Data-driven regularization of Wasserstein barycenters with an application to multivariate density registration

Jérémie Bigot*, Elsa Cazelles & Nicolas Papadakis

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

May 3, 2019

Abstract

We present a framework to simultaneously align and smooth data in the form of multiple point clouds sampled from unknown densities with support in a d-dimensional Euclidean space. This work is motivated by applications in bioinformatics where researchers aim to automatically homogenize large datasets to compare and analyze characteristics within a same cell population. Inconveniently, the information acquired is most certainly noisy due to mis-alignment caused by technical variations of the environment. To overcome this problem, we propose to register multiple point clouds by using the notion of regularized barycenters (or Fréchet mean) of a set of probability measures with respect to the Wasserstein metric. A first approach consists in penalizing a Wasserstein barycenter with a convex functional as recently proposed in [5]. A second strategy is to transform the Wasserstein metric itself into an entropy regularized transportation cost between probability measures as introduced in [12]. The main contribution of this work is to propose data-driven choices for the regularization parameters involved in each approach using the Goldenshluger-Lepski's principle. Simulated data sampled from Gaussian mixtures are used to illustrate each method, and an application to the analysis of flow cytometry data is finally proposed. This way of choosing of the regularization parameter for the Sinkhorn barycenter is also analyzed through the prism of an oracle inequality that relates the error made by such data-driven estimators to the one of an ideal estimator.

1 Introduction

1.1 Motivations

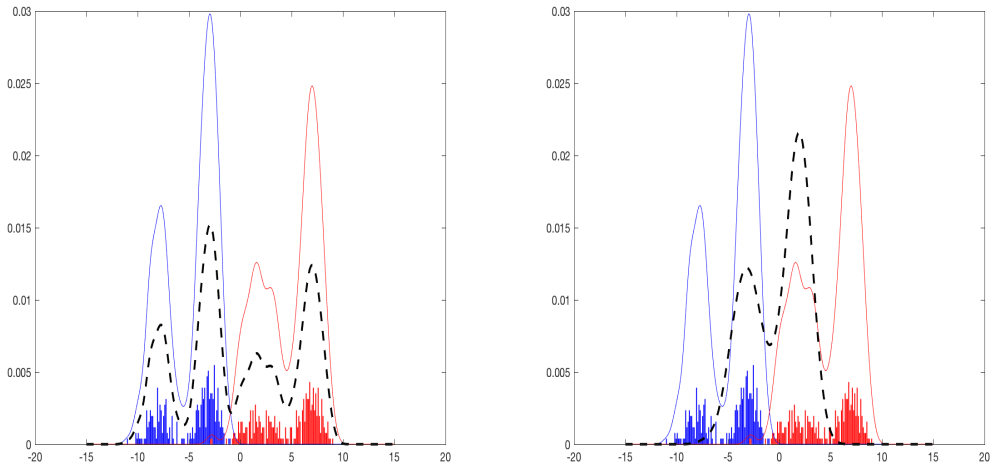
This paper is concerned with the problem of aligning (or registering) elements of a dataset that can be modeled as n random densities, or more generally, probability measures supported on $\mathbb{R}^d$. As raw data in the form of densities are generally not directly available, we focus on the setting where one has access to a set of random vectors $(X_{i,j})_{1\leq j\leq p_i;\, 1\leq i\leq n}$ in $\mathbb{R}^d$ organized in the form of n subjects (or multiple point clouds), such that $X_{i,1},\ldots,X_{i,p_i}$ are iid observations sampled from a random density $\boldsymbol{f}_i$ for each $1\leq i\leq n$. In presence of phase variation in the observations due to mis-alignment in the acquisition process, it is necessary to use a

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

registration step to obtain meaningful notions of mean and variance from the analysis of the dataset. In Figure 1(a), we display a simulated example of $n = 2$ random distributions made of observations sampled from random Gaussian mixtures $\boldsymbol{f}_i$. Certainly, one can estimate a mean density using a preliminary smoothing step (with a kernel K and data-driven choices of the bandwidth parameters $(h_i)_{i=1,\dots,n}$) followed by standard averaging, that is considering

$$\bar{f}_{n,p}(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{p_i h_i}\sum_{j=1}^{p_i} K\left(\frac{x - X_{i,j}}{h_i}\right),\ x \in \mathbb{R}^d. \tag{1.1}$$

Unfortunately this leads to an estimator which is not consistent with the shape of the $\boldsymbol{f}_i$'s. Indeed, the estimator $\bar{f}_{n,p}$ (Euclidean mean) has four modes due to mis-alignment of the data from different subjects.



(a) (b)

Figure 1: A simulated example of $n = 2$ subjects made of $p_1 = p_2 = 300$ observations sampled from Gaussian mixtures with random means and variances. The red and blue bar graphs are histograms with bins of equal and very small size to display the two sets of observations. The red and blue curves represent the kernel density estimators associated to each subject with data-driven choices (using cross-validation) of the bandwidths. (a) The dashed black curve is the Euclidean mean $\bar{f}_{n,p}$ of the red and blue densities. (b) The solid black curve is the entropy regularized Wasserstein barycenter $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ (defined in (1.5)) of the raw data using a Sinkhorn divergence and the numerical approach from [14], with a data-driven choice for $\hat{\varepsilon} = 2.55$.

The need to account for phase variability in the statistical analysis of such datasets is a well-known problem in various scientific fields. For the one-dimensional case ($d = 1$), examples can be found in biodemographic and genomics studies [35], economics [22], and in the analysis of spike trains in neuroscience [34] or functional connectivity between brain regions [27]. For $d \geq 2$ the issue of data registration arises in the statistical analysis of spatial point processes [18, 26] or flow cytometry data [20, 29].

1.2 Related works

In this work, in order to simultaneously align and smooth multiple point clouds (in the idea of recovering the underlying density function), we average the data using the notion of Wasserstein barycenter (as introduced in the seminal work [1]). Surely, this barycenter has been shown to be a relevant tool to account for phase variability in density registration [6, 25, 26]. A Wasserstein barycenter is a Fréchet mean [15] in the space $\mathcal{P}_2(\Omega)$ of probability measures with finite second moment supported on a convex domain $\Omega \subset \mathbb{R}^d$. It is endowed with the Wasserstein metric W_2 defined as

$$W_2(\mu,\nu) = \inf_{\pi\in\Gamma(\mu,\nu)} \left(\iint_{\Omega^2} |x-y|^2 d\pi(x,y) \right)^{1/2}, \qquad \text{for } \mu,\nu \in \mathcal{P}_2(\Omega),$$

where $\Gamma(\mu,\nu)$ is the set of probability measures on the product space $\Omega\times\Omega$ with respective marginals μ and ν, and $|\cdot|$ denotes the usual Euclidean norm on $\mathbb{R}^d$.

In the case of a finite space $\Omega_N = \{x_1,\dots,x_N\} \in (\mathbb{R}^d)^N$ of cardinal N, a discrete probability distribution r (with fixed support included in Ω_N) is identified by a vector in the simplex $\Sigma_N = \{r=(r_1,\dots,r_N)\in\mathbb{R}_+^N \text{ with } \sum_{k=1}^N r_k = 1\}$ such that $r = \sum_{k=1}^N r_k \delta_{x_k}$ where δ_x is the Dirac distribution at x. The Wasserstein distance between two discrete distributions r and q in Σ_N then becomes

$$W_2(r,q) := \min_{U\in U(r,q)} \langle C, U\rangle^{1/2},$$

where the set of couplings is defined as $U(r,q) := \{U\in\mathbb{R}_+^{N\times N} \text{ such that } U\mathbb{1}_N = r,\ U^T\mathbb{1}_N = q\}$ with $\mathbb{1}_N$ the N dimensional vector with all entries equal to 1 and C the cost matrix given by $C_{ml} = |x_m - x_l|^2$, for all $m,l\in\{1,\dots,N\}$.

In what follows, we consider two approaches for the computation of a regularized Wasserstein barycenter of n discrete probability measures given by

$$\hat{\boldsymbol{\nu}}_i^{p_i} = \frac{1}{p_i}\sum_{j=1}^{p_i} \delta_{X_{i,j}} \qquad \text{for} \quad 1\le i\le n, \tag{1.2}$$

from observations $(X_{i,j})_{1\le j\le p_i;\ 1\le i\le n}$.

1.2.1 Penalized Wasserstein barycenters

Adding a convex penalization term to the definition of an empirical Wasserstein barycenter [1] leads to the estimator

$$\hat{\boldsymbol{\mu}}_{n,p}^{\gamma} = \operatorname*{arg\,min}_{\mu\in\mathcal{P}_2(\Omega)} \frac{1}{n}\sum_{i=1}^{n} W_2^2(\mu, \hat{\boldsymbol{\nu}}_i^{p_i}) + \gamma E(\mu), \tag{1.3}$$

where $\gamma>0$ is a regularization parameter, and $E:\mathcal{P}_2(\Omega)\to\mathbb{R}_+$ is a smooth and convex penalty function which enforces the measure $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ to be absolutely continuous. Theoretical properties (such as existence and consistency) of the penalized Wasserstein barycenter $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ have been considered in [5]. In this paper, we discuss the choice of the penalty function E, as well as the numerical computation of $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ (using an appropriate discretization of Ω and a binning of the data), and its benefits for statistical data analysis.

Remark 1.1. *Note that the restriction of the minimization in (1.3) to the set $\mathcal{P}_2(\Omega)$ instead of the whole Wasserstein space $\mathcal{P}_2(\mathbb{R}^d)$ is inconsequential.*

1.2.2 Fréchet mean with respect to a Sinkhorn divergence

Another way to regularize an empirical Wasserstein barycenter is to use the notion of entropy regularized optimal transportation [12, 11] leading to the so-called Sinkhorn divergence

$$W_{2,\varepsilon}^2(\mu,\nu) = \inf_{\pi\in\Gamma(\mu,\nu)} \iint_{\Omega^2} |x-y|^2 d\pi(x,y) - \varepsilon \mathrm{h}(\pi), \tag{1.4}$$

where $\varepsilon > 0$ is a regularization parameter, and h stands for the (negative) entropy of the transport plan π with respect to the Lebesgue measure on $\Omega \times \Omega$. A regularized Wasserstein barycenter [13, 14] is then obtained by considering the estimator

$$\hat{\boldsymbol{r}}_{n,p}^{\varepsilon} = \arg\min_{\mu\in\mathcal{P}_2(\Omega)} \frac{1}{n}\sum_{i=1}^{n} W_{2,\varepsilon}^2(\mu, \hat{\nu}_i^{p_i}), \tag{1.5}$$

that can be interpreted as a Fréchet mean with respect to a Sinkhorn divergence and that we call Sinkhorn barycenter.

1.3 Contributions

The selection of the regularisation parameters γ or ϵ is the main issue for computing adequate penalized or Sinkhorn barycenters in practice. In this paper, we rely on the Goldenshluger-Lepski (GL) principle in order to perform an automatic calibration of such parameters.

1.3.1 Data-driven choice of the regularizing parameters

The main contribution in this paper is to propose a data-driven choice for the regularization parameters γ in (1.3) and ε in (1.5) using the Goldenshluger-Lepski (GL) method (as formulated in [23]), which leans on a bias-variance trade-off function, described in details in Section 4. The method consists in comparing estimators pairwise, for a given range of regularization parameters, with respect to a given loss function. It provides an optimal regularization parameter that minimizes a bias-variance trade-off function. We displayed in Figure 2 this functional for the dataset of Figure 1, which leads to an optimal (in the sense of GL's strategy) parameter choice $\hat{\varepsilon} = 2.55$. The entropy regularized Wasserstein barycenter in Figure 1(b) is thus chosen accordingly.

From the results on simulated data displayed in Figure 1(b), it is clear that computing the regularized Wasserstein barycenter $\hat{\boldsymbol{r}}_{n,p}^{\varepsilon}$ (with an appropriate choice for ε) leads to the estimation of mean density whose shape is consistent with the distribution of the data for each subject. In some sense, the regularization parameters γ and ε may also be interpreted as the usual bandwidth parameter in kernel density estimation, and their choice greatly influences the shape of the estimators $\hat{\boldsymbol{\mu}}_{n,p}^{\gamma}$ and $\hat{\boldsymbol{r}}_{n,p}^{\varepsilon}$ (see Figure 6 and Figure 8 in Section 5).

To choose the optimal parameter, the GL's strategy requires an upper bound on the decay to zero of the expected $\mathbb{L}_2(\Omega)$ distance between a regularized empirical barycenter (computed from the data) and its population counterpart. For penalized barycenters (1.3), adequate bounds have already been provided in [5].

1.3.2 Variance of Sinkhorn estimators

To the best of our knowledge, the automatic selection of ε in the definition of a Sinkhorn divergence has not been considered so far. Another main contribution of this work then

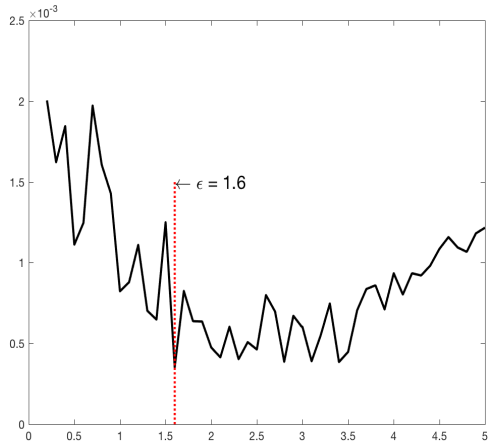


Figure 2: The GL's trade-off function associated to the entropy regularized Wassertein barycenters of the dataset in Figure 1, for ε ranging from 0.2 to 5

consists in derivating upper bounds on the variance for the estimators $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ which explicitly depends on ε, the number n of measures, the number $p = \min_{1\leq i\leq n} p_i$ of observations per measures and the size of their support. Such bounds therefore make possible the application of the GL's strategy.

1.3.3 Theoretical analysis of the GL's strategy for Sinkhorn barycenters

For Sinkhorn barycenters, we show that the GL's principle leads to a data-driven choice $\hat{\varepsilon}$ of the regularization parameter which allows to obtain an estimator $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ that satisfies an oracle inequality implying an optimal trade-off of this estimator between bias and variance terms among a collection of regularization parameters.

1.3.4 Computation issues: binning of the data and discretization of Ω

In our numerical experiments we consider algorithms for computing regularized barycenters from a set of discrete measures (or histograms) defined on possibly different grids of points of $\mathbb{R}^d$ (or different partitions). They are numerical approximations of the regularized Wasserstein barycenters $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ and $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ by a discrete measure of the form $\sum_{k=1}^{N} w_k \delta_{x_k}$ using a fixed grid $\Omega_N = \{x_1, \ldots, x_N\}$ of N equally spaced points $x_k \in \mathbb{R}^d$ (bin locations). For simplicity, we adopt a binning of the data (1.2) on the same grid, leading to a dataset of discrete measures (with supports included in Ω_N) that we denote

$$\tilde{\boldsymbol{q}}^{p_i}_i = \frac{1}{p_i}\sum_{j=1}^{p_i} \delta_{\tilde{X}_{i,j}}, \text{ where } \tilde{X}_{i,j} = \operatorname*{arg\,min}_{x\in\Omega_N} |x - X_{i,j}|, \tag{1.6}$$

for $1 \leq i \leq n$. In this paper, we rely on the smooth dual approach proposed in [14] to compute penalized and Sinkhorn barycenters on a grid of equi-spaced points in Ω (after a proper binning of the data).

