Electronic Journal of Statistics
Vol. 13 (2019) 5120–5150
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1637

Central limit theorems for entropy-regularized optimal transport on finite spaces and statistical applications[*]

Jérémie Bigot, Elsa Cazelles

Université de Bordeaux
Institut de Mathématiques de Bordeaux
Talence, France
e-mail: jeremie.bigot@math.u-bordeaux.fr*;* elsa.cazelles@u-bordeaux.fr

Nicolas Papadakis

CNRS (UMR 5251)
Institut de Mathématiques de Bordeaux
Talence, France
e-mail: nicolas.papadakis@math.u-bordeaux.fr

Abstract: The notion of entropy-regularized optimal transport, also known as Sinkhorn divergence, has recently gained popularity in machine learning and statistics, as it makes feasible the use of smoothed optimal transportation distances for data analysis. The Sinkhorn divergence allows the fast computation of an entropically regularized Wasserstein distance between two probability distributions supported on a finite metric space of (possibly) high-dimension. For data sampled from one or two unknown probability distributions, we derive the distributional limits of the empirical Sinkhorn divergence and its centered version (Sinkhorn loss). We also propose a bootstrap procedure which allows to obtain new test statistics for measuring the discrepancies between multivariate probability distributions. Our work is inspired by the results of Sommerfeld and Munk in [33] on the asymptotic distribution of empirical Wasserstein distance on finite space using unregularized transportation costs. Incidentally we also analyze the asymptotic distribution of entropy-regularized Wasserstein distances when the regularization parameter tends to zero. Simulated and real datasets are used to illustrate our approach.

MSC 2010 subject classifications: 62G20, 62G10, 65C60.

Keywords and phrases: Optimal transport, Sinkhorn divergence, central limit theorem, bootstrap, hypothesis testing.

Received February 2019.

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

1. Introduction

1.1. Motivations

In this paper, we study the convergence (to their population counterparts) of empirical probability measures supported on a finite metric space with respect to entropy-regularized transportation costs. Transport distances are widely employed for comparing probability measures since they capture in a instinctive manner the geometry of distributions (see e.g. [37] for a general presentation on the subject). In particular, the Wasserstein distance is well adapted to deal with discrete probability measures (supported on a finite set), as its computation reduces to solve a linear program. Moreover, since data in the form of histograms may be represented as discrete measures, the Wasserstein distance has been shown to be a relevant statistical measure in various fields such as clustering of discrete distributions [40], nonparametric Bayesian modelling [24], fingerprints comparison [33], unsupervised learning [1] and principal component analysis [3, 31, 5].

However, the computational cost to evaluate a transport distance is generally of order $\mathcal{O}(N^3 \log N)$ for discrete probability distributions with a support of size N. To overcome the computational cost to evaluate a transport distance, Cuturi [7] has proposed to add an entropic regularization term to the linear program corresponding to a standard optimal transport problem, leading to the notion of Sinkhorn divergence between probability distributions. Initially, the purpose of transport plan regularization was to efficiently compute a divergence term close to the Wasserstein distance between two probability measures, through an iterative scaling algorithm where each iteration costs $\mathcal{O}(N^2)$. This proposal has recently gained popularity in machine learning and statistics, as it makes feasible the use of smoothed optimal transportation distance for data analysis. It has found various applications such as generative models [19] and more generally for high dimensional data analysis in multi-label learning [16], dictionary learning [29], image processing [9, 26], text mining via bag-of-words comparison [18], averaging of neuroimaging data [20].

The goal of this paper is to analyze the potential benefits of the Sinkhorn divergence and its centered version [14, 19] for statistical inference from empirical probability measures. We derive novel results on the asymptotic distribution of such divergences for data sampled from (unknown) distributions supported on a finite metric space. The main application is to obtain new test statistics (for one or two samples problems) for the comparison of multivariate probability distributions.

