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@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

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]

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

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])

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

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 $w \in \mathcal{C}(\mathcal{X})$,

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

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

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

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, $\forall \alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$*

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

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

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

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*

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

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 $\nabla_1\mathrm{OT}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})$ 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*

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

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 $\beta_1, \ldots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ and $\omega_1, \ldots, \omega_m \geq 0$ 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

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

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 $k = 0, 1, \ldots$
    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 $j = 1 \ldots m$. 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} = $ SINKHORNKNOPP$(\alpha_k, \alpha_k, \varepsilon)$
     $p(\cdot) = $ SINKHORNGRADIENT$(\mathbf{X}_k, \mathsf{a}_k, \mathsf{p})$
     For $j = 1, \dots m$
         $\mathsf{v}_j = $ SINKHORNKNOPP$(\alpha_k, \beta_j, \varepsilon)$
         $u_j(\cdot) = $ 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} = $ MINIMIZE(φ)
     $\mathbf{X}_{k+1} = [\mathbf{X}_k, x_{k+1}]$ 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 $p_k = \nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \alpha_k)$ 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

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

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 $x \mapsto \sum_{j=1}^m \omega_j \left(u_{jk}(x) - p_k(x) \right)$ 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 $\nabla_1 \mathrm{OT}_\varepsilon$ 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 $x \in \mathcal{X}$

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

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 $\mathrm{OT}_\varepsilon(\alpha_k, \beta_j)$ and $\mathrm{OT}_\varepsilon(\alpha_k, \alpha_k)$ 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 $p = \nabla_1 \mathrm{OT}_\varepsilon(\alpha_j, \alpha_k)$ 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 $\beta_1, \ldots \beta_m \in \mathcal{M}_1^+(\mathcal{X})$ have finite support and let α_k be the k-th iterate of Alg. 2 applied to (9). Then,*

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

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 $\tau \in (0, 1]$*

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

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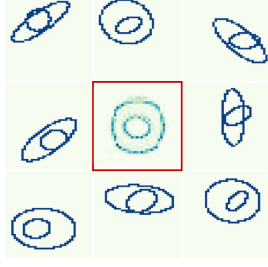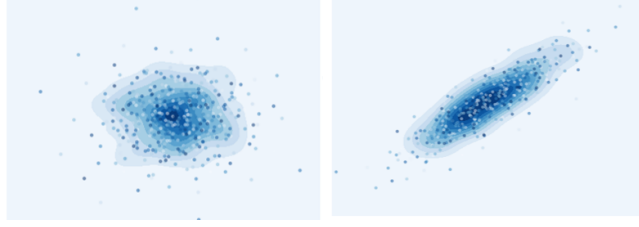
Fig. 1: Barycenter of nested ellipses

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 $\nabla_1 \text{OT}_\varepsilon$ 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 $\nabla_1 \text{OT}_\varepsilon$ 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* $c \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$ *with* $s > d/2$. *Let* $n \in \mathbb{N}$ *and* $\hat{\beta}_1, \ldots, \hat{\beta}_m$ *be empirical distributions with* n *support points, each independently sampled from* $\beta_1, \ldots, \beta_m$. *Let* α_k *be the* k-*th iterate of Alg. 2 applied to* $\hat{\beta}_1, \ldots, \hat{\beta}_m$. *Then for any* $\tau \in (0, 1]$, *the following holds with probability larger than* $1 - \tau$

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

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* $\beta_1, \ldots, \beta_m$ *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.

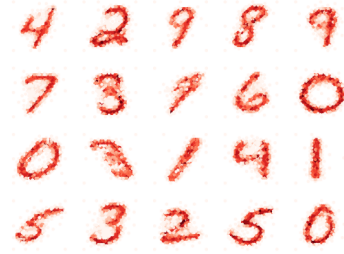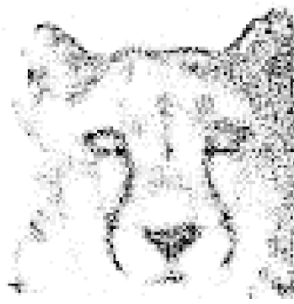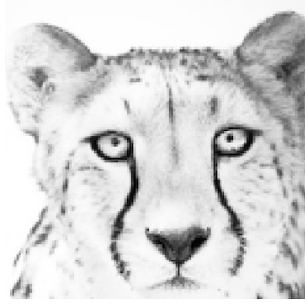
[1] https://github.com/GiulsLu/Sinkhorn-Barycenters

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 $\mathcal{N}(m_i, C_i)$ $i = 1, \ldots, 5$ 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 $\mathcal{N}(m_i, C_i)$, $i = 1, \ldots, 5$. 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.
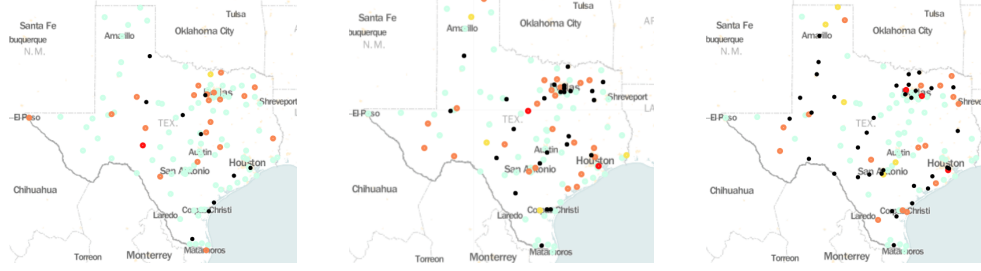
8

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.

Supplementary Material

Below we give an overview of the structure of the supplementary material and highlight the main novel results of this work.

Appendix A: abstract Frank-Wolfe algorithm in dual Banach spaces. This section contains full details on Frank-Wolfe algorithm. The novelty stands in the relaxation of the differentiability assumptions.

Appendix B: DAD problems and convergence of Sinkhorn-Knopp algorithm. This section is a brief review of basic concepts from the nonlinear Perrom-Frobeius theory, DAD problems, and applications to the study of Sinkorn algorithm.

Appendix C: Lipschitz continuitity of the gradient of the Sinkhorn divergence with respect to Total Variation. This section contains one of the main contributions of our work, Theorem C.4, from which we derive Theorem 4 in the main text.

Appendix D: Frank-Wolfe algorithm for Sinkhorn barycenters. This section contains the complete analysis of FW algorithm for Sinkhorn barycenters, which takes into account the error in the computation of Sinkhorn potentials and the error in their minimization. The main result is the convergence of the Frank-Wolfe scheme for finitely supported distributions in Theorem D.2.

Appendix E: Sample complexity of Sinkhorn potential and convergence of Algorithm 2 in case of continuous measures. This section contains the discussion and the proofs of two of main results of the work Theorem 6, Theorem 7.

Appendix F: additional experiments. This section contains additional experiment on barycenters of mixture of Gaussian, barycenter of a mesh in 3D (dinosaur) and additional figures on the experiment on Sinkhorn propagation described in Section 6.

A   The Frank-Wolfe algorithm in dual Banach spaces

In this section we detail the convergence analysis of the Frank-Wolfe algorithm in abstract dual Banach spaces and under mild directional differentiablility assumptions so to cover the setting of Sinkhorn barycenters described in Section 3 of the paper.

Let $\mathcal{W}$ be a real Banach space and let $\mathcal{W}^*$ be its topological dual. Let $\mathcal{D} \subset \mathcal{W}^*$ be a nonempty, closed, convex, and bounded set and let $\mathsf{G} \colon \mathcal{D} \to \mathbb{R}$ be a convex function. We address the following optimization problem

$$\min_{w \in \mathcal{D}} \mathsf{G}(w), \tag{A.1}$$

assuming that the set of solutions is nonemtpy.

We recall the concept of the tangent cone of feasible directions.

Definition A.1. *Let $w \in \mathcal{D}$. Then* the cone of feasible directions *of $\mathcal{D}$ at w is $\mathcal{F}_{\mathcal{D}}(w) = \mathbb{R}_+(\mathcal{D} - w)$ and the* tangent cone *of $\mathcal{D}$ at w is*

$$\mathcal{T}_{\mathcal{D}}(w) = \overline{\mathcal{F}_{\mathcal{D}}(w)} = \big\{ v \in \mathcal{W}^* \,|\, (\exists (t_k)_{k \in \mathbb{N}} \in \mathbb{R}_{++}^{\mathbb{N}})(t_k \to 0)(\exists (w_k)_{k \in \mathbb{N}} \in \mathcal{D}^{\mathbb{N}}) \, t_k^{-1}(w_k - w) \to v \big\}.$$

Remark A.1. $\mathcal{F}_{\mathcal{D}}(w)$ is the cone generated by $\mathcal{D} - w$, and it is a convex cone. Indeed, if $t > 0$ and $v \in \mathcal{F}_{\mathcal{D}}(w)$, then $tv \in \mathcal{F}_{\mathcal{D}}(w)$. Moreover, if $v_1, v_2 \in \mathcal{F}_{\mathcal{D}}(w)$, then there exists $t_1, t_2 > 0$ and $w_1, w_2 \in \mathcal{D}$ such that $v_i = t_i(w_i - w)$, $i = 1, 2$. Thus,

$$v_1 + v_2 = (t_1 + t_2)\Big(\frac{t_1}{t_1 + t_2} w_1 + \frac{t_2}{t_1 + t_2} w_2 - w \Big) \in \mathbb{R}_+(\mathcal{D} - w).$$

So, $\mathcal{T}_{\mathcal{D}}(w)$ is a closed convex cone too.

Definition A.2. *Let $w \in \mathcal{D}$ and $v \in \mathcal{F}_{\mathcal{D}}(w)$. Then,* the directional derivative of G at w in the direction v is

$$\mathsf{G}'(w; v) = \lim_{t \to 0^+} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t} \in [-\infty, +\infty[.$$

Remark A.2. The above definition is well-posed. Indeed, since v is a feasible direction of $\mathcal{D}$ at w, there exists $t_1 > 0$ and $w_1 \in \mathcal{D}$ such that $v = t_1(w_1 - w)$; hence

$$(\forall\, t \in]0, 1/t_1]) \quad w + tv = w + t\, t_1(w_1 - w) = (1 - t\, t_1)w + t\, t_1 w_1 \in \mathcal{D}.$$

Moreover, since G is convex, the function $t \in]0, 1/t_1] \mapsto (\mathsf{G}(w + tv) - \mathsf{G}(w))/t$ is increasing, hence

$$\lim_{t \to 0^+} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t} = \inf_{t \in]0, 1/t_1]} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t}. \tag{A.2}$$

It is easy to prove that the function

$$v \in \mathcal{F}_{\mathcal{D}}(w) \mapsto \mathsf{G}'(w; v) \in [-\infty, +\infty[$$

is positively homogeneous and sublinear (hence convex), that is,

   (i) $(\forall\, v \in \mathcal{F}_{\mathcal{D}}(w))(\forall\, t \in \mathbb{R}_+)\ \mathsf{G}'(w; tv) = t\mathsf{G}'(w; v);$

   (ii) $(\forall\, v_1, v_2 \in \mathcal{F}_{\mathcal{D}}(w))\ \mathsf{G}'(w; v_1 + v_2) \le \mathsf{G}'(w; v_1) + \mathsf{G}'(w; v_2).$

We make the following assumptions about G:

   H1 $(\forall\, w \in \mathcal{D})$ the function $v \mapsto \mathsf{G}'(w; v)$ is finite, that is, $\mathsf{G}'(w; v) \in \mathbb{R}$.

   H2 The *curvature of* G is finite, that is,

$$C_{\mathsf{G}} = \sup_{\substack{w, z \in \mathcal{D} \\ \gamma \in [0,1]}} \frac{2}{\gamma^2} \big(\mathsf{G}(w + \gamma(z - w)) - \mathsf{G}(w) - \gamma\mathsf{G}'(w, z - w)\big) < +\infty. \tag{A.3}$$

Remark A.3. For every $w, z \in \mathcal{D}$, we have

$$\mathsf{G}(z) - \mathsf{G}(w) \ge \mathsf{G}'(w; z - w). \tag{A.4}$$

This follows from (A.2) with $w_1 = z$ and $t = 1$ ($t_1 = 1$).

The (inexact) Frank-Wolfe algorithm is detailed in Algorithm A.1.

Algorithm A.1 Frank-Wolfe in Dual Banach Spaces

Let $(\gamma_k)_{k \in \mathbb{N}} \in \mathbb{R}_{++}^{\mathbb{N}}$ be such that $\gamma_0 = 1$ and, for every $k \in \mathbb{N}$, $1/\gamma_k \le 1/\gamma_{k+1} \le 1/2 + 1/\gamma_k$ (e.g., $\gamma_k = 2/(k+2)$). Let $w_0 \in \mathcal{D}$ and $(\Delta_k)_{k \in \mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ be such that $(\Delta_k/\gamma_k)_{k \in \mathbb{N}}$ is nondecreasing. Then

    for $k = 0, 1, \ldots$

      | find $z_{k+1} \in \mathcal{D}$ is such that $\mathsf{G}'(w_k; z_{k+1} - w_k) \le \inf_{z \in \mathcal{D}} \mathsf{G}'(w_k; z - w_k) + \frac{1}{2}\Delta_k$

      | $w_{k+1} = w_k + \gamma_k(z_{k+1} - w_k)$

Remark A.4.

   (i) Algorithm A.1 does not require the sub-problem $\min_{z \in \mathcal{D}} \mathsf{G}'(w_k; z - w_k)$ to have solutions. Indeed it only requires computing a Δ_k-minimizer of $\mathsf{G}'(w_k; \cdot - w_k)$ on $\mathcal{D}$, which always exists.

   (ii) Since $\mathcal{D}$ is weakly-$*$ compact (by Banach-Alaoglu theorem), if $\mathsf{G}'(w_k; \cdot - w_k)$ is weakly-$*$ continuous on $\mathcal{D}$, then the sub-problem $\min_{z \in \mathcal{D}} \mathsf{G}'(w_k; z - w_k)$ admits solutions. Note that this occurs when the directional derivative $\mathsf{G}'(w; \cdot)$ is linear and can be represented in $\mathcal{W}$. This case is addressed in the subsequent Proposition A.7.

Theorem A.5. *Let $(w_k)_{k \in \mathbb{N}}$ be defined according to Algorithm A.1. Then, for every integer $k \ge 1$,*

$$\mathsf{G}(w_k) - \min_{w \in \mathcal{D}} \mathsf{G}(w) \le C_{\mathsf{G}}\gamma_k + \Delta_k. \tag{A.5}$$

Proof. Let $w_* \in \mathcal{D}$ be a solution of problem (A.1). It follows from H2 and the definition of w_{k+1} in Algorithm A.1, that

$$\mathsf{G}(w_{k+1}) \leq \mathsf{G}(w_k) + \gamma_k \mathsf{G}'(w_k; z_{k+1} - w_k) + \frac{\gamma_k^2}{2} C_\mathsf{G}.$$

Moreover, it follows from the definition of z_{k+1} in Algorithm A.1 and (A.4) that

$$\mathsf{G}'(w_k; z_{k+1} - w_k) \leq \inf_{z \in \mathcal{D}} \mathsf{G}'(w_k; z - w_k) + \frac{1}{2}\Delta_k$$

$$\leq \mathsf{G}'(w_k; w_* - w_k) + \frac{1}{2}\Delta_k$$

$$\leq -(\mathsf{G}(w_k) - \mathsf{G}(w_*)) + \frac{1}{2}\Delta_k.$$

Then,

$$\mathsf{G}(w_{k+1}) - \mathsf{G}(w_*) \leq (1 - \gamma_k)(\mathsf{G}(w_k) - \mathsf{G}(w_*)) + \frac{\gamma_k^2}{2}\left(C_\mathsf{G} + \frac{\Delta_k}{\gamma_k}\right). \tag{A.6}$$

Now, similarly to [28, Theorem 2], we can prove (A.5) by induction. Since $\gamma_0 = 1$, $1/\gamma_1 \leq 1/2 + 1/\gamma_0$, and $\Delta_0/\gamma_0 \leq \Delta_1/\gamma_1$, it follows from (A.6) that

$$\mathsf{G}(w_1) - \mathsf{G}(w_*) \leq \frac{1}{2}\left(C_\mathsf{G} + \frac{\Delta_0}{\gamma_0}\right) \leq \gamma_1\left(C_\mathsf{G} + \frac{\Delta_1}{\gamma_1}\right), \tag{A.7}$$

hence (A.5) is true for $k = 1$. Set, for the sake of brevity, $C_k = C_\mathsf{G} + \Delta_k/\gamma_k$ and suppose that (A.5) holds for $k \in \mathbb{N}$, $k \geq 1$. Then, it follows from (A.6) and the properties of $(\gamma_k)_{k \in \mathbb{N}}$ that

$$\mathsf{G}(w_{k+1}) - \mathsf{G}(w_*) \leq (1 - \gamma_k)\gamma_k C_k + \frac{\gamma_k^2}{2}C_k$$

$$= C_k \gamma_k \left(1 - \frac{\gamma_k}{2}\right)$$

$$\leq C_k \gamma_k \left(1 - \frac{\gamma_{k+1}}{2}\right)$$

$$\leq C_k \frac{1}{1/\gamma_{k+1} - 1/2}\left(1 - \frac{\gamma_{k+1}}{2}\right)$$

$$= C_k \gamma_{k+1}$$

$$\leq C_{k+1}\gamma_{k+1}. \qquad \square$$

Corollary A.6. *Under the assumptions of Theorem A.5, suppose in addition that $\Delta_k = \Delta\gamma_k^\zeta$, for some $\zeta \in [0, 1]$ and $\Delta \geq 0$. Then we have*

$$\mathsf{G}(w_k) - \min_{w \in \mathcal{D}} \mathsf{G}(w) \leq C_\mathsf{G}\gamma_k + \Delta\gamma_k^\zeta. \tag{A.8}$$

Proof. It follows from Theorem A.5 by noting that the sequence $\Delta_k/\gamma_k = 1/\gamma_k^{1-\zeta}$ is nondecreasing.
$\square$

Proposition A.7. *Suppose that there exists a mapping $\nabla\mathsf{G}\colon \mathcal{D} \to \mathcal{W}$ such that[2],*

$$(\forall w \in \mathcal{D})(\forall z \in \mathcal{D}) \quad \langle \nabla\mathsf{G}(w), z - w \rangle = \mathsf{G}'(w; z - w). \tag{A.9}$$

Then the following holds.