Binning (i.e. choosing the grid Ω_N) surely incorporates some sort of additional regularization. A discussion on the influence of the grid size N on the smoothness of the barycenter

is proposed in Section 4 where we describe the GL's strategy. In our simulations, the choice of N is mainly guided by numerical issues on the computational cost of the algorithms used to approximate $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ and $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$.

1.3.5 Registration of flow cytometry data

In biotechnology, flow cytometry is a high-throughput technique that can measure a large number of surface and intracellular markers of single cell in a biological sample. With this technique, one can assess individual characteristics (in the form of multivariate data) at a cellular level to determine the type of cell, their functionality and the way they differ. At the beginning of flow cytometry, the analysis of such data was performed manually by visually separating regions or gates of interest on a series of sequential bivariate projection of the data, a process known as gating. However, the development of this technology now leads to datasets made of multiple measurements (e.g. up to 18) of millions of individuals cells. A significant amount of work has thus been carried out in recent years to propose automatic statistical methods to overcome the limitations of manual gating (see e.g. [20, 21, 24, 29] and references therein).

When analyzing samples in cytometry measured from different patients, a critical issue is data registration across patients. As carefully explained in [20], the alignment of flow cytometry data is a preprocessing step which aims at removing effects coming from technological issues in the acquisition of the data rather than significant biological differences. In this paper, we use data analyzed in [20] that are obtained from a renal transplant retrospective study conducted by the Immune Tolerance Network (ITN). This dataset is freely available from the `flowStats` package of Bioconductor [17] that can be downloaded from `http://bioconductor.org/packages/release/bioc/html/flowStats.html`. It consists of samples from 15 patients.

After an appropriate scaling trough an arcsinh transformation and an initial gating on total lymphocytes to remove artefacts, we focus our analysis on the cell markers FSC (forward-scattered light) and SSC (side-scattered light) which are of interest to measure the volume and morphological complexity of cells. The number of considered cells by patient varies from 88 to 2185. The resulting dataset is displayed in Figure 3. It clearly shows a mis-alignment issue between measurements from different patients.

The last contribution of the paper is thus to demonstrate the usefulness of regularized Wasserstein barycenters to correct mis-alignment effects in the analysis of data produced by flow cytometers.

1.4 Organization of the paper

The analysis of the variance of the regularized Wasserstein barycenters $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ and $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ are detailed in Section 2 and Section 3 respectively. Section 4 contains a description of the Goldenshluger-Lepski principle to choose the regularization parameters γ and ε as well as an oracle inequality justifying this technique for Sinkhorn barycenters. Section 5 finally reports the results from numerical experiments using simulated data and the flow cytometry dataset displayed in Figure 3. We conclude the paper in Section 6 by a brief discussion. Some proofs and technical results are presented in Appendix A, Appendix B, and Appendix C. Algorithmic details on the computation of the estimator $\hat{\boldsymbol{\mu}}^{\gamma}_{n,p}$ are gathered in Appendix D.

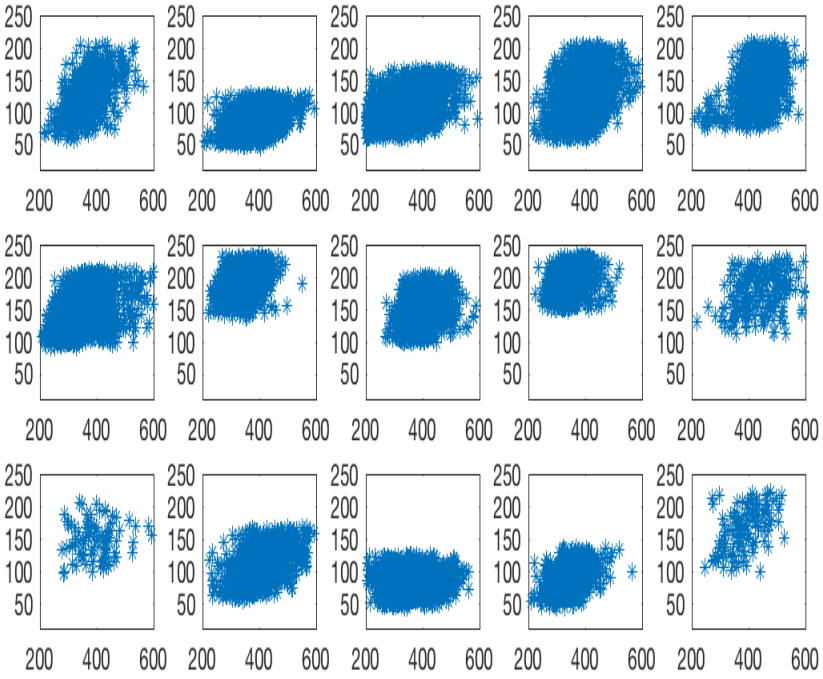


Figure 3: Example of flow cytometry data measured from $n = 15$ patients (restricted to a bivariate projection). The horizontal axis (resp. vertical axis) represent the values of the FSC (resp. SSC) cell marker.

2 Penalized Wasserstein barycenters

In this section, we adopt the framework from [5] in which the Wasserstein barycenter is regularized through a convex penalty function as presented in (1.3).

2.1 Minimization problem and variance properties

Let us remind some of the basic definitions in [5]. We define $W_2(\mathcal{P}_2(\Omega))$ as the space of distributions $\mathbb{P}$ supported on $\mathcal{P}_2(\Omega)$. The penalized Wasserstein barycenter associated to the distribution $\mathbb{P}$ is defined as a solution of the minimization problem

$$\min_{\mu\in\mathcal{P}_2(\Omega)} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu,\nu)d\mathbb{P}(\nu) + \gamma E(\mu)$$

where $\gamma > 0$ is a penalization parameter and the penalty function writes

$$E(\mu) = \begin{cases} \|f\|^2_{H^k(\Omega)}, & \text{if } f = \frac{d\mu}{dx} \text{ and } f \geq \alpha, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.1}$$

where $\|\cdot\|_{H^k(\Omega)}$ denotes the Sobolev norm associated to the $\mathbb{L}^2(\Omega)$ space, $\alpha > 0$ is arbitrarily small and $k > d - 1$. Remark that the function E is strictly convex on its domain

$$\mathcal{D}(E) = \{\mu \in \mathcal{P}_2(\Omega) \text{ such that } E(\mu) < +\infty\}. \tag{2.2}$$

This choice is supported by the discussion in Section 5 of [5], that in particular imposes the barycenter μ to belong to $\mathcal{P}_2^{ac}(\Omega)$, the space of measures in $\mathcal{P}_2(\Omega)$ that are absolutely continuous. It is mainly driven by the need to retrieve an absolutely continuous measure from discrete observations (X_{ij}), as it is often done when approximating data through kernel smoothing in density estimation. Others examples of penalty functions are given in [5], including a class of relative G-functional described in Section 9.4 in [2].

Definition 2.1. Let $\boldsymbol{\nu}_1, \dots, \boldsymbol{\nu}_n \in \mathcal{P}_2(\Omega)$ be iid measures with distribution $\mathbb{P}$ such that $\mathbb{P}(\mathcal{P}_2^{ac}(\Omega)) > 0$. The empirical probability measure defined by $(\boldsymbol{\nu}_i)_{i=1,\dots,n}$ writes $\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{\nu}_i}$. Let us then consider the random measures $(\hat{\boldsymbol{\nu}}_i^{p_i})_{1\leq i\leq n}$ of the form $\hat{\boldsymbol{\nu}}_i^{p_i} = \frac{1}{p_i}\sum_{j=1}^{p_i}\delta_{X_{i,j}}$ where $(X_{i,j})_{1\leq j\leq p_i}$ are iid random variables of law $\boldsymbol{\nu}_i$. From there on we define for $\gamma > 0$ the barycenters:

$$\mu^\gamma = \underset{\mu\in\mathcal{P}_2(\Omega)}{\arg\min} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu,\nu)d\mathbb{P}(\nu) + \gamma E(\mu) = \mathbb{E}_{\boldsymbol{\nu}\sim\mathbb{P}}[W_2^2(\mu,\boldsymbol{\nu})] + \gamma E(\mu) \tag{2.3}$$

$$\hat{\boldsymbol{\mu}}_{n,p}^\gamma = \underset{\mu\in\mathcal{P}_2(\Omega)}{\arg\min} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu,\nu)d\mathbb{P}_n(\nu) + \gamma E(\mu) = \frac{1}{n}\sum_{i=1}^n W_2^2(\mu,\hat{\boldsymbol{\nu}}_i^{p_i}) + \gamma E(\mu) \tag{2.4}$$

called respectively penalized population barycenter (2.3) and penalized empirical barycenter (2.4).

In [5], the existence and uniqueness of these barycenters have been shown for $\gamma > 0$. This holds true for measures defined on $\mathcal{P}_2(\mathbb{R}^d)$. By penalizing the barycenters with function E as in (2.1), we enforce them to be absolutely continuous. Therefore, let $\hat{\boldsymbol{f}}_{n,p}^\gamma$ and $f_{\mathbb{P}}^\gamma$ be the densities associated to $\hat{\boldsymbol{\mu}}_{n,p}^\gamma$ and $\mu_{\mathbb{P}}^\gamma$.

In order to apply the GL's strategy, we now study the expected squared $\mathbb{L}_2(\Omega)$-distance $\mathbb{E}(\|\hat{\boldsymbol{f}}_{n,p}^\gamma - f_{\mathbb{P}}^\gamma\|^2)$ that will be referred to as a "variance term" for $\hat{\boldsymbol{\mu}}_{n,p}^\gamma$.

Theorem 2.2 (Section 5 in [5])**.** *Let $\Omega \subset \mathbb{R}^d$ be compact and $\hat{\boldsymbol{f}}_{n,p}^\gamma$ and $f_{\mathbb{P}}^\gamma$ be the density functions of $\hat{\boldsymbol{\mu}}_{n,p}^\gamma$ and $\mu_{\mathbb{P}}^\gamma$, induced by the choice (2.1) of the penalty function E. Then, provided that $d < 4$, there exists a constant $c > 0$ depending only on Ω such that*

$$\mathbb{E}\left(\|\hat{\boldsymbol{f}}_{n,p}^\gamma - f_{\mathbb{P}}^\gamma\|_{\mathbb{L}_2(\Omega)}^2\right) \leq c\left(\frac{1}{\gamma p^{1/4}} + \frac{1}{\gamma n^{1/2}}\right) \tag{2.5}$$

where $p = \min_{1\leq i\leq n} p_i$.

Thanks to this result, we will be able to automatically calibrate the parameter $\gamma > 0$ by following the GL's parameter selection strategy described in Section 4.

Remark 2.1. *As already mentioned above, in practice, we discretize Ω into a sufficiently fine and fixed grid on which we compute a discretized version the penalized Wasserstein barycenter $\hat{\boldsymbol{f}}_{n,p}^\gamma$. Therefore, the upper bound (2.5) should be considered with some caution as it is not exactly a control of the variance of the discretized estimator that is used in our numerical experiments. Moreover, the upper bound (2.5) involves a constant $c > 0$ whose derivation is guided by theoretical arguments. However, a good calibration of c is of primary importance as discussed in the numerical experiments reported in Section 5.*

2.2 Numerical computation

As a practical complement to [5], we provide in Appendix D efficient minimization algorithms for the computation of $\hat{\boldsymbol{f}}^{\gamma}_{n,p}$, after a binning of the data on a fixed grid Ω_N. For Ω included in the real line, a simple subgradient descent is considered. When data are histograms supported on $\mathbb{R}^d$, $d \geq 2$, we rely on a smooth dual approach based on the work of [14].

3 Sinkhorn barycenters via entropy regularized optimal transport

In this section, we analyze the variance of the Sinkhorn barycenter defined in (1.5).

3.1 Variance properties of the Sinkhorn barycenters

As before we consider a binning of the data on a fixed and finite discrete grid Ω_N. For two discrete measures $r, q \in \Sigma_N$, the Sinkhorn divergence (1.4) reads for $\varepsilon > 0$

$$W^2_{2,\varepsilon}(r,q) := \min_{U \in U(r,q)} \langle C, U \rangle - \varepsilon \mathrm{h}(U). \tag{3.1}$$

where the discrete (negative) entropy for a given coupling $U \in U(r,q)$ is given by $\mathrm{h}(U) := -\sum_{m,\ell} U_{m\ell} \log U_{m\ell}$. We shall then use two key properties to analyze the variance of Sinkhorn barycenters which are the strong convexity (see Theorem 3.4 below) and the Lipschitz continuity (see Lemma 3.5 below) of the mapping $r \mapsto W^2_{2,\varepsilon}(r,q)$ (for a given $q \in \Sigma_N$).

However, to guarantee the Lipschitz continuity of this mapping, it is necessary to restrict the analysis to discrete measures r belonging to the convex set

$$\Sigma^{\rho}_N = \left\{ r \in \Sigma_N \ : \ \min_{1 \leq \ell \leq N} r_\ell \geq \rho \right\},$$

where $0 < \rho < 1$ is an arbitrarily small constant. This means that our theoretical results on the variance of the Sinkhorn barycenters hold for discrete measures with non-vanishing entries. Nevertheless, we obtain upper bounds on these variances which depend explicitly on the constant ρ, allowing to discuss its choice.

Then, as it has been done for the penalized barycenters in Definition 2.1, we introduce the definitions of empirical and population Sinkhorn barycenters.

