Sinkhorn Barycenters with Free Support via Frank-Wolfe Algorithm

Giulia Luise[1], **Saverio Salzo**[2], **Massimiliano Pontil**[1,2], **Carlo Ciliberto**[3]
g.luise.16@ucl.ac.uk, saverio.salzo@iit.it, m.pontil@cs.ucl.ac.uk,c.ciliberto@ic.ac.uk
[1] Department of Computer Science, University College London, UK
[2] CSML, Istituto Italiano di Tecnologia, Genova, Italy
[3] Department of Electrical and Electronic Engineering, Imperial College London, UK

Abstract

We present a novel algorithm to estimate the barycenter of arbitrary probability distributions with respect to the Sinkhorn divergence. Based on a Frank-Wolfe optimization strategy, our approach proceeds by populating the support of the barycenter incrementally, without requiring any pre-allocation. We consider discrete as well as continuous distributions, proving convergence rates of the proposed algorithm in both settings. Key elements of our analysis are a new result showing that the Sinkhorn divergence on compact domains has Lipschitz continuous gradient with respect to the Total Variation and a characterization of the sample complexity of Sinkhorn potentials. Experiments validate the effectiveness of our method in practice.

1 Introduction

Aggregating and summarizing collections of probability measures is a key task in several machine learning scenarios. Depending on the metric adopted, the properties of the resulting average (or *barycenter*) of a family of probability measures vary significantly. By design, optimal transport metrics are better suited at capturing the geometry of the distribution than Euclidean distance or f-divergences [14]. In particular, Wasserstein barycenters have been successfully used in settings such as texture mixing [40], Bayesian inference [49], imaging [26], or model ensemble [18].

The notion of barycenter in Wasserstein space was first introduced by [2] and then investigated from the computational perspective for the original Wasserstein distance [12, 50, 54] as well as its entropic regularizations (e.g. Sinkhorn) [6, 14, 20]. Two main challenges in this regard are: i) how to efficiently identify the support of the candidate barycenter and ii) how to deal with continuous (or infinitely supported) probability measures. The first problem is typically addressed by either fixing the support of the barycenter a-priori [20, 50] or by adopting an alternating minimization procedure to iteratively optimize the support point locations and their weights [12, 14]. While fixed-support methods enjoy better theoretical guarantees, free-support algorithms are more memory efficient and practicable in high dimensional settings. The problem of dealing with continuous distributions has been mainly approached by adopting stochastic optimization methods to minimize the barycenter functional [12, 20, 50].

In this work we propose a novel method to compute the barycenter of a set of probability distributions with respect to the Sinkhorn divergence [25] that does not require to fix the support beforehand. We address both the cases of discrete and continuous probability measures. In contrast to previous free-support methods, our algorithm does not perform an alternate minimization between support and weights. Instead, we adopt a Frank-Wolfe (FW) procedure to populate the support by incrementally adding new points and updating their weights at each iteration, similarly to kernel herding strategies [5]. We prove the convergence of the proposed optimization scheme for both finitely and infinitely

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

supported distribution settings. A central result to our analysis is the characterization of regularity properties of Sinkhorn potentials (i.e., the dual solutions of the Sinkhorn divergence problem), which extends recent work in [21, 23]. We empirically evaluate the performance of the proposed algorithm.

Contributions. The analysis of the proposed algorithm hinges on the following contributions: i) we show that the gradient of the Sinkhorn divergence is Lipschitz continuous on the space of probability measures with respect to the Total Variation. This grants us convergence of the barycenter algorithm in finite settings. ii) We characterize the sample complexity of Sinkhorn potentials of two empirical distributions sampled from arbitrary probability measures. This latter result is interesting on its own but it also enables us to iii) design a concrete optimization scheme to approximately solve the barycenter problem for arbitrary probability measures with convergence guarantees. iv) A byproduct of our analysis is the generalization of the FW algorithm to settings where the objective functional is defined only on a set with empty interior, which is the case for Sinkhorn divergence barycenter problem.

The rest of the paper is organized as follows: Sec. 2 reviews standard notions of optimal transport theory. Sec. 3 introduces the barycenter functional, and analyses the Lipschitz continuity of its gradient. Sec. 4 describes the implementation of our algorithm and Sec. 5 studies its convergence rates. Finally, Sec. 6 evaluates the proposed methods empirically and Sec. 7 provides concluding remarks.

2 Background

The aim of this section is to recall definitions and properties of Optimal Transport theory with entropic regularization. Throughout the work, we consider a compact set $\mathcal{X} \subset \mathbb{R}^d$ and a symmetric cost function $\mathsf{c}\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. We set $\mathsf{D} := \sup_{x,y\in\mathcal{X}} \mathsf{c}(x,y)$ and denote by $\mathcal{M}_1^+(\mathcal{X})$ the space of probability measures on $\mathcal{X}$ (positive Radon measures with mass 1). For any $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$, the Optimal Transport problem with entropic regularization is defined as follow [13, 24, 38]

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \min_{\pi\in\Pi(\alpha,\beta)} \int_{\mathcal{X}^2} \mathsf{c}(x,y)\, d\pi(x,y) + \varepsilon \mathrm{KL}(\pi|\alpha\otimes\beta), \qquad \varepsilon \geq 0 \tag{1}$$

where $\mathrm{KL}(\pi|\alpha\otimes\beta)$ is the *Kullback-Leibler divergence* between the candidate transport plan π and the product distribution $\alpha\otimes\beta$, and $\Pi(\alpha,\beta) = \{\pi \in \mathcal{M}_+^1(\mathcal{X}^2)\colon \mathsf{P}_{1\#}\pi = \alpha,\ \mathsf{P}_{2\#}\pi = \beta\}$, with $\mathsf{P}_i\colon \mathcal{X}\times\mathcal{X} \to \mathcal{X}$ the projector onto the i-th component and $\#$ the push-forward operator. The case $\varepsilon = 0$ corresponds to the classic Optimal Transport problem introduced by Kantorovich [29]. In particular, if $\mathsf{c} = \|\cdot - \cdot\|^p$ for $p \in [1,\infty)$, then OT_0 is the well-known p-Wasserstein distance [52]. Let $\varepsilon > 0$. Then, the dual problem of (1), in the sense of Fenchel-Rockafellar, is (see [10, 21])

$$\mathrm{OT}_\varepsilon(\alpha,\beta) = \max_{u,v\in\mathcal{C}(\mathcal{X})} \int u(x)\, d\alpha(x) + \int v(y)\, d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\alpha(x) d\beta(y), \tag{2}$$

where $\mathcal{C}(\mathcal{X})$ denotes the space of real-valued continuous functions on $\mathcal{X}$, endowed with $\|\cdot\|_\infty$. Let $\mu \in \mathcal{M}_1^+(\mathcal{X})$. We denote by $\mathsf{T}_\mu\colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ the map such that, for any $w \in \mathcal{C}(\mathcal{X})$,

$$\mathsf{T}_\mu(w)\colon x \mapsto -\varepsilon \log \int e^{\frac{w(y)-\mathsf{c}(x,y)}{\varepsilon}}\, d\mu(y). \tag{3}$$

