{\eta,\nu}(x) d\boldsymbol{\mu}(x) = \iint \phi^{\eta,\nu}(x) d\pi^{\boldsymbol{\mu},\nu}(x,y)$. Thus, by using the equalities (B.6), (B.7) and the inequalities (B.8), (B.9), the result (B.5) becomes

$$\begin{aligned} \gamma d_E(\boldsymbol{\mu}, \eta) \leq & -\frac{1}{n} \sum_{i=1}^{n} \iint |x-y|^2 d\pi^{\boldsymbol{\mu},\boldsymbol{\nu}_i}(x,y) + \frac{1}{n} \sum_{i=1}^{n} \iint |x-y|^2 d\pi^{\eta,\boldsymbol{\nu}_i}(x,y) \\ & + \iiint |x-y|^2 d\pi^{\boldsymbol{\mu},\nu}(x,y) d\mathbb{P}(\nu) - \iiint |x-y|^2 d\pi^{\eta,\nu}(x,y) d\mathbb{P}(\nu). \end{aligned}$$

We denote

$$S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}} := \iiint |x-y|^2 d\pi^{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n},\nu}(x,y)d\mathbb{P}(\nu) \; - \frac{1}{n}\sum_{i=1}^n \iint |x-y|^2 d\pi^{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n},\boldsymbol{\nu}_i}(x,y)$$

$$S^n_{\mu^\gamma_{\mathbb{P}}} := \frac{1}{n}\sum_{i=1}^n \iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}_i}(x,y) - \mathbb{E}\left(\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}}(x,y)\right),$$

and finally the last inequality writes

$$\gamma d_E(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}}) \leq S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}} + S^n_{\mu^\gamma_{\mathbb{P}}}. \tag{B.10}$$

Remark. *Since for $i = 1, \dots, n$ the random variables $\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}_i}(x,y)$ are independent and identically distributed. From the law of large numbers, we can notice that $S^n_{\mu^\gamma_{\mathbb{P}}} \longrightarrow 0$ almost surely when $n \to +\infty$.*

Taking the expectation with respect to the random measures, (B.10) implies

$$\gamma^2 \mathbb{E}(d^2_E(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}})) \leq 2\mathbb{E}(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2) + 2\mathbb{E}(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2). \tag{B.11}$$

• *Study of* $\mathbb{E}\left(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2\right)$. Using again the fact that $\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}_i}(x,y)$ are iid, we get

$$\mathbb{E}\left(|S^n_{\mu^\gamma_{\mathbb{P}}}|^2\right) = \frac{1}{n}\mathrm{Var}\left(\iint |x-y|^2 d\pi^{\mu^\gamma_{\mathbb{P}},\boldsymbol{\nu}}(x,y)\right) = \frac{C}{n}. \tag{B.12}$$

Note that since Ω is compact, the above variance term is finite.

• *Study of* $\mathbb{E}\left(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2\right)$. This term can be controlled thanks to the empirical process theory. Define the norm associated to the class of functions $\mathcal{H}$ (4.6) by $\|G\|_{\mathcal{H}} := \sup_{h\in\mathcal{H}} |G(h)|$ where $G : \mathcal{H} \to \mathbb{R}$. Recall that $h_\mu \in \mathcal{H}$ is the function $h_\mu : \nu \in \mathcal{P}_2(\Omega) \mapsto W_2^2(\mu,\nu)$, hence

$$\begin{aligned} S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}} &= \int_{\mathcal{P}_2(\Omega)} h_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}(\nu)d\mathbb{P}(\nu) - \int_{\mathcal{P}_2(\Omega)} h_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}(\nu)d\mathbb{P}_n(\nu) := (\mathbb{P}-\mathbb{P}_n)(h_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}) \\ &\leq \sup_{h\in\mathcal{H}} |(\mathbb{P}-\mathbb{P}_n)(h)| = \frac{1}{\sqrt{n}}\|\mathbb{G}_n\|_{\mathcal{H}} \end{aligned}$$

where $\mathbb{G}_n(h) = \sqrt{n}(\mathbb{P}_n - \mathbb{P})(h)$. We then obtain

$$\mathbb{E}\left(|S^n_{\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}}|^2\right) \leq \frac{1}{n}\mathbb{E}\left(\|\mathbb{G}_n\|^2_{\mathcal{H}}\right) = \frac{1}{n}\| \, \|\mathbb{G}_n\|_{\mathcal{H}}\|^2_{\mathbb{L}_2(\mathbb{P})}. \tag{B.13}$$