1.2. Previous work and main contributions

The derivation of distributional limits of an empirical measure towards its population counterpart in p-Wasserstein distance $W_p(\mu, \nu)$ is well understood for probability measures μ and ν supported on $\mathbb{R}$ [15, 10, 11]. These results have then been extended for specific parametric distributions supported on $\mathbb{R}^d$ belonging to an elliptic class, see [28] and references therein. Recently, a central limit theorem has been established in [12] for empirical transportation cost, and data sampled from absolutely continuous measures on $\mathbb{R}^d$, for any $d \geq 1$. The case of discrete measures supported on a finite metric space has also been considered in [33] with the proof of the convergence (in the spirit of the central limit theorem) of empirical Wasserstein distances toward the optimal value of a linear program. Additionally, Klatt et al. [21] analyzed, in parallel with our results, the distributional limit of regularized optimal transport divergences between empirical distributions. In particular, the work in [21] extends the study of distributional limits of regularized empirical transportation cost to general penalty functions (beyond entropy regularization). The authors of [27] also studied the link between nonparametric tests and the Wasserstein distance, with an emphasis on distributions with support in $\mathbb{R}$.

However, apart from the one-dimensional case ($d = 1$), and the work of [21], these results lead to test statistics whose numerical implementation become prohibitive for empirical measures supported on $\mathbb{R}^d$ with $d \geq 2$. The computational cost required to evaluate a transport distance is indeed only easily tractable in $\mathbb{R}$. It is therefore of interest to propose test statistics based on fast Sinkhorn divergences [7]. In this context, this paper focuses on the study of inference from discrete distributions in terms of entropically regularized transport costs, the link with the inference through unregularized transport, and the construction of tests statistics that are well suited to investigate the equality of two distributions. The results are inspired by the work in [33] on the asymptotic distribution of empirical Wasserstein distance on finite space using unregularized transportation costs.

Our main contributions may be summarized as follows. First, for data sampled from one or two unknown measures μ and ν supported on a finite space, we derive central limit theorems for the Sinkhorn divergence between their empirical counterpart. These results allow to build new test statistics for measuring the discrepancies between multivariate probability distributions. Notice however that the Sinkhorn divergence denoted $W_{p,\varepsilon}^p(\mu, \nu)$ (where $\varepsilon > 0$ is a regularization parameter) is not a distance since $W_{p,\varepsilon}^p(\mu, \mu) \neq 0$. This is a serious drawback for testing the hypothesis of equality between distributions. Thus, as introduced in [14, 19], we further consider the centered version of the Sinkhorn divergence $\overline{W}_{p,\varepsilon}^p(\mu, \nu)$, referred to as Sinkhorn loss, which satisfies $\overline{W}_{p,\varepsilon}^p(\mu, \mu) = 0$. This study thus constitutes an important novel contribution with respect to the work of [21]. We present new results on the asymptotic distributions of the Sinkhorn loss between empirical measures. Interestingly, under the hypothesis that $\mu = \nu$, such statistics do not converge to a Gaussian random variable but to a mixture of chi-squared distributed random variables. To illustrate the applicability of the method to the analysis of real data, we propose a bootstrap procedure to estimate unknown quantities of interest on the distribution of these statistics such as their non-asymptotic variance and quantiles. Simulated and real datasets are used to illustrate our approach. Finally, one may stress that an advantage of the use of test statistics based regularized Wasserstein distance (rather than other losses or divergences) is to allow further statistical inference from the resulting optimal transport plan as demonstrated in [21] for the analysis of protein interaction networks.

1.3. Overview of the paper

In Section 2, we briefly recall the optimal transport problem between probability measures, and we introduce the notions of Sinkhorn divergence and Sinkhorn loss. Then, we derive the asymptotic distributions for the empirical Sinkhorn divergence and the empirical Sinkhorn loss. We also give the behavior of such statistics when the regularization parameter ε tends to zero at a rate depending on the number of available observations. A bootstrap procedure is discussed in Section 3. Numerical experiments are reported in Section 4 for synthetic data and in Section 5 for real data.