The first order optimality conditions for (2) are (see [21] or Appendix B.2)

$$u = \mathsf{T}_\beta(v) \quad \alpha\text{- a.e.} \qquad \text{and} \qquad v = \mathsf{T}_\alpha(u) \quad \beta\text{- a.e.} \tag{4}$$

Pairs (u,v) satisfying (4) exist [30] and are referred to as *Sinkhorn potentials*. They are unique (α,β) - a.e. up to an additive constant, i.e., $(u+t, v-t)$ is also a solution for any $t \in \mathbb{R}$. In line with [21, 23] it will be useful in the following to assume (u,v) to be the Sinkhorn potentials such that: i) $u(x_o) = 0$ for an arbitrary anchor point $x_o \in \mathcal{X}$ and ii) (4) is satisfied pointwise on the entire domain $\mathcal{X}$. Then, u is a fixed point of the map $\mathsf{T}_{\beta\alpha} = \mathsf{T}_\beta \circ \mathsf{T}_\alpha$ (analogously for v). This suggests a fixed point iteration approach to minimize (2), yielding the well-known Sinkhorn-Knopp algorithm which has been shown to converge linearly in $\mathcal{C}(\mathcal{X})$ [30, 41]. See also Thm. B.10 for a precise statement. We recall a key result characterizing the differentiability of OT_ε in terms of the Sinkhorn potentials that will be useful in the following.

Proposition 1 (Prop 2 in [21]). *Let* $\nabla \mathrm{OT}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})^2$ *be such that,* $\forall \alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$

$$\nabla \mathrm{OT}_\varepsilon(\alpha, \beta) = (u, v), \qquad \text{with} \qquad u = \mathsf{T}_\beta(v), \;\; v = \mathsf{T}_\alpha(u) \;\; \text{on } \mathcal{X}, \quad u(x_o) = 0. \tag{5}$$

Then, OT_ε *is directionally differentiable and,* $\forall \alpha, \alpha', \beta, \beta' \in \mathcal{M}_1^+(\mathcal{X})$*, the directional derivative of* OT_ε *at* (α, β) *along the feasible direction* $(\mu, \nu) = (\alpha' - \alpha, \beta' - \beta)$ *is*

$$\mathrm{OT}'_\varepsilon(\alpha, \beta; \mu, \nu) = \langle \nabla \mathrm{OT}_\varepsilon(\alpha, \beta), (\mu, \nu) \rangle = \langle u, \mu \rangle + \langle v, \nu \rangle \,, \tag{6}$$

where $\langle w, \rho \rangle = \int w(x)\, d\rho(x)$ *denotes the canonical pairing between the spaces* $\mathcal{C}(\mathcal{X})$ *and* $\mathcal{M}(\mathcal{X})$*.*

Note that $\nabla \mathrm{OT}_\varepsilon$ is not a gradient in the standard sense. In particular note that the directional derivative in (6) is not defined for any pair of signed measures, but only along *feasible directions* $(\alpha' - \alpha, \beta' - \beta)$.

Sinkhorn Divergence. The fast convergence of Sinkhorn-Knopp algorithm makes OT_ε (with $\varepsilon > 0$) preferable to OT_0 from a computational perspective [13]. However, when $\varepsilon > 0$ the entropic regularization introduces a bias in the optimal transport problem, since in general $\mathrm{OT}_\varepsilon(\mu, \mu) \neq 0$. To compensate for this bias, [25] introduced the Sinkhorn *divergence*

$$\mathsf{S}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X}) \times \mathcal{M}_1^+(\mathcal{X}) \to \mathbb{R}, \qquad (\alpha, \beta) \mapsto \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\mathrm{OT}_\varepsilon(\alpha, \alpha) - \frac{1}{2}\mathrm{OT}_\varepsilon(\beta, \beta), \tag{7}$$

which was shown in [21] to be nonnegative, biconvex and to metrize the convergence in law under mild assumptions. We characterize the gradient of $\mathsf{S}_\varepsilon(\cdot, \beta)$ for a fixed $\beta \in \mathcal{M}_1^+(\mathcal{X})$, which will be key to derive our optimization algorithm for computing Sinkhorn barycenters.

Remark 2. *Let* $\nabla_1 \mathrm{OT}_\varepsilon : \mathcal{M}_1^+(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})$ *denote the first component of* $\nabla \mathrm{OT}_\varepsilon$ *(informally the component* u *of the Sinkhorn potentials* (u, v)*). Then, it follows from Prop. 1 and the definition of Sinkhorn divergence* (7) *that for any* $\beta \in \mathcal{M}_1^+(\mathcal{X})$ *the function* $\mathsf{S}_\varepsilon(\cdot, \beta) \colon \mathcal{M}_1^+(\mathcal{X}) \to \mathbb{R}$ *is directionally differentiable and admits gradient*

$$\nabla[\mathsf{S}_\varepsilon(\cdot, \beta)] \colon \mathcal{M}_1^+(\mathcal{X}) \to \mathcal{C}(\mathcal{X}) \qquad \alpha \mapsto \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \beta) - \frac{1}{2}\nabla_1 \mathrm{OT}_\varepsilon(\alpha, \alpha) = u - p, \tag{8}$$

with $u = \mathsf{T}_{\beta\alpha}(u)$ *and* $p = \mathsf{T}_{\alpha\alpha}(p)$ *the Sinkhorn potentials of* $\mathrm{OT}_\varepsilon(\alpha, \beta)$ *and* $\mathrm{OT}_\varepsilon(\alpha, \alpha)$ *respectively which are zero at* x_o*.*

We refer to Appendix C for an in-depth analysis of the directional differentiability properties of the Sinkorn divergence.