We finally use the following Theorem 2.14.1. of [VDVW96] to control this last expression:

Theorem B.2. *Let $\mathcal{H}$ be a Q-measurable class of measurable functions with measurable envelope function H. Then for $p \geq 1$,*

$$\| \, \|\mathbb{G}_n\|_{\mathcal{H}}\|_{\mathbb{L}_p(Q)} \leq CI(1,\mathcal{H})\|H\|_{\mathbb{L}_{2\vee p}(Q)} \tag{B.14}$$

with C a constant, $I(1,\mathcal{H})$ defined in (4.4) and H an envelope function.

Gathering the results of (B.11), (B.12), (B.13) and (B.14), we get

$$E(d^2_E\left(\boldsymbol{\mu}^\gamma_{\mathbb{P}_n}, \mu^\gamma_{\mathbb{P}}\right)) \leq \frac{1}{\gamma^2 n}\left(C + CI(1,\mathcal{H})\|H\|_{\mathbb{L}_2(\mathbb{P})}\right)$$

which concludes the proof. □

C Algorithmic details

In this section, we describe how the minimization problem

$$\min_{\mu} \frac{1}{n}\sum_{i=1}^{n} W_2^2(\mu,\nu_i) + \gamma E(\mu) \text{ over } \mu \in \mathcal{P}_2(\Omega), \tag{C.1}$$

can be solved numerically by using an appropriate discretization to compute a numerical approximation of a regularized Wasserstein barycenter.

More precisely, given a fixed grid $\{x^k\}_{k=1}^N$ of equally spaced points $x^k \in \mathbb{R}^d$, we approximate $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$ by the discrete measure $\mu_f = \sum_{k=1}^N f^k \delta_{x^k}$ where the f^k are positive weights summing up to one which minimize a discrete version of the optimisation problem (C.1). In what follows, we first describe an algorithm that is specific to the one-dimensional case, and then we propose another algorithm that is valid for any $d \geq 1$.

C.1 Discrete algorithm for $d = 1$ and data defined on the same grid

We first propose to compute a regularized empirical Wasserstein barycenter for a dataset made of discrete measures $\nu_1, \ldots, \nu_n$ (or one-dimensional histograms) defined on the same grid of reals $\{x^k\}_{k=1}^N$ that the one chosen to approximate $\boldsymbol{\mu}_{\mathbb{P}_n}^{\gamma}$. Since the grid is fixed, we identify a discrete measure ν with the vector of weights $\nu = (\nu(x^1), \ldots, \nu(x^N))$ in $\mathbb{R}_+^N$ (with entries that sum up to one) of its values on this grid.

The estimation of the regularized barycenter onto this grid can be formulated as:

$$\min_{f} \frac{1}{n}\sum_{i=1}^{n} W_2^2(f,\nu_i) + \gamma E(f) \text{ s.t } \sum_k f^k = 1, \text{ and } f^k = f(x^k) \geq 0, \tag{C.2}$$

with the obvious abuse of notation $W_2^2(f,\nu_i) = W_2^2(\mu_f,\nu_i)$ and $E(f) = E(\mu_f)$.