2. Distributional limits for entropy-regularized optimal transport

In this section, we give results on the asymptotic distributions of the empirical Sinkhorn divergence and the empirical Sinkhorn loss. The proofs rely on the use of the delta-method and on the property that $W_{p,\varepsilon}^p(\mu,\nu)$ is a differentiable function with respect to μ and ν.

2.1. Notation and definitions

We first introduce various notation and definitions that will be used throughout the paper.

2.1.1. Optimal transport, Sinkhorn divergence and Sinkhorn loss

Let $(\mathcal{X},d)$ be a complete metric space with $d:\mathcal{X}\times\mathcal{X}\to\mathbb{R}_+$. We denote by $\mathcal{P}_p(\mathcal{X})$ the set of Borel probability measures μ supported on $\mathcal{X}$ with finite moment of order $p\geq 1$, in the sense that $\int_{\mathcal{X}} d^p(x,y)d\mu(x)$ is finite for some (and thus for all) $y\in\mathcal{X}$. The p-Wasserstein distance between two measures μ and ν in $\mathcal{P}_p(\mathcal{X})$ is defined by

$$W_p(\mu,\nu)=\left(\inf_{\pi\in\Pi(\mu,\nu)}\iint_{\mathcal{X}^2} d^p(x,y)d\pi(x,y)\right)^{1/p} \tag{1}$$

where the infimum is taken over the set $\Pi(\mu,\nu)$ of probability measures π on the product space $\mathcal{X}\times\mathcal{X}$ with respective marginals μ and ν.

In this work, we consider the specific case where $\mathcal{X}=\{x_1,\dots,x_N\}$ is a finite metric space of size N. In this setting, a measure $\mu\in\mathcal{P}_p(\mathcal{X})$ is discrete, and we write $\mu=\sum_{i=1}^N a_i\delta_{x_i}$ where $(a_1,\dots,a_N)$ is a vector of positive weights belonging to the simplex $\Sigma_N := \{a=(a_i)_{i=1,\dots,N}\in\mathbb{R}_+^N$ such that $\sum_{i=1}^N a_i = 1\}$ and

δ_{x_i} is a Dirac measure at location x_i. Therefore, computing the p-Wasserstein distance between discrete probability measures supported on $\mathcal{X}$ amounts to solve a linear program whose solution is constraint to belong to the convex set $\Pi(\mu, \nu)$. However, the cost of this convex minimization becomes prohibitive for moderate to large values of N. Regularizing a complex problem with an entropy term is a classical approach in optimization to reduce its complexity [39]. This is the approach followed in [7] by adding an entropy regularization to the transport matrix, which yields the strictly convex (primal) problem (2) presented below.

As the space $\mathcal{X}$ is fixed, a probability measure supported on $\mathcal{X}$ is entirely characterized by a vector of weights in the simplex. By a slight abuse of notation, we thus identify a measure $\mu \in \mathcal{P}_p(\mathcal{X})$ by its vector of weights $a = (a_1, \dots, a_n) \in \Sigma_N$ (and we sometimes write $a = \mu$).

Definition 2.1 (Sinkhorn divergence)**.** Let $\varepsilon > 0$ be a regularization parameter. The Sinkhorn divergence [7] between two probability measures $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^N b_i \delta_{x_i}$ in $\mathcal{P}_p(\mathcal{X})$ is defined by

$$W_{p,\varepsilon}^p(a,b) = \min_{T \in U(a,b)} \langle T, C \rangle + \varepsilon H(T | a \otimes b), \text{ with } a \text{ and } b \text{ in } \Sigma_N, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product between matrices, $a \otimes b$ denotes the tensor product $(x_i, x_j) \mapsto a_i b_j$ and

- $U(a,b) = \{T \in \mathbb{R}_+^{N \times N} \, | T \mathbb{1}_N = a, T^T \mathbb{1}_N = b\}$ is the set of transport matrices with marginals a and b (with $\mathbb{1}_N$ denoting the vector of $\mathbb{R}^N$ with all entries equal to one),
- $C \in \mathbb{R}_+^{N \times N}$ is the pairwise cost matrix associated to the metric space (X, d) whose (i,j)-th entry is $c_{ij} = d(x_i, x_j)^p$,
- the regularization function $H(T | a \otimes b) = \sum_{i,j} \log\left(\frac{t_{ij}}{a_i b_j}\right) t_{ij}$ is the relative entropy for a transport matrix $T \in U(a,b)$.