3 Sinkhorn barycenters with Frank-Wolfe

Given $\beta_1, \dots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ and $\omega_1, \dots, \omega_m \geq 0$ a set of weights such that $\sum_{j=1}^m \omega_j = 1$, the main goal of this paper is to solve the following *Sinkhorn barycenter* problem

$$\min_{\alpha \in \mathcal{M}_1^+(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha), \qquad \text{with} \qquad \mathsf{B}_\varepsilon(\alpha) = \sum_{j=1}^m \omega_j \, \mathsf{S}_\varepsilon(\alpha, \beta_j). \tag{9}$$

Although the objective functional B_ε is convex, its domain $\mathcal{M}_+^1(\mathcal{X})$ has *empty* interior in the space of finite signed measure $\mathcal{M}(\mathcal{X})$. Hence standard notions of Fréchet or Gâteaux differentiability do not apply. This, in principle causes some difficulties in devising optimization methods. To circumvent this issue, in this work we adopt the Frank-Wolfe (FW) algorithm. Indeed, one key advantage of this method is that it is formulated in terms of directional derivatives along feasible directions (i.e., directions that locally remain inside the constraint set). Building upon [15, 16, 19], which study the algorithm in Banach spaces, we show that the "weak" notion of directional differentiability of S_ε (and hence of B_ε) in Remark 2 is sufficient to carry out the convergence analysis. While full details are provided in Appendix A, below we give an overview of the main result.

Frank-Wolfe in dual Banach spaces. Let $\mathcal{W}$ be a real Banach space with topological dual $\mathcal{W}^*$ and let $\mathcal{D} \subset \mathcal{W}^*$ be a nonempty, convex, closed and bounded set. For any $w \in \mathcal{W}^*$ denote by $\mathcal{F}_\mathcal{D}(w) = \mathbb{R}_+(\mathcal{D} - w)$ the set of feasible direction of $\mathcal{D}$ at w (namely $s = t(w' - w)$ with $w' \in \mathcal{D}$ and $t > 0$). Let $\mathsf{G} \colon \mathcal{D} \to \mathbb{R}$ be a convex function and assume that there exists a map $\nabla \mathsf{G} \colon \mathcal{D} \to \mathcal{W}$ (not necessarily unique) such that $\langle \nabla \mathsf{G}(w), s \rangle = \mathsf{G}'(w; s)$ for every $s \in \mathcal{F}_\mathcal{D}(w)$. In Alg. 1 we present

Algorithm 1 FRANK-WOLFE IN DUAL BANACH SPACES

Input: initial $w_0 \in \mathcal{D}$, precision $(\Delta_k)_{k\in\mathbb{N}} \in \mathbb{R}^{\mathbb{N}}_{++}$, such that $\Delta_k(k+2)$ is nondecreasing.

For $k = 0, 1, \dots$

Take z_{k+1} such that $\mathsf{G}'(w_k, z_{k+1} - w_k) \leq \min_{z\in\mathcal{D}} \mathsf{G}'(w_k, z - w_k) + \frac{\Delta_k}{2}$

$w_{k+1} = w_k + \frac{2}{k+2}(z_{k+1} - w_k)$

a method to minimize G. The algorithm is structurally equivalent to the standard FW [19, 27] and accounts for possible inaccuracies when computing the conditional gradient (i.e. solving the FW inner minimization). This will be key in Sec. 5 when studying the barycenter problem for β_j with infinite support. The following result (see proof in Appendix A) shows that under the additional assumption that $\nabla\mathsf{G}$ is Lipschitz-continuous and with sufficiently fast decay of the errors, the above procedure converges in value to the minimum of G with rate $O(1/k)$. Here $\operatorname{diam}(\mathcal{D})$ denotes the diameter of $\mathcal{D}$ with respect to the dual norm.

Theorem 3. *Under the assumptions above, suppose in addition that $\nabla\mathsf{G}$ is L-Lipschitz continuous with $L > 0$. Let $(w_k)_{k\in\mathbb{N}}$ and $(\Delta_k)_{k\in\mathbb{N}}$ be defined according to Alg. 1. Then, for every integer $k \geq 1$,*

$$\mathsf{G}(w_k) - \min_{w\in\mathcal{D}} \mathsf{G}(w) \leq \frac{2}{k+2} L \operatorname{diam}(\mathcal{D})^2 + \Delta_k. \tag{10}$$

Frank-Wolfe Sinkhorn barycenters. We show that the barycenter problem (9) satisfies the setting and hypotheses of Thm. 3 and can be thus approached via Alg. 1.

Optimization domain. Let $\mathcal{W} = \mathcal{C}(\mathcal{X})$, with dual $\mathcal{W}^* = \mathcal{M}(\mathcal{X})$. The constraint set $\mathcal{D} = \mathcal{M}_1^+(\mathcal{X})$ is convex, closed, and bounded.

Objective functional. The objective functional $\mathsf{G} = \mathsf{B}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X}) \to \mathbb{R}$, defined in (9), is convex since it is a convex combination of $\mathsf{S}_\varepsilon(\cdot, \beta_j)$, with $j = 1 \dots m$. The gradient $\nabla\mathsf{B}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is $\nabla\mathsf{B}_\varepsilon = \sum_{j=1}^m \omega_j \nabla\mathsf{S}_\varepsilon(\cdot, \beta_j)$, where $\nabla\mathsf{S}_\varepsilon(\cdot, \beta_j)$ is given in Remark 2.

Lipschitz continuity of the gradient. This is the most critical condition and it is studied in the following theorem.

Theorem 4. *The gradient $\nabla\mathrm{OT}_\varepsilon$ defined in Prop. 1 is Lipschitz continuous. In particular, the first component $\nabla_1\mathrm{OT}_\varepsilon$ is $2\varepsilon e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous, i.e., for every $\alpha, \alpha', \beta, \beta' \in \mathcal{M}_1^+(\mathcal{X})$,*

$$\|u - u'\|_\infty = \|\nabla_1\mathrm{OT}_\varepsilon(\alpha, \beta) - \nabla_1\mathrm{OT}_\varepsilon(\alpha', \beta')\|_\infty \leq 2\varepsilon e^{3\mathsf{D}/\varepsilon} \left(\|\alpha - \alpha'\|_{TV} + \|\beta - \beta'\|_{TV}\right), \tag{11}$$

where $\mathsf{D} = \sup_{x,y\in\mathcal{X}} \mathsf{c}(x, y)$, $u = \mathsf{T}_{\beta\alpha}(u), u' = \mathsf{T}_{\beta',\alpha'}(u')$, and $u(x_o) = u'(x_o) = 0$. Moreover, it follows from (8) that $\nabla\mathsf{S}_\varepsilon(\cdot, \beta)$ is $6\varepsilon e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous. The same holds for $\nabla\mathsf{B}_\varepsilon$.