Then, to compute a minimizer of the convex optimization problem (C.2), we perform a subgradient descent. We denote by $(f^{(\ell)})_{\ell\geq 1}$ the resulting sequence of discretized regularized barycenters in $\mathbb{R}^N$ along the descent. Hence, given an initial value $f^{(1)} \in \mathbb{R}_+^N$ and for $\ell \geq 1$, we thus have

$$f^{(\ell+1)} = \Pi_S\left(f^{(\ell)} - \tau^{(\ell)}\left[\gamma \nabla E(f^{(\ell)}) + \frac{1}{n}\sum_{i=1}^{n} \nabla_1 W_2^2(f^{(\ell)},\nu_i)\right]\right) \tag{C.3}$$

where $\tau^{(\ell)}$ is the ℓ-th step time, and Π_S stands for the projection on the simplex $S = \{y \in \mathbb{R}_+^N \text{ such that } \sum_{j=1}^N y^j = 1\}$. Thanks to Proposition 5 in [PFR12], we are able to compute a sub-gradient of the squared Wasserstein distance $W_2^2(f^{(\ell)},\nu_i)$ with respect to its first argument (for discrete distributions). For that purpose, we denote by $R_f(s) = \sum_{x^j \leq s} f(x^j)$ the cdf of $\mu_f = \sum_{k=1}^N f(x^k)\delta_{x^k}$ and by $R_f^-(t) = \inf\{s \in \mathbb{R} : R_f(s) \geq t\}$ its pseudo-inverse.

Proposition C.1 ([PFR12])**.** *Let $f = (f(x^1), f(x^2), \ldots, f(x^N))$ and $\nu = (\nu(x^1), \nu(x^2), \ldots, \nu(x^N))$ be two discrete distributions defined on the same grid of values $x^1, \ldots, x^N$ in $\mathbb{R}$. For $p \geq 1$, the subgradients of $f \mapsto W_p^p(f,\nu)$ can be written as*

$$\nabla_1 W_p^p(f,\nu) : x_j \mapsto \sum_{m\geq j} |x^m - \tilde{x}^m|^p - |x^{m+1} - \tilde{x}^m|^p \tag{C.4}$$

where

$$\begin{cases} \tilde{x}^m = x^k \text{ if } R_g(x^{k-1}) < R_f(x^m) < R_\nu(x^k) \\ \tilde{x}^m \in [x^{k-1}, x^k] \text{ if } R_f(x^m) = R_\nu(x^k) \end{cases}$$

Even if subgradient descent is only shown to converge with diminishing time steps [BM07], we observed that using a small fixed step time (of order 10^{-5}) is sufficient to obtain in practice a convergence of the iterates $(f^{(\ell)})_{\ell \geq 1}$. Moreover, we have noticed that the principles of FISTA (Fast Iterative Soft Thresholding, see e.g. [BT09]) accelerate the speed of convergence of the above described algorithm.

C.2 Discrete algorithm for $d \geq 1$ in the general case

We assume that data $\nu_1, \ldots, \nu_n$ are given in the form of n discrete probability measures (histograms) supported on $\mathbb{R}^d$ (with $d \geq 1$) that are not necessarily defined on the same grid. More precisely, we assume that

$$\nu_i = \sum_{j=1}^{p_i} \nu_i^j \delta_{y_i^j}$$

for $1 \leq i \leq n$ where the y_i^j's are arbitrary locations in $\Omega \subset \mathbb{R}^d$, and the ν_i^j's are positive weights (summing up to one for each i).

The estimation of the regularized barycenter onto a given grid $\{x^k\}_{k=1}^N$ of $\mathbb{R}^d$ can then be formulated as the following minimization problem:

$$\min_f \frac{1}{n} \sum_{i=1}^n W_2^2(f, \nu_i) + \gamma E(f) \text{ s.t } \sum_k f^k = 1, \text{ and } f^k \geq 0, \tag{C.5}$$

with the notation $f = (f^1, f^2, \ldots, f^N)$ and the convention that $W_2^2(f, \nu_i)$ denotes the squared Wasserstein distance between $\mu_f = \sum_{k=1}^N f^k \delta_{x^k}$ and ν_i.