Definition 3.1. Let $0 < \rho < 1/N$, and $\mathbb{P}$ be a probability distribution on Σ^{ρ}_N. Let $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n \in \Sigma^{\rho}_N$ be an iid sample drawn from the distribution $\mathbb{P}$. For each $1 \leq i \leq n$, we assume that $(\tilde{X}_{i,j})_{1 \leq j \leq p_i}$ are iid random variables sampled from $\boldsymbol{q}_i$. For each $1 \leq i \leq n$, let us define the following discrete measures

$$\tilde{\boldsymbol{q}}^{p_i}_i = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\tilde{X}_{i,j}} \quad \text{and} \quad \hat{\boldsymbol{q}}^{p_i}_i = (1 - \rho N) \tilde{\boldsymbol{q}}^{p_i}_i + \rho \mathbb{1}_N,$$

where $\mathbb{1}_N$ is the vector of $\mathbb{R}^N$ with all entries equal to one. Thanks to the condition $0 < \rho < 1/N$, it follows that $\hat{\boldsymbol{q}}^{p_i}_i \in \Sigma^{\rho}_N$ for all $1 \leq i \leq n$, which may not be the case for some

$\tilde{\boldsymbol{q}}_i^{p_i}$, $1 = 1, \ldots, n$. Then, we define

$$r^\varepsilon = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \mathbb{E}_{\boldsymbol{q} \sim \mathbb{P}}[W_{2,\varepsilon}^2(r, \boldsymbol{q})] \qquad \text{the population Sinkhorn barycenter} \tag{3.2}$$

$$\hat{\boldsymbol{r}}_{n,p}^\varepsilon = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \frac{1}{n} \sum_{i=1}^n W_{2,\varepsilon}^2(r, \hat{\boldsymbol{q}}_i^{p_i}) \qquad \text{the empirical Sinkhorn barycenter} \tag{3.3}$$

In the optimisation problem (3.3), we choose to use the discrete measures $\hat{\boldsymbol{q}}_i^{p_i}$ instead of the empirical measures $\tilde{\boldsymbol{q}}_i^{p_i}$ to guarantee the use of discrete measures belonging to Σ_N^ρ in the definition of the empirical Sinkhorn barycenter $\hat{\boldsymbol{r}}_{n,p}^\varepsilon$.

Remark 3.1. *The population and empirical Sinkhorn barycenters in Definition 3.1 are constrained to belong to the set Σ_N^ρ so that the Lipschitz continuity of $r \mapsto W_{2,\varepsilon}^2(r, q)$ holds true.*

The following theorem is the main result of this section which gives an upper bound on $\mathbb{E}(|r^\varepsilon - \hat{r}_{n,p}^\varepsilon|^2)$ which will be referred to as the variance of $\hat{\boldsymbol{r}}_{n,p}^\varepsilon$ in what follows. In particular, the asymptotic regime in which we are interested in is the number n of measures defining the barycenter as well as the number p of observations per measures.

Theorem 3.2. *Recall that $p = \min_{1 \le i \le n} p_i$ and let $\varepsilon > 0$. Then, one has that*

$$\mathbb{E}(|r^\varepsilon - \hat{r}_{n,p}^\varepsilon|^2) \le \frac{16 L_{\rho,\varepsilon}^2}{\varepsilon^2 n} + \frac{2 L_{\rho,\varepsilon}}{\varepsilon} \left(\sqrt{\frac{N}{p}} + 2\rho(N + \sqrt{N}) \right), \tag{3.4}$$

with

$$L_{\rho,\varepsilon} = \left(\sum_{1 \le m \le N} \left(2\varepsilon \log(N) + \sup_{1 \le \ell, k \le N} |C_{m\ell} - C_{k\ell}| - \varepsilon \log(\rho) \right)^2 \right)^{1/2}. \tag{3.5}$$

A few remarks can be made about the above result. The bound in the right-hand side of (3.4) explicitly depends on the size N of the grid. This will be taken into account for the choice of the optimal parameter $\hat{\varepsilon}$ (see Section 4). Moreover, it can be used to discuss the choice of ρ. First, if one takes $\rho = \epsilon^\kappa$, the Lipschitz constant (Lemma 3.5) $L_{\rho,\varepsilon} = L_\varepsilon$ becomes

$$L_\varepsilon = \left(\sum_{1 \le m \le N} \left(\varepsilon(2\log(N) - \kappa \log(\epsilon)) + \sup_{1 \le \ell, k \le N} |C_{m\ell} - C_{k\ell}| \right)^2 \right)^{1/2},$$

which is a constant (not depending on ρ) such that

$$\lim_{\epsilon \to 0} L_\varepsilon = \left(\sum_{1 \le m \le N} \left(\sup_{1 \le \ell, k \le N} |C_{m\ell} - C_{k\ell}| \right)^2 \right)^{1/2}.$$

If we further assume that $\rho = \epsilon^\kappa < \min(1/N, 1/p)$ we obtain the upper bound

$$\mathbb{E}(|r^\varepsilon - \hat{r}_{n,p}^\varepsilon|^2) \le \frac{16 L_\varepsilon^2}{\varepsilon^2 n} + \frac{2 L_\varepsilon}{\varepsilon} \left(\sqrt{\frac{N}{p}} + 2 \left(\frac{N}{p} + \sqrt{\frac{N}{p^2}} \right) \right). \tag{3.6}$$

Therefore, choosing $\varepsilon = \varepsilon(n,p) \underset{\min(n,p) \to \infty}{\longrightarrow}$ such that $1/\varepsilon^2 n \to \infty$ and $1/\varepsilon p \to \infty$, we have that $\mathbb{E}(|r^\varepsilon - \hat{r}_{n,p}^\varepsilon|^2) \underset{n,p \to \infty}{\longrightarrow}$ (take for example $\varepsilon = 1/n^\alpha p^\beta$ with $0 < \alpha < 1/2$ and $0 < \beta < 1$.

Finally, it should be remarked that Theorem 3.2 holds for general cost matrices C that are symmetric and non-negative.

3.2 Proof of Theorem 3.2

The proof of the upper bound (3.4) relies on strong convexity of the functional $r \mapsto W_{2,\varepsilon}^2(r,q)$ for $q \in \Sigma_N$, without constraint on its entries. This property can be derived by studying the Legendre transform of $r \mapsto W_{2,\varepsilon}^2(r,q)$. For a fixed distribution $q \in \Sigma_N$, using the notation in [14], we define the function

$$H_q(r) := W_{2,\varepsilon}^2(r,q), \quad \text{for all } r \in \Sigma_N.$$

Its Legendre transform is given for $g \in \mathbb{R}^N$ by $H_q^*(g) = \max_{r \in \Sigma_N} \langle g, r\rangle - H_q(r)$ and its differentiation properties are presented in the following theorem.

Theorem 3.3 (Theorem 2.4 in [14]). *For $\varepsilon > 0$, the Legendre dual function H_q^* is C^∞. Its gradient function ∇H_q^* is $1/\varepsilon$-Lipschitz. Its value, gradient and Hessian at $g \in \mathbb{R}^N$ are, writing $\alpha = \exp(g/\varepsilon)$ and $K = \exp(-C/\varepsilon)$,*

$$H_q^*(g) = \varepsilon(E(q) + \langle q, \log(K\alpha)\rangle), \ \nabla H_q^*(g) = \operatorname{diag}(\alpha) K \frac{q}{K\alpha} \in \Sigma_N$$
$$\nabla^2 H_q^*(g) = \frac{1}{\varepsilon}\left(\operatorname{diag}\left(\operatorname{diag}(\alpha) K \frac{q}{K\alpha}\right)\right) - \frac{1}{\varepsilon}\operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2}\right) K \operatorname{diag}(\alpha),$$

where the notation $\frac{q}{r}$ stands for the component-wise division of the entries of q and r.

From this result, we can deduce the strong convexity of the dual functional H_q as stated below.

Theorem 3.4. *Let $\varepsilon > 0$. Then, for any $q \in \Sigma_N$, the function H_q is ε-strongly convex for the Euclidean 2-norm.*

The proof of Theorem 3.4 is deferred to Appendix A. We can also ensure the Lipschitz continuity of $H_q(r)$, when restricting our analysis to the set $r \in \Sigma_N^\rho$.

Lemma 3.5. *Let $q \in \Sigma_N$ and $0 < \rho < 1$. Then, one has that $r \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ with $L_{\rho,\varepsilon}$ defined in (3.5).*

The proof of this Lemma is given in Appendix B.

We can now proceed to the proof of Theorem 3.2. Let us introduce the following Sinkhorn barycenter

$$\boldsymbol{r}_n^\varepsilon = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \frac{1}{n}\sum_{i=1}^n W_{2,\varepsilon}^2(r, \boldsymbol{q}_i) = \operatorname*{arg\,min}_{r \in \Sigma_N^\rho} \frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(r), \tag{3.7}$$

of the iid random measures $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$sampled from the distribution $\mathbb{P}$ supported on Σ_N^ρ. By the triangle inequality, we have that

$$\mathbb{E}(|r^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \le 2\mathbb{E}(|r^\varepsilon - \boldsymbol{r}_n^\varepsilon|^2) + 2\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2). \tag{3.8}$$

To control the first term of the right hand side of the above inequality, we use that (for any $q \in \Sigma_N$) $r \mapsto H_q(r)$ is ε-strongly convex by Theorem 3.4 and $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ by Lemma 3.5 where $L_{\rho,\varepsilon}$ is the constant defined by equation (3.5). Under these assumptions, it follows from the arguments in the proof of Theorem 6 in [32] that

$$\mathbb{E}(|r^\varepsilon - \boldsymbol{r}_n^\varepsilon|^2) \le \frac{8L_{\rho,\varepsilon}^2}{\varepsilon^2 n}. \tag{3.9}$$

The strong convexity of H_q has a major role here as it brings out the distance between the empirical minimizer $\boldsymbol{r}_n^\varepsilon$ and any other point in Σ_N. For the second term in the right hand side of (3.8), we obtain by the strong convexity of H_q that

$$\frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\hat{\boldsymbol{r}}_{n,p}^\varepsilon) \geq \frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\boldsymbol{r}_n^\varepsilon) + \frac{1}{n}\sum_{i=1}^n \nabla H_{\boldsymbol{q}_i}(\boldsymbol{r}_n^\varepsilon)^T(\hat{\boldsymbol{r}}_{n,p}^\varepsilon - \boldsymbol{r}_n^\varepsilon) + \frac{\varepsilon}{2}|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2.$$

Theorem 3.1 in [14] ensures that $\frac{1}{n}\sum_i \nabla H_{q_i}(\boldsymbol{r}_n^\varepsilon) = 0$. The same inequality also holds for the terms $H_{\hat{\boldsymbol{q}}_i^{p_i}}$, and we therefore have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\hat{\boldsymbol{r}}_{n,p}^\varepsilon) &\geq \frac{1}{n}\sum_{i=1}^n H_{\boldsymbol{q}_i}(\boldsymbol{r}_n^\varepsilon) + \frac{\varepsilon}{2}|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2,\\
\frac{1}{n}\sum_{i=1}^n H_{\hat{\boldsymbol{q}}_i^{p_i}}(\boldsymbol{r}_n^\varepsilon) &\geq \frac{1}{n}\sum_{i=1}^n H_{\hat{\boldsymbol{q}}_i^{p_i}}(\hat{\boldsymbol{r}}_{n,p}^\varepsilon) + \frac{\varepsilon}{2}|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2.
\end{aligned}$$

Using the symmetry of the Sinkhorn divergence, Lemma 3.5 also implies that the mapping $q \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ for any discrete distribution r. Hence, by summing the two above inequalities, and by taking the expectation on both sides, we obtain that

$$\varepsilon\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leq \frac{2L_{\rho,\varepsilon}}{n}\sum_{i=1}^n \mathbb{E}(|\boldsymbol{q}_i - \hat{\boldsymbol{q}}_i^{p_i}|).$$

Using the inequalities

$$|\boldsymbol{q}_i - \hat{\boldsymbol{q}}_i^{p_i}| \leq |\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| + \rho N|\tilde{\boldsymbol{q}}_i^{p_i}| + \rho|\mathbb{1}_N| \leq |\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| + \rho(N + \sqrt{N}),$$

we finally have that

$$\varepsilon\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leq 2L_{\rho,\varepsilon}\left(\frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}(|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}|^2)} + \rho(N + \sqrt{N})\right). \tag{3.10}$$

Conditionally on $\boldsymbol{q}_i$, one has that $p_i\tilde{\boldsymbol{q}}_i^{p_i}$ is a random vector following a multinomial distribution $\mathcal{M}(p_i, \boldsymbol{q}_i)$. Hence, for each $1 \leq k \leq N$, denoting $\boldsymbol{q}_{i,k}$ (resp. $\tilde{\boldsymbol{q}}_{i,k}^{p_i}$) the k-th coordinate of $\boldsymbol{q}_i$ (resp. $\tilde{\boldsymbol{q}}_i^{p_i}$), one has that

$$\mathbb{E}\left(\tilde{\boldsymbol{q}}_{i,k}^{p_i}|\boldsymbol{q}_i\right) = \boldsymbol{q}_{i,k} \quad \text{and} \quad \mathbb{E}\left[\left(\tilde{\boldsymbol{q}}_{i,k}^{p_i} - \boldsymbol{q}_{i,k}\right)^2|\boldsymbol{q}_i\right] = \frac{\boldsymbol{q}_{i,k}(1-\boldsymbol{q}_{i,k})}{p_i} \leq \frac{1}{4p_i}.$$

Thus, we have

$$\mathbb{E}(|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}|^2) = \sum_{k=1}^N \mathbb{E}(\boldsymbol{q}_{i,k} - \tilde{\boldsymbol{q}}_{i,k}^{p_i})^2 \leq \frac{1}{4}\sum_{k=1}^N p_i^{-1} \leq \frac{N}{4p} \tag{3.11}$$

and we obtain from (3.10) and (3.11) that

$$\mathbb{E}(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2) \leq \frac{L_{\rho,\varepsilon}}{\varepsilon}\left(\sqrt{\frac{N}{p}} + 2\rho(N + \sqrt{N})\right). \tag{3.12}$$

Combining inequalities (3.8), (3.9), and (3.12) concludes the proof of Theorem 3.2.

4 Goldenshluger-Lepski method and oracle inequality

In this section, we present a method to choose, in a data-driven way, the parameters γ in (1.3) and ε in (1.5). By analogy with the work in [23] based on the Goldenshluger-Lepski (GL) principle [19], we propose to compute a bias-variance trade-off functional which will provide an automatic selection method for the regularization parameters within a finite set for either penalized or Sinkhorn barycenters. The method consists in comparing estimators pairwise, for a given range of regularization parameters, with respect to a loss function.

Since the formulation of the GL's principle is similar for both estimators, we only present its principle for the Sinkhorn barycenter described in Section 3. We also show that the formulation of the GL's principle proposed in this paper for the Sinkhorn barycenter leads to a data-driven estimator satisfying an oracle inequality which sheds some light on its theoretical properties.