Remark 1. *This entire section is also valid for symmetric positive cost matrices C for which $C(x_i, x_i) = 0$.*

The dual version of problem (2) is introduced in the following definition.

Definition 2.2 (Dual problem)**.** Following [8], the dual version of the minimization problem (2) is given by

$$W_{p,\varepsilon}^p(a,b) = \max_{u,v \in \mathbb{R}^N} u^T a + v^T b - \varepsilon \sum_{i,j} \left(e^{-\frac{1}{\varepsilon}(c_{ij} - 1 - u_i - v_j)} \right) a_i b_j. \tag{3}$$

We denote by $\mathcal{S}_\varepsilon(a,b)$ the set of optimal solutions of the maximization problem (3).

It is now well known that there exists an explicit relationship between the optimal solutions of primal (2) and dual (3) problems. These solutions can be computed through an iterative method called Sinkhorn's algorithm [8] that is described below and which explicitly gives this relationship.

Proposition 2.1 (Sinkhorn's algorithm)**.** *Let $K = \exp(-C/\varepsilon - \mathbb{1}_{N\times N})$ be the elementwise exponential of the matrix cost C divided by $-\varepsilon$ minus the matrix with all entries equal to 1. Then, there exists a pair of vectors $(\tilde{u}, \tilde{v}) \in \mathbb{R}_+^N \times \mathbb{R}_+^N$ such that the optimal solutions T_ε^* and $(u_\varepsilon^*, v_\varepsilon^*)$ of problems (2) and (3) are respectively given by*

$$T_\varepsilon^* = [\mathrm{diag}(\tilde{u}) K \,\mathrm{diag}(\tilde{v})] \odot (a \otimes b), \ \text{ and } u_\varepsilon^* = \varepsilon \log(\tilde{u}), \ v_\varepsilon^* = \varepsilon \log(\tilde{v}),$$

where $\odot$ is the pointwise multiplication. Moreover, such a pair $(\tilde{u}, \tilde{v})$ is unique up to scalar multiplication (or equivalently $(u_\varepsilon^*, v_\varepsilon^*)$ is unique up to translation), and it can be recovered as a fixed point of the Sinkhorn map

$$(\tilde{u}, \tilde{v}) \in \mathbb{R}^N \times \mathbb{R}^N \mapsto (a/(K\tilde{v}), b/(K^T \tilde{u})), \tag{4}$$

where K^T is the transpose of K and / stands for the component-wise division.

Remark 2. *When the cost matrix C is defined as $c_{ij} = \|x_i - x_j\|^2$ and the grid points x_i are uniformly spread, the matrix vector products involving* $\exp(-C/\varepsilon)$ *within the Sinkhorn algorithm can be efficiently performed via separated one dimensional convolutions [32] without storing C.*

As discussed in the introduction, an important issue regarding the use of Sinkhorn divergence for testing the equality of two distributions is that it leads to a biased statistics in the sense that $W_{p,\varepsilon}^p(a,b)$ is not equal to zero under the null hypothesis $a = b$. A possible alternative to avoid this issue is to consider the so-called notion of Sinkhorn loss [14, 19] as defined below.

Definition 2.3 (Sinkhorn loss)**.** Let $\varepsilon > 0$ be a regularization parameter. The Sinkhorn loss between two probability measures $\mu = \sum_{i=1}^N a_i \delta_{x_i}$ and $\nu = \sum_{i=1}^N b_i \delta_{x_i}$ in $\mathcal{P}_p(\mathcal{X})$ is defined by

$$\overline{W}_{p,\varepsilon}^p(a,b) := W_{p,\varepsilon}^p(a,b) - \frac{1}{2}\left(W_{p,\varepsilon}^p(a,a) + W_{p,\varepsilon}^p(b,b)\right). \tag{5}$$

The Sinkhorn loss is not a distance between probability distributions, but it satisfies various interesting properties for the purpose of this paper, that are summarized below.

