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

July 2, 2018

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 [4]. A second strategy is to transform the Wasserstein metric itself into an entropy regularized transportation cost between probability measures as introduced in [10]. 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.

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 $f_i$ for each $1 \leq i \leq n$. In the presence of phase variation in the observations due to mis-alignment in the acquisition process, it is necessary to use a 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.

*J. Bigot is a member of Institut Universitaire de France.
made of observations sampled from Gaussian mixtures $f_i$ with two components whose means and variances are randomly chosen for each distribution. Certainly, one can estimate a mean density using a preliminary smoothing step (with a kernel $K$ and data-driven choices of the bandwidth parameters $(h_i)_{i=1,...,n}$) followed by standard averaging, that is considering

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

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

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 $\hat{f}_{n,p}$ of the red and blue densities. (b) The solid black curve is the entropy regularized Wasserstein barycenter $\hat{r}_{n,p}$ (defined in (1.5)) of the raw data using a Sinkhorn divergence and the numerical approach from [12], with a data-driven choice for $\hat{\epsilon} = 2.55$.

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 [32], economics [20], and in the analysis of spike trains in neuroscience [31] or functional connectivity between brain regions [25]. For $d \geq 2$ the issue of data registration arises in the statistical analysis of spatial point processes [16, 24] or flow cytometry data [17, 27].
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 [5, 23, 24]. A Wasserstein barycenter is a Fréchet mean [13] 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

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

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

In the case of a finite space $\Omega_N = \{x_1, \ldots, x_N\} \subset (\mathbb{R}^d)^N$ of cardinal $N$, a discrete probability distribution $\rho$ (with fixed support included in $\Omega_N$) is identified by a vector in the simplex $\Sigma_N = \{\rho = (\rho_1, \ldots, \rho_N) \in \mathbb{R}_N^+ \text{ with } \sum_{k=1}^N \rho_k = 1\}$ such that $\rho = \sum_{k=1}^N r_k \delta_{x_k}$ where $\delta_x$ is the Dirac distribution at $x$. The Wasserstein distance between two discrete distributions $\rho$ and $\sigma$ in $\Sigma_N$ then becomes

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

where the set of couplings is defined as $U(\rho, \sigma) := \{U \in \mathbb{R}_N^{N \times N} \text{ such that } U \mathbf{1}_N = \rho, \ U^T \mathbf{1}_N = \sigma\}$ with $\mathbf{1}_N$ 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 $m, l \in \{1, \ldots, N\}$.

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

$$\hat{\nu}_i^{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{X_{i,j}} \quad \text{for } 1 \leq i \leq n,$$  \hspace{1cm} (1.2)

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

1.2.1 Penalized Wasserstein barycenters

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

$$\hat{\mu}^\gamma_{n,p} = \arg \min_{\mu \in \mathcal{P}_2(\Omega)} \frac{1}{n} \sum_{i=1}^n W_2^2(\mu, \hat{\nu}_i^{p_i}) + \gamma E(\mu),$$  \hspace{1cm} (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 $\hat{\mu}^\gamma_{n,p}$ to be absolutely continuous. Theoretical properties (such as existence and consistency) of the penalized Wasserstein barycenter $\hat{\mu}^\gamma_{n,p}$ have been considered in [4]. In this paper, we discuss the choice of the penalty function $E$, as well as the numerical computation of $\hat{\mu}^\gamma_{n,p}$ (using an appropriate discretization of $\Omega$ and a binning of the data), and its benefits for statistical data analysis.
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 \cite{10, 9} leading to the so-called Sinkhorn divergence

\[
W_{2,\varepsilon}^2(\mu, \nu) = \inf_{\pi \in \Gamma(\mu, \nu)} \int_{\Omega^2} |x - y|^2 d\pi(x, y) - \varepsilon h(\pi),
\]

(1.4)

where \(\varepsilon > 0\) is a regularization parameter, and \(h\) stands for the (negative) entropy of the transport plan \(\pi\) with respect to the Lebesgue measure on \(\Omega \times \Omega\). A regularized Wasserstein barycenter \cite{11, 12} is then obtained by considering the estimator

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

(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 \(\gamma\) or \(\varepsilon\) 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 \(\gamma\) in (1.3) and \(\varepsilon\) in (1.5) using the Goldenshluger-Lepski (GL) method (as formulated in \cite{21}), which leans on a bias-variance trade-off function, described in details in Section 4.1. 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 \(\hat{r}_{n,p}^\varepsilon\) (with an appropriate choice for \(\varepsilon\)) 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 \(\gamma\) and \(\varepsilon\) may also be interpreted as the usual bandwidth parameter in kernel density estimation, and their choice greatly influences the shape of the estimators \(\hat{\mu}_{n,p}^\gamma\) and \(\hat{r}_{n,p}^\varepsilon\) (see Figure 7 and Figure 8 in Section 4).

To choose the optimal parameter, the GL’s strategy requires some variance information through the knowledge of an upper bound on the decay to zero of the expected \(L^2(\Omega)\) 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 \cite{4}. 
1.3.2 Variance of Sinkhorn estimators

To the best of our knowledge, the automatic selection of $\varepsilon$ in the definition of a Sinkhorn divergence has not been considered so far. Another main contribution of this work then consists in deriving upper bounds on the variance for the estimators $\hat{\mu}_{\varepsilon}^{n,p}$ which explicitly depends on $\varepsilon$, the number $n$ of measures, the number $p = \min_{1 \leq i \leq n} p_i$ of observations per measures et the size of their support. Such bounds therefore make possible the application of the GL’s strategy.

1.3.3 Computation issues: binning of the data and discretization of $\Omega$

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 $\hat{\mu}_{\varepsilon}^{n,p}$ and $\hat{\mu}_{\varepsilon}^{n,p}$ 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 $\Omega_N$) that we denote

$$\tilde{q}_i^{p_i} = \frac{1}{p_i} \sum_{j=1}^{p_i} \delta_{\tilde{X}_{i,j}},$$

where $\tilde{X}_{i,j} = \arg\min_{x \in \Omega_N} |x - X_{i,j}|$, (1.6)

for $1 \leq i \leq n$. In this paper, we rely on the smooth dual approach proposed in [12] to compute penalized and Sinkhorn barycenters on a grid of equi-spaced points in $\Omega$ (after a proper binning of the data).

Binning (i.e. choosing the grid $\Omega_N$) 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 4.1 where we describe the GL’s strategy. Besides, in our simulations, the
choice of $N$ is mainly guided by numerical issues on the computational cost of the algorithms used to approximate $\hat{\mu}_{n,p}^\gamma$ and $\hat{r}_{n,p}^\varepsilon$.

1.3.4 Registration of flow cytometry data

In biotechnology, flow cytometry is a high-throughput technique that can measures 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 their differentiation. 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. [17, 18, 22, 27] and references therein).