(i) *Let $k \in \mathbb{N}$ and suppose that there exists $u_k \in \mathcal{W}$ such that $\|u_k - \nabla\mathsf{G}(w_k)\| \leq \Delta_{1,k}/4$ and that $z_{k+1} \in \mathcal{D}$ satisfies*

$$\langle u_k, z_{k+1} \rangle \leq \min_{z \in \mathcal{D}} \langle u_k, z \rangle + \frac{\Delta_{2,k}}{2},$$

for some $\Delta_{1,k}, \Delta_{2,k} > 0$. Then

$$\mathsf{G}'(w_k; z_{k+1} - w_k) \leq \min_{z \in \mathcal{D}} \mathsf{G}'(w_k; z - w_k) + \frac{1}{2}(\Delta_{1,k}\mathrm{diam}(\mathcal{D}) + \Delta_{2,k}). \tag{A.10}$$

---

[2]This mapping does not need to be unique.

15

(ii) *Suppose that* $\nabla\mathsf{G}\colon \mathcal{D} \to \mathcal{W}$ *is L-Lipschitz continuous for some $L > 0$. Then, for every* $w, z \in \mathcal{D}$ *and* $\gamma \in [0,1]$,

$$\mathsf{G}(w + \gamma(z - w)) - \mathsf{G}(w) - \gamma \langle z - w, \nabla\mathsf{G}(w)\rangle \leq \frac{L}{2}\gamma^2 \|z - w\|^2$$

and hence $C_\mathsf{G} \leq L\mathrm{diam}(\mathcal{D})^2$.

Proof. (i): We have

$$\langle \nabla\mathsf{G}(w_k), z_{k+1} - w_k \rangle = \langle u_k, z_{k+1} - w_k \rangle + \langle \nabla\mathsf{G}(w_k) - u_k, z_{k+1} - w_k \rangle$$

$$\leq \min_{z \in \mathcal{D}} \langle u_k, z - w_k \rangle + \frac{\Delta_{2,k}}{2} + \frac{\Delta_{1,k}}{4}\mathrm{diam}(\mathcal{D}). \qquad \text{(A.11)}$$

Moreover,

$$(\forall z \in \mathcal{D}) \quad \langle u_k, z - w_k \rangle = \langle \nabla\mathsf{G}(w_k), z - w_k \rangle + \langle u_k - \nabla\mathsf{G}(w_k), z - w_k \rangle$$

$$\leq \langle \nabla\mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_{1,k}}{4}\mathrm{diam}(\mathcal{D}),$$

hence

$$\min_{z \in \mathcal{D}} \langle u_k, z - w_k \rangle \leq \min_{z \in \mathcal{D}} \langle \nabla\mathsf{G}(w_k), z - w_k \rangle + \frac{\Delta_{1,k}}{4}\mathrm{diam}(\mathcal{D}). \qquad \text{(A.12)}$$

Thus, (A.10) follows from (A.11), (A.12), and (A.9).

(ii): Let $w, z \in \mathcal{D}$, and define $\psi\colon [0,1] \to \mathcal{W}^*$ such that, $\forall \gamma \in [0,1]$, $\psi(\gamma) = \mathsf{G}(w + \gamma(z - w))$. Then, it is easy to see that for every $\gamma \in \,]0,1[$, ψ is differentiable at γ and $\psi'(\gamma) = \mathsf{G}'(w + \gamma(z - w); z - w) = \langle \nabla\mathsf{G}(w + \gamma(z - w)), z - w \rangle$. Moreover, ψ is continuous on $[0,1]$. Therefore, the fundamental theorem of calculus yields

$$\psi(\gamma) - \psi(0) = \int_0^\gamma \psi'(t)dt$$

and hence

$$\mathsf{G}(w + \gamma(z - w)) - \mathsf{G}(w) - \langle \nabla\mathsf{G}(w), z - w \rangle = \int_0^\gamma \langle \nabla\mathsf{G}(w + t(z - w)) - \nabla\mathsf{G}(w), z - w \rangle \, dt$$

$$\leq \int_0^\gamma \|\nabla\mathsf{G}(w + t(z - w)) - \nabla\mathsf{G}(w)\| \, \|z - w\| \, dt$$

$$\leq \int_0^\gamma Lt \, \|z - w\|^2 \, dt$$

$$= L\frac{\gamma^2}{2} \|z - w\|^2. \qquad \square$$

The following result is an extension of a classical result on the directional differentiability of a max function [7, Theorem 4.13] which relaxes the inf-compactness condition and allows the parameter space to be a convex set, instead of the entire Banach space. This result provides a prototype of functions (of which the entropic regularization of the Wasserstein distance is an instance) which are directionally differentiable only along the feasible directions of their domain and satisfies the hypotheses of Proposition A.7.

Proposition A.8. *Let Z and $\mathcal{W}$ be real Banach spaces and let $\mathcal{W}^*$ be the topological dual of $\mathcal{W}$. Let $\mathcal{D} \subset \mathcal{W}^*$ be a nonempty closed convex set, and let $g\colon Z \times \mathcal{W}^* \to \mathbb{R}$ be such that*

1) *for every $z \in Z$, $g(z, \cdot)\colon \mathcal{W}^* \to \mathbb{R}$ is Gâteaux differentiable with derivative in $\mathcal{W}$, and the partial derivative with respect to the second variable $D_2g\colon Z \times \mathcal{W}^* \to \mathcal{W}$ is continuous.*

2) *for every $w \in \mathcal{D}$, $S(w) := \mathrm{argmax}_Z \, g(\cdot, w) \neq \varnothing$.*

3) *there exists a continuous mapping $\varphi\colon \mathcal{D} \to Z$ such that, for every $w \in \mathcal{D}$, $\varphi(w) \in S(w)$.*

Let $\mathsf{G}\colon \mathcal{D} \to \mathbb{R}$ *be defined as*

$$\mathsf{G}(w) = \max_{z \in Z} g(z, w). \tag{A.13}$$

Then, G *is continuous, directionally differentiable, and, for every* $w \in \mathcal{D}$ *and* $v \in \mathcal{F}_{\mathcal{D}}(w)$

$$\mathsf{G}'(w; v) = \max_{z \in S(w)} \langle D_2 g(z, w), v \rangle = \langle D_2 g(\varphi(w), w), v \rangle . \tag{A.14}$$

Proof. The function G is well defined, since by assumption 2), for every $w \in \mathcal{D}$, $\operatorname{argmax}_Z g(\cdot, w) \neq \varnothing$. Let $w, u \in \mathcal{D}$ with $w \neq u$. Then, since $\varphi(w) \in S(w)$, we have $\mathsf{G}(w) = g(\varphi(w), w)$ and hence

$$\frac{\mathsf{G}(u) - \mathsf{G}(w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|}$$
$$\geq \frac{g(\varphi(w), u) - g(\varphi(w), w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|} \to 0, \quad \text{(A.15)}$$

since $g(\varphi(w), \cdot)$ is Fréchet differentiable[3] at w with gradient $D_2 g(\varphi(w), w)$. Now, $\varphi(u) \in S(u)$, and hence $\mathsf{G}(u) = g(\varphi(u), u)$. Moreover, $g(\varphi(u), w) \leq \mathsf{G}(w)$. Therefore,

$$\frac{\mathsf{G}(u) - \mathsf{G}(w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|}$$
$$\leq \frac{g(\varphi(u), u) - g(\varphi(u), w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|}. \quad \text{(A.16)}$$

Let $\varepsilon > 0$. Since $D_2 g$ is continuous, there exists $\delta > 0$ such that, for every $z' \in Z$ and $w' \in \mathcal{W}^*$

$$\|z' - \varphi(w)\| \leq \delta \text{ and } \|w' - w\| \leq \delta \implies \|D_2 g(z', w') - D_2 g(\varphi(w), w)\| \leq \varepsilon. \tag{A.17}$$

Moreover, since $\varphi\colon \mathcal{D} \to Z$ is continuous, there exists $\eta > 0$ such that,

$$\|u - w\| \leq \eta \implies \|\varphi(u) - \varphi(w)\| \leq \delta. \tag{A.18}$$

Let $z' \in Z$ and suppose that $\|z' - \varphi(w)\| \leq \delta$ and $\|u - w\| \leq \delta$. Define $\psi\colon [0, 1] \to \mathbb{R}$ such that, for every $s \in [0, 1]$, $\psi(s) = g(z', w + s(u - w))$. Then, ψ is continuously differentiable on $[0, 1]$ and $\psi'(s) = \langle D_2 g(z', w + s(u - w)), u - w \rangle$. Therefore,

$$\psi(1) - \psi(0) = \int_0^1 \psi'(s) ds \tag{A.19}$$

and hence, it follows from (A.17) that

$$|g(z', u) - g(z', w) - \langle D_2 g(\varphi(w), w), u - w \rangle|$$
$$= \left| \int_0^1 \langle D_2 g(z', w + s(u - w)) - D_2 g(\varphi(w), w), u - w \rangle \, ds \right|$$
$$\leq \int_0^1 \|D_2 g(z', w + s(u - w)) - D_2 g(\varphi(w), w)\| \, \|u - w\| \, ds$$
$$\leq \varepsilon \|u - w\| .$$

Therefore, we derive from (A.18), that for every $u \in \mathcal{D}$ such that $\|u - w\| \leq \min\{\eta, \delta\}$, we have

$$\left| \frac{g(\varphi(u), u) - g(\varphi(u), w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|} \right| \leq \varepsilon.$$

This shows that

$$\lim_{\substack{u \in \mathcal{D} \\ u \to w}} \frac{g(\varphi(u), u) - g(\varphi(u), w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|} = 0. \tag{A.20}$$

Then, we derive from (A.15), (A.16), and (A.20) that

$$\lim_{\substack{u \in \mathcal{D} \\ u \to w}} \frac{\mathsf{G}(u) - \mathsf{G}(w) - \langle D_2 g(\varphi(w), w), u - w \rangle}{\|u - w\|} = 0. \tag{A.21}$$

---

[3]continuously Gâteaux differentiable function are Fréchet differentiable [7, pp.34-35].

This implies that $\lim_{u \in \mathcal{D}, u \to w} \mathsf{G}(u) = \mathsf{G}(w)$. Moreover, if $v \in \mathcal{F}_\mathcal{D}(w)$, there exists $\lambda > 0$ and $u \in \mathcal{D}$ such that $v = \lambda(u - w)$ and, for every $t \in \,]0, 1/\lambda]$,

$$
\begin{aligned}
\frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t} &- \langle D_2 g(\varphi(w), w), v \rangle \\
&= \|\lambda(u - w)\| \frac{\mathsf{G}(w + t\lambda(u - w)) - \mathsf{G}(w) - \langle D_2 g(\varphi(w), w), t\lambda(u - w) \rangle}{\|t\lambda(u - w)\|}
\end{aligned} \tag{A.22}
$$

and the right hand side goes to zero as $t \to 0^+$, because of (A.21). Therefore, for every $z \in S(w)$, since $\mathsf{G}(w) = g(z, w)$ and $\mathsf{G}(w + tv) \geq g(z, w + tv)$, we have

$$
\langle D_2 g(\varphi(w), w), v \rangle = \lim_{t \to 0^+} \frac{\mathsf{G}(w + tv) - \mathsf{G}(w)}{t} \geq \lim_{t \to 0^+} \frac{g(z, w + tv) - g(z, w)}{t} = \langle D_2 g(z, w), v \rangle
$$

and (A.14) follows. $\qquad\square$

B DAD problems and convergence of Sinkhorn-Knopp algorithm

In this section we review the basic concepts of the nonlinear Perron-Frobenius theory [32] which provides tools for dealing with DAD problems and ultimately to study the key properties of the Sinkhorn potentials. This analysis will allow us to provide in Appendix C an upper bound estimate for the Lipschitz constant of the gradient of B_ε, which is needed in the Frank-Wolfe algorithm.

B.1 Hilbert's metric and the Birkhoff-Hopf theorem

In the rest of the appendix we will assume $\mathcal{X} \subset \mathbb{R}^d$ to be a compact set. We denote by $\mathcal{C}(\mathcal{X})$ the space of continuous functions on $\mathcal{X}$ endowed with the sup norm, namely $\|f\|_\infty = \sup_{x \in \mathcal{X}} |f(x)|$. Let $\mathcal{C}_+(\mathcal{X})$ be the cone of nonnegative continuous functions, that is, $f \in \mathcal{C}(\mathcal{X})$ such that $f(x) \geq 0$ for every $x \in \mathcal{X}$. Also, we denote by $\mathcal{C}_{++}(\mathcal{X})$ the set of continuous and (strictly) positive functions on $\mathcal{X}$, which turns out to be the interior of $\mathcal{C}_+(\mathcal{X})$.

Let $\mathsf{c} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ be a positive, symmetric, and continuous function and define $\mathsf{k} \colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{++}$ as

$$
(\forall x, y \in \mathcal{X}) \qquad \mathsf{k}(x, y) = e^{-\frac{\mathsf{c}(x,y)}{\varepsilon}}. \tag{B.1}
$$

Set $\mathsf{D} = \sup_{x, y \in \mathcal{X}} \mathsf{c}(x, y)$. Then, we have $\mathsf{k}(x, y) \in [e^{-\mathsf{D}/\varepsilon}, 1]$ for all $x, y \in \mathcal{X}$. Let $\alpha \in \mathcal{M}_1^+(\mathcal{X})$. The operator $\mathsf{L}_\alpha \colon \mathcal{C}(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ is defined as

$$
(\forall f \in \mathcal{C}(\mathcal{X})) \qquad \mathsf{L}_\alpha f \colon x \mapsto \int \mathsf{k}(x, z) f(z) \, d\alpha(z). \tag{B.2}
$$

Note that L_α is linear and continuous. In particular, since $\mathsf{k}(x, y) \in [0, 1]$ for all $x, y \in \mathcal{X}$, we have

$$
(\forall f \in \mathcal{C}_+(\mathcal{X})) \qquad \mathsf{L}_\alpha f \geq 0 \tag{B.3}
$$

and

$$
(\forall f \in \mathcal{C}(\mathcal{X})) \qquad \|\mathsf{L}_\alpha f\|_\infty \leq \|f\|_\infty. \tag{B.4}
$$

Hilbert's Metric. The cone $\mathcal{C}_+(\mathcal{X})$ induces a partial ordering $\leq$ on $\mathcal{C}(\mathcal{X})$, such that

$$
(\forall f, f' \in \mathcal{C}(\mathcal{X})) \qquad f \leq f' \Leftrightarrow f' - f \in \mathcal{C}_+(\mathcal{X}). \tag{B.5}
$$

According to [32], we say that a function $f' \in \mathcal{C}_+(\mathcal{X})$ *dominates* $f \in \mathcal{C}(\mathcal{X})$ if there exist $t, s \in \mathbb{R}$ such that

$$
tf' \leq f \leq sf'. \tag{B.6}
$$

This notion induces an equivalence relation on $\mathcal{C}_+(\mathcal{X})$, denoted $f \sim f'$, meaning that f dominates f' and f' dominates f. The corresponding equivalence classes are called *parts* of $\mathcal{C}_+(\mathcal{X})$. Let $f, f' \in \mathcal{C}_+(\mathcal{X})$ be such that $f \sim f'$. We define

$$
M(f/f') = \inf\{s \in \mathbb{R} \mid f \leq sf'\} \qquad \text{and} \qquad m(f/f') = \sup\{t \in \mathbb{R} \mid tf' \leq f\}. \tag{B.7}
$$

Note that $m(f/f') \leq M(f/f')$. Moreover, for every $f, f' \in \mathcal{C}_+(\mathcal{X})$ such that $f \sim f'$, we have that $\text{supp}(f) = \text{supp}(f')$ and if $f' \neq 0$ (hence $f \neq 0$), then

$$M(f/f') = \max_{x \in \text{supp}(f')} \frac{f(x)}{f'(x)} > 0 \quad \text{and} \quad m(f/f') = \min_{x \in \text{supp}(f')} \frac{f(x)}{f'(x)} > 0. \qquad (\text{B.8})$$

The *Hilbert's metric* is defined as

$$d_H(f, f') = \log \frac{M(f/f')}{m(f/f')}, \qquad (\text{B.9})$$

for all $f \sim f'$ with $f \neq 0$ and $f' \neq 0$, $d_H(0, 0) = 0$ and $d_H(f, f') = +\infty$ otherwise. Direct calculation shows that [33, Proposition 2.1.1]

(i) $d_H(f, f') \geq 0$ and $d_H(f, f') = d_H(f', f)$, for every $f, f' \in \mathcal{C}_+(\mathcal{X})$;

(ii) $d_H(f, f'') \leq d_H(f, f') + d_H(f', f'')$, for every $f, f', f'' \in \mathcal{C}_+(\mathcal{X})$ with $f \sim f'$ and $f' \sim f''$;

(iii) $d_H(sf, tf') = d_H(f, f')$, for every $f, f' \in \mathcal{C}_+(\mathcal{X})$ and $s, t > 0$.