Proposition 2.2. *The Sinkhorn loss satisfies the following three key properties (see Theorem 1 in [14]):*

(i) $\overline{W}_{p,\varepsilon}^p(a,b) \geq 0,$
(ii) $\overline{W}_{p,\varepsilon}^p(a,b) = 0 \Leftrightarrow a = b,$
(iii) $\overline{W}_{p,\varepsilon}^p(a,b) \underset{\varepsilon\to 0}{\longrightarrow} W_p^p(a,b).$

From Proposition 2.2, we have that $a = b$ is equivalent to $\overline{W}_{p,\varepsilon}^p(a,b) = 0$, therefore the function $(a,b) \mapsto \overline{W}_{p,\varepsilon}^p(a,b)$ reaches its global minimum at $a = b$, implying that the gradient of the Sinkhorn loss is zero when $a = b$ which is summarized in the following corollary.

Corollary 2.4. *For any $a \in \Sigma_N$, the gradient of the Sinkhorn loss satisfies $\nabla \overline{W}_{p,\varepsilon}^p(a,a) = 0$.*

2.1.2. Statistical notations

We denote by $\xrightarrow{\mathcal{L}}$ the convergence in distribution of a random variable and $\xrightarrow{\mathbb{P}}$ the convergence in probability. The notation $G \overset{\mathcal{L}}{\sim} a$ means that G is a random variable taking its values in $\mathcal{X}$ with law $a = (a_1, \dots, a_n) \in \Sigma_N$ (namely that $\mathbb{P}(G = x_i) = a_i$ for each $1 \leq i \leq N$). Likewise $G \overset{\mathcal{L}}{\sim} H$ stands for the equality in distribution of the random variables G and H.

Let $a, b \in \Sigma_N$ and $\hat{a}_n$ and $\hat{b}_m$ be the empirical measures respectively generated by iid samples $X_1, \dots, X_n \overset{\mathcal{L}}{\sim} a$ and $Y_1, \dots, Y_m \overset{\mathcal{L}}{\sim} b$, that is

$$\begin{aligned} &\hat{a}_n = (\hat{a}_n^x)_{x \in \mathcal{X}}, \\ &\text{where } \hat{a}_n^{x_i} = \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{X_j = x_i\}} = \frac{1}{n} \#\{j : X_j = x_i\} \text{ for all } 1 \leq i \leq N. \end{aligned} \tag{6}$$

We also define the multinomial covariance matrix

$$\Sigma(a) = \begin{bmatrix} a_{x_1}(1-a_{x_1}) & -a_{x_1}a_{x_2} & \cdots & -a_{x_1}a_{x_N} \\ -a_{x_1}a_{x_2} & a_{x_2}(1-a_{x_2}) & \cdots & -a_{x_2}a_{x_N} \\ \vdots & \vdots & \ddots & \vdots \\ -a_{x_1}a_{x_N} & -a_{x_2}a_{x_N} & \cdots & a_{x_N}(1-a_{x_N}) \end{bmatrix}$$

and the independent Gaussian random vectors $G \sim \mathcal{N}(0, \Sigma(a))$ and $H \sim \mathcal{N}(0, \Sigma(b))$. As classically done in statistics, we say that

$$\begin{cases} H_0 : a = b \text{ is the null hypothesis,} \\ H_1 : a \neq b \text{ is the alternative hypothesis.} \end{cases}$$

Remark 3. *As stated in Proposition 2.1, the dual variables $(u_\varepsilon^*, v_\varepsilon^*)$ solutions of (3) for a and b in the simplex are unique up to a scalar addition. Hence for any $t \in \mathbb{R}$,*

$$\langle G, u_\varepsilon^* + t\mathbb{1}_N \rangle \overset{\mathcal{L}}{\sim} \langle G, u_\varepsilon^* \rangle,$$

since G is centered in 0 and $\mathbb{1}_N' \Sigma(a) \mathbb{1}_N = 0$ for a in the simplex.