When analyzing samples in cytometry measured from different patients, a critical issue is data registration across patients. As carefully explained in [17], 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 [17] 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 [15] 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 through 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 $\hat{\mu}_{n,p}^\gamma$ and $\hat{r}_{n,p}^\varepsilon$ are detailed in Section 2 and Section 3 respectively. Section 4 contains a description of the Goldenshluger-Lepski principle to choose the regularization parameters $\gamma$ and $\varepsilon$. It also reports the results from numerical experiments using simulated data and the flow cytometry dataset displayed in Figure 3. Some proofs and technical results are presented in Appendix A and Appendix B. Algorithmic details on the computation of the estimator $\hat{\mu}_{n,p}$ are gathered in Appendix C.
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 [4] in which the Wasserstein barycenter is regularized through a convex penalty function as presented in (1.3).

2.1 The minimization problem

Let us remind some of the basic definitions in [4]. We define $W^2(P_2(\Omega))$ as the space of distributions $P$ supported on $\mathcal{P}_2(\Omega)$. The penalized Wasserstein barycenter associated to the distribution $P$ 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

$$E : \mathcal{P}_2(\Omega) \rightarrow \mathbb{R}_+$$

$$\mu \quad \mapsto \quad \left\{ \begin{array}{ll}
\int_{\Omega} G \left( \frac{d\mu}{d\lambda}(x) \right) d\lambda(x), & \text{if } \mu \ll \lambda \\
+\infty & \text{otherwise,}
\end{array} \right. \quad (2.1)$$

for $\lambda$ a given positive measure on $\Omega$ and $G : [0, +\infty) \rightarrow [0, +\infty]$ a proper, lower semicontinuous and strictly convex function with superlinear growth. Remark that the function $E$ is strictly
convex on its domain
\[ \mathcal{D}(E) = \{ \mu \in \mathcal{P}_2(\Omega) \text{ such that } E(\mu) < +\infty \}. \] (2.2)

A typical example consists in taking \( G \) as the square function and \( \lambda \) as the Lebesgue measure on \( \Omega \), which imposes the barycenter \( \mu \) to belong to \( \mathcal{P}_2^{ac}(\Omega) \), the space of measures in \( \mathcal{P}_2(\Omega) \) that are absolutely continuous. This choice is driven by the need to retrieve an absolutely continuous measure from discrete observations \( (X_{ij})_i \), as it is often done when approximating data through kernel smoothing in density estimation. Others examples of penalty functions are given in [4].

**Definition 2.1.** Let \( \nu_1, \ldots, \nu_n \in \mathcal{P}_2(\Omega) \) be iid measures with distribution \( P \) such that \( P(\mathcal{P}_2^{ac}(\Omega)) > 0 \). The empirical probability measure defined by \( (\nu_1, i=1, \ldots, n) \) writes \( P_n = 1/n \sum_{i=1}^n \delta_{\nu_i} \). Let us then consider the random measures \( \hat{\nu}_i^{\nu} = 1/p_i \sum_{j=1}^{p_i} \delta_{X_{ij}} \) where \( (X_{ij})_{1 \leq i \leq n} \) are iid random variables of law \( \nu_i \). From there on we define for \( \gamma > 0 \) the barycenters:

\[
\hat{\mu}^\gamma = \arg \min_{\mu \in \mathcal{P}_2(\Omega)} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) dP(\nu) + \gamma E(\mu) = E_{\nu \sim P}[W_2^2(\mu, \nu)] + \gamma E(\mu) \tag{2.3}
\]

\[
\hat{\mu}_{n,p}^\gamma = \arg \min_{\mu \in \mathcal{P}_2(\Omega)} \int_{\mathcal{P}_2(\Omega)} W_2^2(\mu, \nu) dP_n(\nu) + \gamma E(\mu) = 1/n \sum_{i=1}^n W_2^2(\mu, \hat{\nu}_i^{\nu}) + \gamma E(\mu) \tag{2.4}
\]

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

In [4], the existence and uniqueness of these barycenters have been shown for \( \gamma > 0 \). By penalizing the barycenters with function \( E \) of the form (2.1), we enforce them to be absolutely continuous. Therefore, let \( \tilde{f}_{n,p}^\gamma \) and \( f_p^\gamma \) be the densities associated to \( \hat{\mu}_{n,p}^\gamma \) and \( \hat{\mu}^\gamma \).

In order to apply the GL’s strategy, we now analyze the variance of \( \hat{\mu}_{n,p}^\gamma \) by studying the expected squared \( L_2(\Omega) \)-distance \( E(\|\tilde{f}_{n,p}^\gamma - f_p^\gamma\|^2) \).

### 2.2 Variance properties of the penalized empirical barycenter \( \hat{\mu}_{n,p}^\gamma \)

From the discussion in Section 5 of [4], we choose the penalization function
\[
E(\mu) = \begin{cases} \|f\|^2_{H^k(\Omega)} & \text{if } f = \frac{d\mu}{dx} \text{ and } f \geq \alpha, \\ +\infty & \text{otherwise.} \end{cases} \tag{2.5}
\]

where \( \| \cdot \|_{H^k(\Omega)} \) denotes the Sobolev norm associated to the \( L_2(\Omega) \) space, \( \alpha > 0 \) is arbitrarily small and \( k > d - 1 \).

**Theorem 2.2** (Section 5 in [4]). Let \( \Omega \subset \mathbb{R}^d \) be compact and \( \tilde{f}_{n,p}^\gamma \) and \( f_p^\gamma \) be the density functions of \( \hat{\mu}_{n,p}^\gamma \) and \( \hat{\mu}^\gamma \), induced by the choice (2.5) of the penalty function \( E \). Then there exists a constant \( c > 0 \) depending only on \( \Omega \), such that
\[
E\left(\|\tilde{f}_{n,p}^\gamma - f_p^\gamma\|^2_{L_2(\Omega)}\right) \leq c \left( \frac{1}{\gamma p^{1/3}} + \frac{1}{\gamma n^{1/2}} \right) \tag{2.6}
\]

where \( p = \min_{1 \leq i \leq n} p_i \) and provided that \( d < 4 \) and \( E(f_{\Omega} |x|^q d\nu_1(x)) < +\infty \) for some \( q > 4 \).

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.
2.3 Numerical computation

As a practical complement to [4], we provide in Appendix C efficient minimization algorithms for the computation of $\hat{f}_{n,p}^{\gamma}$, after a binning of the data on a fixed grid $\Omega_N$. For $\Omega$ 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 [12].

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 $\Omega_N$. For two discrete measures $r,q \in \Sigma_N$, the Sinkhorn divergence (1.4) reads for $\varepsilon > 0$

$$W_{2,\varepsilon}^2(r,q) := \min_{U \in U(r,q)} \langle C,U \rangle - \varepsilon h(U).$$

(3.1)

where the discrete (negative) entropy for a given coupling $U \in U(r,q)$ is given by $h(U) := -\sum_{m,t} U_{mt} \log U_{mt}$. 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 $r \mapsto W_{2,\varepsilon}^2(r,q)$ (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 $\rho$, 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 (constrained to belong to the set $\Sigma^\rho_N$).