Note that d_H is not a metric on the parts of $\mathcal{C}_+(\mathcal{X})$. However the set $\mathcal{C}_{++}(\mathcal{X}) \cap \partial B_1(0) = \{f \in \mathcal{C}_{++}(\mathcal{X}) \mid \|f\|_\infty = 1\}$ equipped with d_H is a complete metric space [36]. Also, d_H induces a metric on the rays of the parts of $\mathcal{C}_+(\mathcal{X})$ [33, Lemma 2.1].

We now focus on $\mathcal{C}_{++}(\mathcal{X})$. A direct consequence of Hilbert's metric properties is the following.

Lemma B.1 (Hilbert's Metric on $\mathcal{C}_{++}(\mathcal{X})$). *The interior of $\mathcal{C}_+(\mathcal{X})$ corresponds to the set of (strictly) positive functions $\mathcal{C}_{++}(\mathcal{X})$ and is a part of $\mathcal{C}_+(\mathcal{X})$ with respect to the equivalence relation induced by dominance. For every $f, f' \in \mathcal{C}_{++}(\mathcal{X})$,*

$$M(f/f') = \max_{x \in \mathcal{X}} \frac{f(x)}{f'(x)} \qquad m(f/f') = \min_{x \in \mathcal{X}} \frac{f(x)}{f'(x)}, \qquad (\text{B.10})$$

and $M(f/f') \geq m(f/f') > 0$. Therefore

$$d_H(f, f') = \log \max_{x, y \in \mathcal{X}} \frac{f(x) \, f'(y)}{f(y) \, f'(x)}. \qquad (\text{B.11})$$

Proof. Since $\mathcal{X}$ is compact it is straightfoward to see that $\mathcal{C}_{++}(\mathcal{X})$ is the interior of $\mathcal{C}_+(\mathcal{X})$. By applying [32, Lemma 1.2.2] we have that $\mathcal{C}_{++}(\mathcal{X})$ is a part of $\mathcal{C}_+(\mathcal{X})$. The characterization of $M(f/f')$ and $m(f/f')$ follow by direct calculation from the definition using the fact that $\inf_\mathcal{X} h = \min_\mathcal{X} h > 0$ for any $h \in \mathcal{C}_{++}(\mathcal{X})$ since $\mathcal{X}$ is compact. Finally, the characterization of Hilbert's metric on $\mathcal{C}_{++}(\mathcal{X})$ is obtained by recalling that $(\min_{x \in \mathcal{X}} h(x))^{-1} = \max_{x \in \mathcal{X}} h(x)^{-1}$ for every $h \in \mathcal{C}_{++}(\mathcal{X})$. $\qquad\square$

Lemma B.2 (Ordering properties of L_α). *Let $\alpha \in \mathcal{M}_1^+(\mathcal{X})$. Then the following holds:*

(i) *the operator L_α is order-preserving (with respect to the cone $\mathcal{C}_+(\mathcal{X})$), that is,*

$$(\forall f, f' \in \mathcal{C}(\mathcal{X})) \qquad f \leq f' \Rightarrow \mathsf{L}_\alpha f \leq \mathsf{L}_\alpha f'; \qquad (\text{B.12})$$

(ii) *L_α maps parts of $\mathcal{C}_+(\mathcal{X})$ into parts of $\mathcal{C}_+(\mathcal{X})$, that is,*

$$(\forall f, f' \in \mathcal{C}_+(\mathcal{X})) \qquad f \sim f' \Rightarrow \mathsf{L}_\alpha f \sim \mathsf{L}_\alpha f'; \qquad (\text{B.13})$$

(iii) *$\mathsf{L}_\alpha(\mathcal{C}_+(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$ and $\mathsf{L}_\alpha(\mathcal{C}_{++}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X})$.*

Proof. (i): Let $f, f' \in \mathcal{C}(\mathcal{X})$ with $f \leq f'$. Then $f' - f \in \mathcal{C}_+(\mathcal{X})$ and by linearity of L_α combined with (B.3), we have $\mathsf{L}_\alpha f' - \mathsf{L}_\alpha f = \mathsf{L}_\alpha(f - f') \geq 0$.

(ii): Let $f, f' \in \mathcal{C}_+(\mathcal{X})$ with $f \sim f'$. Then there exist $t, s \in \mathbb{R}$ and $s', t' \in \mathbb{R}$ such that $tf' \leq f \leq sf'$ and $t'f \leq f' \leq s'f$. Since L_α is linear and order-preserving, we have $\mathsf{L}_\alpha f \sim \mathsf{L}_\alpha f'$.

(iii): Let $f \in \mathcal{C}_+(\mathcal{X})$. By (B.3) and (B.4), for any $x \in \mathcal{X}$

$$0 \leq (\mathsf{L}_\alpha f)(x) \leq \|\mathsf{L}_\alpha f\|_\infty \leq \int f(x) \, d\alpha(x) = \|f\|_{L^1(\mathcal{X}, \alpha)}. \qquad (\text{B.14})$$

19

Moreover,

$$\mathsf{L}_\alpha f(x) = \int k(y, x) f(y) \, d\alpha(y) \geq e^{-\mathsf{D}/\varepsilon} \, \|f\|_{L^1(\mathcal{X}, \alpha)}. \tag{B.15}$$

Therefore, if $\|f\|_{L^1(\mathcal{X},\alpha)} = 0$ then by (B.14) $\mathsf{L}_\alpha f = 0$ while, if $\|f\|_{L^1(\mathcal{X},\alpha)} > 0$ then by (B.15) $\mathsf{L}_\alpha f \in \mathcal{C}_{++}(\mathcal{X})$. We conclude that the operator L_α maps $\mathcal{C}_+(\mathcal{X})$ in $\mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$. Moreover, $\mathsf{L}_\alpha(\mathcal{C}_{++}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X})$, since for every $f \in \mathcal{C}_{++}(\mathcal{X})$ we have $\|f\|_{L^1(\mathcal{X},\alpha)} \geq \min_\mathcal{X} f > 0$. □

Following [32, Section A.4] we now introduce a quantity which plays a central role in our analysis.

Definition B.1 (Projective Diameter of L_α). *Let* $\alpha \in \mathcal{M}_1^+(\mathcal{X})$. *The* projective diameter of L_α *is*

$$\Delta(\mathsf{L}_\alpha) = \sup\{d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \mid f, f' \in \mathcal{C}_+(\mathcal{X}), \ \mathsf{L}_\alpha f \sim \mathsf{L}_\alpha f'\}. \tag{B.16}$$

The following result shows that it is possible to find a finite upper bound on $\Delta(\mathsf{L}_\alpha)$ that is independent on α.

Proposition B.3 (Upper bound on the Projective Diameter of L_α). *Let* $\alpha \in \mathcal{M}_1^+(\mathcal{X})$. *Then*

$$\Delta(\mathsf{L}_\alpha) \leq 2\mathsf{D}/\varepsilon. \tag{B.17}$$

Proof. Let $f, f' \in \mathcal{C}_+(\mathcal{X})$. Recall that L_α maps $\mathcal{C}_+(\mathcal{X})$ into $\mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$ (see Lemma B.2 (iii)) and that $\{0\}$ and $\mathcal{C}_{++}(\mathcal{X})$ are two parts of $\mathcal{C}_+(\mathcal{X})$ with respect to the relation $\sim$ (see [32, Lemma 1.2.2]). Now, if $\mathsf{L}_\alpha f = \mathsf{L}_\alpha f' = 0$, then we have $d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') = d_H(0, 0) = 0$. Therefore it is sufficient to study the case that $\mathsf{L}_\alpha f, \mathsf{L}_\alpha f' \in \mathcal{C}_{++}(\mathcal{X})$. Following the characterization of Hilbert's metric on $\mathcal{C}_{++}(\mathcal{X})$ given in Lemma B.1, we have

$$
\begin{aligned}
d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') &= \log \ \max_{x,y \in \mathcal{X}} \ \frac{(\mathsf{L}_\alpha f)(x) \, (\mathsf{L}_\alpha f')(y)}{(\mathsf{L}_\alpha f)(y) \, (\mathsf{L}_\alpha f')(x)} \\
&= \log \ \max_{x,y \in \mathcal{X}} \ \frac{\int \mathsf{k}(x, z) f(z) \, d\alpha(z) \ \int \mathsf{k}(y, w) f'(w) \, d\alpha(w)}{\int \mathsf{k}(y, z) f(z) \, d\alpha(z) \ \int \mathsf{k}(x, w) f'(w) \, d\alpha(w)} \\
&= \log \ \max_{x,y \in \mathcal{X}} \ \frac{\int \mathsf{k}(x, z) \mathsf{k}(y, w) \, f(z) f'(w) \, d\alpha(z) d\alpha(w)}{\int \mathsf{k}(y, z) \mathsf{k}(x, w) \, f(z) f'(w) \, d\alpha(z) d\alpha(w)} \\
&= \log \ \max_{x,y \in \mathcal{X}} \ \frac{\int \frac{\mathsf{k}(x,z)\mathsf{k}(y,w)}{\mathsf{k}(y,z)\mathsf{k}(x,w)} \, \mathsf{k}(y, z) \mathsf{k}(x, w) \, f(z) f'(w) \, d\alpha(z) d\alpha(w)}{\int \mathsf{k}(y, z) \mathsf{k}(x, w) \, f(z) f'(w) \, d\alpha(z) d\alpha(w)} \\
&\leq \log \ \max_{x,y,z,w \in \mathcal{X}} \ \frac{\mathsf{k}(x, z) \mathsf{k}(y, w)}{\mathsf{k}(y, z) \mathsf{k}(x, w)}.
\end{aligned}
$$

Since, for every $x, y \in \mathcal{X}$, $\mathsf{c}(x, y) \in [0, \mathsf{D}]$, we have $\mathsf{k}(x, y) \in [e^{-\mathsf{D}/\varepsilon}, 1]$ and hence

$$d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq 2\mathsf{D}/\varepsilon. \qquad \square$$

A consequence of Proposition B.3 is a special case of Birkhoff-Hopf theorem.

Theorem B.4 (Birkhoff-Hopf Theorem). *Let* $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$ *and* $\alpha \in \mathcal{M}_1^+(\mathcal{X})$. *Then, for every* $f, f' \in \mathcal{C}_+(\mathcal{X})$ *such that* $f \sim f'$, *we have*

$$d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq \lambda \, d_H(f, f'). \tag{B.18}$$

Proof. The statement is a direct application of the Birkhoff-Hopf theory [32, Sections A.4 and A.7] The *Birkhoff contraction ratio* of L_α is defined as

$$\kappa(\mathsf{L}_\alpha) = \inf\left\{\hat{\lambda} \in \mathbb{R}_+ \mid d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \leq \hat{\lambda} d_H(f, f') \ \forall f, f' \in \mathcal{C}_+(\mathcal{X}), \ f \sim f'\right\}.$$

Then it follows from Birkhoff-Hopf theorem [32, Theorem A.4.1] that

$$\kappa(\mathsf{L}_\alpha) = \tanh\left(\frac{1}{4}\Delta(\mathsf{L}_\alpha)\right). \tag{B.19}$$

Recalling the upper bound on the projective diameter f L_α given in Proposition B.3, we have

$$\kappa(\mathsf{L}_\alpha) \leq \tanh\left(\frac{\mathsf{D}}{2\varepsilon}\right) = \frac{e^{\mathsf{D}/\varepsilon} - 1}{e^{\mathsf{D}/\varepsilon} + 1} = \lambda,$$

and (B.18) follows. □

20

B.2 DAD problems

The map A_α. Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$. We define the map $A_\alpha \colon \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$, such that

$$(\forall\, f \in \mathcal{C}_{++}(\mathcal{X})) \qquad A_\alpha(f) = R \circ L_\alpha(f) = 1/(L_\alpha f), \tag{B.20}$$

where $R \colon \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ is defined by $R(f) = 1/f$ with

$$(1/f) \colon x \mapsto \frac{1}{f(x)}. \tag{B.21}$$

Note that A_α is well defined since, by Lemma B.2 (iii), $L_\alpha(\mathcal{C}_{++}(\mathcal{X})) \subset \mathcal{C}_{++}(\mathcal{X})$ and, for every $f \in \mathcal{C}_{++}(\mathcal{X})$, $\min_{\mathcal{X}} f > 0$, being $\mathcal{X}$ compact. Moreover, it follows from (B.11) in Lemma B.1, that, for any two $f, f' \in \mathcal{C}_{++}(\mathcal{X})$

$$d_H(1/f, 1/f') = \log \max_{x,y \in \mathcal{X}} \frac{f(y)f'(x)}{f(x)f'(y)} = d_H(f, f'). \tag{B.22}$$

We highlight here the connection between T_α introduced in the main text in (3) and A_α, namely for any $\alpha \in \mathcal{M}_1^+(\mathcal{X})$ and $u \in \mathcal{C}(\mathcal{X})$

$$T_\alpha(u) = \varepsilon \log(A_\alpha(e^{u/\varepsilon})). \tag{B.23}$$

Dual OT_ε Problem. We focus on the dual problem (2) of the optimal transport problem with entropic regularization. Let $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ and $\varepsilon > 0$, we consider

$$\max_{u,v \in \mathcal{C}(\mathcal{X})} \int u(x)\, d\alpha + \int v(y)\, d\beta(y) - \varepsilon \int e^{\frac{u(x)+v(y)-c(x,y)}{\varepsilon}}\, d\alpha(x)d\beta(y). \tag{B.24}$$

The optimality conditions for problem (B.24) are

$$\begin{cases} e^{-\frac{u(x)}{\varepsilon}} = \displaystyle\int_{\mathcal{X}} e^{\frac{v(y)-c(x,y)}{\varepsilon}}\, d\beta(y) & (\forall\, x \in \mathrm{supp}(\alpha)) \\[2mm] e^{-\frac{v(y)}{\varepsilon}} = \displaystyle\int_{\mathcal{X}} e^{\frac{u(x)-c(x,y)}{\varepsilon}}\, d\alpha(x) & (\forall\, y \in \mathrm{supp}(\beta)), \end{cases} \tag{B.25}$$

which are equivalent to

$$\begin{cases} g(y)^{-1} = \displaystyle\int_{\mathcal{X}} e^{\frac{-c(x,y)}{\varepsilon}} f(x)\, d\alpha(x) & (\forall\, y \in \mathrm{supp}(\beta)) \\[2mm] f(x)^{-1} = \displaystyle\int_{\mathcal{X}} e^{\frac{-c(x,y)}{\varepsilon}} g(y)\, d\beta(y) & (\forall\, x \in \mathrm{supp}(\alpha)), \end{cases} \tag{B.26}$$

where $f = e^{u/\varepsilon} \in \mathcal{C}_{++}(\mathcal{X})$ and $g = e^{v/\varepsilon} \in \mathcal{C}_{++}(\mathcal{X})$. In the rest of the section we will consider the following *DAD problem* [32, 37]: find $f, g \in \mathcal{C}_{++}(\mathcal{X})$ such that

$$(\forall\, y \in \mathcal{X})\ \int_{\mathcal{X}} f(x)k(x,y)g(y)\, d\alpha(x) = 1 \ \text{ and } \ (\forall\, x \in \mathcal{X})\ \int_{\mathcal{X}} f(x)k(x,y)g(y)\, d\beta(y) = 1, \tag{B.27}$$

where k is defined in (B.1). It is clear that a solution of (B.27) is also a solution of (B.26). However, the vice versa is in general not true, even though there is a canonical way to build solutions of (B.27) starting from solutions of (B.26): indeed if (f, g) is a solution of (B.26), then the functions $\bar{f}, \bar{g} \colon \mathcal{X} \to \mathbb{R}$ defined through $\bar{f}(x)^{-1} = \int_{\mathcal{X}} k(x,y)g(y)\, d\beta(y)$ and $\bar{g}(y)^{-1} = \int_{\mathcal{X}} k(x,y)f(x)\, d\alpha(x)$ provide a solution of (B.27). So, the dual OT_ε problem (B.24) admits a solution if and only if the corresponding DAD problem (B.27) admits a solution. Recalling the definition of A_α in (B.20), problem (B.27) can be more compactly written as

$$f = A_\beta(g) \qquad \text{and} \qquad g = A_\alpha(f), \tag{B.28}$$

or equivalently, by setting $A_{\beta\alpha} = A_\beta \circ A_\alpha$ and $A_{\alpha\beta} = A_\alpha \circ A_\beta$,

$$f = A_{\beta\alpha}(f) \qquad \text{and} \qquad g = A_{\alpha\beta}(g). \tag{B.29}$$

This shows that the solutions of the DAD problem (B.27) are the fixed points of $\mathsf{A}_{\alpha\beta}$ and $\mathsf{A}_{\beta\alpha}$ respectively. Note that the operators $\mathsf{A}_{\beta\alpha}$ and $\mathsf{A}_{\alpha\beta}$ are positively homogeneous, that is, for every $t \in \mathbb{R}_{++}$ and $f \in \mathcal{C}_{++}(\mathcal{X})$, $\mathsf{A}_{\beta\alpha}(tf) = t\mathsf{A}_{\beta\alpha}(f)$ and $\mathsf{A}_{\alpha\beta}(tf) = t\mathsf{A}_{\alpha\beta}(f)$. Thus, if f is a fixed point of $\mathsf{A}_{\beta\alpha}$, then tf is also a fixed point of $\mathsf{A}_{\beta\alpha}$, for every $t > 0$. If (f, g) is a solution of the DAD problem (B.27), then the pair (u, v), with $u = \varepsilon \log f$ and $v = \varepsilon \log g$ is a solution of (B.24). We refer to these solutions as *Sinkhorn potentials* of the pair (α, β). Finally, note that, it follows from (B.25) that solutions of (B.24) are determined (α, β)-a.e. on $\mathcal{X}$ and up to a translation of the form $(u + t, v - t)$, for some $t \in \mathbb{R}$.