Thm. 4 is one of the main contributions of this paper. It can be rephrased by saying that the operator that maps a pair of distributions to their Sinkhorn potentials is Lipschitz continuous. This result is significantly deeper than the one given in [20, Lemma 1], which establishes the Lipschitz continuity of the gradient in the *semidiscrete* case. The proof (given in Appendix D) relies on non-trivial tools from Perron-Frobenius theory for Hilbert's metric [32], which is a well-established framework to study Sinkhorn potentials [38]. We believe this result is interesting not only for the application of FW to the Sinkhorn barycenter problem, but also for further understanding regularity properties of entropic optimal transport.

4 Algorithm: practical Sinkhorn barycenters

According to Sec. 3, FW is a valid approach to tackle the barycenter problem (9). Here we describe how to implement in practice the abstract procedure of Alg. 1 to obtain a sequence of distributions $(\alpha_k)_{k\in\mathbb{N}}$ minimizing B_ε. A main challenge in this sense resides in finding a minimizing feasible direction for $\mathsf{B}'_\varepsilon(\alpha_k; \mu - \alpha_k) = \langle \nabla\mathsf{B}_\varepsilon(\alpha_k), \mu - \alpha_k \rangle$. According to Remark 2, this amounts to solve

$$\mu_{k+1} \in \operatorname*{argmin}_{\mu\in\mathcal{M}_1^+(\mathcal{X})} \sum_{j=1}^m \omega_j \ \langle u_{jk} - p_k, \mu \rangle \qquad \text{where} \qquad u_{jk} - p_k = \nabla\mathsf{S}_\varepsilon[(\cdot, \beta_j)](\alpha_k), \tag{12}$$

Algorithm 2 SINKHORN BARYCENTER

Input: $\beta_j = (\mathbf{Y}_j, \mathsf{b}_j)$ with $\mathbf{Y}_j \in \mathbb{R}^{d \times n_j}, \mathsf{b}_j \in \mathbb{R}^{n_j}, \omega_j > 0$ for $j = 1, \dots, m$, $x_0 \in \mathbb{R}^d$, $\varepsilon > 0$, $K \in \mathbb{N}$.

Initialize: $\alpha_0 = (\mathbf{X}_0, \mathsf{a}_0)$ with $\mathbf{X}_0 = x_0$, $\mathsf{a}_0 = 1$.

For $k = 0, 1, \dots, K-1$

$\mathsf{p} = \text{SINKHORNKNOPP}(\alpha_k, \alpha_k, \varepsilon)$

$p(\cdot) = \text{SINKHORNGRADIENT}(\mathbf{X}_k, \mathsf{a}_k, \mathsf{p})$

For $j = 1, \dots m$

$\mathsf{v}_j = \text{SINKHORNKNOPP}(\alpha_k, \beta_j, \varepsilon)$

$u_j(\cdot) = \text{SINKHORNGRADIENT}(\mathbf{Y}_j, \mathsf{b}_j, \mathsf{v}_j)$

Let $\varphi \colon x \mapsto \sum_{j=1}^m \omega_j \, u_j(x) - p(x)$

$x_{k+1} = \text{MINIMIZE}(\varphi)$

$\mathbf{X}_{k+1} = [\mathbf{X}_k, x_{k+1}]$ and $\mathsf{a}_{k+1} = \frac{1}{k+2}\,[k\,\mathsf{a}_k, 2]$

$\alpha_{k+1} = (\mathbf{X}_{k+1}, \mathsf{a}_{k+1})$

Return: α_K

with $p_k = \nabla_1 \text{OT}_\varepsilon(\alpha_k, \alpha_k)$ not depending on j. In general (12) would entail a minimization over the set of all probability distributions on $\mathcal{X}$. However, since the objective functional is linear in μ and $\mathcal{M}_1^+(\mathcal{X})$ is a weakly-$*$ compact convex set, we can apply Bauer maximum principle (see e.g., [3, Thm. 7.69]). Hence, solutions are achieved at the extreme points of the optimization domain. These correspond to Dirac's deltas in the case of $\mathcal{M}_1^+(\mathcal{X})$ [11, p. 108]. Denote by $\delta_x \in \mathcal{M}_1^+(\mathcal{X})$ the Dirac's delta centered at $x \in \mathcal{X}$. We have $\langle w, \delta_x \rangle = w(x)$ for every $w \in \mathcal{C}(\mathcal{X})$. Hence (12) is equivalent to

$$\mu_{k+1} = \delta_{x_{k+1}} \qquad \text{with} \qquad x_{k+1} \in \operatorname*{argmin}_{x \in \mathcal{X}} \sum_{j=1}^m \omega_j \, \big(u_{jk}(x) - p_k(x)\big). \tag{13}$$

Once the new support point x_{k+1} has been obtained, the update in Alg. 1 corresponds to

$$\alpha_{k+1} = \alpha_k + \frac{2}{k+2}(\delta_{x_{k+1}} - \alpha_k) = \frac{k}{k+2}\alpha_k + \frac{2}{k+2}\delta_{x_{k+1}}. \tag{14}$$

If FW is initialized with a Dirac's delta $\alpha_0 = \delta_{x_0}$ for some $x_0 \in \mathcal{X}$, then every further iterate α_k will have at most $k+1$ support points. According to (13), the inner optimization for FW consists in minimizing the functional $x \mapsto \sum_{j=1}^m \omega_j \, \big(u_{jk}(x) - p_k(x)\big)$ over $\mathcal{X}$. In practice, having access to such functional poses already a challenge, since it requires computing the Sinkhorn potentials u_{jk} and p_k, which are infinite dimensional objects. Below we discuss how to estimate these potentials when the β_j have finite support. We then address the general setting.