**Definition 3.1.** Let $0 < \rho < 1/N$, and $\mathbb{P}$ be a probability distribution on $\Sigma^\rho_N$. Let $q_1, \ldots, q_n \in \Sigma^\rho_N$ be an iid sample drawn from the distribution $\mathbb{P}$. For each $1 \leq i \leq n$, we assume that $(X_{i,j})_{1 \leq j \leq s_i}$ are iid random variables sampled from $q_i$. For each $1 \leq i \leq n$, let us define the following discrete measures

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

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 $q_i^{\rho} \in \Sigma^\rho_N$ for all $1 \leq i \leq n$, which may not be the case for some
\( \hat{q}^p_i, 1 = 1, \ldots, n \). Then, we define

\[
\hat{r}^\varepsilon = \arg \min \mathbb{E}_{q \sim \mathbb{P}}[W^2_2(\hat{r}, q)] \quad \text{the population Sinkhorn barycenter (3.2)}
\]

\[
\hat{r}^\varepsilon_{n,p} = \arg \min_{r \in \Sigma^\rho_N} \frac{1}{n} \sum_{i=1}^n W^2_2(\hat{r}, \hat{q}^\varepsilon_i) \quad \text{the empirical Sinkhorn barycenter (3.3)}
\]

In the optimisation problem (3.3), we choose to use the discrete measures \( \hat{q}^\varepsilon_i \) instead of the empirical measures \( \tilde{q}^\varepsilon_i \) to guarantee the use of discrete measures belonging to \( \Sigma^\rho_N \) in the definition of the empirical Sinkhorn barycenter \( \hat{r}^\varepsilon_{n,p} \).

The following theorem is the main result of this section which gives an upper bound on the variance of \( \hat{r}^\varepsilon_{n,p} \) in terms of the expected squared Euclidean norm between elements of \( \Sigma_N \).

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

\[
\mathbb{E}(|r^\varepsilon - \hat{r}^\varepsilon_{n,p}|^2) \leq \frac{32L^2_{\rho,\varepsilon}}{\varepsilon^2 n} + \frac{2L_{\rho,\varepsilon}}{\varepsilon} \left( \frac{N}{p} + 2 \rho (N + \sqrt{N}) \right), \quad (3.4)
\]

with

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

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 4.1). Moreover, it can be used to discuss the choice of \( \rho \). First, if one take \( \rho = \epsilon^\kappa \), the Lipschitz constant (Lemma 3.5) \( L_{\rho,\varepsilon} = L_\varepsilon \) becomes

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

which is a constant (not depending on \( \rho \)) such that

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

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

\[
\mathbb{E}(|r^\varepsilon - \hat{r}^\varepsilon_{n,p}|^2) \leq \frac{32L^2_{\rho,\varepsilon}}{\varepsilon^2 n} + \frac{2L_{\rho,\varepsilon}}{\varepsilon} \left( \frac{N}{p} + 2 \left( \frac{N}{p} + \sqrt{\frac{N}{p^2}} \right) \right), \quad (3.6)
\]

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 the use of the strong convexity of the functional \( r \mapsto W^2_{2,\varepsilon}(r, q) \) for \( q \in \Sigma_N \), without constraint on its entries. This property can be derived by studying the Legendre transform of \( r \mapsto W^2_{2,\varepsilon}(r, q) \). For a fixed distribution \( q \in \Sigma_N \), using the notation in [12], we define the function
\[
H_q(r) := W^2_{2,\varepsilon}(r, q), \quad \text{for all } r \in \Sigma_N.
\]
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 [12]). For \( \varepsilon > 0 \), the Fenchel-Legendre dual function \( H^*_q \) is \( C^\infty \). Its gradient function \( \nabla H^*_q(g) \) is \( \frac{1}{\varepsilon} \)-Lipschitz. Its value, gradient and Hessian at \( g \in \mathbb{R}^N \) are, writing \( \alpha = \exp(g/\varepsilon) \) and \( K = \exp(-C/\varepsilon) \),
\[
\begin{align*}
H^*_q(g) &= \varepsilon(E(q) + \langle q, log(K\alpha) \rangle), \\
\nabla H^*_q(g) &= \text{diag}(\alpha)K\frac{q}{K\alpha} \in \Sigma_N \\
\nabla^2 H^*_q(g) &= \frac{1}{\varepsilon} \left( \text{diag} \left( \text{diag}(\alpha)K\frac{q}{K\alpha} \right) \right) - \frac{1}{\varepsilon} \text{diag}(\alpha)K \text{diag} \left( \frac{q}{(K\alpha)^2} \right) K \text{diag}(\alpha),
\end{align*}
\]
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 \( \varepsilon \)-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 \( r \in \Sigma_N^0 \).

**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 \( \Sigma_N^0 \) 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
\[
r^\varepsilon_n = \text{arg min}_{r \in \Sigma_N^0} \frac{1}{n} \sum_{i=1}^{n} W^2_{2,\varepsilon}(r, q_i) = \text{arg min}_{r \in \Sigma_N^0} \frac{1}{n} \sum_{i=1}^{n} H_{q_i}(r),
\]
of the iid random measures \( q_1, \ldots, q_n \) (assumed to belong to \( \Sigma_N^0 \)). By the triangle inequality, we have that
\[
\mathbb{E}(|r^\varepsilon - \hat{r}^\varepsilon_{n,p}|^2) \leq 2\mathbb{E}(|r^\varepsilon - r^\varepsilon_n|^2) + 2\mathbb{E}(|r^\varepsilon_n - \hat{r}^\varepsilon_{n,p}|^2).
\]
(3.7)

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 \( \varepsilon \)-strongly convex by Theorem 3.4 and \( L_{\rho,\varepsilon} \)-Lipschitz on \( \Sigma_N^0 \) by Lemma 3.5 where \( L_{\rho,\varepsilon} \) is the constant defined by equation (3.5). Under these assumptions, it follows from Theorem 6 in [30] that
\[
\mathbb{E}(|r^\varepsilon - r^\varepsilon_n|^2) \leq \frac{16L_{\rho,\varepsilon}^2}{\varepsilon^2 n}.
\]
(3.8)
For the second term in the right hand side of (3.7), we obtain by the strong convexity of \( H_q \) that

\[
\frac{1}{n} \sum_{i=1}^{n} H_q(\hat{r}_n^e) \geq \frac{1}{n} \sum_{i=1}^{n} H_q(r_n^e) + \frac{1}{n} \sum_{i=1}^{n} \nabla H_q(r_n^e)^T (\hat{r}_n^e - r_n^e) + \frac{\varepsilon}{2} |r_n^e - \hat{r}_n^e|^2.
\]