The following result is essentially the specialization of [32, Thm. 7.1.4] to the case of the map $\mathsf{A}_{\beta\alpha}$. We report the proof here for completeness and the reader's convenience.

Theorem B.5 (Hilbert's metric contraction for $\mathsf{A}_{\beta\alpha}$). *The map $\mathsf{A}_{\beta\alpha} : \mathcal{C}_{++}(\mathcal{X}) \to \mathcal{C}_{++}(\mathcal{X})$ has a unique fixed point up to positive scalar multiples. Moreover, let $\lambda = \frac{e^{\mathsf{D}/\varepsilon}-1}{e^{\mathsf{D}/\varepsilon}+1}$. Then, for every $f, f' \in \mathcal{C}_{++}(\mathcal{X})$,*

$$d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta\alpha}(f')) \le \lambda^2 \, d_H(f, f'). \tag{B.30}$$

Proof. By combining (B.22) with Theorem B.4 we obtain that, for any $f, f' \in \mathcal{C}_{++}(\mathcal{X})$

$$d_H(\mathsf{A}_\alpha(f), \mathsf{A}_\alpha(f')) = d_H(1/(\mathsf{L}_\alpha f), 1/(\mathsf{L}_\alpha f')) = d_H(\mathsf{L}_\alpha f, \mathsf{L}_\alpha f') \le \lambda \, d_H(f, f'). \tag{B.31}$$

Since the same holds for A_β then (B.30) is satisfied. Now, let $C = \mathcal{C}_{++}(\mathcal{X}) \cap \partial B_1(0)$. Let $\overline{\mathsf{A}}_{\beta\alpha} : C \to C$ be the map such that

$$(\forall f \in C) \qquad \overline{\mathsf{A}}_{\beta\alpha}(f) = \frac{\mathsf{A}_{\beta\alpha}(f)}{\|\mathsf{A}_{\beta\alpha}(f)\|_\infty}. \tag{B.32}$$

Then, since $d_H(sf, tf') = d_H(f, f')$ for any $s, t > 0$ and $f, f' \in C$, we have

$$d_H(\overline{\mathsf{A}}_{\beta\alpha}(f), \overline{\mathsf{A}}_{\beta\alpha}(f')) = d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta\alpha}(f')) \le \lambda^2 \, d_H(f, f'). \tag{B.33}$$

Since (C, d_H) is a complete metric space [36, Theorem 1.2] and $\overline{\mathsf{A}}_{\beta\alpha}$ is a contraction, we can apply Banach's contraction theorem and conclude that there exists a unique fixed point of $\overline{\mathsf{A}}_{\beta\alpha}$, namely a function $\bar{f} \in C$ such that

$$\bar{f} = \overline{\mathsf{A}}_{\beta\alpha}(\bar{f}) = \frac{\mathsf{A}_{\beta\alpha}(\bar{f})}{\|\mathsf{A}_{\beta\alpha}(\bar{f})\|_\infty}. \tag{B.34}$$

Hence $\bar{f}$ is an eigenvector for $\mathsf{A}_{\beta\alpha}$ with eigenvalue $t = \|\mathsf{A}_{\beta\alpha}(\bar{f})\|_\infty > 0$. Now, we note that

$$(\forall \, f, g \in \mathcal{C}_{++}(\mathcal{X})) \quad \langle g\mathsf{L}_\alpha f, \beta \rangle = \langle f\mathsf{L}_\beta g, \alpha \rangle = \int_{\mathcal{X} \times \mathcal{X}} f(x)k(x, y)g(y)d(\alpha \otimes \beta)(x, y). \tag{B.35}$$

Set $\bar{g} = \mathsf{A}_\alpha(\bar{f})$, so that $\mathsf{A}_\beta(\bar{g}) = t\bar{f}$. Then, recalling the definitions of A_α and A_β, we have $\bar{g}\mathsf{L}_\alpha\bar{f} \equiv 1$ and $t^{-1} \equiv \bar{f}\mathsf{L}_\beta\bar{g}$. Hence $t^{-1} = \langle \bar{f}\mathsf{L}_\beta\bar{g}, \alpha \rangle = \langle \bar{g}\mathsf{L}_\alpha\bar{f}, \beta \rangle = 1$. Therefore $\bar{f}$ is a fixed point of $\mathsf{A}_{\beta\alpha}$. Finally, if $\bar{f}' \in \mathcal{C}_{++}(\mathcal{X})$ is a fixed point of $\mathsf{A}_{\beta\alpha}$, then, since $\mathsf{A}_{\beta\alpha}$ is positively homogeneous, we have

$$\overline{\mathsf{A}}_{\beta\alpha}(\bar{f}'/\|\bar{f}'\|_\infty) = \frac{\mathsf{A}_{\beta\alpha}(\bar{f}'/\|\bar{f}'\|_\infty)}{\|\mathsf{A}_{\beta\alpha}(\bar{f}'/\|\bar{f}'\|_\infty)\|_\infty} = \frac{\mathsf{A}_{\beta\alpha}(\bar{f}')}{\|\mathsf{A}_{\beta\alpha}(\bar{f}')\|_\infty} = \frac{\bar{f}'}{\|\bar{f}'\|_\infty}, \tag{B.36}$$

that is, $\bar{f}'/\|\bar{f}'\|_\infty$ is a fixed point of $\overline{\mathsf{A}}_{\beta\alpha}$. Thus, $\bar{f}'/\|\bar{f}'\|_\infty = \bar{f}$ and hence $\bar{f}'$ is a multiple of $\bar{f}$. $\square$

Corollary B.6 (Existence and uniqueness of Sinkhorn potentials). *Let $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$. Then, the DAD problem (B.27) admits a solution (f, g) and every other solution is of type $(tf, t^{-1}g)$, for some $t > 0$. Moreover, there exists a pair $(u, v) \in \mathcal{C}(\mathcal{X})^2$ of Sinkhorn potentials and every other pair of Sinkhorn potentials is of type $(u + s, v - s)$, for some $s \in \mathbb{R}$. In particular, for every $x_o \in \mathcal{X}$, there exist a unique pair (u, v) of Sinkhorn potentials such that $u(x_0) = 0$.*

Proof. It follows from Theorem B.5 and the discussion after (B.29). $\square$

Bounding (f, g) point-wise. We conclude this section by providing additional properties of the solutions (f, g) of the DAD problem (B.28). In particular, we show that there exists one such solution for which it is possible to provide a point-wise upper and lower bound independent on α and β.

Remark B.7. *Let $f \in \mathcal{C}_{++}(\mathcal{X})$ and set $g = A_\alpha(f)$. Then, recalling (B.20) and (B.4), we have that, for every $x \in \mathcal{X}$,*

$$1 = g(x)(L_\alpha\, f)(x) \leq g(x)\,\|L_\alpha f\|_\infty \leq g(x)\,\|f\|_\infty$$

and

$$1 = g(x)(L_\alpha\, f)(x) \geq g(x)(\min_{\mathcal{X}} f)\int k(x,z)\,d\alpha(z) \geq g(x)(\min_{\mathcal{X}} f)e^{-\mathsf{D}/\varepsilon}.$$

Therefore,

$$\min_{\mathcal{X}} g \geq \frac{1}{\|f\|_\infty} \quad and \quad \|g\|_\infty \leq \frac{e^{\mathsf{D}/\varepsilon}}{\min_{\mathcal{X}} f}. \tag{B.37}$$

Lemma B.8. *(Auxiliary Cone) Consider the set*

$$K = \{f \in \mathcal{C}_+(\mathcal{X}) \mid f(x) \leq f(y)\, e^{\mathsf{D}/\varepsilon} \ \forall x,y \in \mathcal{X}\}. \tag{B.38}$$

Let $\alpha \in \mathcal{M}_+^1(\mathcal{X})$. Then the following holds.

(i) K *is a closed convex cone and* $K \subset \mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$;

(ii) $L_\alpha(\mathcal{C}_+(\mathcal{X})) \subset K$;

(iii) $R(K) \subset K$;

(iv) $\mathrm{Ran}(A_\alpha) \subset K$;

(v) *If* $f \in K$ *and* $g = A_\alpha f$, *then* $g \in K$ *and* $1 \leq (\min_{\mathcal{X}} g)\,\|f\|_\infty \leq \|g\|_\infty\,\|f\|_\infty \leq e^{2\mathsf{D}/\varepsilon}$.

(vi) *If* $f \in K$ *is such that* $f(x_o) = 1$ *for some* $x_o \in \mathcal{X}$, *then* $\|\varepsilon \log f\|_\infty \leq \mathsf{D}$.

Proof. (i): We see that for any $f \in K$,

$$\max_{\mathcal{X}} f \leq (\min_{\mathcal{X}} f)\, e^{\mathsf{D}/\varepsilon}, \tag{B.39}$$

so, if $f(x) = 0$ for some $x \in \mathcal{X}$, then $f(x) = 0$ on all $\mathcal{X}$. Hence $K \subseteq \mathcal{C}_{++}(\mathcal{X}) \cup \{0\}$. It is straightforward to verify that K is a convex cone. Moreover K is also closed. Indeed if $(f_n)_{n \in \mathbb{N}}$ is a sequence in K which converges uniformly to $f \in \mathcal{C}(\mathcal{X})$, then, for every $x, y \in \mathcal{X}$ and every $n \in \mathbb{N}$, $f_n(x) \leq f_n(y)e^{\mathsf{D}/\varepsilon}$ and hence, letting $n \to +\infty$, we have $f(x) \leq f(y)e^{\mathsf{D}/\varepsilon}$.

(ii): For every $f \in \mathcal{C}_+(\mathcal{X})$ and $x, y \in \mathcal{X}$, we have

$$\begin{aligned}
(L_\alpha f)(x) &= \int k(x,z)f(z)\,d\alpha(z) \\
&= \int \frac{k(x,z)}{k(y,z)}\, k(y,z)f(z)\,d\alpha(z) \\
&\leq e^{\mathsf{D}/\varepsilon}\int k(y,z)f(z)\,d\alpha(z) \\
&= e^{\mathsf{D}/\varepsilon}(L_\alpha f)(y).
\end{aligned}$$

(iii): For every $f \in K$,

$$(\forall\, x, y \in \mathcal{X}) \qquad f(x) \leq f(y)\, e^{\mathsf{D}/\varepsilon} \ \Leftrightarrow\ \frac{1}{f(y)} \leq \frac{1}{f(x)}\, e^{\mathsf{D}/\varepsilon}.$$

(iv) It follows from (ii) and (iii) and the definitions of A_α.

(v): It follows from (iv), (B.37), and (B.39).

(vi): Let $f \in K$ be such that $f(x_o) = 1$. Then $\min_{\mathcal{X}} f \leq 1 \leq \max_{\mathcal{X}} f$. Thus, it follows from (B.39) that

$$\max_{\mathcal{X}} f \leq e^{\mathsf{D}/\varepsilon} \quad and \quad \min_{\mathcal{X}} f \geq e^{-\mathsf{D}/\varepsilon} \tag{B.40}$$

and hence, for every $x \in \mathcal{X}$, $-\mathsf{D} \leq \varepsilon \log f(x) \leq \mathsf{D}$. $\qquad\square$

As a direct consequence of Lemma B.8 we can establish a uniform point-wise upper and lower bound for the value of DAD solutions.

Corollary B.9. *Let $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$. Let $x_o \in \mathcal{X}$ and let (f, g) be the solution of* (B.28) *such that $f(x_o) = 1$. Then $\|f\|_\infty \leq e^{D/\varepsilon}$ and $\|g\|_\infty \leq e^{2D/\varepsilon}$. Moreover, the corrisponding pair (u, v) of Sinkhorn potentials satifies $\|u\|_\infty \leq D$ and $\|v\|_\infty \leq 2D$.*

Proof. Since f and g are fixed points of $A_{\beta\alpha}$ and $A_{\alpha\beta}$ respectively, it follows from Lemma B.8 (iv) that $f, g \in K$. Then, Lemma B.8 (vi) yields $\|f\|_\infty \leq e^{D/\varepsilon}$, whereas by the second of (B.37) and (B.40) we derive that $\|g\|_\infty \leq e^{2D/\varepsilon}$. $\qquad \square$

B.3 Sinkhorn-Knopp algorithm in infinite dimension

In the context of optimal transport, Sinkhorn-Knopp algorithm is often presented and studied in finite dimension [13, 38]. The algorithm originates from so called *matrix scaling problems*, also called *DAD problems*, which consists in finding, for a given matrix A with nonnegative entries, two diagonal matrices D_1, D_2 such that $D_1 A D_2$ is doubly stochastic [41]. In our setting it is crucial to analyze the algorithm in infinite dimension.

Theorem B.5 shows that $A_{\beta\alpha}$ is a contraction with respect to the Hilbert's metric. This suggests a direct approach to find the solutions of the DAD problem by adopting a fixed-point strategy, which turns out to applying the operators A_α and A_β alternatively, starting from some $f^{(0)} \in \mathcal{C}_{++}(\mathcal{X})$. This is exactly the approach to the Sinkhorn algorithm pioneered by [22, 34] and further developed in an infinite dimensional setting in [37]. In this section we review the algorithm and give the convergence properties for the special kernel k in (B.1). In particular we provide rate of convergence in the sup norm $\|\cdot\|_\infty$.

Algorithm B.1 Sinkhorn-Knopp algorithm (infinite dimensional case)

Let $\alpha, \beta \in \mathcal{M}_+^1(\mathcal{X})$. Let $f^{(0)} \in \mathcal{C}_{++}(\mathcal{X})$ and define,

$$
\begin{aligned}
&\text{for } \ell = 0, 1, \dots \\
&\quad \left|
\begin{aligned}
g^{(\ell+1)} &= A_\alpha(f^{(\ell)}) \\
f^{(\ell+1)} &= A_\beta(g^{(\ell+1)})
\end{aligned}
\right.
\end{aligned}
$$

Theorem B.10 (Convergence of Sinkhorn-Knopp algorithm). *Let $(f^{(\ell)})_{\ell \in \mathbb{N}}$ be defined according to Algorithm B.1. Let $x_o \in \mathcal{X}$ and let (f, g) be the solution of the DAD problem* (B.26) *such that $f(x_o) = 1$. Then, defining λ according to Theorem B.5 and, for every $\ell \in \mathbb{N}$, $\tilde{f}^{(\ell)} = f^{(\ell)}/f^{(\ell)}(x_o)$ and $\tilde{g}^{(\ell+1)} = g^{(\ell+1)} f^{(\ell)}(x_o)$, we have*

$$
\begin{cases}
\|\log \tilde{f}^{(\ell)} - \log f\|_\infty \leq \lambda^{2\ell} \left(\dfrac{D}{\varepsilon} + \log \dfrac{\|f^{(0)}\|_\infty}{\min_{\mathcal{X}} f^{(0)}} \right) \\[2ex]
\|\log \tilde{g}^{(\ell+1)} - \log g\|_\infty \leq e^{3D/\varepsilon} \|\log \tilde{f}^{(\ell)} - \log f\|_\infty.
\end{cases}
\tag{B.41}
$$

Moreover, let the potentials $(u, v) = (\varepsilon \log f, \varepsilon \log g)$ and, for every $\ell \in \mathbb{N}$, $(\tilde{u}^{(\ell)}, \tilde{v}^{(\ell)}) = (\varepsilon \log \tilde{f}^{(\ell)}, \varepsilon \log \tilde{g}^{(\ell)})$. Then we have

$$
\|\tilde{u}^{(\ell)} - u\|_\infty \leq \lambda^{2\ell} \left(\frac{D + \max_{\mathcal{X}} u^{(0)} - \min_{\mathcal{X}} u^{(0)}}{\varepsilon} \right).
\tag{B.42}
$$

Proof. Let $\mathcal{A}$ be the set in Lemma C.1. Clearly, for every $\ell \in \mathbb{N}$, we have $f^{(\ell+1)} = A_{\beta\alpha}(f^{(\ell)})$ and $\bar{f}, \tilde{f}^\ell \in \mathcal{A}$. Thus, it follows from Theorem B.5 and (C.2) in Lemma C.1 that, for every $\ell \in \mathbb{N}$,

$$
\|\log \tilde{f}^{(\ell)} - \log f\|_\infty \leq d_H(\tilde{f}^\ell, f) = d_H(A_{\beta\alpha}^{(\ell)}(f^{(0)}), f) \leq \lambda^{2\ell} d_H(f^{(0)}, f).
$$