Computing $\nabla_1 \text{OT}_\varepsilon$ for probability distributions with finite support. Let $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$, where $\beta = \sum_{i=1}^n b_i \delta_{y_i}$ and $\mathsf{b} = (b_i)_{i=1}^n$ nonnegative weights summing up to 1. It is useful to identify β with the pair $(\mathbf{Y}, \mathsf{b})$, where $\mathbf{Y} \in \mathbb{R}^{d \times n}$ is the matrix with i-th column equal to y_i. Let $(u, v) \in \mathcal{C}(\mathcal{X})^2$ be the pair of Sinkhorn potentials associated to α and β in Prop. 1, recall that $u = \mathsf{T}_\beta(v)$. Denote by $\mathsf{v} \in \mathbb{R}^n$ the *evaluation vector* of the Sinkhorn potential v, with i-th entry $\mathsf{v}_i = v(y_i)$. According to the definition of T_β in (3), for any $x \in \mathcal{X}$

$$[\nabla_1 \text{OT}_\varepsilon(\alpha, \beta)](x) = u(x) = [\mathsf{T}_\beta(v)](x) = -\varepsilon \log \sum_{i=1}^n e^{(\mathsf{v}_i - \mathsf{c}(x, y_i))/\varepsilon} \, b_i, \tag{15}$$

since the integral $\mathsf{T}_\beta(v)$ reduces to a sum over the support of β. Hence, the gradient of OT_ε (i.e. the potential u), *is uniquely characterized in terms of the finite dimensional vector* v *collecting the values of the potential* v *on the support of* β . We refer as SINKHORNGRADIENT to the routine which associates to each triplet $(\mathbf{Y}, \mathsf{b}, \mathsf{v})$ the map $x \mapsto -\varepsilon \, \log \sum_{i=1}^n \, e^{(\mathsf{v}_i - \mathsf{c}(x, y_i))/\varepsilon} \, b_i$.

Sinkhorn barycenters: finite case. Alg. 2 summarizes FW applied to the barycenter problem (9) when the β_j's have finite support. Starting from a Dirac's delta $\alpha_0 = \delta_{x_0}$, at each iteration $k \in \mathbb{N}$ the algorithm proceeds by: i) finding the corresponding evaluation vectors v_j's and p of the Sinkhorn potentials for $\text{OT}_\varepsilon(\alpha_k, \beta_j)$ and $\text{OT}_\varepsilon(\alpha_k, \alpha_k)$ respectively, via the routine SINKHORNKNOPP (see [13, 21] or Alg. B.2). This is possible since both β_j and α_k have finite support and therefore the

problem of approximating the evaluation vectors v_j and p reduces to an optimization problem over finite vector spaces that can be efficiently solved [13]; $ii)$ obtain the gradients $u_j = \nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \beta_j)$ and $p = \nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \alpha_k)$ via SINKHORNGRADIENT; $iii)$ minimize $\varphi : x \mapsto \sum_{j=1}^n \omega_j\, u_j(x) - p(x)$ over $\mathcal{X}$ to find a new point x_{k+1} (we comment on this meta-routine MINIMIZE below); $iv)$ finally update the support and weights of α_k according to (14) to obtain the new iterate α_{k+1}.

A key feature of Alg. 2 is that the support of the candidate barycenter is updated *incrementally* by adding at most one point at each iteration, a procedure similar in flavor to the kernel herding strategy in [5, 31] and conditional gradient for sparse inverse problem [8, 9]. This contrasts with previous methods for barycenter estimation [6, 14, 20, 50], which require the support set, or at least its cardinality, to be fixed beforehand. However, indentifying the new support point requires solving the nonconvex problem (13), a task addressed by the meta-routine MINIMIZE. This problem is typically smooth (e.g., a linear combination of Gaussians when $\mathsf{c}(x, y) = \|x - y\|^2$) and first or second order nonlinear optimization methods can be adopted to find stationary points. We note that all free-support methods in the literature for barycenter estimation are also affected by nonconvexity since they typically require solving a biconvex problem (alternating minimization between support points and weights) which is not jointly convex [12, 14]. We conclude by observing that if we restrict to the setting of [20, 50] with fixed finite support set, then MINIMIZE can be solved exactly by evaluating the functional in (13) on each candidate support point.

Sinkhorn barycenters: general case. When the β_j's have infinite support, it is not possible to apply Sinkhorn-Knopp in practice. In line with [23, 50], we can randomly sample empirical distributions $\hat{\beta}_j = \frac{1}{n}\sum_{i=1}^n \delta_{x_{ij}}$ from each β_j and apply Sinkhorn-Knopp to $(\alpha_k, \hat{\beta}_j)$ in Alg. 1 rather than to the ideal pair (α_k, β_j). This strategy is motivated by [21, Prop 13], where it was shown that Sinkhorn potentials vary continuously with the input measures. However, it opens two questions: $i)$ whether this approach is theoretically justified (consistency) and $ii)$ how many points should we sample from each β_j to ensure convergence (rates). We answer these questions in Thm. 7 in the next section.

5 Convergence analysis

We finally address the convergence of FW applied to both the finite and infinite settings discussed in Sec. 4. We begin by considering the finite setting.

Theorem 5. *Suppose that $\beta_1, \dots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ have finite support and let α_k be the k-th iterate of Alg. 2 applied to (9). Then,*

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{M}_1^+(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{48\,\varepsilon\, e^{3\mathsf{D}/\varepsilon}}{k+2}. \tag{16}$$

The result follows by the convergence result of FW in Thm. 3 applied with the Lipschitz constant from Thm. 4, and recalling that $\mathrm{diam}(\mathcal{M}_1^+(\mathcal{X})) = 2$ with respect to the Total Variation. Note that Thm. 5 assumes SINKHORNKNOPP and MINIMIZE in Alg. 2 to yield exact solutions. In Appendix D we extend of Alg. 2 and Thm. 5 which account for approximation errors in the above routines.

General setting. As mentioned in Sec. 4, when the β_j's are not finitely supported we adopt a sampling approach. More precisely we propose to *replace* in Alg. 2 the ideal Sinkhorn potentials of the pairs (α, β_j) with those of $(\alpha, \hat{\beta}_j)$, where each $\hat{\beta}_j$ is an empirical measure randomly sampled from β_j. In other words we are performing the FW algorithm with a (possibly rough) approximation of the correct gradient of B_ε. According to Thm. 3, FW allows errors in the gradient estimation (which are captured into the precision Δ_k in the statement). To this end, the following result *quantifies* the approximation error between $\nabla_1 \mathrm{OT}_\varepsilon(\cdot, \beta)$ and $\nabla_1 \mathrm{OT}_\varepsilon(\cdot, \hat{\beta})$ in terms of the sample size of $\hat{\beta}$.