The key point in the GL method is the definition of a data-driven trade-off functional that is composed of a term measuring the disparity between two estimators and of a penalty term that is chosen according to the upper bounds on the variance of the Sinkhorn barycenter given in Section 3. More precisely, we assume that we have at our disposal a collection of estimators $(\hat{\boldsymbol{r}}^{\varepsilon}_{n,p})_{\varepsilon}$ for ε ranging in a finite set $\Lambda \subset \mathbb{R}_+$ depending on the data at hand. The GL method consists in choosing a value $\hat{\varepsilon}$ which minimizes the following bias-variance trade-off function:

$$\hat{\varepsilon} = \underset{\varepsilon \in \Lambda}{\arg\min} \;\; B(\varepsilon) + 3V(\varepsilon) \tag{4.1}$$

for which we set the "bias term" as

$$B(\varepsilon) = \sup_{\tilde{\varepsilon} \leq \varepsilon} \left[|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - \hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p}|^2 - 3V(\tilde{\varepsilon}) \right]_+ \tag{4.2}$$

where $x_+ = \max(x, 0)$ denotes the positive part, and the "variance term" V is chosen accordingly to the oracle inequality in the Theorem (4.1) below as follows

$$V(\varepsilon) = V_{b_1,b_2}(\varepsilon) := b_1 \frac{8L^2_{\rho,\varepsilon}}{\varepsilon^2 n} + \frac{2L_{\rho,\varepsilon}}{\varepsilon} \left(\sqrt{b_2 \frac{N}{p}} + \rho(N + \sqrt{N}) \right), \tag{4.3}$$

where $b_1 > 0$ and $b_2 > 0$ are constants whose choice is discussed below. Note that, in definition (4.2) of the bias term, it is implicitly understood that the supremum is restricted to the regularization parameters $\tilde{\varepsilon} \leq \varepsilon$ such that $\tilde{\varepsilon} \in \Lambda$. To stress the dependence of the variance term on b_1 and b_2, we sometimes write $V(\varepsilon) = V_{b_1,b_2}(\varepsilon)$.

Following [23], we propose to show that, under an appropriate choice of the constants b_1 and b_2, the selected estimator $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ satisfies an oracle inequality which represents an optimal bias-variance tradeoff (depending on the set Λ) for its risk $|\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - r^0|^2$, where

$$r^0 \in \underset{r \in \Sigma^{\rho}_N}{\arg\min}\, \mathbb{E}_{\boldsymbol{q} \sim \mathbb{P}}[W^2_{2,0}(r, \boldsymbol{q})] \quad \text{with} \quad W^2_{2,0}(r, q) := \min_{U \in U(r,q)} \langle C, U \rangle. \tag{4.4}$$

Remark 4.1. *It should be remarked that we chose to refer to $\mathbb{E}(|r^{\varepsilon} - \hat{r}^{\varepsilon}_{n,p}|^2)$ as a "variance term". This is somewhat imprecise as the estimator $\hat{r}^{\varepsilon}_{n,p}$ is certainly such that $\mathbb{E}(\hat{r}^{\varepsilon}_{n,p}) \neq r^{\varepsilon}$. Therefore, $\mathbb{E}(|r^{\varepsilon} - \hat{r}^{\varepsilon}_{n,p}|^2)$ is not the usual statistical notion of variance for $\hat{r}^{\varepsilon}_{n,p}$. Similarly, we have chosen to implicitly referred to $|r^{\varepsilon} - r^0|$ as a bias term which is rather an approximation error term. Nevertheless, we prefer to keep this terminology of bias and variance as it is*

consistent with the one used to present the GL's principle in [23] for the classical problem of kernel density estimation for which the standard notions of a bias term and an approximation error coincide.

Now, as in [23], we introduce the following generalized approximation error

$$D(\varepsilon) := \max\left(\sup_{\tilde{\varepsilon}\leq\varepsilon} |r^{\tilde{\varepsilon}} - r^{\varepsilon}|, |r^{\varepsilon} - r^{0}|\right)$$

which satisfies $D(\varepsilon) \leq 2\sup_{\tilde{\varepsilon}\leq\varepsilon} |r^{\tilde{\varepsilon}} - r^0|$. The following result shows that, under an appropriate choice of b_1 and b_2, the data-driven choice $\hat{\varepsilon}$ of the regularization parameter by the GL method leads to a Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ satisfying an oracle inequality leading to an optimal bias-variance tradeoff within the collection Λ of regularization parameters.

Theorem 4.1. *Assume that the constants b_1 and b_2 in the calibration of the variance term $V(\varepsilon) = V_{b_1,b_2}(\varepsilon)$ are such that*

$$b_1 > (1 + \sqrt{\log(|\Lambda|/2)})^2 \quad \textit{and} \quad b_2 > \log(2) + \frac{\log(n) + \sqrt{\log(|\Lambda|/2)}}{N}. \tag{4.5}$$

where $|\Lambda|$ denotes the cardinal of Λ. Then, one has that

$$|\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - r^0| \leq (1 + 2\sqrt{3}) \inf_{\varepsilon\in\Lambda} \left\{ D(\varepsilon) + \sqrt{V_{b_1,b_2}(\varepsilon)} \right\}, \tag{4.6}$$

with probability larger than $1 - |\Lambda| \left(e^{-(\sqrt{b_1}-1)^2} + e^{N(\log(2)-b_2)+\log(n)} \right)$.

Proof. We first start as in the proof of Proposition 1 in [23] by showing that for any $\varepsilon \in \Lambda$

$$|\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - r^0| \leq \sqrt{2B(\varepsilon) + 6V(\varepsilon)} + |\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| + D(\varepsilon). \tag{4.7}$$

For completeness, we repeat the arguments in [23] yielding to inequality (4.7) as they allow to shed some lights on the basic principles of the GL method. For any fixed $\varepsilon \in \Lambda$ one has that

$$\begin{aligned} |\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - r^0| &\leq |\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}| + |\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^0| \leq |\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}| + |\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| + |r^{\varepsilon} - r^0| \\ &\leq |\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}| + |\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| + D(\varepsilon). \end{aligned} \tag{4.8}$$

Then, for any $\tilde{\varepsilon} \leq \varepsilon$, the definition (4.2) of the bias term implies that $|\hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}|^2 \leq B(\varepsilon) + 3V(\tilde{\varepsilon})$ which can also be written as $|\hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}|^2 \leq B(\max(\varepsilon,\tilde{\varepsilon})) + 3V(\min(\varepsilon,\tilde{\varepsilon}))$ for all $\varepsilon, \tilde{\varepsilon} \in \Lambda$. Therefore, by definition (4.1) of $\hat{\varepsilon}$, one obtains that

$$|\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}|^2 \leq B(\max(\varepsilon,\hat{\varepsilon})) + 3V(\min(\varepsilon,\hat{\varepsilon})) \leq B(\varepsilon) + 3V(\varepsilon) + \max(B(\varepsilon), 3V(\varepsilon)). \tag{4.9}$$

Hence, inserting inequality (4.9) into (4.8) finally yields inequality (4.7).

Now that inequality (4.7) has been established, the main steps in the proof are the control of the stochastic terms $B(\varepsilon)$ and $|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}|$. First, using the triangle inequality

$$|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| \leq |\hat{\boldsymbol{r}}^{\varepsilon}_{n} - r^{\varepsilon}| + |\hat{\boldsymbol{r}}^{\varepsilon}_{n} - \hat{\boldsymbol{r}}^{\varepsilon}_{n,p}|$$

we obtain by Proposition C.1 and Proposition C.2 that are presented in the appendix, that for any $u_1 > 0$ and $u_2 > 0$,

$$\mathbb{P}\left(|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| > (1+u_1)\frac{2\sqrt{2}L_{\rho,\varepsilon}}{\varepsilon\sqrt{n}} + \sqrt{\frac{2L_{\rho,\varepsilon}}{\varepsilon}\left(u_2 + \rho(N+\sqrt{N})\right)}\right) \leq \exp\left(-u_1^2\right) + 2^N n \exp\left(-p u_2^2\right)$$

where $p = \min_{1\leq i\leq n} p_i$. Hence, choosing $u_1 > \sqrt{b_1} - 1$ and $u_2 > \sqrt{b_2 \frac{N}{p}}$, implies that

$$|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| \leq \sqrt{V(\varepsilon)}, \quad \text{for all} \quad \varepsilon \in \Lambda, \tag{4.10}$$

with probability larger than $1 - |\Lambda|\left(e^{-(\sqrt{b_1}-1)^2} + e^{N(\log(2)-b_2)+\log(n)}\right)$. To control $B(\varepsilon)$ with a bias term, we use the upper bound

$$|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - \hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p}|^2 \leq \left(|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - r^{\varepsilon}| + |\hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p} - r^{\tilde{\varepsilon}}| + |r^{\varepsilon} - r^{\tilde{\varepsilon}}|\right)^2$$

combined with inequality (4.10) to obtain that

$$|\hat{\boldsymbol{r}}^{\varepsilon}_{n,p} - \hat{\boldsymbol{r}}^{\tilde{\varepsilon}}_{n,p}|^2 \leq \left(\sqrt{V(\varepsilon)} + \sqrt{V(\tilde{\varepsilon})} + |r^{\varepsilon} - r^{\tilde{\varepsilon}}|\right)^2 \leq 3\left(V(\varepsilon) + V(\tilde{\varepsilon}) + |r^{\varepsilon} - r^{\tilde{\varepsilon}}|^2\right),$$

for all ε and $\tilde{\varepsilon}$ belonging to Λ, with probability $1 - |\Lambda|\left(e^{-(\sqrt{b_1}-1)^2} + e^{N(\log(2)-b_2)+\log(n)}\right)$, which finally implies that (with the same probability)

$$B(\varepsilon) \leq 3V(\varepsilon) + 3D^2(\varepsilon). \tag{4.11}$$

Combining inequalities (4.7), (4.10) and (4.11), we obtain that

$$\begin{aligned} |\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p} - r^0| &\leq \sqrt{6D^2(\varepsilon) + 12V(\varepsilon)} + \sqrt{V(\varepsilon)} + D(\varepsilon), \\ &\leq (1+2\sqrt{3})\left(D(\varepsilon) + \sqrt{V(\varepsilon)}\right), \quad \text{for all} \quad \varepsilon \in \Lambda, \end{aligned}$$

with probability larger than $1 - |\Lambda|\left(e^{-(\sqrt{b_1}-1)^2} + e^{N(\log(2)-b_2)+\log(n)}\right)$, which completes the proof of Theorem 4.1. □

5 Numerical experiments

In this section, we first illustrate with one-dimensional datasets the performances of the Goldenshluger-Lepski method to choose the regularization parameters γ and ε. Then, we report the results from numerical experiments on simulated Gaussian mixtures and flow cytometry dataset in $\mathbb{R}^2$.

Unfortunately, in numerical experiments, we have found that the Lipschitz constant (3.5) leads to a too rough estimate of the variance term $\mathbb{E}(|r^{\varepsilon} - \hat{r}^{\varepsilon}_{n,p}|^2)$ which has a magnitude much smaller than the upper bound (3.4). Using the conditions of Theorem 4.1 on b_1 and b_2 leads to a quantity $V(\varepsilon)$ which is much larger than the bias term $B(\varepsilon)$ leading to always choosing the smallest value of ε in Λ. To overcome this problem, it is necessary to either scale the magnitude of $V(\varepsilon)$ by choosing small values of b_1 and b_2, or to improve the upper bound (3.4) using numerical methods. To this end, we use Monte-Carlo simulations to obtain the right order for the variance term, which allows to show that the Goldenshluger-Lepski principle leads to satisfactory choices of ε in this setting.

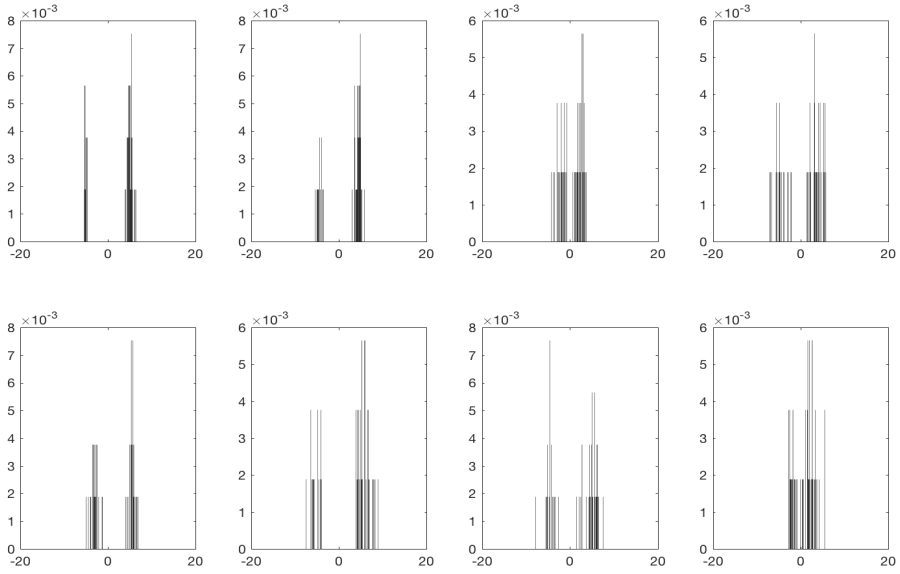


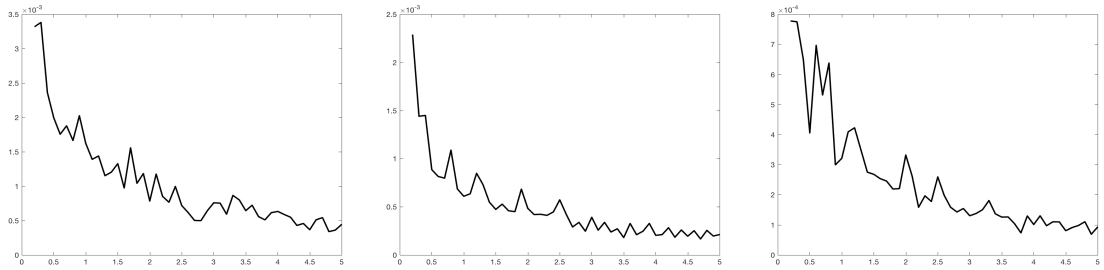
Figure 4: A subset of 8 histograms (out of $n = 15$) obtained with random variables sampled from one-dimensional Gaussian mixtures distributions $\boldsymbol{\nu}_i$ (with random means and variances). Histograms are constructed by binning the data $(X_{i,j})_{1\leq i\leq n\ ;1\leq p}$ on a grid Ω_N of size $N = 2^8$.

5.1 Simulated data: one-dimensional Gaussian mixtures

We illustrate GL's principle for the one-dimensional example of random Gaussian mixtures that is displayed in Figure 4. Our dataset consists of observations $(X_{i,j})_{1\leq i\leq n\ ;1\leq p}$ sampled from $n = 15$ random distributions $\boldsymbol{\nu}_i$ that are mixtures of two Gaussian distributions with weights $(0.35, 0.65)$, random means respectively belonging to the intervals $[-6, -2]$ and $[2, 6]$ and random variances both belonging to the interval $(0, 2]$. For each random mixture distribution, we sample $p = 50$ observations. A first step is to perform Monte-Carlo experiments over 20 regularized barycenters $\hat{r}^{\varepsilon}_{n,p}$ for each different values of ε in the interval $[0.2, 5]$. The results are presented in Figure 5 where we plot the estimation of the variance $\mathbb{E}(|r^{\varepsilon} - \hat{r}^{\varepsilon}_{n,p}|^2)$ with respect to the regularization parameter $\varepsilon \in [0.2, 5]$ for different values of $N = 2^6, 2^7, 2^8$. A first aspect is that the variance term estimated by Monte Carlo simultions decreases as the regularization parameter increases.