Moreover, recalling (B.11), we have

$$
d_H(f^{(0)}, f) = d_H(1/f^{(0)}, L_\beta g) = \log \max_{x, y \in \mathcal{X}} \frac{f^{(0)}(y) L_\beta g(y)}{f^{(0)}(x) L_\beta g(x)} \leq \log \left[e^{D/\varepsilon} \max_{x, y \in \mathcal{X}} \frac{f^{(0)}(y)}{f^{(0)}(x)} \right]
$$

where we used the fact that $\mathsf{L}_\beta(\mathcal{C}_{++}(\mathcal{X})) \subset K$ and the definition (B.38). Thus, the first inequality in (B.41) follows. The second inequality in (B.41) and (B.42) follow directly from Lemma C.3 and the fact that $u^{(0)} = \varepsilon \log f^{(0)}$. $\qquad\square$

Algorithm B.2 Sinkhorn-Knopp algorithm (finite dimensional case)

Let $\mathsf{M} \in \mathbb{R}_{++}^{n_1 \times n_2}$, $\mathsf{a} \in \mathbb{R}_+^{n_1}$, with $\mathsf{a}^\top 1_{n_1} = 1$, and $\mathsf{b} \in \mathbb{R}_+^{n_2}$, with $\mathsf{b}^\top 1_{n_2} = 1$. Let $\mathsf{f}^{(0)} \in \mathbb{R}_{++}^{n_1}$ and define

$$\text{for } \ell = 0, 1, \dots$$
$$\left\lfloor \begin{aligned} \mathsf{g}^{(\ell+1)} &= \frac{\mathsf{b}}{\mathsf{M}^\top \mathsf{f}^{(\ell)}} \\ \mathsf{f}^{(\ell+1)} &= \frac{\mathsf{a}}{\mathsf{M}\mathsf{g}^{(\ell+1)}}. \end{aligned} \right.$$

Proposition B.11. *Suppose that α and β are probability measures with finite support. Then Algorithm B.1 can be reduced to the finite dimensional Algorithm B.2. More specifically, suppose that $\alpha = \sum_{i=1}^{n_1} a_i \delta_{x_i}$, and $\beta = \sum_{i=1}^{n_2} b_i \delta_{y_i}$, where $\mathsf{a} = (a_i)_{1 \le i \le n_1} \in \mathbb{R}_+^{n_1}$, $\sum_{i=1}^n a_i = 1$ and $\mathsf{b} = (b_i)_{1 \le i \le n_2} \in \mathbb{R}_+^{n_2}$, $\sum_{i=1}^n b_i = 1$. Let $\mathsf{K} \in \mathbb{R}^{n_1 \times n_2}$ be such that $\mathsf{K}_{i_1, i_2} = \mathsf{k}(x_{i_1}, y_{i_2})$ and let $\mathsf{M} = \mathrm{diag}(\mathsf{a})\mathsf{K}\mathrm{diag}(\mathsf{b}) \in \mathbb{R}^{n_1 \times n_2}$. Let $(\mathsf{f}^{(\ell)})_{\ell \in \mathbb{N}}$ and $(f^{(\ell)})_{\ell \in \mathbb{N}}$ be defined according to Algorithm B.2 and Algorithm B.1 respectively, with $\mathsf{f}^{(0)} = (f^{(0)}(x_i))_{1 \le i \le n_1}$. Then, for every $\ell \in \mathbb{N}$,*

$$(\forall x \in \mathcal{X})(\forall y \in \mathcal{X})\, g^{(\ell+1)}(y)^{-1} = \sum_{i_1=1}^{n_1} k(x_{i_1}, y) a_{i_1} \mathsf{f}_{i_1}^{(\ell)} \text{ and } f^{(\ell+1)}(x)^{-1} = \sum_{i_2=1}^{n_2} k(x, y_{i_2}) b_{i_2} \mathsf{g}_{i_2}^{(\ell+1)}.$$

Moreover, setting $u^{(\ell)} = \varepsilon \log f^{(\ell)}$, $v^{(\ell)} = \varepsilon \log g^{(\ell)}$, $\mathsf{u}^{(\ell)} = \varepsilon \log \mathsf{f}^{(\ell)}$, and $\mathsf{v}^{(\ell)} = \varepsilon \log \mathsf{g}^{(\ell)}$, we have

$$\left\{ \begin{aligned} (\forall y \in \mathcal{X}) \quad v^{(\ell+1)}(y) &= -\varepsilon \log \sum_{i_1=1}^{n_1} \exp(\mathsf{u}_{i_1}^{(\ell)} - \mathsf{c}(x_{i_1}, y)) a_{i_1} \\ (\forall x \in \mathcal{X}) \quad u^{(\ell+1)}(x) &= -\varepsilon \log \sum_{i_2=1}^{n_2} \exp(\mathsf{v}_{i_2}^{(\ell+1)} - \mathsf{c}(x, y_{i_2})) b_{i_2}. \end{aligned} \right. \tag{B.43}$$

Proof. Since α and β have finite support, we derive from the definitions of $f^{(\ell+1)}$ and $g^{(\ell+1)}$ in Algorithm B.1 and that of A_α and A_β that

$$\left\{ \begin{aligned} (\forall x \in \mathcal{X}) \quad g^{(\ell+1)}(y)^{-1} &= (\mathsf{L}_\alpha f^{(\ell)})(y) = \sum_{i_1=1}^{n_1} a_{i_1} \mathsf{k}(x_{i_1}, y) f^{(\ell)}(x_{i_1}) \\ (\forall y \in \mathcal{X}) \quad f^{(\ell+1)}(x)^{-1} &= (\mathsf{L}_\beta g^{(\ell+1)})(x) = \sum_{i_2=1}^{n_2} \mathsf{k}(x, y_{i_2}) b_{i_2} g^{(\ell+1)}(y_{i_2}). \end{aligned} \right.$$

Now, multiplying the above equations by b_{i_2} and a_{i_1} respectively, and recalling that $\mathsf{M}_{i_1, i_2} = a_{i_1} \mathsf{k}(x_{i_1}, y_{i_2}) b_{i_2}$, we have

$$\begin{bmatrix} b_1 g^{(\ell+1)}(y_1)^{-1} \\ \vdots \\ b_{n_2} g^{(\ell+1)}(y_{n_2})^{-1} \end{bmatrix} = \mathsf{M}^\top \begin{bmatrix} f^{(\ell)}(x_1) \\ \vdots \\ f^{(\ell)}(x_{n_1}) \end{bmatrix}, \quad \begin{bmatrix} a_1 f^{(\ell+1)}(x_1)^{-1} \\ \vdots \\ a_{n_1} f^{(\ell+1)}(x_{n_1})^{-1} \end{bmatrix} = \mathsf{M} \begin{bmatrix} g^{(\ell+1)}(y_1) \\ \vdots \\ g^{(\ell+1)}(y_{n_2}) \end{bmatrix},$$

and hence

$$\begin{bmatrix} g^{(\ell+1)}(y_1) \\ \vdots \\ g^{(\ell+1)}(y_{n_2}) \end{bmatrix} = \mathsf{b} \Big/ \mathsf{M}^\top \begin{bmatrix} f^{(\ell)}(x_1) \\ \vdots \\ f^{(\ell)}(x_{n_1}) \end{bmatrix}, \quad \begin{bmatrix} f^{(\ell+1)}(x_1) \\ \vdots \\ f^{(\ell+1)}(x_{n_1}) \end{bmatrix} = \mathsf{a} \Big/ \mathsf{M} \begin{bmatrix} g^{(\ell+1)}(y_1) \\ \vdots \\ g^{(\ell+1)}(y_{n_2}) \end{bmatrix}.$$

Therefore, since $\mathsf{f}^{(0)} = (f^{(0)}(x_i))_{1 \le i \le n_1}$, recalling Algorithm B.2, it follows by induction that, for every $\ell \in \mathbb{N}$, $\mathsf{f}^{(\ell)} = (f^{(\ell)}(x_i))_{1 \le i \le n_1}$ and $\mathsf{g}^{(\ell)} = (g^{(\ell)}(x_i))_{1 \le i \le n_1}$. Thus, the first part of the statement follows. The second part follows directly from the definitions of $u^{(\ell)}$, $v^{(\ell)}$, $\mathsf{u}^{(\ell)}$, and $\mathsf{v}^{(\ell)}$. $\qquad\square$

Remark B.12.

   (i) Algorithm B.2 is the classical (discrete) Sinkhorn algorithm which was recently studied in several papers [13]. It follows from Theorem B.10 that considering the solution (f, g) of the DAD problem such that $f(x_1) = 1$ and defining $\tilde{\mathsf{f}}^{(\ell)} = \mathsf{f}^{(\ell)}/\mathsf{f}_0^{(\ell)}$ and $\tilde{\mathsf{g}}^{(\ell)} = \mathsf{g}^{(\ell)}\mathsf{f}_0^{(\ell)}$, and $\mathsf{f}_i = f(x_i)$ and $\mathsf{g}_j = g(y_j)$, we have

$$\|\log \tilde{\mathsf{f}}^{(\ell)} - \log \mathsf{f}\|_\infty \leq \lambda^{2\ell}\left(\frac{\mathsf{D}}{\varepsilon} + \log \frac{\max_i \mathsf{f}_i^{(0)}}{\min_i \mathsf{f}_i^{(0)}}\right).$$

  (ii) The procedure SINKHORNKNOPP discussed in the paper and called in Algorithm 2, actually output the vector $\mathsf{v} = \varepsilon \log \mathsf{g}^{(\ell)}$ for sufficiently large ℓ.

 (iii) Referring to Section 4 in the paper, we recognize that the expressions on the right hand side of (B.43) are precisely $\mathsf{T}_\alpha(u^{(\ell)})(x)$ and $\mathsf{T}_\beta(v^{(\ell+1)})(x)$ respectively.

C   Lipschitz continuity of the gradient of Sinkhorn divergence with respect to the Total Variation

In this section we show that the gradient of the Sinkhorn divergence is Lipschitz continuous with respect to the Total Variation on $\mathcal{M}_1^+(\mathcal{X})$.

We start by characterizing the relation between the Hilbert's metric and the metric induced by the norm $\|\cdot\|_\infty$.

Lemma C.1. *Let $f, f' \in \mathcal{C}_{++}(\mathcal{X})$ and set $u = \varepsilon \log f$ and $u' = \varepsilon \log f'$. Then*

$$d_H(f, f') \leq 2\|\log f - \log f'\|_\infty \quad \text{or, equivalently} \quad d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \frac{2}{\varepsilon}\|u - u'\|_\infty. \quad \text{(C.1)}$$

Moreover, let $x_o \in \mathcal{X}$, consider the sets $\mathcal{A} = \{h \in \mathcal{C}_{++}(\mathcal{X}) \mid h(x_o) = 1\}$ and $\mathcal{B} = \{w \in \mathcal{C}(\mathcal{X}) \mid w(x_o) = 0\}$. Suppose that $f, f' \in \mathcal{A}$ (or equivalently that $u, u' \in \mathcal{B}$). Then

$$\frac{1}{2}d_H(f, f') \leq \|\log f - \log f'\|_\infty \leq d_H(f, f'). \quad \text{(C.2)}$$

and

$$\frac{\varepsilon}{2}d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}) \leq \|u - u'\|_\infty \leq \varepsilon\, d_H(e^{u/\varepsilon}, e^{u'/\varepsilon}). \quad \text{(C.3)}$$

Proof. We have

$$\begin{aligned}
d_H(f, f') &= \log \max_{x,y\in\mathcal{X}} \frac{f(x)f'(y)}{f(y)f'(x)} \\
&= \log \max_{x\in\mathcal{X}} \frac{f(x)}{f'(x)} + \log \max_{y\in\mathcal{X}} \frac{f'(y)}{f(y)} \\
&= \max_{x\in\mathcal{X}} \log \frac{f(x)}{f'(x)} + \max_{y\in\mathcal{X}} \log \frac{f'(y)}{f(y)} \\
&\leq 2\max_{x\in\mathcal{X}} \left|\log \frac{f(x)}{f'(x)}\right| \\
&= 2\|\log(f/f')\|_\infty \\
&= 2\|\log f - \log f'\|_\infty
\end{aligned}$$

and (C.1) follows. Suppose that $f, f' \in \mathcal{A}$. Then

$$\begin{aligned}
\|\log f - \log f'\|_\infty &= \max\left\{\log \max_{x\in\mathcal{X}} \frac{f(x)}{f'(x)}, \log \max_{x\in\mathcal{X}} \frac{f'(x)}{f(x)}\right\} \\
&= \max\left\{\log \max_{x\in\mathcal{X}} \frac{f(x)f'(\bar{x})}{f(\bar{x})f'(x)}, \log \max_{x\in\mathcal{X}} \frac{f(\bar{x})f'(x)}{f(x)f'(\bar{x})}\right\} \\
&\leq \max\left\{\log \max_{x,y\in\mathcal{X}} \frac{f(x)f'(y)}{f(y)f'(x)}, \log \max_{x,y\in\mathcal{X}} \frac{f(y)f'(x)}{f(x)f'(y)}\right\} \\
&= d_H(f, f'),
\end{aligned}$$

26

since $f(x_o)/f'(x_o) = f'(x_o)/f(x_o) = 1$. Therefore, (C.2) follows. $\qquad\square$

Lemma C.2. *For every $x, y \in \mathbb{R}_{++}$ we have*

$$|\log x - \log y| \le \max\left\{x^{-1}, y^{-1}\right\}|x - y|. \tag{C.4}$$

The following result allows to extend the previous observations on a pair f, f' to the corresponding $g = \mathsf{A}_\alpha f$ and $g' = \mathsf{A}_\alpha f'$.

Lemma C.3. *Let $x_o \in \mathcal{X}$ and $K \subset \mathcal{C}_+(\mathcal{X})$ the cone from Lemma B.8. Let $f, f' \in K$ be such that $f(x_o) = f'(x_o) = 1$, and set $g = \mathsf{A}_\alpha f$ and $g' = \mathsf{A}_\alpha f'$. Then,*

$$\|\log g - \log g'\|_\infty \le e^{3\mathsf{D}/\varepsilon} \|\log f - \log f'\|_\infty. \tag{C.5}$$

Proof. It follows from (B.20) and Lemma C.2 that

$$|\log g - \log g'| = \left|\log \frac{g}{g'}\right| = \left|\log \frac{\mathsf{L}_\alpha f'}{\mathsf{L}_\alpha f}\right| \le \max\{g', g\}|\mathsf{L}_\alpha f - \mathsf{L}_\alpha f'|.$$

Therefore, since $1 \le \|f\|_\infty, \|f'\|_\infty$, and recalling Lemma B.8 (v) and (B.4), we have

$$\begin{aligned}
\|\log g - \log g'\|_\infty &\le \max\{\|g\|_\infty, \|g'\|_\infty\} \|\mathsf{L}_\alpha f - \mathsf{L}_\alpha f'\|_\infty \\
&\le \max\{\|f\|_\infty \|g\|_\infty, \|f'\|_\infty \|g'\|_\infty\} \|\mathsf{L}_\alpha f - \mathsf{L}_\alpha f'\|_\infty \\
&\le e^{2\mathsf{D}/\varepsilon} \|f - f'\|_\infty \\
&= e^{2\mathsf{D}/\varepsilon}\|e^{\log f} - e^{\log f'}\|_\infty.
\end{aligned}$$

Now, since $f, f' \le e^{\mathsf{D}/\varepsilon}$, we have $\log f, \log f' \le \mathsf{D}/\varepsilon$. Thus, the statement follows by noting that the exponential function is Lipschitz continuous on $]-\infty, \mathsf{D}/\varepsilon]$ with constant $e^{\mathsf{D}/\varepsilon}$. $\qquad\square$

We are ready to prove the main result of the section.

Theorem C.4 (Lipschitz continuity of the Sinkhorn potentials with respect to the total variation). *Let $\alpha, \beta, \alpha', \beta' \in \mathcal{M}_1^+(\mathcal{X})$ and let $x_o \in \mathcal{X}$. Let $(u, v), (u', v') \in \mathcal{C}(\mathcal{X})^2$ be the two pairs of Sinkhorn potentials corresponding to the solution of the regularized OT problem in (B.24) for (α, β) and (α', β') respectively such that $u(x_o) = u'(x_o) = 0$. Then*

$$\|u - u'\|_\infty \le 2\varepsilon e^{3\mathsf{D}/\varepsilon} \|(\alpha - \alpha', \beta - \beta')\|_{TV}. \tag{C.6}$$

Hence, the map which, for each pair of probability distributions $(\alpha, \beta) \in \mathcal{M}_1^+(\mathcal{X})^2$ associates the component u of the corresponding Sinkhorn potentials is $2\varepsilon e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous with respect to the total variation.