Theorem 6 (Sample Complexity of Sinkhorn Potentials)**.** *Suppose that $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ with $s > d/2$. Then, there exists a constant $\bar{\mathsf{r}} = \bar{\mathsf{r}}(\mathcal{X}, \mathsf{c}, d)$ such that for any $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ and any empirical measure $\hat{\beta}$ of a set of n points independently sampled from β, we have, for every $\tau \in (0, 1]$*

$$\|u - u_n\|_\infty = \|\nabla_1 \mathrm{OT}_\varepsilon(\alpha, \beta) - \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \hat{\beta})\|_\infty \leq \frac{8\varepsilon\, \bar{\mathsf{r}} e^{3\mathsf{D}/\varepsilon} \log \frac{3}{\tau}}{\sqrt{n}} \tag{17}$$

with probability at least $1 - \tau$, where $u = \mathsf{T}_{\beta\alpha}(u), u_n = \mathsf{T}_{\hat{\beta}\alpha}(u_n)$ and $u(x_o) = u_n(x_o) = 0$.

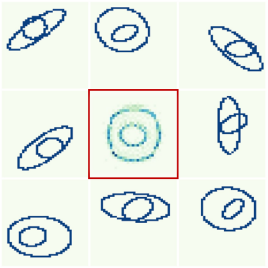

Fig. 1: Barycenter of nested ellipses

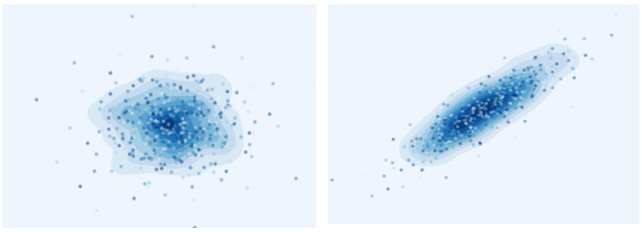

Fig. 2: Barycenters of Gaussians (see text)

The result in Thm. 6 is of central importance in this work. We point out that it *cannot* be obtained by means of the Lipschitz continuity of $\nabla_1 \mathrm{OT}_\varepsilon$ in Thm. 4, since empirical measures do not converge in $\|\cdot\|_{TV}$ to their target distribution [17]. Instead, the proof consists in considering the weaker *Maximum Mean Discrepancy (MMD)* metric associated to a universal kernel [46], which metrizes the topology of the convergence in law of $\mathcal{M}_1^+(\mathcal{X})$ [47]. Empirical measures converge in MMD metric to their target distribution [46]. Therefore, by proving the Lipschitz continuity of $\nabla_1 \mathrm{OT}_\varepsilon$ with *respect to* MMD (see Prop. E.5) we are able to conclude that (17) holds. This latter result relies on regularity properties of Sinkhorn potentials, which have been recently shown [23, Thm.2] to be uniformly bounded in Sobolev spaces under the additional assumption $c \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$. For sufficiently large s, the Sobolev norm is in duality with the MMD [35] and allows us to derive the required Lipschitz continuity. We conclude noting that while [23] studied the sample complexity of the Sinkhorn *divergence*, Thm. 6 is a sample complexity result for Sinkhorn *potentials*. In this sense, we observe that the constants appearing in the bound are tightly related to those in [23, Thm.3] and have similar behavior with respect to ε. We can now study the convergence of FW in continuous settings.

Theorem 7. *Suppose that* $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ *with* $s > d/2$. *Let* $n \in \mathbb{N}$ *and* $\hat{\beta}_1, \dots, \hat{\beta}_m$ *be empirical distributions with* n *support points, each independently sampled from* $\beta_1, \dots, \beta_m$. *Let* α_k *be the* k*-th iterate of Alg. 2 applied to* $\hat{\beta}_1, \dots, \hat{\beta}_m$. *Then for any* $\tau \in (0, 1]$, *the following holds with probability larger than* $1 - \tau$

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{M}_1^+(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \leq \frac{64 \bar{r} \varepsilon e^{3\mathsf{D}/\varepsilon} \log \frac{3m}{\tau}}{\min(k, \sqrt{n})}. \tag{18}$$

The proof is shown in Appendix E. A consequence of Thm. 7 is that the accuracy of FW depends simultaneously on the number of iterations and the sample size used in the approximation of the gradients: by choosing $n = k^2$ we recover the $O(1/k)$ rate of the finite setting, while for $n = k$ we have a rate of $O(k^{-1/2})$, which is reminiscent of typical sample complexity results, highlighting the statistical nature of the problem.

Remark 8 (Incremental Sampling)**.** *The above strategy requires sampling the empirical distributions for* $\beta_1, \dots, \beta_m$ *beforehand. A natural question is whether it is be possible to do this* incrementally, *sampling new points and updating* $\hat{\beta}_j$ *accordingly, as the number of* FW *iterations increase. To this end, one can perform an intersection bound and see that this strategy is still consistent, but the bound in Thm. 7 worsens the logarithmic term, which becomes* $\log(3mk/\tau)$.

6 Experiments

In this section we show the performance of our method in a range of experiments. Additional experiments are provided in the supplementary material. Code has been made publicly available[1].

Discrete measures: barycenter of nested ellipses. We compute the barycenter of 30 randomly generated nested ellipses on a 50×50 grid similarly to [14]. We interpret each image as a probability distribution in 2D. The cost matrix is given by the squared Euclidean distances between pixels. Fig. 1 reports 8 samples of the input ellipses and the barycenter obtained with Alg. 2. It shows qualitatively that our approach captures key geometric properties of the input measures.