From now on, we fix the size $N = 2^8$ of the grid, and we comment on the choice of the parameters $\hat{\varepsilon}$ and $\hat{\gamma}$. To obtain this data-driven choice of regularization, we use the Goldenshluger-Lepskii principle where the variance term $V(\varepsilon)$ is given by the variance function displayed in Figure 5 obtained via Monte-Carlo simulations. We display in Figure 6(a) the trade-off function $B(\varepsilon) + 3V(\varepsilon)$, and we discuss the influence of ε on the smoothness the Sinkhorn barycenter.

From Figure 6(b), we observe that, when the parameter $\varepsilon = 0.2$ is small (dotted blue curve), then the corresponding Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\varepsilon}_{n,p}$ is irregular, and it presents spurious peaks. On the contrary, too much regularization, e.g. $\varepsilon = 5$, implies that the barycenter (dashed green curve) is flattened and its mass is spread out. The optimal barycenter (solid red curve), that is $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ for $\hat{\varepsilon} = 3.8$ minimizing the trade-off function (4.1), gives here a good



(a) $N = 2^6$ (b) $N = 2^7$ (c) $N = 2^8$

Figure 5: A Monte-Carlo experiments for estimating the variance $\mathbb{E}(|r^{\varepsilon} - \hat{r}^{\varepsilon}_{n,p}|^2)$ for three values of the grid N.

compromise between under and over-smoothing.

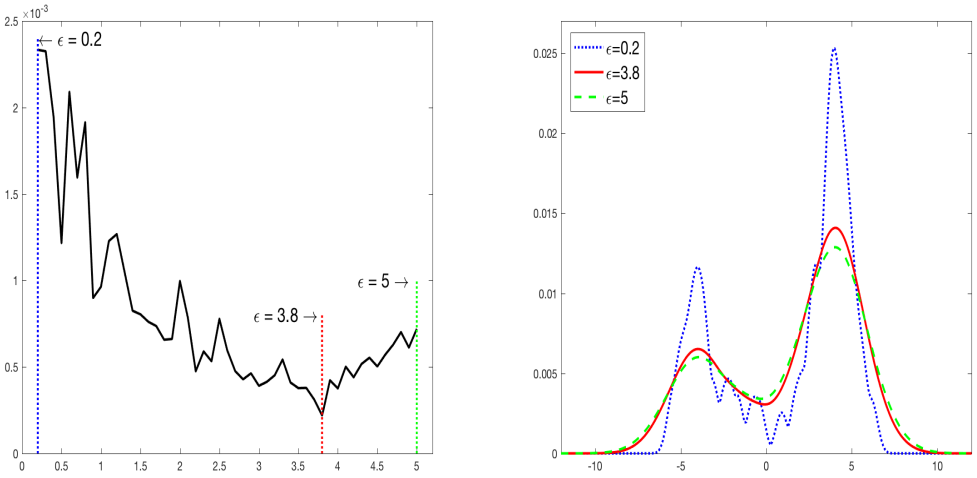
We repeat the same experiment for the penalized barycenter $\hat{\boldsymbol{f}}^{\gamma}_{n,p}$ of Section 2 with the Sobolev norm $H^1(\Omega)$ in the penalization function E (2.1) . In particular, we remark that the size of the grid does not appear explicitly in the the upper bound of the variance function in inequality (2.5). The Monte-Carlo experiments are presented in Figure 7 for $N = 2^8$. This gives us an approximation for the variance term in Goldenshluger-Lepskii.

The results of the Goldenshluger-Lepskii technique are displayed in Figure 8. The advantage of choosing a Sobolev penalty function over an entropy term is that the mass of the barycenter is overall less spread out and the spikes are sharper. However, for a small regularization parameter $\gamma = 20$ (dotted blue curve), the barycenter $\boldsymbol{f}^{\gamma}_{n,p}$ presents a lot of irregularities as the penalty function tries to minimize its $\mathbb{L}_2$-norm. When the regularization parameter increases in a significant way ($\gamma = 1000$ associated to the dashed green curve), the irregularities disappear and the support of the penalized barycenter becomes wider. The GL's principle leads to the choice $\hat{\gamma} = 840$ which corresponds to a penalized barycenter (solid red curve) that is satisfactory.

We compare these Wasserstein barycenters to the Euclidean mean $\bar{f}_{n,p}$ (1.1), obtained after a pre-smoothing step of the data for each subject using the kernel method. From Figure 9, the density $\bar{f}_{n,p}$ is very irregular and it suffers from mis-alignment issues. The irregularity of this estimator mainly comes from the low sample size per subject ($p = 50$).

5.2 Sinkhorn versus penalized barycenters

To conclude these numerical experiments with one-dimensional simulated data, we would like to point out that computing the Sinkhorn barycenter is much faster than computing the penalized barycenter. Indeed, entropy regularization of the transport plan in the Wasserstein distance has been first introduced in order to reduce the computational cost of a transport distance. The computation of an unregularized transport distance requires $\mathcal{O}(N^3 \log N)$ operations for discrete probability measures with a support of size N when the computation of a Sinkhorn divergence only takes $\mathcal{O}(N^2)$ operations at each iteration of a gradient descent (see e.g. [12]). We have also found that the Sinkhorn barycenter yields more satisfying estimators in terms of smoothness. Therefore, in the rest of this section, we do not consider the penalized barycenter anymore.



(a) (b)

Figure 6: One dimensional Gaussian mixtures dataset and Sinkhorn barycenters. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon) + 3V(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 3.8$. (b) Three Sinkhorn barycenters $\hat{\boldsymbol{r}}_{n,p}^{\hat{\varepsilon}}$ associated to the parameters $\varepsilon = 0.2, 3.8, 5$.

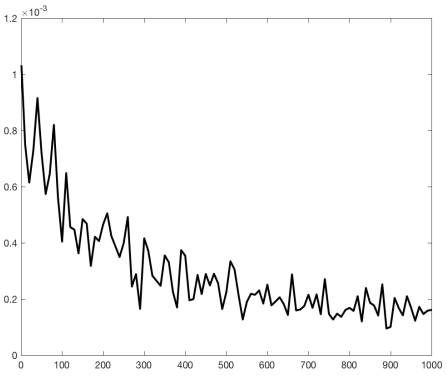
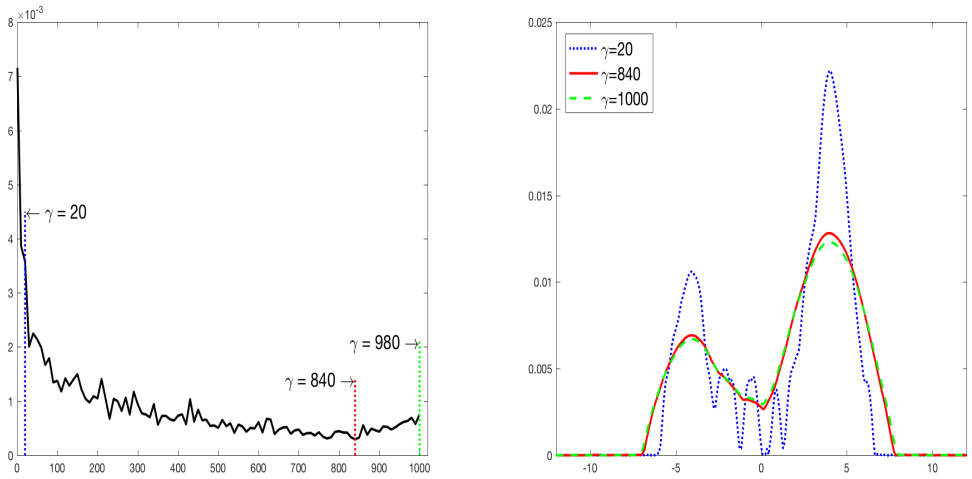


Figure 7: A Monte-Carlo experiments for estimating the variance $\mathbb{E}\left(\|\hat{\boldsymbol{f}}_{n,p}^{\gamma} - f_{\mathbb{P}}^{\gamma}\|_{\mathbb{L}_2(\Omega)}^2\right)$.



(a) (b)

Figure 8: One dimensional Gaussian mixtures dataset and penalized barycenters. (a) The trade-off function $\gamma \mapsto \quad B(\gamma) + 3V(\gamma)$ which attains its optimum at $\hat{\gamma} = 840$. (b) Three penalized barycenters $\boldsymbol{f}_{n,p}^{\gamma}$ associated to the parameters $\gamma = 20, 840, 100$.

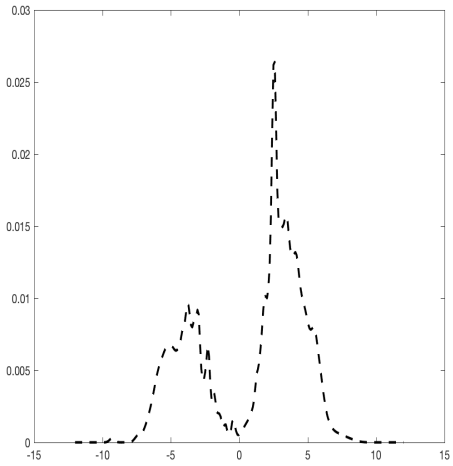


Figure 9: Euclidean mean density $\bar{f}_{n,p}$ of the one dimensional Gaussian mixtures dataset using a preliminary smoothing step of each subject with a Gaussian kernel.

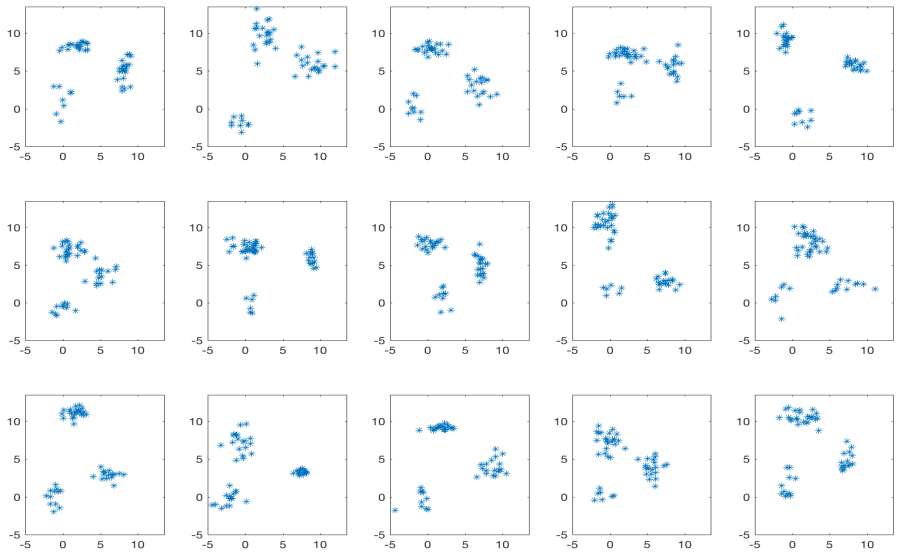


Figure 10: Dataset $(X_{i,j})_{1\leq j\leq p\ ;1\leq i\leq n}$ generated from $n = 15$ two-dimensional random Gaussian mixtures $\boldsymbol{\nu}_i$ with $p = 50$.

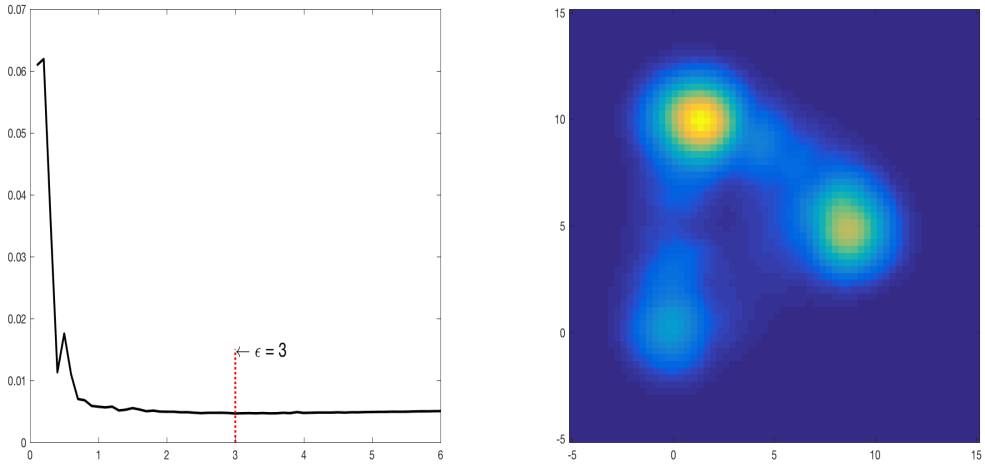
5.3 Simulated data: two-dimensional Gaussian mixtures

In this section, we illustrate our methods for two-dimensional data. We consider a simulated example of observations $(X_{i,j})_{1\leq i\leq n\ ;1\leq j\leq p}$ sampled from $n = 15$ random distributions $\boldsymbol{\nu}_i$ that are a mixture of three multivariate Gaussian distributions $\boldsymbol{\nu}_i = \sum_{j=1}^{3} \theta_j \mathcal{N}(\boldsymbol{m}_j^i, \boldsymbol{\Gamma}_j^i)$ with fixed weights $\theta = (1/6, 1/3, 1/2)$. The means $\boldsymbol{m}_j^i$ and covariance matrices $\boldsymbol{\Gamma}_j^i$ are random variables with expectation given by (for $j = 1, 2, 3$)

$$m_1 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad m_2 = \begin{pmatrix} 7 \\ 4 \end{pmatrix}, \quad m_3 = \begin{pmatrix} 1 \\ 9 \end{pmatrix}, \quad \text{and} \quad \Gamma_1 = \Gamma_2 = \Gamma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The covariances Γ_i are chosen diagonal to ensure that the perturbed random covariances that form our dataset are positive semi-definite matrices to properly define the Gaussian distributions associated. For each $i = 1, \ldots, n$, we simulate a sequence $(X_{i,j})_{1\leq j\leq p}$ of $p = 50$ iid random variables sampled from $\boldsymbol{\nu}_i = \sum_{j=1}^{3} \theta_j \mathcal{N}(\boldsymbol{m}_j^i, \boldsymbol{\Gamma}_j^i)$ where $\boldsymbol{m}_j^i$ (resp. $\boldsymbol{\Gamma}_j^i$) are random vectors (resp. matrices) such that each of their coordinate follows a uniform law centered in m_j with amplitude ± 2 (resp. each of their diagonal elements follows a uniform law centered in Γ_j with amplitude ± 0.95). We display in Figure 10 the dataset $(X_{i,j})_{1\leq j\leq p\ ;1\leq i\leq n}$. Each $X_{i,j}$ is then binned on a grid of size 64×64 (thus $N = 4096$).