Proof. The functions $f = e^{u/\varepsilon}$ and $f' = e^{u'/\varepsilon}$ are fixed points of the maps $\mathsf{A}_{\beta\alpha}$ and $\mathsf{A}_{\beta'\alpha'}$ respectively. Then, it follows from Theorem B.5 that

$$\begin{aligned}
d_H(f, f') = d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f')) &\le d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) + d_H(\mathsf{A}_{\beta'\alpha'}(f), \mathsf{A}_{\beta'\alpha'}(f')) \\
&\le d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) + \lambda^2 d_H(f, f'),
\end{aligned}$$

hence,

$$d_H(f, f') \le \frac{1}{1 - \lambda^2} d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)). \tag{C.7}$$

Moreover, using (C.1), we have

$$\begin{aligned}
d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) &\le d_H(\mathsf{A}_{\beta\alpha}(f), \mathsf{A}_{\beta'\alpha}(f)) + d_H(\mathsf{A}_{\beta'\alpha}(f), \mathsf{A}_{\beta'\alpha'}(f)) \\
&\le d_H(\mathsf{A}_\beta(g), \mathsf{A}_{\beta'}(g)) + \lambda d_H(\mathsf{A}_\alpha(f), \mathsf{A}_{\alpha'}(f)) \\
&\le 2\left\|\log \frac{\mathsf{A}_\beta(g)}{\mathsf{A}_{\beta'}(g)}\right\|_\infty + 2\lambda \left\|\log \frac{\mathsf{A}_\alpha(f)}{\mathsf{A}_{\alpha'}(f)}\right\|_\infty. \tag{C.8}
\end{aligned}$$

Now, note that by Lemma C.2

$$\left| \log \frac{\mathsf{A}_\beta(g)}{\mathsf{A}_{\beta'}(g)} \right| = \left| \log \frac{\mathsf{L}_{\beta'} g}{\mathsf{L}_\beta g} \right| \leq \max\{1/\mathsf{L}_\beta g, 1/\mathsf{L}_{\beta'} g\} |(\mathsf{L}_{\beta'} - \mathsf{L}_\beta) g| \tag{C.9}$$

and that, for every $x \in \mathcal{X}$,

$$[(\mathsf{L}_{\beta'} - \mathsf{L}_\beta) g](x) = \int \mathsf{k}(x, z) g(z) \, d(\beta - \beta')(z) \tag{C.10}$$
$$= \langle \mathsf{k}(x, \cdot) g, \beta - \beta' \rangle \leq \|g\|_\infty \|\beta - \beta'\|_{TV},$$

and, similarly, $[(\mathsf{L}_\beta - \mathsf{L}_{\beta'}) g](x) \leq \|g\|_\infty \|\beta - \beta'\|_{TV}$. Therefore, since $1/(\mathsf{L}_\beta g) = \mathsf{A}_\beta(g) = f$ and $\mathsf{L}_{\beta'} g \geq e^{-\mathsf{D}/\varepsilon} \min g$, it follows from Lemma B.8 (v) and (B.39) (applied to g) that

$$\left\| \log \frac{\mathsf{A}_\beta(g)}{\mathsf{A}_{\beta'}(g)} \right\|_\infty \leq \max \left\{ \|f\|_\infty, \frac{e^{\mathsf{D}/\varepsilon}}{\min g} \right\} \|g\|_\infty \|\beta - \beta'\|_{TV} \leq e^{2\mathsf{D}/\varepsilon} \|\beta - \beta'\|_{TV}. \tag{C.11}$$

Analogously, we have

$$\left\| \log \frac{\mathsf{A}_\alpha(f)}{\mathsf{A}_{\alpha'}(f)} \right\|_\infty \leq e^{2\mathsf{D}/\varepsilon} \|\alpha - \alpha'\|_{TV}. \tag{C.12}$$

Putting (C.7), (C.8), (C.11), and (C.12) together, we have

$$d_H(f, f') \leq \frac{2e^{2\mathsf{D}/\varepsilon}}{1 - \lambda^2} \left(\lambda \|\alpha - \alpha'\|_{TV} + \|\beta - \beta'\|_{TV} \right). \tag{C.13}$$

Now, note that since $e^{\mathsf{D}/\varepsilon} \geq 1$

$$\frac{1}{1 - \lambda^2} = \frac{(e^{\mathsf{D}/\varepsilon} + 1)^2}{4e^{\mathsf{D}/\varepsilon}} \leq e^{\mathsf{D}/\varepsilon}. \tag{C.14}$$

Finally, recalling (C.3), we have

$$\|u - u'\|_\infty \leq 2\varepsilon e^{3\mathsf{D}/\varepsilon} \|(\alpha - \alpha', \beta - \beta')\|_{TV}, \tag{C.15}$$

where $\|(\alpha - \alpha', \beta - \beta')\|_{TV} = \|\alpha - \alpha'\|_{TV} + \|\beta - \beta'\|_{TV}$ is the total variation norm on $\mathcal{M}(\mathcal{X})^2$. $\qquad\square$

Corollary C.5. *Under the assumption of Theorem C.4, we have*

$$\|u - u'\|_\infty + \|v - v'\|_\infty \leq 2\varepsilon e^{3\mathsf{D}/\varepsilon}(1 + \varepsilon e^{3\mathsf{D}/\varepsilon}) \|(\alpha - \alpha', \beta - \beta')\|_{TV}. \tag{C.16}$$

Proof. It follows from Theorem C.4 and Lemma C.3. $\qquad\square$

We finally address the issue of the differentiability of the Sinkhorn divergence. We first recall a few facts about the directional differentiability of OT_ε briefly recalled in Section 2 of the main text. For a more in-depth analysis on this topic we refer to [21, Proposition 2]. See also Proposition C.10.

Fact C.6. *Let $x_o \in \mathcal{X}$, $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ and $(u, v) \in \mathcal{C}(\mathcal{X})^2$ be the pair of corresponding Sinkhorn potentials with $u(x_o) = 0$. The function OT_ε is directionally differentiable and the directional derivative of OT_ε in (α, β) along a feasible direction $(\mu, \nu) \in \mathcal{F}_{\mathcal{M}_1^+(\mathcal{X})^2}\big((\alpha, \beta)\big)$ (see Definition A.2) is*

$$\mathrm{OT}_\varepsilon'(\alpha, \beta; \mu, \nu) = \int u(x) \, d\mu(x) + \int v(y) \, d\nu(y) = \langle (u, v), (\mu, \nu) \rangle. \tag{C.17}$$

Let $\nabla \mathrm{OT}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})^2$ be the operator that maps every pair of probability distributions $(\alpha, \beta) \in \mathcal{M}_1^+(\mathcal{X})^2$ to the corresponding pair of Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ with $u(x_o) = 0$. Then (C.17) can be written as

$$\mathrm{OT}_\varepsilon'(\alpha, \beta; \mu, \nu) = \langle \nabla \mathrm{OT}_\varepsilon(\alpha, \beta), (\mu, \nu) \rangle. \tag{C.18}$$

Remark C.7. In Fact C.6, the requirement $u(x_o) = 0$ is only a convention to remove ambiguities. Indeed, for every $t \in \mathbb{R}$, replacing the Sinkhorn potential $(u + t, u - t)$ in Definition A.1 does not affect (C.17).

Fact C.8. *Let $\beta \in \mathcal{M}_+^1(\mathcal{X})$ and let $\nabla_1 \mathrm{OT}_\varepsilon$ be the first component of the gradient operator defined in Fact C.6. Then the Sinkhorn divergence function $S_\varepsilon(\cdot, \beta) \colon \mathcal{M}_+^1(\mathcal{X}) \to \mathbb{R}$ in (7) is directionally differentiable and, for every $\alpha \in \mathcal{M}_+^1(\mathcal{X})$ and every $\mu \in \mathcal{F}_{\mathcal{M}_+^1(\mathcal{X})}(\alpha)$,*

$$[S_\varepsilon(\cdot, \beta)]'(\alpha; \mu) = \langle \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \beta) - \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \alpha), \mu \rangle.$$

So, one can define $\nabla S_\varepsilon(\cdot, \beta) \colon \mathcal{M}_1^+(\mathcal{X}) \to \mathcal{C}(\mathcal{X})$ such that, for every $\alpha \in \mathcal{M}_+^1(\mathcal{X})$, $\nabla[S_\varepsilon(\cdot, \beta)](\alpha) = \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \beta) - \nabla_1 \mathrm{OT}_\varepsilon(\alpha, \alpha)$ and we have

$$[S_\varepsilon(\cdot, \beta)]'(\alpha; \mu) = \langle \nabla S_\varepsilon(\cdot, \beta), \mu \rangle. \tag{C.19}$$

Finally, if k in (B.1) is a positive definite kernel, then the Sinkhorn divergence $S_\varepsilon(\cdot, \beta)$ is convex.

We are now ready to prove Theorem 4 in the paper. We recall also the statement for reader's convenience.

Theorem 4. *The gradient $\nabla \mathrm{OT}_\varepsilon$ defined in Proposition 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 $\nabla S_\varepsilon(\cdot, \beta)$ is $6\varepsilon e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous. The same holds for ∇B_ε.

Proof. The first part is just a consequence of Theorem C.4 and Fact C.6. The second part, follows from the first part and Fact C.8. $\qquad\square$

Remark C.9. It follows from the optimality conditions (B.25) that, for every $x \in \mathrm{supp}(\alpha)$ and $y \in \mathrm{supp}(\beta)$,

$$1 = \int_\mathcal{X} e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}} d\beta(y) \quad \text{and} \quad 1 = \int_\mathcal{X} e^{\frac{u(x)+v(y)-\mathsf{c}(x,y)}{\varepsilon}} d\alpha(x),$$

hence,

$$\int_\mathcal{X} e^{\frac{u \oplus v - \mathsf{c}}{\varepsilon}} d\alpha \otimes \beta = 1. \tag{C.20}$$

Then, recalling the definition of OT_ε in (2) and that of its gradient, given above, we have

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \langle \nabla \mathrm{OT}_\varepsilon(\alpha, \beta), (\alpha, \beta) \rangle - \varepsilon. \tag{C.21}$$

Since, $\nabla \mathrm{OT}_\varepsilon$ is bounded and Lipschitz continuous, it follows that OT_ε is Lipschitz continuous with respect to the total variation.

We end the section by providing an independent proof of Fact C.6, which is based on Proposition A.8 and Corollary C.5.

Proposition C.10. *The function $\mathrm{OT}_\varepsilon \colon \mathcal{M}_+^1(\mathcal{X})^2 \to \mathbb{R}$, defined in (2), is continuous with respect to the total variation, directionally differentiable, and, for every $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X})^2$ and every feasible direction $(\mu, \nu) \in \mathcal{F}_{\mathcal{M}_+^1(\mathcal{X})^2}(\alpha, \beta)$, we have*

$$\mathrm{OT}_\varepsilon'(\alpha, \beta; \mu, \nu) = \langle (u, v), (\mu, \nu) \rangle, \tag{C.22}$$

where $(u, v) \in \mathcal{C}(\mathcal{X})^2$ is any solution of problem (2).

Proof. Let $g \colon \mathcal{C}(\mathcal{X})^2 \times \mathcal{M}(\mathcal{X})^2 \to \mathbb{R}$ be such that,

$$g((u, v), (\alpha, \beta)) = \langle u, \alpha \rangle + \langle v, \beta \rangle - \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \alpha \otimes \beta \rangle. \tag{C.23}$$

Then, for every $(\alpha, \beta) \in \mathcal{M}_+^1(\mathcal{X})^2$,

$$\mathrm{OT}_\varepsilon(\alpha, \beta) = \max_{(u,v) \in \mathcal{C}(\mathcal{X})^2} g((u, v), (\alpha, \beta)). \tag{C.24}$$

Thus, OT_ε is of the type considered in Proposition A.8. Let $(u, v) \in \mathcal{C}(\mathcal{X})$. Then the function $g((u, v), \cdot)$ admits directional derivatives and, for every $(\alpha, \beta), (\mu, \nu) \in \mathcal{M}(\mathcal{X})^2$, we have

$$[g((u, v), \cdot)]'((\alpha, \beta); (\mu, \nu))$$
$$= \left\langle u - \varepsilon e^{\frac{u}{\varepsilon}} \int_{\mathcal{X}} e^{\frac{v - c(\cdot, y)}{\varepsilon}} d\beta(y), \mu \right\rangle + \left\langle v - \varepsilon e^{\frac{v}{\varepsilon}} \int_{\mathcal{X}} e^{\frac{u - c(x, \cdot)}{\varepsilon}} d\alpha(x), \nu \right\rangle. \quad \text{(C.25)}$$

Indeed, for every $t > 0$,

$$\frac{1}{t} \left[g((u, v), (\alpha, \beta) + t(\mu, \nu)) - g((u, v), (\alpha, \beta)) \right]$$
$$= \frac{1}{t} \big[\langle u, \alpha + t\mu \rangle + \langle v, \beta + t\nu \rangle - \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), (\alpha + t\mu) \otimes (\beta + t\nu) \rangle$$
$$- \langle u, \alpha \rangle - \langle v, \beta \rangle + \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \alpha \otimes \beta \rangle \big]$$
$$= \langle u, \mu \rangle + \langle v, \nu \rangle - \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \alpha \otimes \nu \rangle - \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \mu \otimes \beta \rangle$$
$$- t\varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \mu \otimes \nu \rangle,$$

hence

$$[g((u, v), \cdot)]'((\alpha, \beta); (\mu, \nu))$$
$$= \langle u, \mu \rangle + \langle v, \nu \rangle - \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \alpha \otimes \nu \rangle - \varepsilon \langle \exp((u \oplus v - \mathsf{c})/\varepsilon), \mu \otimes \beta \rangle$$

and (C.25) follows. Thus, the function g is Gâteaux differentiable with respect to the second variable, with derivative

$$D_2 g((u, v), (\alpha, \beta)) = \left(u - \varepsilon e^{\frac{u}{\varepsilon}} \int_{\mathcal{X}} e^{\frac{v - c(\cdot, y)}{\varepsilon}} d\beta(y), v - \varepsilon e^{\frac{v}{\varepsilon}} \int_{\mathcal{X}} e^{\frac{u - c(x, \cdot)}{\varepsilon}} d\alpha(x) \right)$$
$$= (u, v) - \varepsilon (e^{\frac{u}{\varepsilon}} \mathsf{L}_\beta e^{\frac{v}{\varepsilon}}, e^{\frac{v}{\varepsilon}} \mathsf{L}_\alpha e^{\frac{u}{\varepsilon}}) \in \mathcal{C}(\mathcal{X})^2,$$

which is jointly continuous, since the maps $(u, \alpha) \mapsto \mathsf{L}_\alpha e^{u/\varepsilon}$ and $(v, \beta) \mapsto \mathsf{L}_\beta e^{v/\varepsilon}$ are continuous. Moreover, it follows from Corollary C.5 that there exists a continuous selection of Sinkhorn potentials. Therefore, it follows from Proposition A.8 that OT_ε is directionally differentiable and

$$OT_\varepsilon'((\alpha, \beta); (\mu, \nu)) = \max_{(u, v) \text{ solution of (C.24)}} \langle D_2 g((u, v), (\alpha, \beta)), (\mu, \nu) \rangle. \quad \text{(C.26)}$$

However, if (u, v) is a solution of (C.24), it follows from the optimality conditions (B.25) that

$$e^{\frac{u}{\varepsilon}} \int_{\mathcal{X}} e^{\frac{v - c(\cdot, y)}{\varepsilon}} d\beta(y) = 1 \quad \text{and} \quad e^{\frac{v}{\varepsilon}} \int_{\mathcal{X}} e^{\frac{u - c(x, \cdot)}{\varepsilon}} d\alpha(x) = 1, \quad \text{(C.27)}$$

hence

$$\langle D_2 g((u, v), (\alpha, \beta)), (\mu, \nu) \rangle = \langle (u - \varepsilon, v - \varepsilon), (\mu, \nu) \rangle = \langle (u, v), (\mu, \nu) \rangle, \quad \text{(C.28)}$$

where we used the fact that, since $(\mu, \nu) = t(\mu_1 - \mu_2, \nu_1 - \nu_2)$ for some $t > 0$ and $\mu_1, \mu_2, \nu_1, \nu_2 \in \mathcal{M}_+^1(\mathcal{X})$, we have $\langle 1, \mu \rangle = t \langle 1, \mu_1 - \mu_2 \rangle = 0$ and $\langle 1, \nu \rangle = t \langle 1, \nu_1 - \nu_2 \rangle = 0$. $\qquad\square$

D The Frank-Wolfe algorithm for Sinkhorn barycenters

In this section we finally analyze the Frank-Wolfe algorithm for the Sinkhorn barycenters and give convergence results. The following result is a direct consequence of Theorem B.10 and Fact C.6.

Theorem D.1. *Let* $(\tilde{u}^{(\ell)})_{\ell \in \mathbb{N}}$ *be generated through Algorithm B.1 as in Theorem B.10. Then,*

$$(\forall \ell \in \mathbb{N}) \quad \|\tilde{u}^{(\ell)} - \nabla_1 OT_\varepsilon(\alpha, \beta)\|_\infty \leq \lambda^{2\ell} \left(\frac{\mathsf{D} + \max_{\mathcal{X}} u^{(0)} - \min_{\mathcal{X}} u^{(0)}}{\varepsilon} \right), \quad \text{(D.1)}$$

where $u^{(\ell)} = \varepsilon \log f^{(\ell)}$ *and* $\tilde{u}^{(\ell)} = u^{(\ell)} - u^{(\ell)}(x_o)$.