[1] https://github.com/GiulsLu/Sinkhorn-Barycenters

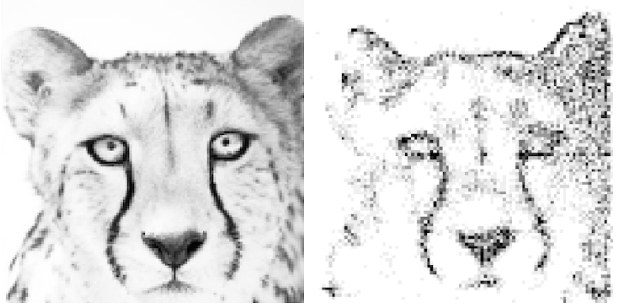

Fig. 3: Matching of a 140x140 image. 5000 FW iterations

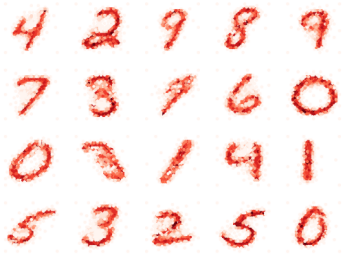

Fig. 4: MNIST k-means (20 centers)

Continuous measures: barycenter of Gaussians. We compute the barycenter of 5 Gaussian distributions $\mathcal{N}(m_i, C_i)$ $i = 1, \dots, 5$ in $\mathbb{R}^2$, with mean $m_i \in \mathbb{R}^2$ and covariance C_i randomly generated. We apply Alg. 2 to empirical measures obtained by sampling $n = 500$ points from each $\mathcal{N}(m_i, C_i)$, $i = 1, \dots, 5$. Since the (Wasserstein) barycenter of Gaussian distributions can be estimated accurately (see [2]), in Fig. 2 we report both the output of our method (as a scatter plot) and the true Wasserstein barycenter (as level sets of its density). We observe that our estimator recovers both the mean and covariance of the target barycenter. See the supplementary material for additional experiments also in the case of mixtures of Gaussians.

Image "compression" via distribution matching. Similarly to [12], we test Alg. 2 in the special case of computing the "barycenter" of a single measure $\beta \in \mathcal{M}_+^1(\mathcal{X})$. While the solution of this problem is the distribution β itself, we can interpret the intermediate iterates α_k of Alg. 2 as compressed version of the original measure. In this sense k would represent the level of compression since α_k is supported on *at most* k points. Fig. 3 (Right) reports iteration $k = 5000$ of Alg. 2 applied to the 140×140 image in Fig. 3 (Left) interpreted as a probability measure β in 2D. We note that the number of points in the support is ~ 3900: indeed, Alg. 2 selects the most relevant support points points multiple times to accumulate the right amount of mass on each of them (darker color = higher weight). This shows that FW tends to greedily search for the most relevant support points, prioritizing those with higher weight.

k-means on MNIST digits. We tested our algorithm on a k-means clustering experiment. We consider a subset of 500 random images from the MNIST dataset. Each image is suitably normalized to be interpreted as a probability distribution on the grid of 28×28 pixels with values scaled between 0 and 1. We initialize 20 centroids according to the k-means++ strategy [4]. Fig. 4 depicts the 20 centroids obtained by performing k-means with Alg. 2. We see that the structure of the digits is successfully detected, recovering also minor details (e.g. note the difference between the 2 centroids).

Real data: Sinkhorn propagation of weather data. We consider the problem of Sinkhorn *propagation* similar to the one in [45]. The goal is to predict the distribution of missing measurements for weather stations in the state of Texas, US by "propagating" measurements from neighboring stations in the network. The problem can be formulated as minimizing the functional $\sum_{(v,u)\in\mathcal{V}} \omega_{uv} \mathsf{S}_\varepsilon(\rho_v, \rho_u)$ over the set $\{\rho_v \in \mathcal{M}_1^+(\mathbb{R}^2) | v \in \mathcal{V}_0\}$ with: $\mathcal{V}_0 \subset \mathcal{V}$ the subset of stations with missing measurements, $G = (\mathcal{V}, \mathcal{E})$ the whole graph of the stations network, ω_{uv} a weight inversely proportional to the geographical distance between two vertices/stations $u, v \in \mathcal{V}$. The variable $\rho_v \in \mathcal{M}_1^+(\mathbb{R}^2)$ denotes the distribution of measurements at station v of daily *temperature* and *atmospheric pressure* over one year. This is a generalization of the barycenter problem (9) (see also [38]).

From the total $|\mathcal{V}| = 115$, we randomly select $10\%, 20\%$ or 30% to be *available* stations, and use Alg. 2 to propagate their measurements to the remaining "missing" ones. We compare our approach (FW) with the Dirichlet (DR) baseline in [45] in terms of the error $d(C_T, \hat{C})$ between the covariance matrix C_T of the groundtruth distribution and that of the predicted one. Here $d(A, B) = \|\log(A^{-1/2} B A^{-1/2})\|$ is the geodesic distance on the cone of positive definite matrices. The average prediction errors are: 2.07 (FW), 2.24 (DR) for 10%, 1.47 (FW), 1.89(DR) for 20% and 1.3 (FW), 1.6 (DR) for 30%. Fig. 5 qualitatively reports the improvement $\Delta = d(C_T, C_{DR}) - d(C_T, C_{FW})$ of our method on individual stations: a higher color intensity corresponds to a wider gap in our favor between prediction errors, from light green ($\Delta \sim 0$) to red ($\Delta \sim 2$). Our approach tends to propagate the distributions to missing locations with higher accuracy.

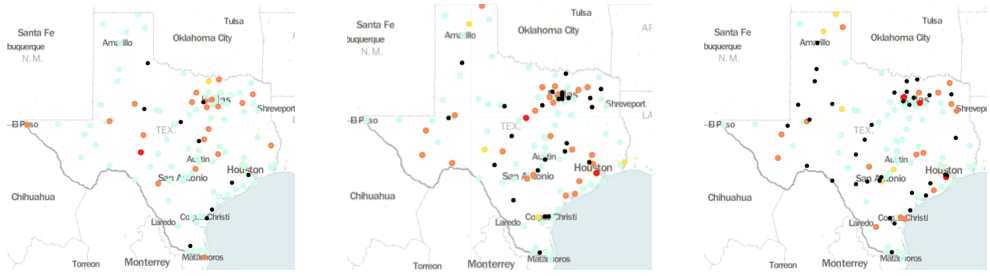

Fig. 5: From Left to Right: propagation of weather data with $10\%, 20\%$ and 30% stations with available measurements (see text).

7 Conclusion

We proposed a Frank-Wolfe-based algorithm to find the Sinkhorn barycenter of probability distributions with either finitely or infinitely many support points. Our algorithm belongs to the family of barycenter methods with free support since it adaptively identifies support points rather than fixing them a-priori. In the finite settings, we were able to guarantee convergence of the proposed algorithm by proving the Lipschitz continuity of gradient of the barycenter functional in the Total Variation sense. Then, by studying the sample complexity of Sinkhorn potential estimation, we proved the convergence of our algorithm also in the infinite case. We empirically assessed our method on a number of synthetic and real experiments, showing that it exhibits good qualitative and quantitative performance. While in this work we have considered FW iterates that are a convex combination of Dirac's delta, models with higher regularity (e.g. mixture of Gaussians) might be more suited to approximate the barycenter of distributions with smooth density. Hence, in the future we plan to investigate whether the perspective adopted in this work could be extended also to other barycenter estimators.