First, we compute 60 Sinkhorn barycenters by letting ε ranging from 0.1 to 6. We draw in Figure 11(a) the trade-off function that is also based on 10 Monte-Carlo experiments to approximate the term V, with its minimizer at $\hat{\varepsilon} = 3$. The corresponding Sinkhorn barycenter $\hat{\boldsymbol{r}}_{n,p}^{\hat{\varepsilon}}$ is displayed in Figure 11(b). We also present the Euclidean mean $\bar{f}_{n,p}$ (after a preliminary smoothing step) in Figure 12(b). The Sinkhorn barycenter has three distinct modes. Hence, this approach handles in a very efficient way the scaling and translation variations in the



(a) (b)

Figure 11: Two-dimensional Gaussian mixtures dataset. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon) + 3V(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 3$. (b) The Sinkhorn barycenter $\hat{\boldsymbol{r}}^{\hat{\varepsilon}}_{n,p}$ for $\hat{\varepsilon} = 3$ chosen by the GL's principle.

dataset (which corresponds to the correction of the mis-alignment issue). On the other hand, the Euclidean mean mixes the distinct modes of the Gaussian mixtures. It is thus less robust to outliers since the support of the barycenter is significantly spread out.

5.4 Real data: flow cytometry

We have at our disposal data from flow cytometry that have been described in Section 1.3.5, and we focus on the FSC and SSC cell markers resulting in the dataset that is displayed in Figure 3. We again apply a binning of the data on a two-dimensional grid of size $N = 64 \times 64$. In Figure 13(a) we plot the trade-off function related to the Sinkhorn barycenters. To that end, we use the results on the Monte-Carlo estimation of the variance in the gaussian case (see the previous Section 5.3). Indeed, as the upper-bound of the variance in (3.4) depends explicitly on parameters of the problem, we can compare the parameters from the 2D gaussian case to the parameters of the real cytometry data case. Let us first remark that in both experiments, the size of the grid is chosen as $N = 64^2$, the number of distributions is equal to $n = 15$, the minimum number of observations per measure is approximately $p \sim 50$, and the collection of parameters $\varepsilon = [0.1, 6]$ that we test is the same. However, the grid differs in these two experiments. For the gaussian case, the grid is taken as the square $[-5, 15]^2$; in the cytometry case, the measurements of the cells are within the rectangle $[200, 600] \times [0, 250]$. Therefore, in the Lipschitz constant term, the quantity

$$(\Delta_m)_m := (\sup_{1 \leq \ell, k \leq N} |C_{m\ell} - C_{k\ell}|)_m,$$

which is a vector of length 64, involving the cost function will differ. For the gaussian case grid we have $\max_m \Delta_m = 800$, for the cytometry case grid we have $\max_m \Delta_m = 217600$. For the reason, we thus choose to scale the variance term in the GL method by $800/217600 = 1/272$.

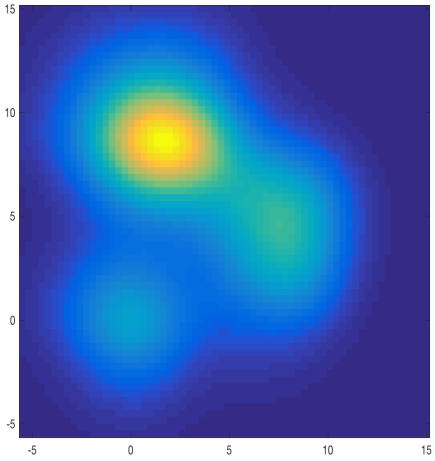


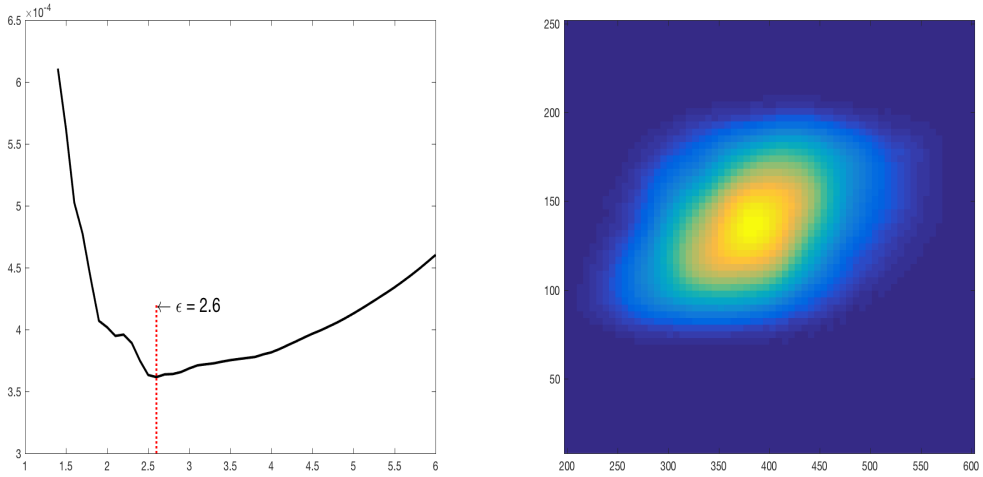
Figure 12: Two-dimensional Gaussian mixtures dataset. The Euclidean mean $\bar{f}_{n,p}$ (after a preliminary smoothing step).

The regularization parameter ε still ranges from 1 to 6. The minimum of the trade-off function is reached for $\hat{\varepsilon} = 2.6$. We display the corresponding Sinkhorn barycenter in Figure 13(b). This barycenter clearly corrects mis-alignment issues of the data.

To analyze the relevance of this result, we present in Figure 14(a) the Euclidean mean $\bar{f}_{n,p}$ of this dataset (after kernel smoothing of the data for each patient). The support of this estimator is again spread out due to the presence of a strong translation variance in the data which clearly need to be registered. We also compare our method to the more relevant one proposed in [20] which consists in approximating each of the 15 subjects with a two dimensional kernel density estimate (with an automatic choice of the bandwidth parameter). The densities obtained are then projected onto a one dimensional space. Landmarks are estimated by identifying peaks of the resulting one-dimensional densities. Then, these landmarks are registered across the whole dataset in order to finally align the densities. The $\mathbb{L}_2$-mean density obtained after this pre-processing step is displayed in Figure 14(b). This method leads to results that are similar to the one obtained with a data-driven Sinkhorn barycenter. However, contrary to regularized Wasserstein barycenters that can handle automatically non registered multi-modal densities, the procedure in [20] suffers from two main drawbacks: (i) the registration of densities is performed by one-dimensional projections, (ii) the number of peaks to align is chosen manually. Notice that we have also conducted experiments for Sinkhorn barycenters with non-equal weights, corresponding to the proportion of measurements for each patient. The result being analogous, we do not report them.

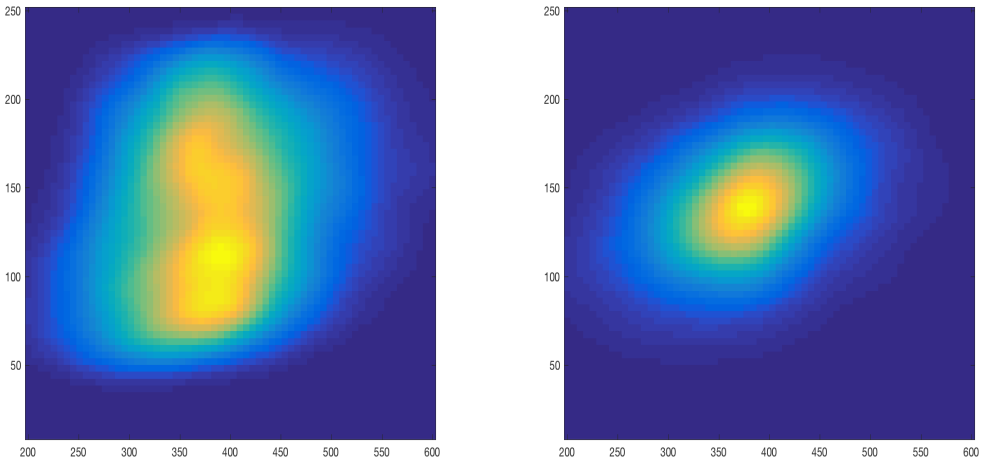
6 Conclusion and perspectives

It would be interesting to derive more accurate upper bounds of the variance term for the Sinkhorn barycenter that would be of a magnitude of the same order than the one found in numerical experiments using Monte Carlo simulations. An additional difficulty is the possibility of using Sinkhorn barycenters with data-driven choice of the regularization parameter



(a) (b)

Figure 13: Two dimensional flow cytometry dataset and Sinkhorn barycenter. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon)+3V(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 2.6$. (b) Sinkhorn barycenter $\hat{\boldsymbol{r}}_{n,p}^{\hat{\varepsilon}}$ associated to the parameter $\hat{\varepsilon} = 2.6$.



(a) (b)

Figure 14: Two dimensional flow cytometry dataset. (a) Euclidean mean $\bar{f}_{n,p}$ of the data (after smoothing but without registration), (b) $\mathbb{L}_2$-mean of pre-processed data using kernel smoothing and density registration by landmark alignment with the method in [20].

for the registration of multiple point clouds beyond the dimension $d \geq 3$, e.g. for applications in flow cytometry where the dimension d can be 30 or 40. This is clearly a challenging task.

Acknowlegements. 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).

A Strong convexity of the Sinkhorn divergence - Proof of Theorem 3.4

The proof of Theorem 3.4 relies on the analysis of the eigenvalues of the Hessian matrix $\nabla^2 H_q^*(g)$ of the functional H_q^*.

Proposition A.1. *For all $g \in \mathbb{R}^N$, $\nabla^2 H_q^*(g)$ admits $\lambda_N = 0$ as eigenvalue with its associated normalized eigenvector $v_N := \frac{1}{\sqrt{N}} \mathbb{1}_N \in \mathbb{R}^N$, which means that rank$(\nabla^2 H_q^*(g)) \leq N-1$ for all $g \in \mathbb{R}^N$ and $q \in \Sigma_N$.*

Proof. Let $g \in \mathbb{R}^N$, then by Theorem 3.3

$$\begin{aligned}\nabla^2 H_q^*(g) v_N &= \frac{1}{\varepsilon} \operatorname{diag}(\alpha) K \frac{q}{K\alpha} - \frac{1}{\varepsilon} \operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2}\right) K\alpha \\ &= \frac{1}{\varepsilon} \operatorname{diag}(\alpha) K \frac{q}{K\alpha} - \frac{1}{\varepsilon} \operatorname{diag}(\alpha) K \frac{q}{K\alpha} = 0,\end{aligned}$$

and $\lambda_N = 0$ is an eigenvalue of $\nabla^2 H_q^*(g)$. □

Let $(v_k)_{1 \leq k \leq N}$ be the eigenvectors of $\nabla^2 H_q^*(g)$, depending on both q and g, with their respective eigenvalues $(\lambda_k)_{1 \leq k \leq N}$. As the Hessian matrix is symmetric and diagonalisable, let us now prove that the eigenvalues associated to the eigenvectors $(v_k)_{1 \leq k \leq N-1}$ of $\nabla^2 H_q^*(g)$ are all positive.

Proposition A.2. *For all $q \in \Sigma_N$ and $g \in \mathbb{R}^N$, we have that*

$$0 = \lambda_N < \lambda_k \qquad \text{for all} \quad 1 \leq k \leq N-1.$$

Proof. The eigenvalue $\lambda_N = 0$ associated to v_N has been treated in Proposition A.1. Let $v \in V = (\operatorname{Vect}(v_N))^\perp$ (i.e. v does not have constant coordinates) an eigenvector of $\nabla^2 H_q^*(g)$. Hence we can suppose that, let say $v^{(j)}$, is its larger coordinate, and that their exists $i \neq j$ such that $v^{(j)} > v^{(i)}$. Without loss of generality, we can assume that $v^{(j)} > 0$. Then

$$\begin{aligned}[\nabla^2 H_q^*(g) v]_j &= \left[\frac{1}{\varepsilon}\left(\operatorname{diag}\left(\operatorname{diag}(\alpha) K \frac{q}{K\alpha}\right)\right) v\right]_j - \left[\frac{1}{\varepsilon} \operatorname{diag}(\alpha) K \operatorname{diag}\left(\frac{q}{(K\alpha)^2}\right) K \operatorname{diag}(\alpha) v\right]_j \\ &= \frac{1}{\varepsilon} \alpha_j v^{(j)} \sum_{i=1}^N K_{ji} \frac{q_i}{[K\alpha]_i} - \frac{1}{\varepsilon} \sum_{i=1}^N \sum_{m=1}^N \alpha_j K_{jm} \frac{q_m}{[K\alpha]_m^2} \alpha_i K_{mi} v^{(i)} \\ &> \frac{1}{\varepsilon} \alpha_j v^{(j)} \sum_{i=1}^N K_{ji} \frac{q_i}{[K\alpha]_i} - \frac{1}{\varepsilon} \sum_{i=1}^N \sum_{m=1}^N \alpha_j K_{jm} \frac{q_m}{[K\alpha]_m^2} \alpha_i K_{mi} v^{(j)} \text{ since } v^{(j)} \geq v^{(i)}, \forall i \\ &= 0 \qquad \text{since} \quad \sum_{i=1}^N \alpha_i K_{im} = [K\alpha]_m.\end{aligned}$$

Thus $\lambda v^{(j)} = [\nabla^2 H_q^*(g)v]_j > 0$, and we necessarily have that $\lambda > 0$. □

The set of eigenvalues of $\nabla^2 H_q^*(g)$ is also bounded from above.

Proposition A.3. *For all $q \in \Sigma_N$ and $g \in \mathbb{R}^N$ we have that $Tr(\nabla^2 H_q^*(g)) \leq \frac{1}{\varepsilon}$ and thus $\lambda_k \leq 1/\varepsilon$ for all $k = 1, \dots, N$.*

Proof. We directly get from Theorem 3.3 that

$$\mathrm{Tr}(\nabla^2 H_q^*(g)) \leq \frac{1}{\varepsilon}\mathrm{Tr}\left(\mathrm{diag}\underbrace{\left(\mathrm{diag}(\alpha)K\frac{q}{K\alpha}\right)}_{\in \Sigma_N}\right) = \frac{1}{\varepsilon}.$$

□

We can now provide the proof of Theorem 3.4. Since H_q is convex, proper and lower-semicontinuous, we know by the Fenchel-Moreau theorem that $H_q^{**} = H_q$. Hence by Corollary 12.A in the Rockafellar's book [30], we have that

$$\nabla H_q = (\nabla H_q^*)^{-1}, \tag{A.1}$$

in the sense that $\nabla H_q^* \circ \nabla H_q(r) = r$ for any $r \in \Sigma_N$.