2.1.3. Notations for differentiation

For a sufficiently smooth function $f : (x, y) \in \mathbb{R}^N \times \mathbb{R}^N \longmapsto \mathbb{R}$, we denote by ∇f and $\nabla^2 f$ the gradient and the hessian of the function f. In particular, the gradient of f at the point $(x, y) \in \mathbb{R}^N \times \mathbb{R}^N$ in the direction $(h_1, h_2) \in \mathbb{R}^N \times \mathbb{R}^N$ is denoted by $\nabla f(x, y)(h_1, h_2)$ (this notation also holds for the hessian). Moreover, the first-order partial derivative with respect to the first variable x (resp. y) is given by $\partial_1 f$ (resp. $\partial_2 f$). Equivalently, the second-order partial derivative is denoted $\partial_{ij}^2 f$, with $i \in \{1, 2\}, j \in \{1, 2\}$.

2.2. Differentiability of $W^p_{p,\varepsilon}$

As stated at the beginning of the section, the differentiability of $W^p_{p,\varepsilon}$ (in the usual Fréchet sense) is needed in order to apply the delta-method. This is proved in the following proposition.

Proposition 2.3. *The functional $(a,b) \mapsto W^p_{p,\varepsilon}(a,b)$ is differentiable in $\Sigma_N \times \Sigma_N$ with gradient*

$$\nabla W^p_{p,\varepsilon}(a,b)(h_1,h_2) = \langle u_\varepsilon, h_1\rangle + \langle v_\varepsilon, h_2\rangle,$$

where $(u_\varepsilon, v_\varepsilon) \in \mathcal{S}_\varepsilon(a,b)$, the set of optimal solutions of (3).

Proof. From Proposition 2 in [14], $W^p_{p,\varepsilon}$ is Gâteaux differentiable and its derivative reads

$$\nabla W^p_{p,\varepsilon}(a,b)(h_1,h_2) = \langle u_\varepsilon, h_1\rangle + \langle v_\varepsilon, h_2\rangle,$$

for $(u_\varepsilon, v_\varepsilon) \in \mathcal{S}_\varepsilon(a,b)$. In order to prove its differentiability (or Fréchet differentiability, since $\mathbb{R}^N$ is a finite dimensional space) at the point (a,b), we only need to prove that the operator $\nabla W^p_{p,\varepsilon}$ is continuous (see *e.g.* Prop. 3.2.3. in [41]) in (a,b). Suppose that (a^n,b^n) tends to (a,b) when n tends to infinity. Therefore, this convergence also holds in the weak* topology for the probability measures $\mu^n = \sum_{i=1}^N a_i^n \delta_{x_i}, \nu^n = \sum_{i=1}^N b_i^n \delta_{x_i}$ and $\mu = \sum_{i=1}^N a_i \delta_{x_i}, \nu = \sum_{i=1}^N b_i \delta_{x_i}$. We denote by (u^n, v^n) the unique couple in $\mathcal{S}_\varepsilon(a^n,b^n)$ such that for an arbitrary $i_0 \in \{1,\dots,N\}, u^n_{i_0} = 0$. Then, we can apply Cauchy-Schwarz inequality and then use Proposition 13 in [14] on the convergence of the pair (u^n,v^n) of dual variables towards $(u,v) \in \mathcal{S}_\varepsilon(a,b)$ (such that $u_{i_0} = 0$), to obtain that

$$\begin{aligned}
\lim_{(a^n,b^n)\to(a,b)} &\|\nabla W^p_{p,\varepsilon}(a^n,b^n) - \nabla W^p_{p,\varepsilon}(a,b)\| \\
&= \lim_{(a^n,b^n)\to(a,b)} \sup_{\|(h_1,h_2)\|\le 1} |\langle u^n - u, h_1\rangle + \langle v^n - v, h_2\rangle| \\
&\le \lim_{(a^n,b^n)\to(a,b)} \sup_{\|(h_1,h_2)\|\le 1} \|u^n - u\|\;\|h_1\| + \|v^n - v\|\;\|h_2\| \\
&\underset{(a^n,b^n)\to(a,b)}{\longrightarrow} 0,
\end{aligned}$$

which concludes the proof. □

2.3. Distributional limits for the empirical Sinkhorn divergence

2.3.1. Convergence in distribution

The following theorem is our main result on distributional limits of empirical Sinkhorn divergences.