Theorem 3.1 in [12] ensures that \( \frac{1}{n} \sum_i \nabla H_q(r_n^e) = 0 \). The same inequality also holds for the terms \( H_q^{\hat{p}_i} \), and we therefore have

\[
\frac{1}{n} \sum_{i=1}^{n} H_q(\hat{r}_n^{\hat{p}_i}) \geq \frac{1}{n} \sum_{i=1}^{n} H_q(r_n^{\hat{p}_i}) + \frac{\varepsilon}{2} |r_n^{\hat{p}_i} - \hat{r}_n^{\hat{p}_i}|^2,
\]

\[
\frac{1}{n} \sum_{i=1}^{n} H_q^{\hat{p}_i}(\hat{r}_n^{\hat{p}_i}) \geq \frac{1}{n} \sum_{i=1}^{n} H_q^{\hat{p}_i}(\hat{r}_n^{\hat{p}_i}) + \frac{\varepsilon}{2} |r_n^{\hat{p}_i} - \hat{r}_n^{\hat{p}_i}|^2.
\]

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 \( \Sigma_N^q \) for any discrete distribution \( r \). Hence, by summing the two above inequalities, and by taking the expectation on both sides, we obtain that

\[
\varepsilon \mathbb{E}(|r_n^e - \hat{r}_n^e|^2) \leq \frac{2L_{\rho,\varepsilon}}{n} \sum_{i=1}^{n} \mathbb{E}(|q_i - \hat{q}_i^{p_i}|).
\]

Using the inequalities

\[
|q_i - \hat{q}_i^{p_i}| \leq |q_i - \hat{q}_i^{p_i}| + \rho N|\hat{q}_i^{p_i}| + \rho N \leq |q_i - \hat{q}_i^{p_i}| + \rho (N + \sqrt{N}),
\]

we finally have that

\[
\varepsilon \mathbb{E}(|r_n^e - \hat{r}_n^e|^2) \leq \frac{2L_{\rho,\varepsilon}}{n} \left( \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}(|q_i - \hat{q}_i^{p_i}|^2) + \rho (N + \sqrt{N}) \right).
\]

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

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

Thus, we have

\[
\mathbb{E}(|q_i - \hat{q}_i^{p_i}|^2) = \sum_{k=1}^{N} \mathbb{E}(q_{i,k} - \hat{q}_{i,k}^{p_i})^2 \leq \frac{1}{4} \sum_{k=1}^{N} p_i^{-1} \leq \frac{N}{4p}
\]

and we obtain from (3.9) and (3.10) that

\[
\mathbb{E}(|r_n^e - \hat{r}_n^{e,p_i}|^2) \leq \frac{L_{\rho,\varepsilon}}{\varepsilon} \left( \frac{N}{p} + 2\rho (N + \sqrt{N}) \right).
\]

Combining inequalities (3.7), (3.8), and (3.11) concludes the proof of Theorem 3.2.
4 Numerical experiments

In this section, we first present a method to automatically choose the parameters $\gamma$ in (1.3) and $\varepsilon$ in (1.5), that we illustrate with one-dimensional datasets. Then, we report the results from numerical experiments on simulated Gaussian mixtures and flow cytometry dataset in $\mathbb{R}^2$.

4.1 Goldenshluger-Lepski method

By analogy with the work in [21] based on the Goldenshluger-Lepski (GL) principle, we propose to compute a bias-variance trade-off functional which will provide an automatic selection method for the regularization parameters 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 the theory for the Sinkhorn barycenter described in Section 3. The trade-off functional is composed of a term measuring the disparity between two estimators and of a penalty term that is chosen according to the upper bounds of the variance in Section 3. More precisely, assume that we dispose of a finite collection of estimators $(\hat{r}_{n,p})_{\varepsilon}$ for $\varepsilon$ ranging in a space $\Lambda$ 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:

$$\hat{\varepsilon} = \arg \min_{\varepsilon \in \Lambda} B(\varepsilon) + bV(\varepsilon)$$

for which we set the “bias term” as

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

where $x_{+} = \max(x, 0)$ denotes the positive part. The authors in [21] propose a few suggestions to properly choose both the parameter $b > 0$ and the functional $V$. This leads to a “variance term” $V$ chosen proportional to the right-hand side of (3.4). Following [21] and from our numerical experiments, we observed that $bV$ has to depend on the size $N$ of the grid $\Omega_N$ in order to fit the scaling of the disparity term $|\hat{r}_{n,p} - \tilde{r}_{n,p}|^2$.

4.2 Simulated data: one-dimensional Gaussian mixtures

We illustrate GL’s principle as well as the choice of the parameter $b$ for the one-dimensional example of Gaussian mixtures that is displayed in Figure 4. Our dataset consists of observations $(X_{i,j})_{1 \leq i \leq n ; 1 \leq j \leq p}$ sampled from $n = 15$ random distributions $\nu_i$ 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.

Thanks to inequality (3.4), we choose to take the function $V$ defined by

$$V(\varepsilon) = \frac{32L_{\hat{r}}}{{\varepsilon^2}n} + \frac{2L_{\hat{r}}}{\varepsilon} \left( \sqrt{\frac{N}{p}} + 2\rho(N + \sqrt{N}) \right),$$

and $\rho = \min(1/N, 1/p)$. 

13
Figure 4: A subset of 8 histograms (out of \( n = 15 \)) obtained with random variables sampled from one-dimensional Gaussian mixtures distributions \( \nu_i \) (with random means and variances). Histograms are constructed by binning the data \((X_{i,j})_{1 \leq i \leq n ; 1 \leq j \leq p}\) on a grid \( \Omega_N \) of size \( N = 2^8 \).

By definition, see equation (3.5), the Lipschitz constant \( L \) is of order \( \sqrt{N} \) since the term
\[
\sup_{1 \leq \ell, k \leq N} |C_{m \ell} - C_{k \ell}| = \sup_{1 \leq \ell, k \leq N} |x_m - x_\ell|^2 - |x_k - x_\ell|^2 |
\]
is not influenced by the size of the grid but rather by the largest distances between points in the support. Hence the variance function \( V \) clearly scales polynomially fast with \( N \). To compensate this scaling effect, we choose the parameter \( b = a/N^2 \) for some constant \( a > 0 \), as our experiments suggest. Using this choice for \( b \), we obtain data-driven regularization parameters \( \hat{\epsilon} \) that are of the same order for different grid size \( N \) as it can be seen from Figure 5, where we display the function \( \varepsilon \mapsto B(\varepsilon) + bV(\varepsilon) \) for different values of \( a = bN^2 \) ranging from \( 10^{-9} \) to \( 10^{-6} \) and grid sizes \( N = 2^6, 2^8, 2^{10} \) (using the same data sampled from random Gaussian mixtures before binning). For a better representation, we normalize the trade-off functions since we are only interested in their minimizer. We also present in Figure 6 the Sinkhorn barycenters associated to the regularization parameters \( \hat{\epsilon} \) that minimize the trade-off functions displayed in Figure 5. Note that for a better visualization, we have again normalized these barycenters with respect to the grid of size \( N = 2^8 \). The shapes of these barycenters are very similar despite the change of grid size. Finally, we suggest to choose \( a \) such that the trade-off curve has a minimum that is roughly in the center of the parameter’s range of interest.