To continue the proof, we focus on a definition of the function H_q restricted to the linear subspace V. Let $(v_1, \dots, v_{N-1})$ be an orthonormal basis of $V = (\mathrm{Vect}(v_N))^\perp$ and $P = [v_1 \ \cdots \ v_{N-1}] \in \mathbb{R}^{N\times(N-1)}$ the matrix of the basis. Remark that PP^T is the matrix of the orthogonal projection onto V, and that $PP^T = I_N - v_N v_N^T$. If we define $\tilde{\Sigma}_{N-1} := P^T\Sigma_N \in \mathbb{R}^{N-1}$, then for $r \in \Sigma_N$, there exists $\tilde{r} \in \tilde{\Sigma}_{N-1}$ such that $r = P\tilde{r} + \frac{1}{\sqrt{N}}v_N$. Hence we can introduce the functional $\tilde{H}_q : \tilde{\Sigma}_{N-1} \to \mathbb{R}$ defined by

$$\tilde{H}_q(\tilde{r}) := H_q\left(P\tilde{r} + \frac{1}{\sqrt{N}}v_N\right).$$

For $\tilde{g} \in \mathbb{R}^{N-1}$ we have that

$$\begin{aligned}
\tilde{H}_q^*(\tilde{g}) &= \max_{\tilde{r}\in\tilde{\Sigma}_{N-1}} \langle \tilde{g}, \tilde{r}\rangle - \tilde{H}_q(\tilde{r}) \\
&= \max_{r\in\Sigma_N} \langle \tilde{g}, P^T r - u_N\rangle - H_q(r) \quad \text{where } u_N = \frac{1}{N}\left(\sum_{i=1}^N v_1^{(i)}, \dots, \sum_{i=1}^N v_{N-1}^{(i)}\right) \\
&= H_q^*(P\tilde{g}) - \langle \tilde{g}, u_N\rangle.
\end{aligned}$$

Since H_q^* is C^∞ (see Theorem 3.3), we can differentiate $\tilde{H}_q^*$ with respect to $\tilde{g}$ to obtain that

$$\begin{aligned}
\nabla\tilde{H}_q^*(\tilde{g}) &= P^T\nabla H_q^*(P\tilde{g}) - u_N \\
\nabla^2\tilde{H}_q^*(\tilde{g}) &= P^T\nabla^2 H_q^*(P\tilde{g})P.
\end{aligned}$$

By Proposition A.2, we know that $\nabla^2 H_q^*(P\tilde{g}) \in \mathbb{R}^{N\times N}$ admits a unique eigenvalue equals to 0 which is associated to the eigenvector v_N. All other eigenvalues are positive (Proposition

A.2) and bounded from above by $1/\varepsilon$ (Proposition A.3). Since $\nabla \tilde{H}_q^* : \mathbb{R}^{(N-1)} \to \mathbb{R}^{(N-1)}$ is a C^∞-diffeomorphism, using equality (A.1) (that is also valid for $\tilde{H}_q$), we have that

$$\begin{aligned}\nabla^2 \tilde{H}_q(\tilde{r}) &= \nabla\left((\nabla \tilde{H}_q^*)^{-1}(\tilde{r})\right) \\ &= [\nabla^2 \tilde{H}_q^*((\nabla \tilde{H}_q^*)^{-1}(\tilde{r}))]^{-1} \\ &= [\nabla^2 \tilde{H}_q^*(\nabla \tilde{H}_q(\tilde{r}))]^{-1},\end{aligned}$$

where the second equality follows from the global inversion theorem, and the last one again uses equality (A.1). Thus we get

$$\lambda_{\min}(\nabla^2 \tilde{H}_q(\tilde{r})) \geq \varepsilon.$$

The above inequality implies the strong convexity of $\tilde{H}_q$ which reads for $\tilde{r}_0, \tilde{r}_1 \in \tilde{\Sigma}_{n-1}$

$$\tilde{H}_q(\tilde{r}_1) \geq \tilde{H}_q(\tilde{r}_0) + \nabla \tilde{H}_q(\tilde{r}_0)^T(\tilde{r}_1 - \tilde{r}_0) + \frac{\varepsilon}{2}\|\tilde{r}_1 - \tilde{r}_0\|^2,$$

and this translates for H_q and $r_0, r_1 \in \Sigma_N$ to

$$H_q(r_1) \geq H_q(r_0) + \nabla H_q(r_0)^T PP^T(r_1 - r_0) + \frac{\varepsilon}{2}\|PP^T(r_1 - r_0)\|^2.$$

To conclude, we remark that $(r_1 - r_0) \in V$ (indeed one has that $r_1 - r_0 = \sum_{j=1}^{N-1}\langle v_j, r_1 - r_0\rangle v_j$ since $\langle v_N, r_1 - r_0\rangle = 0$ and thus $PP^T(r_1 - r_0) = r_1 - r_0$). Hence, we finally obtain the strong convexity of H_q

$$H_q(r_1) \geq H_q(r_0) + \nabla H_q(r_0)^T(r_1 - r_0) + \frac{\varepsilon}{2}\|(r_1 - r_0)\|^2.$$

This completes the proof of Theorem 3.4.

B Lipschitz constant of H_q - Proof of Lemma 3.5

The dual version of the minimization problem (3.1) is given in [13] by

$$W_{2,\varepsilon}^2(r, q) = \max_{\alpha,\beta \in \mathbb{R}^N} \alpha^T r + \beta^T q - \sum_{1 \leq m,\ell \leq N} \varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m - \beta_\ell)} \tag{B.1}$$

where $C_{m\ell}$ are the entries of the matrix cost C. We recall the notation

$$\Sigma_N^\rho = \left\{ r \in \Sigma_N \ : \ \min_{1 \leq \ell \leq N} r_\ell \geq \rho \right\} \text{ for some } 0 < \rho < 1.$$

We now recall the Lemma 3.5.

Lemma B.1. *Let $q \in \Sigma_N$ and $0 < \rho < 1$. Then, one has that $r \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ with*

$$L_{\rho,\varepsilon} = \left(\sum_{1 \leq m \leq N} \left(2\varepsilon \log(N) + \sup_{1 \leq \ell,k \leq N} |C_{m\ell} - C_{k\ell}| - \varepsilon \log(\rho) \right)^2 \right)^{1/2}. \tag{B.2}$$

Proof. Let $r, s, q \in \Sigma_N$. We denote by $(\alpha^{q,r}, \beta^{q,r})$ a pair of optimal dual variables in the problem (B.1). Then, we have that

$$
\begin{aligned}
|H_q(r) - H_q(s)| &= (H_q(r) - H_q(s))\mathbb{1}_{H_q(r)\geq H_q(s)} + (H_q(s) - H_q(r))\mathbb{1}_{H_q(r)\leq H_q(s)} \\
&\leq \left(\langle \alpha^{q,r}, r\rangle + \langle \beta^{q,r}, q\rangle - \sum_{m,\ell} \varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,r} - \beta_\ell^{q,r})} - \langle \alpha^{q,r}, s\rangle - \langle \beta^{q,r}, q\rangle + \right. \\
&\qquad \left. \sum_{m,\ell} \varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,r} - \beta_\ell^{q,r})} \right) \mathbb{1}_{(H_q(r)\geq H_q(s))} \\
&+ \left(\langle \alpha^{q,s}, s\rangle + \langle \beta^{q,s}, q\rangle - \sum_{m,\ell} \varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,s} - \beta_\ell^{q,s})} - \langle \alpha^{q,s}, r\rangle - \langle \beta^{q,s}, q\rangle + \right. \\
&\qquad \left. \sum_{m,\ell} \varepsilon e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,s} - \beta_\ell^{q,s})} \right) \mathbb{1}_{(H_q(r)\leq H_q(s))} \\
&\leq \sup_{\alpha\in\{\alpha^{q,r},\alpha^{q,s}\}} |\langle \alpha, r - s\rangle| \leq \sup_{\alpha\in\{\alpha^{q,r},\alpha^{q,s}\}} |\alpha|\ |r - s|. \qquad \text{(B.3)}
\end{aligned}
$$

Let us now prove that the norm of the dual variable $\alpha^{q,r}$ (resp. $\alpha^{q,s}$) is bounded by a constant not depending on q and r (resp. q and s). To this end, we follow some of the arguments in the proof of Proposition A.1 in [16]. Since the dual variable $\alpha^{q,r}$ achieves the maximum in equation (B.1), we have that for any $1 \leq m \leq N$

$$
r_m - \sum_{1\leq\ell\leq N} e^{-\frac{1}{\varepsilon}(C_{m\ell} - \alpha_m^{q,r} - \beta_\ell^{q,r})} = 0.
$$

Let $r \in \Sigma_N^\rho$. Hence, $r_m \neq 0$, and thus one may define $\lambda_m = \varepsilon \log(r_m)$. Then, it follows from the above equality that $\sum_{1\leq\ell\leq N} e^{-\frac{1}{\varepsilon}(C_{m\ell} + \lambda_m - \alpha_m^{q,r} - \beta_\ell^{q,r})} = 1$ which implies that

$$
\alpha_m^{q,r} = -\varepsilon \log \left(\sum_{1\leq\ell\leq N} e^{-\frac{1}{\varepsilon}(C_{m\ell} + \lambda_m - \beta_\ell^{q,r})} \right).
$$

Now, for each $1 \leq m \leq N$, we define

$$
\tilde{\alpha}_m^{q,r} = \min_{1\leq\ell\leq N} \{C_{m\ell} + \lambda_m - \beta_\ell^{q,r}\} = \min_{1\leq\ell\leq N} \{C_{m\ell} - \beta_\ell^{q,r}\} + \lambda_m, \qquad \text{(B.4)}
$$

and we consider the inequality

$$
|\alpha_m^{q,r} - \alpha_k^{q,r}| \leq |\alpha_m^{q,r} - \tilde{\alpha}_m^{q,r}| + |\tilde{\alpha}_m^{q,r} - \tilde{\alpha}_k^{q,r}| + |\tilde{\alpha}_k^{q,r} - \alpha_k^{q,r}|. \qquad \text{(B.5)}
$$

By equation (B.4) one has that $\tilde{\alpha}_m^{q,r} + \beta_\ell^{q,r} - C_{m\ell} - \lambda_m \leq 0$. Hence we get

$$
-\alpha_m^{q,r} = \varepsilon \log \left(\sum_{1\leq\ell\leq N} e^{-\frac{1}{\varepsilon}\tilde{\alpha}_m^{q,r}} e^{\frac{1}{\varepsilon}(\tilde{\alpha}_m^{q,r} + \beta_\ell^{q,r} - C_{m\ell} - \lambda_m)} \right) \leq -\tilde{\alpha}_m^{q,r} + \varepsilon \log(N). \qquad \text{(B.6)}
$$

On the other hand, using the inequality

$$
\sum_{1\leq\ell\leq N} e^{-\frac{1}{\varepsilon}(C_{m\ell} + \lambda_m - \beta_\ell^{q,r})} \geq e^{-\frac{1}{\varepsilon}(C_{m\ell_*} + \lambda_m - \beta_{\ell_*}^{q,r})} = e^{-\frac{1}{\varepsilon}\tilde{\alpha}_m^{q,r}},
$$

where ℓ_* is a value of $1 \leq \ell \leq N$ achieving the minimum in (B.4), we obtain that

$$-\alpha_m^{q,r} \geq -\tilde{\alpha}_m^{q,r}. \tag{B.7}$$

By combining inequalities (B.6) and (B.7), we finally have

$$|\tilde{\alpha}_m^{q,r} - \alpha_m^{q,r}| \leq \varepsilon \log(N). \tag{B.8}$$

To conclude, it remains to remark that, by equation (B.4), the vector $(\tilde{\alpha}_m^{q,r} - \lambda_m)_{1\leq m\leq N}$ is the c-transform of the vector $(\beta_\ell^{q,r})_{1\leq \ell\leq N}$ for the cost matrix C. Therefore, by using standard results in optimal transport which relate c-transforms to the modulus of continuity of the cost (see e.g. [31], p. 11) one obtains that

$$|\tilde{\alpha}_m^{q,r} - \tilde{\alpha}_k^{q,r} + \lambda_k - \lambda_m| \leq \sup_{1\leq \ell\leq N} |C_{m\ell} - C_{k\ell}|,$$

which implies that

$$|\tilde{\alpha}_m^{q,r} - \tilde{\alpha}_k^{q,r}| \leq \sup_{1\leq \ell\leq N} |C_{m\ell} - C_{k\ell}| + \varepsilon |\log(r_m) - \log(r_k)|. \tag{B.9}$$

By combining the upper bounds (B.8) and (B.9) with the decomposition (B.5) we finally come to the inequality

$$|\alpha_m^{q,r} - \alpha_k^{q,r}| \leq 2\varepsilon \log(N) + \sup_{1\leq \ell\leq N} |C_{m\ell} - C_{k\ell}| + \varepsilon |\log(r_m) - \log(r_k)|.$$

Since the dual variables achieving the maximum in equation (B.1) are defined up to an additive constant, one may assume that $\alpha_k^{q,r} = 0$. Under such a condition, we finally obtain that

$$|\alpha| \leq \left(\sum_{1\leq m\leq N} \left(2\varepsilon \log(N) + \sup_{1\leq k\leq N} \left\{ \sup_{1\leq \ell\leq N} |C_{m\ell} - C_{k\ell}| + \varepsilon |\log(r_m) - \log(r_k)| \right\} \right)^2 \right)^{1/2}.$$

Using inequality (B.3) and the assumption that $r \in \Sigma_N^\rho$ (in particular $\rho \leq r_m/r_k \leq 1/\rho$), we can thus conclude that $r \mapsto H_q(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on Σ_N^ρ for

$$L_{\rho,\varepsilon} = \left(\sum_{1\leq m\leq N} \left(2\varepsilon \log(N) + \sup_{1\leq \ell,k\leq N} |C_{m\ell} - C_{k\ell}| - \varepsilon \log(\rho) \right)^2 \right)^{1/2}. \tag{B.10}$$

□

C Useful concentration inequalities

We state and prove below the various concentration inequalities that have been used in the proof of Theorem 4.1.

Proposition C.1. *For any $u > 0$*

$$\mathbb{P}\left(|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon| > (1+u) \frac{2\sqrt{2} L_{\rho,\varepsilon}}{\varepsilon \sqrt{n}} \right) \leq \exp\left(-u^2\right), \tag{C.1}$$

where $\boldsymbol{r}_n^\varepsilon$ is the Sinkhorn barycenter of the iid random measures $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ (assumed to belong to Σ_N^ρ) as defined by (3.7).