Therefore, in view of Fact C.8, Theorem D.1, and Proposition A.7, we can address the problem of the Sinkhorn barycenter (9) via the Frank-Wolfe Algorithm A.1. Note that, according to Proposition A.7(ii), since the diameter of $\mathcal{M}_1^+(\mathcal{X})$ with respect to $\|\cdot\|_{TV}$ is 2, we have that the curvature of B_ε is upper bounded by

$$C_{\mathsf{B}_\varepsilon} \leq 24\varepsilon e^{3\mathrm{D}/\varepsilon}. \tag{D.2}$$

Let $k \in \mathbb{N}$ and α_k be the current iteration. For every $j \in \{1, \dots, m\}$, we can compute $\nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \beta_j)$ and $\nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \alpha_k)$ by the Sinkhorn-Knopp algorithm. Thus, by (D.1), we find $\ell \in \mathbb{N}$ large enough so that $\|\tilde{u}_j^{(\ell)} - \nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \beta_q)\|_\infty \leq \Delta_{1,k}/8$ and $\|\tilde{p}^{(\ell)} - \nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \alpha_k)\|_\infty \leq \Delta_{1,k}/8$ and we set

$$\tilde{u}^{(\ell)} := \sum_{j=1}^m \omega_j \tilde{u}_j^{(\ell)} - \tilde{p}^{(\ell)}. \tag{D.3}$$

Then,

$$\|\tilde{u}^{(\ell)} - \nabla \mathsf{B}_\varepsilon(\alpha_k)\|_\infty \leq \frac{\Delta_{1,k}}{4}. \tag{D.4}$$

Now, Frank-Wolf Algorithm A.1 (in the version considered in Proposition A.7(i)) requires finding

$$\eta_{k+1} \in \operatorname*{argmin}_{\eta \in \mathcal{M}_1^+(\mathcal{X})} \langle \tilde{u}^{(\ell)}, \eta - \alpha_k \rangle \tag{D.5}$$

and make the update

$$\alpha_{k+1} = (1 - \gamma_k)\alpha_k + \gamma_k \eta_{k+1}. \tag{D.6}$$

Since the solution of (D.5) is a Dirac measure (see Section 4 in the paper), the algorithm reduces to

$$\begin{cases} \text{find } x_{k+1} \in \mathcal{X} \text{ such that } \tilde{u}^{(\ell)}(x_{k+1}) \leq \min_{x \in \mathcal{X}} \tilde{u}^{(\ell)}(x) + \dfrac{\Delta_{2,k}}{2} \\ \alpha_{k+1} = (1 - \gamma_k)\alpha_k + \gamma_k \delta_{x_{k+1}}. \end{cases} \tag{D.7}$$

So, if we initialize the algorithm with $\alpha_0 = \delta_{x_0}$, then any α_k will be a discrete probability measure with support contained in $\{x_0, \dots, x_k\}$. This implies that if all the β_j's are probability measures with finite support, the computation of $\nabla_1 \mathrm{OT}_\varepsilon(\alpha_k, \beta_j)$ by the Sinkhorn algorithm can be reduced to a fully discrete algorithm, as showed in Proposition B.11. More precisely, assume that

$$(\forall j = 1, \dots, m) \quad \beta_j = \sum_{i_2=0}^n b_{j,i_2} \delta_{y_{j,i_2}}. \tag{D.8}$$

and that at iteration k we have

$$\alpha_k = \sum_{i_1=0}^k a_{k,i_1} \delta_{x_{i_1}}. \tag{D.9}$$

Set

$$\mathsf{a}_k = \begin{bmatrix} a_{k,0} \\ \vdots \\ a_{k,k} \end{bmatrix} \in \mathbb{R}^{k+1}, \quad \mathsf{M}_{0,k} = \begin{bmatrix} a_{k,0}\mathsf{k}(x_0,x_0)a_{k,0} & \dots & a_{k,0}\mathsf{k}(x_0,x_k)a_{k,k} \\ \vdots & \ddots & \vdots \\ a_{k,k}\mathsf{k}(x_k,x_0)a_{k,0} & \dots & a_{k,k}\mathsf{k}(x_k,x_k)a_{k,k} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(k+1)} \tag{D.10}$$

and, for every $j = 1 \dots, m$,

$$\mathsf{b}_j = \begin{bmatrix} b_{j,0} \\ \vdots \\ b_{j,n} \end{bmatrix} \in \mathbb{R}^{n+1}, \quad \mathsf{M}_{j,k} = \begin{bmatrix} a_{k,0}\mathsf{k}(x_0,y_{j,0})b_{j,0} & \dots & a_{k,0}\mathsf{k}(x_0,y_{j,n})b_{j,n} \\ \vdots & \ddots & \vdots \\ a_{k,k}\mathsf{k}(x_k,y_{j,0})b_{j,0} & \dots & a_{k,n}\mathsf{k}(x_k,y_{j,n})b_{j,n} \end{bmatrix} \in \mathbb{R}^{(k+1)\times(n+1)}. \tag{D.11}$$

Then, run Algorithm B.2, with input a_k, a_k, and $\mathsf{M}_{0,k}$ to get $(\mathsf{e}^{(\ell)}, \mathsf{h}^{(\ell)})$, and, for every $j = 1, \dots, m$, with input a_k, b_j, and $\mathsf{M}_{j,k}$ to get $(\mathsf{f}_j^{(\ell)}, \mathsf{g}_j^{(\ell)})$. So, we have,

$$(\forall \ell \in \mathbb{N}) \quad \begin{cases} \mathsf{h}^{(\ell+1)} = \dfrac{\mathsf{a}_k}{\mathsf{M}_{0,k}^\top \mathsf{e}^{(\ell)}}, \quad \mathsf{e}^{(\ell+1)} = \dfrac{\mathsf{a}_k}{\mathsf{M}_{0,k}\mathsf{h}^{(\ell+1)}} \\ (\forall j = 1, \dots, m) \; \mathsf{g}_j^{(\ell+1)} = \dfrac{\mathsf{b}_j}{\mathsf{M}_{j,k}^\top \mathsf{f}_j^{(\ell)}}, \quad \mathsf{f}_j^{(\ell+1)} = \dfrac{\mathsf{a}_k}{\mathsf{M}_{j,k}\mathsf{g}_j^{(\ell+1)}}. \end{cases} \tag{D.12}$$

Then, according to Proposition B.11, for every $\ell \in \mathbb{N}$, we have

$$(\forall\, x \in \mathcal{X}) \quad \begin{cases} e^{(\ell)}(x)^{-1} = \displaystyle\sum_{i_2=0}^{k} \mathsf{k}(x, x_{i_2})\mathsf{h}_{i_2}^{(\ell-1)} a_{k,i_2}, \\[2mm] p^{(\ell)}(x) = \varepsilon \log e^{(\ell)}(x) = -\varepsilon \log \displaystyle\sum_{i_2=0}^{k} \mathsf{k}(x, x_{i_2})\mathsf{h}_{i_2}^{(\ell-1)} a_{k,i_2} \\[2mm] \tilde{p}^{(\ell)}(x) = p^{(\ell)}(x) - p^{(\ell)}(x_o). \end{cases} \tag{D.13}$$

and, for every $j = 1, \ldots, m$,

$$(\forall\, x \in \mathcal{X}) \quad \begin{cases} f_j^{(\ell)}(x)^{-1} = \displaystyle\sum_{i_2=0}^{n} \mathsf{k}(x, y_{i_2})\mathsf{g}_{j,i_2}^{(\ell-1)} b_{j,i_2}, \\[2mm] u_j^{(\ell)}(x) = \varepsilon \log f_j^{(\ell)}(x) = -\varepsilon \log \displaystyle\sum_{i_2=0}^{n} \mathsf{k}(x, y_{i_2})\mathsf{g}_{j,i_2}^{(\ell-1)} b_{j,i_2} \\[2mm] \tilde{u}_j^{(\ell)}(x) = u_j^{(\ell)}(x) - u_j^{(\ell)}(x_o). \end{cases} \tag{D.14}$$

Since the $\tilde{u}_j^{(\ell)}$'s and $u_j^{(\ell)}$'s, and $\tilde{p}^{(\ell)}$ and $p^{(\ell)}$, differ for a constant only, the final algorithm can be written as in Algorithm D.1. We stress that this algorithm is even more general than Algorithm 2 since, in the computation of the Sinkhorn potentials and in their minimization, errors have been taken into account.

Algorithm D.1 Frank-Wolfe algorithm for Sinkhorn barycenter

Let $\alpha_0 = \delta_{x_0}$ for some $x_0 \in \mathcal{X}$. Let $(\Delta_{1,k})_{k\in\mathbb{N}}, (\Delta_{2,k})_{k\in\mathbb{N}} \in \mathbb{R}_+^{\mathbb{N}}$ be such that $(2\Delta_{1,k} + \Delta_{2,k})/\gamma_k$ is nondecreasing. Define

> for $k = 0, 1, \ldots$
>> run Algorithm B.2 with input $\mathsf{a}_k, \mathsf{a}_k, \mathsf{M}_{0,k}$ till $\lambda^{2\ell}\mathsf{D}/\varepsilon \le \frac{\Delta_{1,k}}{8}$ $\rightarrow$ $\mathsf{h} \in \mathbb{R}^{k+1}$
>> compute p via (D.13) with h
>> for $j = 1, \ldots m$
>>> run Algorithm B.2 with input $\mathsf{a}_k, \mathsf{b}_j, \mathsf{M}_{j,k}$ till $\lambda^{2\ell}\mathsf{D}/\varepsilon \le \frac{\Delta_{1,k}}{8}$ $\rightarrow$ $\mathsf{g}_j \in \mathbb{R}^{n+1}$
>>> compute u_j via (D.14) with g_j
>> set $u = \sum_{j=1}^{m} \omega_j u_j - p$
>> find $x_{k+1} \in \mathcal{X}$ such that $u(x_{k+1}) \le \min_{x\in\mathcal{X}} u(x) + \dfrac{\Delta_{2,k}}{2}$
>> $\alpha_{k+1} = (1 - \gamma_k)\alpha_k + \gamma_k \delta_{x_{k+1}}.$

We now give a final converge theorem, of which Theorem 5 in the paper is a special case.

Theorem D.2. *Suppose that $\beta_1, \ldots, \beta_m \in \mathcal{M}_+^1(\mathcal{X})$ are probability measures with finite support, each of cardinality $n \in \mathbb{N}$. Let $(\alpha_k)_{k\in\mathbb{N}}$ be generated by Algorithm D.1. Then, for every $k \in \mathbb{N}$,*

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\alpha \in \mathcal{M}_+^1(\mathcal{X})} \mathsf{B}_\varepsilon(\alpha) \le \gamma_k 24\varepsilon e^{3\mathsf{D}/\varepsilon} + 2\Delta_{1,k} + \Delta_{2,k} \tag{D.15}$$

Proof. It follows from Theorem A.5, (D.2), and Proposition A.7, recalling that $\mathrm{diam}(\mathcal{M}_+^1(\mathcal{X})) = 2.$ $\qquad\square$

E Sample complexity of Sinkhorn potential

In the following we will denote by $\mathcal{C}^s(\mathcal{X})$ the space of s-differentiable functions with continuous derivatives and by $W^{s,p}(\mathcal{X})$ the Sobolev space of functions $f\colon \mathcal{X} \to \mathbb{R}$ with p-summable weak derivatives up to order s [1]. We denote by $\|\cdot\|_{s,p}$ the corresponding norm.

The following result shows that under suitable smoothness assumptions on the cost function c, the Sinkhorn potentials are uniformly bounded as functions in a suitable Sobolev space of corresponding smoothness. This fact will play a key role in approximating the Sinkhorn potentials of general distributions in practice.

Theorem E.1 (Proposition 2 in [23]). *Let $\mathcal{X}$ be a closed bounded domain with Lipschitz boundary in $\mathbb{R}^d$ ([1, Definition 4.9]) and let $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$. Then for every $(\alpha, \beta) \in \mathcal{M}_1^+(\mathcal{X})^2$, the associated Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ are functions in $W^{s,\infty}(\mathcal{X})$. Moreover, let $x_o \in \mathcal{X}$. Then there exists a constant $\mathsf{r} > 0$, depending only on ε, s and $\mathcal{X}$, such that for every $(\alpha, \beta) \in \mathcal{M}_1^+(\mathcal{X})^2$ the associated Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ with $u(x_o) = 0$ satisfies $\|u\|_{s,\infty}, \|v\|_{s,\infty} \leq \mathsf{r}$.*

In the original statement of [23, Proposition 2] the above result is formulated for $\mathsf{c} \in \mathcal{C}^\infty(\mathcal{X})$ for simplicity. However, as clarified by the authors, it holds also for the more general case $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X})$.

Lemma E.2. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a closed bounded domain with Lipschitz boundary and let $u, u' \in W^{s,\infty}(\mathcal{X})$. Then the following holds*

    (i) $\|uu'\|_{s,\infty} \leq \mathsf{m}_1 \|u\|_{s,\infty} \|u'\|_{s,\infty}$,

    (ii) $\|e^u\|_{s,\infty} \leq \|e^u\|_\infty (1 + \mathsf{m}_2 \|u\|_{s,\infty})$,

where $\mathsf{m}_1 = \mathsf{m}_1(s, d)$ and $\mathsf{m}_2 = \mathsf{m}_2(s, d) > 0$ depend only on the dimension d and the order of differentiability s but not on u and u'.

Proof. (i) follows directly from Leibniz formula. To see (ii), let $\mathbf{i} = (i_1, \ldots, i_d) \in \mathbb{N}^d$ be a multi-index with $|\mathbf{i}| = \sum_{\ell=1}^d i_\ell \leq s$ and note that by chain rule the derivatives of e^u

$$D^{\mathbf{i}} e^u = e^u P_{\mathbf{i}}\left((D^{\mathbf{j}}u)_{\mathbf{j} \leq \mathbf{i}}\right),$$

where $P_{\mathbf{i}}$ is a polynomial of degree $|\mathbf{i}|$ and $\mathbf{j} \leq \mathbf{i}$ is the ordering associated to the cone of non-negative vectors in $\mathbb{R}^d$. Note that $P_0 = 1$, while for $|\mathbf{i}| > 0$, the associated polyomial $P_{\mathbf{i}}$ has a root in zero (i.e. it does not have constant term). Hence

$$\|e^u\|_{s,\infty} \leq \|e^u\|_\infty \left(1 + |P|\left((\|D^{\mathbf{i}}u\|_\infty)_{|\mathbf{i}| \leq s}\right)\right),$$

where we have denoted by $P = \sum_{0 < |\mathbf{i}| \leq s} P_{\mathbf{i}}$ and by $|P|$ the polynomial with coefficients corresponding to the absolute value of the coefficients of P. Therefore, since $\|D^{\mathbf{i}}u\|_\infty \leq \|u\|_{s,\infty}$ for any $|\mathbf{i}| \leq s$, by taking

$$\mathsf{m}_2 = |P|\left((1)_{|\mathbf{i}| \leq s}\right),$$

namely the sum of all the coefficients of $|P|$, we obtain the desired result. Indeed note that the coefficients of P do not depend on u but only on the smoothness s and dimension d. $\square$

Lemma E.3. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a closed bounded domain with Lipschitz boundary and let $x_o \in \mathcal{X}$. Let $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$, for some $s \in \mathbb{N}$. Then for any $\alpha, \beta \in \mathcal{M}_1^+(\mathcal{X})$ and corresponding pair of Sinkhorn potentials $(u, v) \in \mathcal{C}(\mathcal{X})^2$ with $u(x_o) = 0$, the functions $\mathsf{k}(x, \cdot)e^{u/\varepsilon}$ and $\mathsf{k}(x, \cdot)e^{v/\varepsilon}$ belong to $W^{s,2}(\mathcal{X})$ for every $x \in \mathcal{X}$. Moreover, they admit an extension to $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$ and there exists a constant $\bar{\mathsf{r}}$ independent on α and β, such that for every $x \in \mathcal{X}$*

$$\left\|\mathsf{k}(x, \cdot)e^{u/\varepsilon}\right\|_{\mathcal{H}}, \left\|\mathsf{k}(x, \cdot)e^{v/\varepsilon}\right\|_{\mathcal{H}} \leq \bar{\mathsf{r}} \qquad (\text{E.1})$$

(with some abuse of notation, we have identified $\mathsf{k}(x, \cdot)e^{u/\varepsilon}$ and $\mathsf{k}(x, \cdot)e^{v/\varepsilon}$ with their extensions to $\mathbb{R}^d$).