To define the variance function in the case of penalized barycenters \( \hat{f}^{\gamma}_{n,p} \) of Section 2, we use the upper bound in inequality (2.6) leading to the choice
\[
V(\gamma) = \frac{1}{\gamma p^{1/4}} + \frac{1}{\gamma n^{1/2}}
\]
We remark that the size of the grid does not appear in the above variance function. Thus, the parameter \( b \) is chosen independent of \( N \) in the trade-off function \( \gamma \mapsto B(\gamma) + bV(\gamma) \).
Figure 5: Influence of the parameter \( b = a/N^2 \) on the shape of the bias-variance trade-off function \( \varepsilon \mapsto B(\varepsilon) + bV(\varepsilon) \) for Sinkhorn barycenters of the dataset in Figure 4. The range of \( \varepsilon \) is the interval \([0, 10]\). A zoom is performed for the two figures (c) and (d). From left to right and top to bottom, \( a = 10^{-9}, 10^{-8}, 10^{-7}, 10^{-6} \). The dotted red curve corresponds to a grid size \( N = 2^6 \), the dashed green curve to \( N = 2^8 \) and the solid blue curve to \( N = 2^{10} \).
Figure 6: Sinkhorn barycenters of the dataset in Figure 4 associated to the regularization parameters $\hat{\epsilon}$ minimizing the trade-off functions displayed in Figure 5 for different values of $a$ and different grid sizes $N$. 

(a) $a = 10^{-9}$

(b) $a = 10^{-8}$

(c) $a = 10^{-7}$

(d) $a = 10^{-6}$
Figure 7: One dimensional Gaussian mixtures dataset and Sinkhorn barycenters. (a) The trade-off function \( \varepsilon \mapsto B(\varepsilon) + bV(\varepsilon) \) which attains its optimum at \( \hat{\varepsilon} = 1.94 \). (b) Three Sinkhorn barycenters \( \hat{r}_{n,p}^{\varepsilon} \) associated to the parameters \( \varepsilon = 0.18, 1.94, 9.5 \).

From now on, we fix the size \( N = 2^8 \) of the grid, and we comment the choice of the parameters \( \hat{\varepsilon} \) and \( \hat{\gamma} \). We display in Figure 7(a) the trade-off function \( B(\varepsilon) + bV(\varepsilon) \), and we discuss the influence of \( \varepsilon \) on the smoothness and support of the Sinkhorn barycenter. From Figure 7(b), we observe that, when the parameter \( \varepsilon = 0.18 \) is small (dotted blue curve), then the corresponding Sinkhorn barycenter \( \hat{r}_{n,p}^{\varepsilon} \) is irregular, and it presents spurious peaks. On the contrary, too much regularization, e.g. \( \varepsilon = 9.5 \), implies that the barycenter (dashed green curve) is flattened and its mass is spread out. Here, the optimal barycenter (solid red curve), that is \( \hat{r}_{n,p}^{\varepsilon} \) for \( \varepsilon = 1.94 \) minimizing the trade-off function (4.1), seems to be a good compromise between under and over-smoothing.

We repeat the same experiment for the penalized barycenter of Section 2 with the Sobolev norm \( H^1(\Omega) \) in the penalization function \( E \) (2.5). The results are displayed in Figure 8. The advantage of choosing a Sobolev penalty function 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 \( \tilde{f}_{n,p}^{\gamma} \) presents a lot of irregularities as the penalty function tries to minimize its \( L_2 \)-norm. When the regularization parameter increases in a significant way (\( \gamma = 980 \) 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 \( \hat{\gamma} = 520 \) 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-dimensional sampling per subject (\( p = 50 \)).
Figure 8: One dimensional Gaussian mixtures dataset and penalized barycenters. (a) The trade-off function $\gamma \mapsto B(\gamma) + bV(\gamma)$ which attains its optimum at $\hat{\gamma} = 520$. (b) Three penalized barycenters $f^\gamma_{n,p}$ associated to the parameters $\gamma = 20, 520, 980$.

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.
4.3 Simulated data: two-dimensional Gaussian mixtures

In this section, we illustrate the validity of 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 \(\nu_i = \sum_{j=1}^{3} \theta_j N(m^i_j, \Gamma^i_j)\) with fixed weights \(\theta = (1/6, 1/3, 1/2)\). The means \(m^i_j\) and covariance matrices \(\Gamma^i_j\) are random variables with expectation given by (for \(j = 1, 2, 3\))

\[
\begin{align*}
m_1 &= \begin{pmatrix} 0 \\ 0 \end{pmatrix}, & m_2 &= \begin{pmatrix} 7/4 \\ 0 \end{pmatrix}, & m_3 &= \begin{pmatrix} 1/9 \\ 0 \end{pmatrix}, & \text{and} & & \Gamma_1 = \Gamma_2 = \Gamma_3 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.
\end{align*}
\]

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 \(\nu_i = \sum_{j=1}^{3} \theta_j N(m^i_j, \Gamma^i_j)\) where \(m^i_j\) (resp. \(\Gamma^i_j\)) are random vectors (resp. matrices) such that each of their coordinate follows a uniform law centered in \(m^i_j\) with amplitude \(\pm 2\) (resp. each of their diagonal elements follows a uniform law centered in \(\Gamma^i_j\) with amplitude \(\pm 0.95\)). 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 \(64 \times 64\) (thus \(N = 4096\)).

First, we compute 60 Sinkhorn barycenters by letting \(\varepsilon\) ranging from 0.1 to 6. We draw in Figure 11(a) the trade-off function, and we plot a zoom of this function in Figure 11(b) around the minimizer \(\hat{\varepsilon} = 3.6\). The corresponding Sinkhorn barycenter \(\hat{r}^{\hat{\varepsilon}}_{n,p}\) is displayed in Figure 12(a). 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 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.

Finally, we display the penalized barycenter of this dataset in Figure 13(b) for the Sobolev penalty function (2.5) and the data-driven choice of $\gamma$. In Figure 13(a), we plot the trade-off function for $\gamma$ ranging from 1 to 140. This curve suggests to choose $\hat{\gamma} = 80$. From Figure 13(b), we observe that the mass of $\hat{f}_{\gamma}^{\varepsilon,n,p}$ is concentrated on three main modes. The approach thus manages to keep the geometry of the underlying Gaussian mixtures.