Problem (C.5) could be exactly solved by considering the discrete $p_i \times N$ transport matrices S_i between the barycenter μ_f to estimate and the data ν_i. Indeed, problem (C.5) is equivalent to the convex problem

$$\min_f \min_{S_1 \cdots S_n} \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^{p_i} \sum_{k=1}^N ||y_i^j - x^k||^2 S_i^{j,k} + \gamma E(f) \tag{C.6}$$

under the linear constraints

$$\forall i = 1, \ldots, n, \sum_{j=1}^{p_i} S_i^{j,k} = f^k, \sum_{k=1}^N S_i^{j,k} = \nu_i^j, \text{ and } S_i^{j,k} \geq 0.$$

However, optimizing over the $p_i \times N$ transport matrices S_i for $1 \leq i \leq n$ involves memory issues when using an accurate discretization grid $\{x^k\}_{k=1}^N$ with a large value of N. For this reason, we consider subgradient descent algorithms that allow dealing directly with problem (C.5).

To this end, we rely on the dual approach introduced in [COO15] and the numerical optimisation scheme proposed in [CP16]. Following these works, one can show that the dual problem of (C.5) with a regularization of the form $E(Kf)$ and K a discrete linear operator reads as

$$\min_{\phi_0, \cdots \phi_n} \sum_{i=1}^n H_{\nu_i}(\phi_i) + E_\gamma^*(\phi_0) \text{ s.t } K^T \phi_0 + \sum_{i=1}^n \phi_i = 0, \tag{C.7}$$

where the ϕ_i's are dual variables (vectors in $\mathbb{R}^N$) defined on the discrete grid $\{x^k\}_{k=1}^N$, E_γ^* is the Legendre transform of γE and $H_{\nu_i}(.)$ is the Legendre transform of $W_2^2(., \nu_i)$ that reads:

$$H_{\nu_i}(\phi_i) = \sum_{j=1}^{p_i} \nu_i^j \min_{k=1\cdots N} \left(\frac{1}{2}||y_i^j - x^k||^2 - \phi_i^k \right).$$

Barycenter estimations f_i can finally be recovered from the optimal dual variables ϕ_i solution of (C.7) as:

$$f_i \in \partial H_{\nu_i}(\phi_i), \text{ for } i = 1 \cdots n. \tag{C.8}$$

Following [COO15], one value of the above subgradient can be obtained at point x^k as:

$$\partial H_{\nu_i}(\phi_i)_k = \sum_{j=1}^{p_i} \nu_i^j S_i^{j,k}, \tag{C.9}$$

where $S_i^{j,k}$ is any row stochastic matrix of size $p_i \times N$ checking:

$$S_i^{j,k} \neq 0 \text{ iff } k \in \underset{k=1\cdots N}{argmin} \left(\frac{1}{2}||y_i^j - x^k||^2 - \phi_i^k \right).$$

From the previous expressions, we see that $f_i^k = \sum_{j=1}^{p_i} \nu_i^j S_i^{j,k}$ corresponds to the discrete pushforward of data ν_i with the transport matrix S_i with the associated cost:

$$H_{\nu_i}(\phi_i) = \sum_{j=1}^{p_i} \sum_{k=1}^{N} \left(\frac{1}{2}||y_i^j - x^k||^2 - \phi_i^k \right) S_i^{j,k} \nu_i^j.$$

Numerical optimization Following [CP16], the dual problem (C.7), can be simplified by removing one variable and thus discarding the linear constraint $K^T\phi_0 + \sum_{i=1}^n \phi_i = 0$. In order to inject the regularity given by ϕ_0 in all the reconstructed barycenters obtained by ϕ_i, $i = 1 \cdots n$, we modified the change of variables of [CP16] by setting $\psi_i = \phi_i + K^T\phi_0/n$ for $i = 1 \cdots n$ and $\psi_0 = \phi_0$, leading to $\sum_{i=1}^n \psi_i = 0$. One variable, say ψ_n, can then be directly obtained from the other ones. Observing that $\phi_n = -K^T\psi_0 - \sum_{i=1}^{n-1} \psi_i/n$, we thus obtain:

$$\min_{\psi_0, \cdots \psi_{n-1}} \sum_{i=1}^{n-1} H_{\nu_i}(\psi_i - K^T\psi_0/n) + H_{\nu_n}(-K^T\psi_0 - \sum_{i=1}^{n-1} \psi_i/n) + E_\gamma^*(\psi_0). \tag{C.10}$$