Proof. In the following we denote by $\|\cdot\|_{s,2} = \|\cdot\|_{s,2,\mathcal{X}}$ the norm of $W^{s,2}(\mathcal{X})$ and by $\|\cdot\|_{\mathcal{H}} = \|\cdot\|_{s,2,\mathbb{R}^d}$ the norm of $\mathcal{H} = W^{s,2}(\mathbb{R})$. Let $x \in \mathcal{X}$. Then, since $u - \mathsf{c}(x, \cdot) \in W^{s,\infty}(\mathcal{X})$ and

33

$\|u\|_{s,\infty} \leq \mathsf{r}$, it follows from Lemma E.2 that

$$
\begin{aligned}
\left\|\mathsf{k}(x,\cdot)e^{u/\varepsilon}\right\|_{s,\infty} &= \left\|e^{(u-\mathsf{c}(x,\dot{\cdot}))/\varepsilon}\right\|_{s,\infty} \\
&\leq \left\|e^{(u-\mathsf{c}(x,\dot{\cdot}))/\varepsilon}\right\|_{\infty}(1+\mathsf{m}_2\|u-\mathsf{c}(x,\cdot)\|_{s,\infty}) \\
&= \left\|\mathsf{k}(x,\cdot)e^{u/\varepsilon}\right\|_{\infty}(1+\mathsf{m}_2\|u-\mathsf{c}(x,\cdot)\|_{s,\infty}) \\
&\leq \left\|e^{u/\varepsilon}\right\|_{\infty}(1+\mathsf{m}_2(\mathsf{r}+\|\mathsf{c}\|_{s,\infty})) \\
&\leq e^{\mathsf{D}/\varepsilon}(1+\mathsf{m}_2(\mathsf{r}+\|\mathsf{c}\|_{s,\infty})),
\end{aligned}
$$

where we used the fact that $D^{\mathbf{i}}[\mathsf{c}(x,\cdot)] = (D^{\mathbf{i}}\mathsf{c})(x,\cdot)$. This implies

$$
\left\|\mathsf{k}(x,\cdot)e^{u/\varepsilon}\right\|_{s,2} \leq |\mathcal{X}|^{1/2}e^{\mathsf{D}/\varepsilon}(1+\mathsf{m}_2(\mathsf{r}+\|\mathsf{c}\|_{s,\infty}))
$$

where $|\mathcal{X}|$ is the Lebesgue measure of $\mathcal{X}$. Now, we can proceed analogously to [23, Proposition 2], and use Stein's Extension Theorem [1, Theorem 5.24],[51, Chapter 6], to guarantee the existence of a *total extension operator* [1, Definition 5.17]. In particular, there exists a constant $\mathsf{m}_3 = \mathsf{m}_3(s,2,\mathcal{X})$ such that for any $\varphi \in W^{s,2}(\mathcal{X})$ there exists $\tilde{\varphi} \in W^{s,2}(\mathbb{R}^d)$ such that

$$
\|\tilde{\varphi}\|_{\mathcal{H}} = \|\tilde{\varphi}\|_{s,2,\mathbb{R}^d} \leq \mathsf{m}_3 \|\varphi\|_{s,2,\mathcal{X}} = \mathsf{m}_3 \|\varphi\|_{s,2}. \tag{E.2}
$$

Therefore, we conclude

$$
\left\|\mathsf{k}(x,\cdot)e^{u/\varepsilon}\right\|_{\mathcal{H}} \leq \mathsf{m}_3|\mathcal{X}|^{1/2}e^{\mathsf{D}/\varepsilon}(1+\mathsf{m}_2(\mathsf{r}+\|\mathsf{c}\|_{s,\infty})) =: \bar{\mathsf{r}}. \tag{E.3}
$$

The same argument applies to $\mathsf{k}(x,\cdot)e^{v/\varepsilon}$ with the only exception that now, in virtue of Corollary B.9, we have $\|e^{v/\varepsilon}\|_{\infty} \leq e^{2\mathsf{D}/\varepsilon}$. Note that $\bar{\mathsf{r}}$ is a constant depending only on $\mathcal{X}$, c, s and d but it is independent on the probability distributions α and β. $\qquad\square$

Sobolev spaces and reproducing kernel Hilbert spaces. Recall that for $s > d/2$ the space $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$, is a reproducing kernel Hilbert space (RKHS) [53, Chapter 10]. In this setting we denote by $\mathsf{h} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ the associated reproducing kernel, which is continuous and bounded and satisfies the reproducing property

$$
(\forall x \in \mathcal{X})(\forall f \in \mathcal{H}) \qquad \langle f, \mathsf{h}(x,\cdot)\rangle_{\mathcal{H}} = f(x). \tag{E.4}
$$

We can also assume that h is *normalized*, namely, $\|\mathsf{h}(x,\cdot)\|_{\mathcal{H}} = 1$ for all $x \in \mathcal{X}$ [53, Chapter 10].

Kernel mean embeddings. For every $\beta \in \mathcal{M}_1^+(\mathcal{X})$, we denote by $\mathsf{h}_\beta \in \mathcal{H}$ the *Kernel Mean Embedding* of β in $\mathcal{H}$ [35, 43], that is, the vector

$$
\mathsf{h}_\beta = \int \mathsf{h}(x,\cdot)\, d\beta(x). \tag{E.5}
$$

In other words, the kernel mean embedding of a distribution β corresponds to the expectation of $\mathsf{h}(x,\cdot)$ with respect to β. By the linearity of the inner product and the integral, for every $f \in \mathcal{H}$, the inner product

$$
\langle f, \mathsf{h}_\beta\rangle_{\mathcal{H}} = \int \langle f, \mathsf{h}(x,\cdot)\rangle \, d\beta(x) = \int f(x)\, d\beta(x), \tag{E.6}
$$

corresponds to the expectation of $f(x)$ with respect to β. The *Maximum Mean Discrepancy (MMD)* [35, 46, 47] between two probability distributions $\beta, \beta' \in \mathcal{M}_1^+(\mathcal{X})$ is defined as

$$
\mathrm{MMD}(\beta, \beta') = \|\mathsf{h}_\beta - \mathsf{h}_{\beta'}\|_{\mathcal{H}}. \tag{E.7}
$$

In the case of the Sobolev space $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$, the MMD metrizes the weak-$*$ topology of $\mathcal{M}_1^+(\mathcal{X})$ [47, 48].

A well-established approach to approximate a distribution $\beta \in \mathcal{M}_1^+(\mathcal{X})$ is to independently sample a set of points $x_1,\ldots,x_n \in \mathcal{X}$ from β and consider the empirical distribution $\beta_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. The following result shows that β_n converges to β in MMD with high probability. The original version of this result can be found in [46], we report an independent proof for completeness.

Lemma E.4. *Let $\beta \in \mathcal{M}_1^+(\mathcal{X})$. Let $x_1, \ldots, x_n \in \mathcal{X}$ be indepedently sampled according to β and denote by $\beta_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$. Then, for any $\tau \in (0, 1]$, we have*

$$\mathrm{MMD}(\beta_n, \beta) \leq \frac{4 \log \frac{3}{\tau}}{\sqrt{n}} \tag{E.8}$$

with probability at least $1 - \tau$.

Proof. The proof follows by applying Pinelis' inequality [39, 42, 55] for random vectors in Hilbert spaces. More precisely, for $i = 1, \ldots, n$, denote by $\zeta_i = \mathsf{h}(x_i, \cdot) \in \mathcal{H}$ and recall that $\|\zeta_i\| = \|\mathsf{h}(x, \cdot)\| = 1$ for all $x \in \mathcal{X}$. We can therefore apply [42, Lemma 2] with constants $\widetilde{M} = 1$ and $\sigma^2 = \sup_i \mathbb{E}\|\zeta_i\|^2 \leq 1$, which guarantees that, for every $\tau \in (0, 1]$

$$\left\| \frac{1}{n} \sum_{i=1}^n \left[\zeta_i - \mathbb{E}\, \zeta_i \right] \right\|_{\mathcal{H}} \leq \frac{2 \log \frac{2}{\tau}}{n} + \sqrt{\frac{2 \log \frac{2}{\tau}}{n}} \leq \frac{4 \log \frac{3}{\tau}}{\sqrt{n}}, \tag{E.9}$$

holds with probability at least $1 - \tau$. Here, for the second inequality we have used the fact that $\log \frac{2}{\tau} \leq \log \frac{3}{\tau}$ and $\log \frac{3}{\tau} \geq 1$ for every $\tau \in (0, 1]$. The desired result follows by observing that

$$\mathsf{h}_\beta = \int \mathsf{h}(x, \cdot)\, d\beta(x) = \mathbb{E}\, \zeta_i \tag{E.10}$$

for all $i = 1, \ldots, n$, and

$$\mathsf{h}_{\beta_n} = \frac{1}{n} \sum_{i=1}^n \mathsf{h}(x_i, \cdot) = \frac{1}{n} \sum_{i=1}^n \zeta_i. \tag{E.11}$$

Therefore,

$$\mathrm{MMD}(\beta_n, \beta) = \|\mathsf{h}_{\beta_n} - \mathsf{h}_\beta\|_{\mathcal{H}} = \left\| \frac{1}{n} \sum_{i=1}^n \left[\zeta_i - \mathbb{E}\, \zeta_i \right] \right\|_{\mathcal{H}}, \tag{E.12}$$

which combined with (E.9) leads to the desired result. $\qquad\square$

Proposition E.5 (Lipschitz continuity of the Sinkhorn Potentials with respect to the MMD). *Let $\mathcal{X} \subset \mathbb{R}^d$ be a compact Lipschitz domain and $\mathsf{c} \in \mathcal{C}^{s+1}(\mathcal{X} \times \mathcal{X})$, with $s > d/2$. Let $\alpha, \beta, \alpha', \beta' \in \mathcal{M}_1^+(\mathcal{X})$. Let $x_o \in \mathcal{X}$ and let $(u, v), (u', v') \in \mathcal{C}(\mathcal{X})^2$ be the two Sinkhorn potentials corresponding to the solution of the regularized OT problem in (B.24) for (α, β) and (α', β') respectively such that $u(x_o) = u'(x_o) = 0$. Then*

$$\|u - u'\|_\infty \leq 2\varepsilon \bar{r} e^{3\mathsf{D}/\varepsilon} \left(\mathrm{MMD}(\alpha, \alpha') + \mathrm{MMD}(\beta, \beta') \right), \tag{E.13}$$

with $\bar{r}$ from Lemma E.3. In other words, the operator $\nabla_1 \mathrm{OT}_\varepsilon \colon \mathcal{M}_1^+(\mathcal{X})^2 \to \mathcal{C}(\mathcal{X})$, defined in Fact C.6, is $2\varepsilon\bar{r}e^{3\mathsf{D}/\varepsilon}$-Lipschitz continuous with respect to the MMD.

Proof. Let $f = e^{u/\varepsilon}$ and $g = e^{v/\varepsilon}$. By relying on Lemma E.3 we can now refine the analysis in Theorem C.4. More precisely, we observe that in (C.10) we have

$$
\begin{aligned}
[(\mathsf{L}_{\beta'} - \mathsf{L}_\beta)g](x) &= \int \mathsf{k}(x, z)g(z)\, d(\beta - \beta')(z) \\
&= \int \langle \mathsf{k}(x, \cdot)g, \mathsf{h}(z, \cdot) \rangle_{\mathcal{H}}\, d(\beta - \beta')(z) \\
&= \langle \mathsf{k}(x, \cdot)g, \mathsf{h}_\beta - \mathsf{h}_{\beta'} \rangle_{\mathcal{H}} \\
&\leq \|\mathsf{k}(x, \cdot)g\|_{\mathcal{H}} \, \|\mathsf{h}_\beta - \mathsf{h}_{\beta'}\|_{\mathcal{H}} \\
&\leq \bar{r}\, \mathrm{MMD}(\beta, \beta'),
\end{aligned}
$$

where in the first equality, with some abuse of notation, we have implicitly considered the extension of $\mathsf{k}(x, \cdot)g$ to $\mathcal{H} = W^{s,2}(\mathbb{R}^d)$ as discussed in Lemma E.3. The rest of the analysis in Theorem C.4 remains invaried, eventually leading to (E.13). $\qquad\square$

It is now clear that Theorem 6 in the paper is just a consequence of Lemma E.4 and Proposition E.5. We give the statement of the theorem for reader's convenience.

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* $\tau \in (0, 1]$

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

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

We finally provide the proof of Theorem 7 in the paper.

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* $\hat{\beta}_1, \ldots, \hat{\beta}_m$ *be empirical distributions with* n *support points, each independently sampled from* $\beta_1, \ldots, \beta_m$. *Let* α_k *be the* k-*th iterate of Algorithm 2 applied to* $\hat{\beta}_1, \ldots, \hat{\beta}_m$. *Then for any* $\tau \in (0, 1]$, *the following holds with probability larger than* $1 - \tau$

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

Proof. Let $\widehat{\mathsf{B}_\varepsilon}(\alpha) = \sum_{j=1}^m \omega_j \mathsf{S}_\varepsilon(\alpha, \hat{\beta}_j)$. We apply Theorem 6 independently for each distribution $\hat{\beta}_j$ and then take the intersection bound between all these separate events. Then, for every $k \in \mathbb{N}$, and with probability larger than $1 - \tau$, we have

$$\begin{aligned}
\|\nabla\widehat{\mathsf{B}_\varepsilon}(\alpha_k) - \nabla\mathsf{B}_\varepsilon(\alpha_k)\|_\infty &\leq \sum_{j=1}^m \omega_j \|\nabla[\mathsf{S}_\varepsilon(\cdot, \hat{\beta}_j)](\alpha_k) - \mathsf{S}_\varepsilon(\cdot, \beta_j)](\alpha_k)\|_\infty \\
&= \sum_{j=1}^m \omega_j \|\nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \hat{\beta}_j) - \nabla_1 \mathsf{OT}_\varepsilon(\alpha_k, \beta_j)\|_\infty \\
&\leq \frac{8\varepsilon \, \bar{\mathsf{r}} e^{3\mathsf{D}/\varepsilon} \log \frac{3m}{\tau}}{\sqrt{n}} \\
&= \frac{\Delta_1}{4},
\end{aligned}$$

where

$$\Delta_1 := \frac{32\varepsilon \, \bar{\mathsf{r}} e^{3\mathsf{D}/\varepsilon} \log \frac{3m}{\tau}}{\sqrt{n}}.$$

Now, let $\gamma_k = 2/(k + 2)$. Since Algorithm 2 is applied to $\hat{\beta}_1, \ldots \hat{\beta}_m$, we have

$$\delta_{x_{k+1}} \in \operatorname*{argmin}_{\mathcal{M}_+^1(\mathcal{X})} \langle \nabla\widehat{\mathsf{B}_\varepsilon}(\alpha_k), \cdot \rangle \quad \text{and} \quad \alpha_{k+1} = (1 - \gamma_k)\alpha_k + \gamma_k \delta_{x_{k+1}}.$$

Therefore, it follows from Theorem A.5, Proposition A.7 (with $\Delta_{1,k} = \Delta_1$ and $\Delta_{2,k} = 0$), and Theorem 4 that, with probability larger than $1 - \tau$, we have

$$\mathsf{B}_\varepsilon(\alpha_k) - \min_{\mathcal{M}_+^1(\mathcal{X})} \mathsf{B}_\varepsilon \leq 6\varepsilon\bar{\mathsf{r}}e^{3\mathsf{D}/\varepsilon}\operatorname{diam}(\mathcal{M}_+^1(\mathcal{X}))^2\gamma_k + \Delta_1\operatorname{diam}(\mathcal{M}_+^1(\mathcal{X})).$$

The statement follows by noting that $\operatorname{diam}(\mathcal{M}_+^1(\mathcal{X})) = 2$. $\qquad \square$

F Additional experiments

Sampling of continuous measures: mixture of Gaussians. We perform the barycenter of 5 mixtures of two Gaussians μ_j, centered at $(j/2, 1/2)$ and $(j/2, 3/2)$ for $j-0, \dots, 4$ respectively. Samples are provided in Figure 6. We use different relative weights pairs in the mixture of Gaussians, namely $(1/10, 9/10), (1/4, 3/4), (1/2, 1/2)$. At each iteration, a sample of $n = 500$ points is drawn from $\mu_j, j = 0 \dots, 4$. Results are reported in Figure 7.
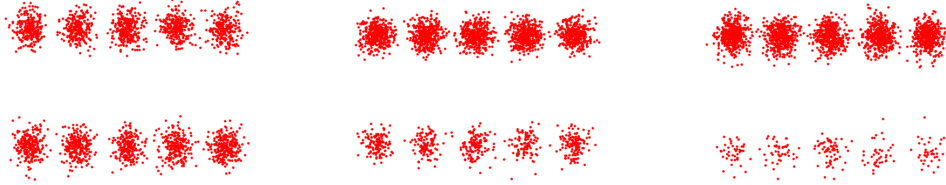

Fig. 6: Samples of input measures

Fig. 7: Barycenters of Mixture of Gaussians

Large scale discrete measures: meshes. We perform the barycenter of two discrete measures with support in $\mathbb{R}^3$. Meshes of the dinosaur are taken from [44] and rescaled by a 0.5 factor. The internal problem in Frank-Wolfe algorithm is solved using L-BFGS-B SciPy optimizer. Formula of the Jacobian is passed to the method. The result is displayed in Figure 8.

Propagation. We extend the description on the experiment about propagation in Section 6. Edges $\mathcal{E}$ are selected as follows: we created a matrix D such that D_{ij} contains the distance between station at vertex i and station at vertex j, computed using the geographical coordinates of the stations. Each node v in $\mathcal{V}$, is connected to those nodes $u \in \mathcal{V}$ such that $D_{vu} \leq 3$. If the number of nodes u that meet this condition is *less* than 5, we connect v with its 5 nearest nodes. If the number of nodes u that meet this condition is *more* than 10, we connect v with its 10 nearest nodes. Each edge e_{uv} is weighted with $\omega_{uv} := D_{uv}$. Since intuitively we may expect that nearer nodes should have more influence in the construction of the histograms of unknown nodes, in the propagation functional we weight $S_\varepsilon(\rho_v, \rho_u)$ with use $\exp(-\omega_{uv}/\sigma)$ or $1/\omega_{vu}$ suitably normalized.
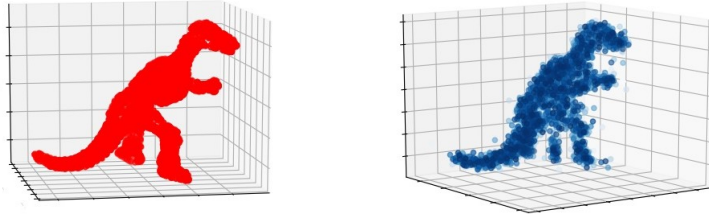

Fig. 8: True barycenter (left), result of our algorithm (right)