Theorem 2.5. *For $a,b \in \Sigma_N$, let $(u_\varepsilon, v_\varepsilon) \in \mathcal{S}_\varepsilon(a,b)$ be an optimal solution of the dual problem (3) and $\hat{a}_n, \hat{b}_m$ be the empirical measures defined in (6). Then, the following central limit theorems hold for empirical Sinkhorn divergences.*

1. *One sample. As $n \to +\infty$, one has that*

$$\sqrt{n}(W_{p,\varepsilon}^{p}(\hat{a}_n, b) - W_{p,\varepsilon}^{p}(a,b)) \xrightarrow{\mathcal{L}} \langle G, u_\varepsilon \rangle. \tag{7}$$

2. *Two samples. For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \gamma \in (0,1)$ as $\min(n,m) \to +\infty$, one has that*

$$\rho_{n,m}(W_{p,\varepsilon}^{p}(\hat{a}_n, \hat{b}_m) - W_{p,\varepsilon}^{p}(a,b)) \xrightarrow{\mathcal{L}} \sqrt{\gamma}\langle G, u_\varepsilon \rangle + \sqrt{1-\gamma}\langle H, v_\varepsilon \rangle. \tag{8}$$

Proof. Following the proof of Theorem 1 in [33], we have that (see e.g. Theorem 14.6 in [38])

$$\sqrt{n}(\hat{a}_n - a) \xrightarrow{\mathcal{L}} G, \text{ where } G \overset{\mathcal{L}}{\sim} \mathcal{N}(0, \Sigma(a)),$$

since $n\hat{a}_n$ is a sample of a multinomial probability measure with probability a. For the two samples case, we use that

$$\rho_{n,m}((\hat{a}_n, \hat{b}_m) - (a,b)) \xrightarrow{\mathcal{L}} (\sqrt{\gamma}G, \sqrt{1-\gamma}H),$$

where $\rho_{n,m}$ and γ are the quantities defined in the statement of Theorem 2.5. From Proposition 2.3, we can directly apply the delta-method:

$$\sqrt{n}(W_{p,\varepsilon}^{p}(\hat{a}_n, b) - W_{p,\varepsilon}^{p}(a,b)) \xrightarrow{\mathcal{L}} \langle G, u_\varepsilon \rangle, \text{ as } n \to +\infty, \tag{9}$$

while, for n and m tending to infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \gamma \in (0,1)$, we obtain that

$$\rho_{n,m}(W_{p,\varepsilon}^{p}(\hat{a}_n, \hat{b}_m) - W_{p,\varepsilon}^{p}(a,b)) \xrightarrow{\mathcal{L}} \sqrt{\gamma}\langle G, u_\varepsilon \rangle + \sqrt{1-\gamma}\langle H, v_\varepsilon \rangle. \tag{10}$$

This completes the proof of Theorem 2.5. □

2.3.2. *Convergence in probability*

Distributional limits of empirical Sinkhorn divergences may also be characterized by a convergence in probability by the following result which directly follows from the delta-method (see *e.g.* Theorem 3.9.4 in [36]).