The subgradient (C.9) can then be used in a descent algorithm over the dual problem (C.10). For differentiable penalizers E, we consider the L-BFGS algorithm [ZBLN97, Bec11] that integrates a line search method (see e.g. [BV04]) to select the best time step $\tau^{(\ell)}$ at each iteration ℓ of the subgradient descent:

$$\left\{ \begin{array}{ll} \psi_0^{(\ell+1)} & = \psi_0^{(\ell)} - \tau^{(\ell)}(\nabla E_\gamma^*(\psi_0^{(\ell)}) + d_0^\ell) \\ \psi_i^{(\ell+1)} & = \psi_i^{(\ell)} - \tau^{(\ell)} d_i^\ell \qquad\qquad i = 1 \cdots n-1, \end{array} \right. \tag{C.11}$$

where:

$$\begin{array}{ll} d_0^\ell & = K\left(\partial H_{\nu_n}\left(-K^T\psi_0^{(\ell)}/n - \sum_{i=1}^{n-1} \psi_i^{(\ell)} \right) - \sum_{i=1}^{n-1} \partial H_{\nu_i}\left(\psi_i^{(\ell)} - K^T\psi_0^{(\ell)}/n \right) \right) \\ d_i^\ell & = \partial H_{\nu_i}\left(\psi_i^{(\ell)} - K^T\psi_0^{(\ell)}/n \right) - \partial H_{\nu_n}\left(-K^T\psi_0^{(\ell)}/n - \sum_{i=1}^{n-1} \psi_i^{(\ell)} \right). \end{array}$$

The barycenter is finally given by (C.8), taking $\phi_i = \psi_i - K^T\psi_0/n$. Even if we only treated differentiable functions E in the theoretical part of this paper, we can numerically consider non differentiable penalizers E, such as Total Variation ($K = \nabla$, $E = |.|_1$). In this case, we make use of the Fista algorithm. This just modifies the update of ψ_0 in (C.11), by changing the explicit scheme involving ∇E_γ^* onto an implicit one through the proximity operator of E_γ^*:

$$\psi_0^{(\ell+1)} = \mathrm{Prox}_{\tau^{(\ell)} E_\gamma^*}\left(\psi_0^{(\ell)} - \tau^{(\ell)} d_0^\ell\right) = \underset{\psi}{argmin}\ \frac{1}{2\tau^{(\ell)}}||\psi_0^{(\ell)} - \tau^{(\ell)} d_0^\ell - \psi||^2 + E_\gamma^*(\psi).$$

Algorithmic issues and stabilization As detailed in [COO15], the computation of one subgradient in (C.9) relies on the look for Euclidean nearest neighbors between vectors $(y_i^j, 0)$ and $(x^k, \sqrt{c - \phi_i^k})$, with $c = \max_k \phi_i^k$. Selecting only one nearest neighbor leads to bad numerical results in practice as subgradient descent may not be stable. For this reason, we considered the $K = 10$ nearest neighbors for each j to build the row stochastic matrices S_i at each iteration as: $S_i^{j,k} = w_i^{jk}/\sum_{k'} w_i^{jk'}$, with $w_i^{jk} = \exp(-(\frac{1}{2}\|y_i^j - x^k\|^2 - \phi_i^k)/\epsilon)$ if k is within the K nearest neighbors for j and data i and $w_i^{jk} = 0$ otherwise.

References

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

[2] C. D. Aliprantis and K. Border. *Infinite dimensional analysis: a Hitchhiker's guide*. Springer Science & Business Media, 2006.

[3] P. Álvarez-Esteban, E. del Barrio, J. Cuesta-Albertos, and C. Matrán. Wide consensus for parallelized inference. *arXiv e-prints 1511.05350*, 2015.

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

[5] F. Bauer and A. Munk. Optimal regularization for ill-posed problems in metric spaces. *Journal of Inverse and Ill-posed Problems*, 15(2):137–148, 2007.

[6] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[7] S. Becker. Matlab wrapper and c implementation of L-BFGS-B-C, 2011. https://github.com/stephenbeckr/L-BFGS-B-C.