### 4.4 Sinkhorn versus penalized barycenters

To conclude these numerical experiments with 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. Its computation requires $O(N^3 \log N)$ operations for discrete probability measures with a support of size $N$ when the computation of a Sinkhorn divergence only takes $O(N^2)$ operations at each iteration of a gradient descent (see e.g. [10]). 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.

### 4.5 Real data: flow cytometry

We have at our disposal data from flow cytometry that have been described in Section 1.3.4, 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 14(a) we plot the trade-off function related to the Sinkhorn barycenters for the parameter $\varepsilon$ ranging from 1 to 6. Its minimum is attained for $\hat{\varepsilon} = 3.1$. We display the
Figure 12: Two-dimensional Gaussian mixtures dataset. (a) The Sinkhorn barycenter $\hat{\mathbf{r}}_{n,p}$ for $\hat{\varepsilon} = 3.6$ chosen by the GL’s principle. (b) The Euclidean mean $\bar{f}_{n,p}$ (after a preliminary smoothing step).

Figure 13: Two-dimensional Gaussian mixtures dataset. (a) The trade-off function $\gamma \mapsto B(\gamma) + bV(\gamma)$ which attains its optimum at $\hat{\gamma} = 80$. (b) Penalized barycenter $\hat{f}_{n,p}$ associated to the parameter $\hat{\gamma} = 80$. 

21
Figure 14: Two dimensional flow cytometry dataset and Sinkhorn barycenter. (a) The trade-off function $\varepsilon \mapsto B(\varepsilon) + bV(\varepsilon)$ which attains its optimum at $\hat{\varepsilon} = 3.1$. (b) Sinkhorn barycenter $\tilde{r}_{n,p}^{\hat{\varepsilon}}$ associated to the parameter $\hat{\varepsilon} = 3.1$.

corresponding Sinkhorn barycenter in Figure 14(b). This barycenter clearly allows to correct mis-alignment issues of the data.

To analyze the relevance of this result, we present in Figure 15(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 [17] 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. The authors in [17] estimate landmarks by identifying peaks of the resulting one-dimensional densities. Then, they register these landmarks across the whole dataset in order to finally align these densities. The $L_2$-mean density obtained after this pre-processing step is displayed in Figure 15(b). This method leads to results that are very 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, this procedure 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.

Acknowledgements. 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).
Figure 15: Two dimensional flow cytometry dataset. (a) Euclidean mean \( \bar{f}_{n,p} \) of the data (after smoothing but without registration), (b) \( L_2 \)-mean of pre-processed data using kernel smoothing and density registration by landmark alignment with the method in [17].

### 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 \( v_N := \frac{1}{\sqrt{N}} \mathbf{1}_N \in \mathbb{R}^N \), which means that \( \text{rank}(\nabla^2 H_q^*(g)) \leq N - 1 \) for all \( g \in \mathbb{R}^N \) and \( q \in \Sigma_N \).

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

\[
\nabla^2 H_q^*(g) v_N = \frac{1}{\varepsilon} \text{diag}(\alpha) K \frac{q}{K\alpha} - \frac{1}{\varepsilon} \text{diag}(\alpha) K \text{diag} \left( \frac{q}{(K\alpha)^2} \right) K\alpha = \frac{1}{\varepsilon} \text{diag}(\alpha) K \frac{q}{K\alpha} - \frac{1}{\varepsilon} \text{diag}(\alpha) K \frac{q}{K\alpha} = 0,
\]

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

\[
0 = \lambda_N < \lambda_k \quad \text{for all} \quad 1 \leq k \leq N - 1.
\]
Proof. The eigenvalue $\lambda_N = 0$ associated to $v_N$ has been treated in Proposition A.1. Let $v \in V = (\text{Vect}(v_N))^\perp$ (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 there exists $i \neq j$ such that $v^{(j)} > v^{(i)}$. Without loss of generality, we can assume that $v^{(j)} > 0$. Then

$$
[\nabla^2 H_q^*(g)v]_{i} = \left[\frac{1}{\varepsilon} \left( \text{diag} \left( \text{diag}(\alpha) K \frac{q}{K \alpha} \right) \right) \right]_{i} - \left[\frac{1}{\varepsilon} \text{diag}(\alpha) K \text{diag} \left( \frac{q}{(K \alpha)^2} \right) K \text{diag}(\alpha) v \right]_{i}
$$

$$
= \frac{1}{\varepsilon} \alpha_j v^{(j)} \sum_{i=1}^{N} \alpha_j [K\alpha]_{i} - \frac{1}{\varepsilon} \sum_{i=1}^{N} \alpha_j [K\alpha]_{i} \alpha_i K_{mi} v^{(i)}
$$

$$
> \frac{1}{\varepsilon} \alpha_j v^{(j)} \sum_{i=1}^{N} \alpha_j [K\alpha]_{i} - \frac{1}{\varepsilon} \sum_{i=1}^{N} \alpha_j [K\alpha]_{i} \alpha_i K_{mi} v^{(j)} \text{ since } v^{(j)} \geq v^{(i)}, \forall i
$$

$$
= 0 \text{ since } \sum_{i=1}^{N} \alpha_i K_{im} = [K\alpha]_{m}.
$$

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 $\text{Tr}(\nabla^2 H_q^*(g)) \leq \frac{1}{\varepsilon}$ and thus $\lambda_k \leq 1/\varepsilon$ for all $k = 1, \ldots, N$.

Proof. We directly get from Theorem 3.3 that

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

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 [28], we have that

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

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 $(v_1, \ldots, v_{N-1})$ be an orthonormal basis of $V = (\text{Vect}(v_N))^\perp$ 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 $r \in \Sigma_N$, their 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
\[
\hat{H}_q^* (\tilde{g}) = \max_{\tilde{r} \in \Sigma_{N-1}} \langle \tilde{g}, \tilde{r} \rangle - \hat{H}_q (\tilde{r}) 
= \max_{r \in \Sigma_N} \langle \tilde{g}, P^T r - u_N \rangle - H_q (r) \quad \text{where } u_N = \frac{1}{N} \left( \sum_{i=1}^{N} v_1^{(i)} , \ldots , \sum_{i=1}^{N} v_{N-1}^{(i)} \right) 
= H_q^* (P \tilde{g}) - \langle \tilde{g}, u_N \rangle.
\]
Since \( H_q^* \) is \( C^\infty \) (see Theorem 3.3), we can differentiate \( \hat{H}_q^* \) with respect to \( \tilde{g} \) to obtain that
\[
\nabla \hat{H}_q^* (\tilde{g}) = P^T \nabla H_q^* (P \tilde{g}) - u_N \nabla^2 \hat{H}_q^* (\tilde{g}) = P^T \nabla^2 H_q^* (P \tilde{g}) P.
\]
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 \( \nabla H_q^* : \mathbb{R}^{(N-1)} \rightarrow \mathbb{R}^{(N-1)} \) is a \( C^\infty \)-diffeomorphism, using equality (A.1) (that is also valid for \( \hat{H}_q \)), we have that
\[
\nabla^2 \hat{H}_q^* (\tilde{r}) = \nabla \left( \left( \nabla \hat{H}_q^* \right)^{-1} (\tilde{r}) \right) 
= \left[ \nabla^2 \hat{H}_q^* (\nabla \hat{H}_q^* (\tilde{r})) \right]^{-1} 
= \left[ \nabla^2 H_q^* (\nabla H_q (\tilde{r})) \right]^{-1},
\]
where the second equality follows from the global inversion theorem, and the last one again uses equality (A.1). Thus we get
\[
\lambda_{\min}(\nabla^2 \hat{H}_q^* (\tilde{r})) \geq \varepsilon.
\]
The above inequality implies the strong convexity of \( \hat{H}_q \) which reads for \( \tilde{r}_0, \tilde{r}_1 \in \Sigma_{n-1} \)
\[
\hat{H}_q (\tilde{r}_1) \geq \hat{H}_q (\tilde{r}_0) + \nabla \hat{H}_q (\tilde{r}_0)^T (\tilde{r}_1 - \tilde{r}_0) + \frac{\varepsilon}{2} \| \tilde{r}_1 - \tilde{r}_0 \|_2,
\]
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 [11] by