Proof. To derive a concentration inequality for the random variable

$$Z = |\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon|,$$

we write it as $Z = f(\boldsymbol{q}_1, \dots, \boldsymbol{q}_n)$ where $f : \Sigma_N^\rho \times \dots \times \Sigma_N^\rho \to \mathbb{R}$ is a measurable function, and we shall prove that f satisfies the following bounded difference inequality

$$|f(\boldsymbol{q}_1, \dots, \boldsymbol{q}_i, \dots, \boldsymbol{q}_n) - f(\boldsymbol{q}_1, \dots, \boldsymbol{q}_i', \dots, \boldsymbol{q}_n)| \leq \frac{4L_{\rho,\varepsilon}^2}{\varepsilon n}, \quad \text{for any } 1 \leq i \leq n, \tag{C.2}$$

where $\boldsymbol{q}_i'$ denotes an independent random measure of the sample $\boldsymbol{q}_1, \dots, \boldsymbol{q}_n$ that is distributed as $\boldsymbol{q}$. To this end, we introduce the following empirical versions of the function

$$r \mapsto F(r) := \mathbb{E}_{\boldsymbol{q}\sim\mathbb{P}}[W_{2,\varepsilon}^2(r, \boldsymbol{q})] \tag{C.3}$$

that are defined as

$$\hat{F}(r) = \frac{1}{n}\sum_{i=1}^n W_{2,\varepsilon}^2(r, \boldsymbol{q}_i) \quad \text{and} \quad \hat{F}^{(i)}(r) = \frac{1}{n}\left(\sum_{j=1, j\neq i}^n W_{2,\varepsilon}^2(r, \boldsymbol{q}_j) + W_{2,\varepsilon}^2(r, \boldsymbol{q}_i')\right),$$

and we introduce the quantities

$$\boldsymbol{r}_n^{\varepsilon,(i)} = \operatorname*{arg\,min}_{r\in\Sigma_N^\rho} \hat{F}^{(i)}(r) \quad \text{and} \quad Z^{(i)} = |\boldsymbol{r}_n^{\varepsilon,(i)} - r^\varepsilon|.$$

Then, we proceed as in the proof of Theorem 6 in [32]. First, we remark that

$$\begin{aligned}
\hat{F}(\boldsymbol{r}_n^{\varepsilon,(i)}) - \hat{F}(\boldsymbol{r}_n^\varepsilon) &= \frac{W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i) - W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, \boldsymbol{q}_i)}{n} + \frac{1}{n}\sum_{j=1, j\neq i}^n W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_j) - W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, \boldsymbol{q}_j) \\
&= \frac{W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i) - W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, \boldsymbol{q}_i)}{n} + \frac{W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, \boldsymbol{q}_i') - W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i')}{n} \\
&\quad + \hat{F}^{(i)}(\boldsymbol{r}_n^{\varepsilon,(i)}) - \hat{F}^{(i)}(\boldsymbol{r}_n^\varepsilon) \\
&\leq \frac{|W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i) - W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, \boldsymbol{q}_i)|}{n} + \frac{|W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, \boldsymbol{q}_i') - W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i')|}{n} \\
&\leq \frac{2L_{\rho,\varepsilon}}{n}|\boldsymbol{r}_n^{\varepsilon,(i)} - \boldsymbol{r}_n^\varepsilon|,
\end{aligned} \tag{C.4}$$

where the first inequality follows from the fact that $\boldsymbol{r}_n^{\varepsilon,(i)}$ is a minimizer of $\hat{F}^{(i)}$, and the second one from Lemma 3.5 on the Lipschitz continuity of $r \mapsto W_{2,\varepsilon}^2(r, q)$. Now, using Theorem 3.4, one has that the function $\hat{F}$ is ε-strongly convex, which implies that

$$\hat{F}(\boldsymbol{r}_n^{\varepsilon,(i)}) - \hat{F}(\boldsymbol{r}_n^\varepsilon) \geq \frac{\varepsilon}{2}|\boldsymbol{r}_n^{\varepsilon,(i)} - \boldsymbol{r}_n^\varepsilon|^2. \tag{C.5}$$

Combining (C.4) and (C.5), we obtain that $|\boldsymbol{r}_n^{\varepsilon,(i)} - \boldsymbol{r}_n^\varepsilon| \leq \frac{4L_{\rho,\varepsilon}}{\varepsilon n}$, and note that by the Lipschitz continuity of $r \mapsto W_{2,\varepsilon}^2(r, q)$, it follows that, for any $q \in \Sigma_N^\rho$,

$$|W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, q) - W_{2,\varepsilon}^2(\boldsymbol{r}_n^\varepsilon, q)| \leq \frac{4L_{\rho,\varepsilon}^2}{\varepsilon n}. \tag{C.6}$$

By the triangle inequality we finally obtain that

$$|Z - Z^{(i)}| = ||\boldsymbol{r}_n^\varepsilon - r^\varepsilon| - |\boldsymbol{r}_n^{\varepsilon,(i)} - r^\varepsilon|| \leq |\boldsymbol{r}_n^{\varepsilon,(i)} - \boldsymbol{r}_n^\varepsilon| \leq \frac{4L_{\rho,\varepsilon}}{\varepsilon n},$$

which proves that inequality (C.2) holds. Now, using concentration of measure for a function of random variables satisfying the bounded difference inequality (C.2), we obtain that, for any $t > 0$ (see e.g. Theorem 6.2 in [7])

$$\mathbb{P}\left(|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon| - \mathbb{E}[|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon|] > t\right) \leq \exp\left(-\frac{n\varepsilon^2 t^2}{8L_{\rho,\varepsilon}^2}\right). \tag{C.7}$$

To conclude the proof it remains to obtain an upper bound on $\mathbb{E}[|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon|]$. We start with the basic inequality

$$\mathbb{E}[|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon|] \leq \sqrt{\mathbb{E}[|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon|^2]}. \tag{C.8}$$

By Theorem 3.4, it follows that the function F defined in (C.3) is ε-strongly convex for the Euclidean 2-norm. Hence, as r^ε is by definition a minimizer of F, one has that

$$|\hat{\boldsymbol{r}}_n^\varepsilon - r^\varepsilon|^2 \leq \frac{2}{\varepsilon}\left(F(\hat{\boldsymbol{r}}_n^\varepsilon) - F(r^\varepsilon)\right), \tag{C.9}$$

Then, for any $1 \leq i \leq n$, since $\boldsymbol{q}_i'$ is an independent copy of $\boldsymbol{q}_i$, we clearly have that

$$\mathbb{E}[F(\hat{\boldsymbol{r}}_n^\varepsilon)] = \mathbb{E}[F(\boldsymbol{r}_n^{\varepsilon,(i)})] = \mathbb{E}[W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i)]$$

Hence, we may write $\mathbb{E}[F(\hat{\boldsymbol{r}}_n^\varepsilon)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[W_{2,\varepsilon}^2(\boldsymbol{r}_n^{\varepsilon,(i)}, \boldsymbol{q}_i)]$, and since one also has that $\mathbb{E}[\hat{F}(\hat{\boldsymbol{r}}_n^\varepsilon)] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[W_{2,\varepsilon}^2(\hat{\boldsymbol{r}}_n^\varepsilon, \boldsymbol{q}_i)]$, inequality (C.6) yields

$$\mathbb{E}[F(\hat{\boldsymbol{r}}_n^\varepsilon) - \hat{F}(\hat{\boldsymbol{r}}_n^\varepsilon)] \leq \frac{4L_{\rho,\varepsilon}^2}{\varepsilon n}.$$

Finally, as $F(r^\varepsilon) = \mathbb{E}[\hat{F}(r^\varepsilon)] \geq \mathbb{E}[\hat{F}(\hat{\boldsymbol{r}}_n^\varepsilon)]$ we obtain from the above inequality that

$$\mathbb{E}[F(\hat{\boldsymbol{r}}_n^\varepsilon)] - F(r^\varepsilon) \leq \frac{4L_{\rho,\varepsilon}^2}{\varepsilon n}. \tag{C.10}$$

Combining inequalities (C.7), (C.8) (C.9) and (C.10) allows to conclude the proof of Proposition C.1.

□

Proposition C.2. *For any $u > 0$*

$$\mathbb{P}\left(|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2 > \frac{2L_{\rho,\varepsilon}}{\varepsilon}\left(u + \rho(N + \sqrt{N})\right)\right) \leq 2^N \sum_{i=1}^n \exp\left(-p_i u^2\right), \tag{C.11}$$

where $\boldsymbol{r}_n^\varepsilon$ is the Sinkhorn barycenter of the iid random measures $\boldsymbol{q}_1, \ldots, \boldsymbol{q}_n$ (assumed to belong to Σ_N^ρ) as defined by (3.7).

Proof. By arguing as in the proof of Theorem 3.2, we have the following inequality

$$|\boldsymbol{r}_n^\varepsilon - \hat{\boldsymbol{r}}_{n,p}^\varepsilon|^2 \leq \frac{2L_{\rho,\varepsilon}}{\varepsilon}\left(\frac{1}{n}\sum_{i=1}^n |\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| + \rho(N+\sqrt{N})\right). \tag{C.12}$$

Then, conditionally on $\boldsymbol{q}_i$, we recall that $p_i\tilde{\boldsymbol{q}}_i^{p_i}$ is a random vector following a multinomial distribution $\mathcal{M}(p_i, \boldsymbol{q}_i)$. Hence, using the fact that the Euclidean 2-norm satisfies $|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| \leq \sum_{k=1}^N |\boldsymbol{q}_{i,k} - \tilde{\boldsymbol{q}}_{i,k}^{p_i}|$, it follows from the so-called Bretagnolle-Huber-Carol inequality (see Proposition A6.6 in [33]) and by conditioning on $\boldsymbol{q}_i$ that, for any $u > 0$,

$$\mathbb{P}\left(|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| \geq u\right) \leq 2^N \exp\left(-\frac{p_i u^2}{2}\right). \tag{C.13}$$

Hence, combining inequalities (C.12) and (C.13) with the union bound $\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n |\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| \geq u\right) \leq \sum_{i=1}^n \mathbb{P}\left(|\boldsymbol{q}_i - \tilde{\boldsymbol{q}}_i^{p_i}| \geq u\right)$ allows to complete the proof of Proposition C.2.

□

D Algorithms to compute penalized Wasserstein barycenters of Section 2

In this section we describe how the minimization problem

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

can be solved numerically by using an appropriate discretization to compute a numerical approximation of a regularized Wasserstein barycenter and the work of [14]. In our 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. A discretization of the minimization problem (D.1) 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 optimization problem (D.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 $d \geq 1$.

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:

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

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 (D.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

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

where $\tau^{(\ell)}$ is the ℓ-th step time, and Π_S stands for the projection on the simplex $S = \{y \in \mathbb{R}_+^N$ such that $\sum_{j=1}^N y^j = 1\}$. Thanks to Proposition 5 in [28], we are able to compute a subgradient 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 D.1 ([28]). *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*

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

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 [8], 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. [3]) accelerate the speed of convergence of the above described algorithm.

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:

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

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 (D.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 (D.5) is equivalent to the convex problem

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

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 (D.5).

To this end, we rely on the dual approach introduced in [10] and the numerical optimisation scheme proposed in [14]. Following these works, one can show that the dual problem of (D.5) with a regularization of the form $E(Kf)$ and K a discrete linear operator reads as

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

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 (D.7) as:

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

Following [10], one value of the above subgradient can be obtained at point x^k as:

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

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

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

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 [14], the dual problem (D.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 [14] 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:

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

The subgradient (D.9) can then be used in a descent algorithm over the dual problem (D.10). For differentiable penalizers E, we consider the L-BFGS algorithm [36, 4] that integrates a line search method (see e.g. [9]) to select the best time step $\tau^{(\ell)}$ at each iteration ℓ of the subgradient descent:

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

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 (D.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 (D.11), by changing the explicit scheme involving ∇E^*_γ onto an implicit one through the proximity operator of E^*_γ:

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

Algorithmic issues and stabilization As detailed in [10], the computation of one subgradient in (D.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 $w_i^{jk} = \exp(-(\frac{1}{2}\|y_i^j - x^k\|^2 - \phi_i^k)/\varepsilon)$ 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] L. Ambrosio, N. Gigli, and G. Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

[3] 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.

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

[5] J. Bigot, E. Cazelles, and N. Papadakis. Penalization of barycenters in the Wasserstein space. *SIAM J. Math. Analysis*, To be published, 2019.

[6] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. *Electron. J. Stat.*, 12(2):2253–2289, 2018.

[7] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

[8] S. Boyd and A. Mutapcic. Subgradient methods. *Lecture notes of EE364, Stanford University*, 2007, 2006.

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

[10] 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.

[11] Guillaume Carlier, Vincent Duval, Gabriel Peyré, and Bernhard Schmitzer. Convergence of entropic schemes for optimal transport and gradient flows. *SIAM J. Math. Analysis*, 49(2):1385–1418, 2017.

[12] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc., 2013.

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

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

[15] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Annales de l'Institut H.Poincaré, Sect. B, Probabilités et Statistiques*, 10:235–310, 1948.

[16] A. Genevay, G. Cuturi, M. andPeyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Proc. NIPS'16*. Curran Associates, Inc., 2016.

[17] Robert C. Gentleman, Vincent J. Carey, Douglas M. Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J. Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth,

Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, Sep 2004.

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

[19] Alexander Goldenshluger, Oleg Lepski, et al. Universal pointwise selection rule in multivariate function estimation. *Bernoulli*, 14(4):1150–1190, 2008.

[20] F. Hahne, A.H. Khodabakhshi, A. Bashashati, C.-J. Wong, R.D. Gascoyne, A.P. Weng, V. Seyfert-Margolis, K. Bourcier, A. Asare, T. Lumley, R. Gentleman, and R.R. Brinkman. Per-channel basis normalization methods for flow cytometry data. *Cytometry Part A*, 77(2):121–131, 2010.

[21] B.P. Hejblum, C. Alkhassim, R. Gottardo, F. Caron, and R. Thiébaut. Sequential dirichlet process mixtures of multivariate skew t-distributions for model-based clustering of flow cytometry data. *ArXiv preprint: 1702.04407*, 2017.

[22] A. Kneip and K.J. Utikal. Inference for density families using functional principal component analysis. *Journal of the American Statistical Association*, 96(454):519–542, 2001.

[23] Claire Lacour and Pascal Massart. Minimal penalty for goldenshluger–lepski method. *Stochastic Processes and their Applications*, 126(12):3774–3789, 2016.

[24] S.X. Lee, G.J. McLachlan, and S. Pyne. Modeling of inter-sample variation in flow cytometric data with the joint clustering and matching procedure. *Cytometry Part A*, 89(1):30–43, 2016.

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

[26] V. M. Panaretos and Y. Zemel. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 2017.

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

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

[29] S. Pyne, S.X. Lee, K. Wang, J. Irish, P. Tamayo, M.-D. Nazaire, T. Duong, S.-K. Ng, D. Hafler, R. Levy, G.P. Nolan, J. Mesirov, and G.J. McLachlan. Joint modeling and registration of cell populations in cohorts of high-dimensional flow cytometric data. *PloS one*, 9(7), 2014.

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

[31] F. Santambrogio. *Optimal Transport for Applied Mathematicians - Calculus of Variations, PDEs, and Modeling*. Springer Verlag Italia, 2015.

[32] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[33] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

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

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

[36] 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, 1997.