[8] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

[9] S. Bobkov and M. Ledoux. *One-dimensional empirical measures, order statistics and Kantorovich transport distances*. To appear in the Memoirs of the Amer. Math. Soc., 2014. Available at http://perso.math.univ-toulouse.fr/ledoux/files/2013/11/Order.statistics.10.pdf.

[10] N. Bonneel, J. Rabin, G. Peyré, and H. Pfister. Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45, 2015.

[11] A. W. Bowman and A. Azzalini. *Applied smoothing techniques for data analysis : the kernel approach with S-Plus illustrations*. Clarendon Press ; Oxford University Press, 1997.

[12] S. Boyd and A. Mutapcic. Subgradient methods, Winter 2006-07. Notes for EE364 Standford University.

[13] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[14] A. Braides. A handbook of γ-convergence. *Handbook of Differential Equations: stationary partial differential equations*, 3:101–213, 2006.

[15] M. Burger, M. Franek, and C.-B. Schönlieb. Regularized regression and density estimation based on optimal transport. *Applied Mathematics Research eXpress*, 2012(2):209–253, 2012.

[16] L. A. Caffarelli. The regularity of mappings with a convex potential. *Journal of the American Mathematical Society*, 5(1):99–104, 1992.

[17] L. A. Caffarelli. Boundary regularity of maps with convex potentials–ii. *Annals of mathematics*, 144(3):453–496, 1996.

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

[19] F. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*, volume 264 of *Graduate Texts in Mathematics*. Springer - Verlag London, 2013.

[20] M. Cuturi and A. Doucet. Fast computation of Wasserstein barycenters. In *International Conference on Machine Learning 2014, JMLR W&CP*, volume 32, pages 685–693, 2014.

[21] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016.

[22] G. De Philippis and A. Figalli. The monge–ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[23] A. Dessein, N. Papadakis, and J.-L. Rouas. Regularized Optimal Transport and the Rot Mover's Distance. *ArXiv e-prints 1610.06447*, Oct. 2016.

[24] P. Dupuis and R. S. Ellis. *A weak convergence approach to the theory of large deviations*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, 1997. A Wiley-Interscience Publication.

[25] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[26] N. Fournier and A. Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707, Aug. 2015.

[27] D. Gervini. Independent component models for replicated point processes. *Spatial Statistics*, 18, Part B:474 – 488, 2016.

[28] Y. H. Kim and B. Pass. Wasserstein barycenters over Riemannian manifolds. *arXiv e-prints 1412.7726*, 2014.

[29] A. Kneip and K. J. Utikal. Inference for density families using functional principal component analysis. *J. Amer. Statist. Assoc.*, 96(454):519–542, 2001. With comments and a rejoinder by the authors.

[30] A. N. Kolmogorov and V. M. Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

[31] T. Le Gouic and J.-M. Loubes. Existence and Consistency of Wasserstein Barycenters. *Probability Theory and Related Fields*, pages 1–17, 2016.

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

[33] B. Pass. Optimal transportation with infinitely many marginals. *Journal of Functional Analysis*, 264(4):947–963, 2013.

[34] K. Petersen and H.-G. Müller. Functional data analysis for density functions by transformation to a hilbert space. *Annals of Statistics*, To be published, 2015.

[35] G. Peyré, J. Fadili, and J. Rabin. Wasserstein active contours. In *IEEE International Conference on Image Processing (ICIP)*, pages 2541–2544. IEEE, 2012.

[36] R. Rockafellar. *Conjugate duality and optimization*, volume 16. Siam, 1974.

[37] A. W. Van Der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[38] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

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

[40] 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, Nov. 2011.

[41] C. Zalinescu. *Convex analysis in general vector spaces*. World Scientific, 2002.

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

[43] C. Zhu, R. H. Byrd, P. Lu, and J. Nocedal. Algorithm 778: L-BFGS-B: Fortran subroutines for large-scale bound-constrained optimization. *ACM Trans. Math. Softw.*, 23(4):550–560, Dec. 1997.