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

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

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

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

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

$$|H_q(r) - H_q(s)| = (H_q(r) - H_q(s)) 1_{H_q(r) \geq H_q(s)} + (H_q(s) - H_q(r)) 1_{H_q(r) \leq H_q(s)}$$

\leq \left( \langle \alpha^{q,r}, r \rangle + \langle \beta^{q,r}, q \rangle - \sum_{m, \ell} \varepsilon e^{-\frac{1}{2}(C_{m\ell} - \alpha^{q,r}_m - \beta^{q,r}_\ell)} - \langle \alpha^{q,r}, s \rangle - \langle \beta^{q,r}, q \rangle + \sum_{m, \ell} \varepsilon e^{-\frac{1}{2}(C_{m\ell} - \alpha^{q,r}_m - \beta^{q,r}_\ell)} 1_{(H_q(r) \geq H_q(s))} \right)$$

\quad + \left( \langle \alpha^{q,s}, s \rangle + \langle \beta^{q,s}, q \rangle - \sum_{m, \ell} \varepsilon e^{-\frac{1}{2}(C_{m\ell} - \alpha^{q,s}_m - \beta^{q,s}_\ell)} - \langle \alpha^{q,s}, r \rangle - \langle \beta^{q,s}, q \rangle + \sum_{m, \ell} \varepsilon e^{-\frac{1}{2}(C_{m\ell} - \alpha^{q,s}_m - \beta^{q,s}_\ell)} 1_{(H_q(r) \leq H_q(s))} \right)$$

\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|. \hspace{1cm} (B.3)

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 [14]. Since the dual variable $\alpha^{q,r}$ achieves the maximum in equation (B.1), we have that for any $1 \leq m \leq N$

$$r_m - \sum_{1 \leq \ell \leq N} e^{-\frac{1}{2}(C_{m\ell} - \alpha^{q,r}_m - \beta^{q,r}_\ell)} = 0.$$

Let $r \in \Sigma^\rho_N$. Hence, $r_m \neq 0$, and thus one may define $\lambda_m = \varepsilon \log(r_m)$. Then, it follows from the above equality that $\sum_{1 \leq \ell \leq N} e^{-\frac{1}{2}(C_{m\ell} + \lambda_m - \alpha^{q,r}_m - \beta^{q,r}_\ell)} = 1$ which implies that

$$\alpha^{q,r}_m = -\varepsilon \log \left( \sum_{1 \leq \ell \leq N} e^{-\frac{1}{2}(C_{m\ell} + \lambda_m - \beta^{q,r}_\ell)} \right).$$
Now, for each $1 \leq m \leq N$, we define

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

and we consider the inequality

$$|\alpha^{q,r}_m - \alpha^{q,r}_k| \leq |\alpha^{q,r}_m - \tilde{\alpha}^{q,r}_m| + |\tilde{\alpha}^{q,r}_m - \tilde{\alpha}^{q,r}_k| + |\tilde{\alpha}^{q,r}_k - \alpha^{q,r}_k|. \hspace{1cm} \text{(B.5)}$$

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

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

On the other hand, using the inequality

$$\sum_{1 \leq \ell \leq N} e^{-\frac{1}{2}(C_{m\ell} + \lambda_m - \beta^{q,r}_\ell)} \geq e^{-\frac{1}{2}(C_{m\ell} + \lambda_m - \beta^{q,r}_\ell)} = e^{-\frac{1}{2} \tilde{\alpha}^{q,r}_m},$$

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

$$-\alpha^{q,r}_m \geq -\tilde{\alpha}^{q,r}_m.$$  \hspace{1cm} \text{(B.7)}

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

$$|\tilde{\alpha}^{q,r}_m - \alpha^{q,r}_m| \leq \varepsilon \log(N).$$  \hspace{1cm} \text{(B.8)}

To conclude, it remains to remark that, by equation (B.4), the vector $(\tilde{\alpha}^{q,r}_m - \lambda_m)_{1 \leq m \leq N}$ is the c-transform of the vector $(\beta^{q,r}_\ell)_{1 \leq \ell \leq N}$ 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. [29], p. 11) one obtains that

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

which implies that

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

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

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

Since the dual variables achieving the maximum in equation (B.1) are defined up to an additive constant, one may assume that $\alpha^{q,r}_k = 0$. 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 $r \in \Sigma^g_N$, we can thus conclude that $r \mapsto H_\rho(r)$ is $L_{\rho,\varepsilon}$-Lipschitz on $\Sigma^g_N$, for

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

$\Box$
C Algorithms to compute penalized Wasserstein barycenters of Section 2

In this section we describe how the minimization problem

$$\min_{\mu} \frac{1}{n} \sum_{i=1}^{n} W_2^2(\mu, \nu_i) + \gamma E(\mu) \text{ over } \mu \in P_2(\Omega),$$  \tag{C.1}$$
can be solved numerically by using an appropriate discretization to compute a numerical approximation of a regularized Wasserstein barycenter and the work of [12]. In our numerical experiments, we focus on the case where $E(\mu) = +\infty$ if $\mu$ is not a.c. to enforce the regularized Wasserstein barycenter to have a smooth pdf (we write $E(f) = E(\mu_f)$ if $\mu$ 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 (C.1) is used to compute a numerical approximation of a regularized Wasserstein barycenter $\mu^\gamma_{\nu_n}$. 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 $\mu^\gamma_{\nu_n}$ 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 (C.1).

In what follows, we first describe an algorithm that is specific to the one-dimensional case, and then we propose another algorithm that is valid for any $d \geq 1$.