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*. Elsevier, 2003.

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

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

[4] David Arthur and Sergei Vassilvitskii. K-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '07, pages 1027–1035, Philadelphia, PA, USA, 2007. Society for Industrial and Applied Mathematics.

[5] Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. *arXiv preprint arXiv:1203.4523*, 2012.

[6] Jean-David Benamou, Guillaume Carlier, Marco Cuturi, Luca Nenna, and Gabriel Peyré. Iterative bregman projections for regularized transportation problems. *SIAM J. Scientific Computing*, 37(2), 2015.

[7] J Frédéric Bonnans and Alexander Shapiro. *Perturbation analysis of optimization problems*. Springer Science & Business Media, 2013.

[8] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

[9] Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

[10] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[11] G. Chouquet. *Lectures on Analysis, Vol. II*. W. A. Bejamin, Inc., Reading, MA, USA., 1969.

[12] S. Claici, E. Chien, and J. Solomon. Stochastic Wasserstein Barycenters. *ArXiv e-prints*, February 2018.

[13] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 2292–2300, 2013.

[14] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 685–693, Bejing, China, 22–24 Jun 2014. PMLR.

[15] V. F. Demyanov and A. M. Rubinov. The minimization of smooth convex functional on a convex set. *J. SIAM Control.*, 5(2):280–294, 1967.

[16] V. F. Demyanov and A. M. Rubinov. Minimization of functionals in normed spaces. *J. SIAM Control.*, 6(1):73–88, 1968.

[17] Luc Devroye, Laszlo Gyorfi, et al. No empirical probability measure can converge in the total variation sense for all distributions. *The Annals of Statistics*, 18(3):1496–1499, 1990.

[18] Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross, Cicero Dos Santos, and Tom Sercu. Wasserstein barycenter model ensembling. *arXiv preprint arXiv:1902.04999*, 2019.

[19] Joseph C Dunn and S Harshbarger. Conditional gradient algorithms with open loop step size rules. *Journal of Mathematical Analysis and Applications*, 62(2):432–444, 1978.

[20] Pavel Dvurechenskii, Darina Dvinskikh, Alexander Gasnikov, Cesar Uribe, and Angelia Nedich. Decentralize and randomize: Faster algorithm for wasserstein barycenters. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 10760–10770. Curran Associates, Inc., 2018.

[21] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-Ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and mmd using sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics (AIStats)*, 2019.

[22] J. Franklin and J. Lorenz. On the scaling of multidimensional matrices. *Linear Algebra and its Applications*, 114:717–735, 1989.

[23] Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of sinkhorn divergences. *International Conference on Artificial Intelligence and Statistics (AIStats)*, 2018.

[24] Aude Genevay, Marco Cuturi, Gabriel Peyré, and Francis Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 3440–3448. Curran Associates, Inc., 2016.

[25] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning generative models with sinkhorn divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

[26] Alexandre Gramfort, Gabriel Peyré, and Marco Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

[27] Martin Jaggi. Revisiting frank-wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

[28] Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *International Conference on Machine Learning*, pages 427–435, 2013.

[29] L Kantorovich. On the transfer of masses (in russian). *Doklady Akademii Nauk USSR*, 1942.

[30] Paul Knopp and Richard Sinkhorn. A note concerning simultaneous integral equations. *Canadian Journal of Mathematics*, 20:855–861, 1968.

[31] Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-wolfe optimization for particle filtering. *arXiv preprint arXiv:1501.02056*, 2015.

[32] Bas Lemmens and Roger Nussbaum. *Nonlinear Perron-Frobenius Theory*, volume 189. Cambridge University Press, 2012.

[33] Bas Lemmens and Roger Nussbaum. Birkhoff's version of Hilbert's metric and its applications in analysis. *arXiv preprint arXiv:1304.7921*, 2013.

[34] M. V Menon. Reduction of a matrix with positive elements to a doubly stochastic matrix. *Proc. Amer. Math. Soc.*, 18:244–247, 1967.

[35] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Bernhard Schölkopf, et al. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141, 2017.

[36] Roger Nussbaum. Hilbert's projective metric and iterated nonlinear maps. *Mem. Amer. Math. Soc.*, 391:1–137, 1988.

[37] Roger Nussbaum. Entropy minimization, Hilbert's projective metric and scaling integral kernels. *Journal of Functional Analysis*, 115:45–99, 1993.

[38] Gabriel Peyré and Marco Cuturi. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

[39] Iosif Pinelis. Optimum bounds for the distributions of martingales in banach spaces. *The Annals of Probability*, pages 1679–1706, 1994.

[40] Julien Rabin, Gabriel Peyré, Julie Delon, and Marc Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.

[41] Richard Sinkhorn and Paul Knopp. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.*, 21(2):343–348, 1967.

[42] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive approximation*, 26(2):153–172, 2007.

[43] Alex Smola, Arthur Gretton, Le Song, and Bernhard Schölkopf. A Hilbert space embedding for distributions. In *International Conference on Algorithmic Learning Theory*, pages 13–31. Springer, 2007.

[44] Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (TOG)*, 34(4):66, 2015.

[45] Justin Solomon, Raif M. Rustamov, Leonidas Guibas, and Adrian Butscher. Wasserstein propagation for semi-supervised learning. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, pages I–306–I–314. JMLR.org, 2014.

[46] Le Song. Learning via Hilbert space embedding of distributions. 2008.

[47] Bharath K Sriperumbudur, Kenji Fukumizu, and Gert RG Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(Jul):2389–2410, 2011.

[48] Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

[49] Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

[50] Matthew Staib, Sebastian Claici, Justin M Solomon, and Stefanie Jegelka. Parallel streaming wasserstein barycenters. In *Advances in Neural Information Processing Systems*, pages 2647–2658, 2017.

[51] Elias M Stein. *Singular integrals and differentiability properties of functions (PMS-30)*, volume 30. Princeton university press, 2016.

[52] C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.

[53] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[54] J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using wasserstein barycenter with sparse support. *IEEE Transactions on Signal Processing*, 65(9):2317–2332, May 2017.

[55] VV Yurinskii. Exponential inequalities for sums of random vectors. *Journal of multivariate analysis*, 6(4):473–499, 1976.