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 $\mu_{\nu_n}$. Since the grid is fixed, we identify a discrete measure $\nu$ with the vector of weights $\nu = (\nu(x^1), \ldots, \nu(x^N))$ in $\mathbb{R}^N$ (with entries that sum up to one) of its values on this grid.

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

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

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

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

where $\tau^{(\ell)}$ is the $\ell$-th step time, and $\PiS$ stands for the projection on the simplex $S = \{y \in \mathbb{R}^N_+ \text{ such that } \sum_{j=1}^{N} y^j = 1\}$. Thanks to Proposition 5 in [26], 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 $R_f(s) = \sum_{x^k \leq s} f(x^k)$ the cdf of $\mu_f = \sum_{k=1}^{N} f(x^k) \delta_{x^k}$ and by $R_f(t) = \inf\{s \in \mathbb{R} : R_f(s) \geq t\}$ its pseudo-inverse.
Proposition C.1 ([26]). 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 $f \mapsto W^p(f, \nu)$ can be written as

$$\nabla W^p(f, \nu) : x_j \mapsto \sum_{m \geq j} |x^m - \tilde{x}^m|^p - |x^{m+1} - \tilde{x}^{m+1}|^p$$

where

$$\tilde{x}^m = x^k \text{ if } R_f(x^{k-1}) < R_f(x^m) < R_f(x^k)$$

$$\tilde{x}^m \in [x^{k-1}, x^k] \text{ if } R_f(x^m) = R_f(x^k)$$

Even if subgradient descent is only shown to converge with diminishing time steps [6], we observed that using a small fixed step time (of order $10^{-5}$) is sufficient to obtain in practice a convergence of the iterates $(f^{(t)})_{t \geq 1}$. Moreover, we have noticed that the principles of FISTA (Fast Iterative Soft Thresholding, see e.g. [2]) 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 $\nu_i^j$'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:

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

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

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

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

under the linear constraints

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

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

To this end, we rely on the dual approach introduced in [8] and the numerical optimisation scheme proposed in [12]. Following these works, one can show that the dual problem of (C.5) with a regularization of the form $E(Kf)$ and $K$ a discrete linear operator reads as

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

where the $\phi_i$’s are dual variables (vectors in $\mathbb{R}^N$) defined on the discrete grid $\{x^k\}_{k=1}^N$. $E_\gamma^*$ is the Legendre transform of $\gamma E$ and $H_{\nu_i}(.)$ is the Legendre transform of $W^2_2(., \nu_i)$ that reads:

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

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

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

Following [8], one value of the above subgradient can be obtained at point $x^k$ as:

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

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

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

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

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

### Numerical optimization

Following [12], the dual problem (C.7), can be simplified by removing one variable and thus discarding the linear constraint $K^T \phi_0 + \sum_{i=1}^{n} \phi_i = 0$. In order to inject the regularity given by $\phi_0$ in all the reconstructed barycenters obtained by $\phi_i$, $i = 1 \cdots n$, we modified the change of variables of [12] 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 $\psi_n$, 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:

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

The subgradient (C.9) can then be used in a descent algorithm over the dual problem (C.10).

For differentiable penalizers $E$, we consider the L-BFGS algorithm [33, 3] that integrates a
The barycenter is finally given by (C.8), taking non differentiable penalizers differentiable functions nearest neighbors for the explicit scheme involving make use of the Fista algorithm. This just modifies the update of

\[
\begin{array}{l}
\psi_0^{(\ell+1)} = \psi_0^{(\ell)} - \tau^{(\ell)} (\nabla E_0^*(\psi_0^{(\ell)}) + d_0^\ell) \\
\psi_i^{(\ell+1)} = \psi_i^{(\ell)} - \tau^{(\ell)} d_i^\ell \\
\end{array}
\]

(C.11)

where:

\[
\begin{align*}
d_0^\ell &= K \left( \partial H_{\nu_0} \left( -K^T \psi_0^{(\ell)} / n - \sum_{i=1}^{n-1} \psi_i^{(\ell)} \right) - \sum_{i=1}^{n-1} \partial H_{\nu_i} \left( \psi_i^{(\ell)} - K^T \psi_0^{(\ell)} / n \right) \right) \\
d_i^\ell &= \partial H_{\nu_i} \left( \psi_i^{(\ell)} - K^T \psi_0^{(\ell)} / n \right) - \partial H_{\nu_0} \left( -K^T \psi_0^{(\ell)} / n - \sum_{i=1}^{n-1} \psi_i^{(\ell)} \right)
\end{align*}
\]

The barycenter is finally given by (C.8), taking \( \phi_i = \psi_i - K^T \psi_0 / n \). Even if we only treated differentiable functions \( E \) in the theoretical part of this paper, we can numerically consider non differentiable penalizers \( E \), such as Total Variation \( (K = \nabla, E = ||.||_1) \). In this case, we make use of the Fista algorithm. This just modifies the update of \( \psi_0 \) in (C.11), by changing the explicit scheme involving \( \nabla E_0^* \) onto an implicit one through the proximity operator of \( E_0^* \):

\[
\psi_0^{(\ell+1)} = \text{Prox}_{\tau^{(\ell)} E_0^*} \left( \psi_0^{(\ell)} - \tau^{(\ell)} d_0^\ell \right) = \arg \min_{\psi} \frac{1}{2\tau^{(\ell)}} ||\psi^{(\ell)} - \tau^{(\ell)} d_0^\ell - \psi||^2 + E_0^*(\psi).
\]

Algorithmic issues and stabilization As detailed in [8], the computation of one subgradient in (C.9) relies on the look for Euclidean nearest neighbors between vectors \( (y_i^j, 0) \) and \( (x^j, \sqrt{c - \phi^k}) \), with \( c = \max_k \phi^k \). 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:

\[
S_i^{j,k} = w_i^{j,k} / \sum_k w_i^{j,k}, \quad \text{with} \quad w_i^{j,k} = \exp(-\frac{1}{2} ||y_i^j - x^k||^2 - \phi^k) / \varepsilon \text{ if } k \text{ is within the } K \text{ nearest neighbors for } j \text{ and data } i \text{ and } w_i^{j,k} = 0 \text{ otherwise.}
\]

References

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

[2] 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.

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

[4] J. Bigot, E. Cazelles, and N. Papadakis. Penalized barycenters in the Wasserstein space. ArXiv preprint: 1606.01025v2, 2018.

[5] J. Bigot, R. Gouet, T. Klein, and A. Lopez. Upper and lower risk bounds for estimating the Wasserstein barycenter of random measures on the real line. Electronic Journal of Statistics, 2018.

[6] S. Boyd and A. Mutapcic. Subgradient methods. Lecture notes of EE364, Stanford University, 2007, 2006.
[7] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 2013.
[20] 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.

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

[22] 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.

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

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

[25] 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.

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

[27] 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.

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

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

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

[31] 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.

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

[33] 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.