Theorem 2.6. *The following asymptotic results hold for empirical Sinkhorn divergences, for any $(u_\varepsilon, v_\varepsilon) \in \mathcal{S}_\varepsilon(a,b)$.*

1. *One sample. As $n \to +\infty$, one has that*

$$\sqrt{n}\left(W_{p,\varepsilon}^{p}(\hat{a}_n, b) - W_{p,\varepsilon}^{p}(a,b) - \langle \hat{a}_n - a, u_\varepsilon \rangle\right) \xrightarrow{\mathbb{P}} 0.$$

2. *Two samples – For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \gamma \in (0,1)$ as $\min(n,m) \to +\infty$, one has that*

$$\rho_{n,m}\left(W_{p,\varepsilon}^{p}(\hat{a}_n, \hat{b}_m) - W_{p,\varepsilon}^{p}(a,b) - (\langle \hat{a}_n - a, u_\varepsilon \rangle + \langle \hat{b}_m - b, v_\varepsilon \rangle)\right) \xrightarrow{\mathbb{P}} 0.$$

Proof. As the map $(h_1, h_2) \mapsto \nabla W_{p,\varepsilon}^{p}(a,b)(h_1,h_2)$ is defined, linear and continuous on $\mathbb{R}^N \times \mathbb{R}^N$, Theorem 3.9.4 in [36] allows us to conclude. □

2.4. Distributional limits for the empirical Sinkhorn loss

2.4.1. Convergence in distribution

The following theorems are our main results on distributional limits of the empirical Sinkhorn loss, for which we now distinguish the cases $a \neq b$ (alternative hypothesis) and $a = b$ (null hypothesis).

Theorem 2.7. *Let $a \neq b$ be two probability distributions in Σ_N. Let us denote by $\hat{a}_n, \hat{b}_m$ their empirical counterparts and by $(u_\varepsilon^{a,b}, v_\varepsilon^{a,b}) \in \mathcal{S}_\varepsilon(a,b)$ the dual variables which are the optimal solutions of the dual problem (3). Then, the following asymptotic results hold.*

1. *One sample. As $n \to +\infty$, one has that*

$$\sqrt{n}(\overline{W}_{p,\varepsilon}^p(\hat{a}_n, b) - \overline{W}_{p,\varepsilon}^p(a,b)) \xrightarrow{\mathcal{L}} \langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a})\rangle. \tag{11}$$

2. *Two samples. For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \gamma \in (0,1)$ as $\min(n,m) \to +\infty$, one has that*

$$\rho_{n,m}(\overline{W}_{p,\varepsilon}^p(\hat{a}_n, \hat{b}_m) - \overline{W}_{p,\varepsilon}^p(a,b)) \xrightarrow{\mathcal{L}} \sqrt{\gamma}\langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a})\rangle$$
$$+\sqrt{1-\gamma}\langle H, v_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{b,b} + v_\varepsilon^{b,b})\rangle.$$

Proof. The only difference with the proof of Theorem 2.5 is the computation of the gradient of $\overline{W}_{p,\varepsilon}^p$, which is given by

$$\nabla\overline{W}_{p,\varepsilon}^p(a,b)(h_1,h_2) = \langle u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a}), h_1\rangle + \langle v_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{b,b} + v_\varepsilon^{b,b}), h_2\rangle. \tag{12}$$

The proof of Theorem 2.7 then follows from the same arguments as those used in the proof of Theorem 2.5. □

Under the null hypothesis $a = b$, the derivation of the distributional limit of either $\overline{W}_{p,\varepsilon}^p(\hat{a}_n, a)$ or $\overline{W}_{p,\varepsilon}^p(\hat{a}_n, \hat{b}_m)$ requires further attention. Indeed, thanks to Proposition 2.2, one has that the function $(a,b) \mapsto \overline{W}_{p,\varepsilon}^p(a,b)$ reaches its global minimum at $a = b$, and therefore the gradient of the Sinkhorn loss satisfies $\nabla\overline{W}_{p,\varepsilon}^p(a,a) = 0$. Hence, to obtain the distributional limit of the empirical Sinkhorn loss it is necessary to apply a second-order delta-method yielding an asymptotic distribution which is not Gaussian anymore.

Theorem 2.8. *Let $a = b$ be a probability distribution on Σ_N, and denote by $\hat{a}_n$ an empirical measures obtained by independent sampling data from a. Then, as n tends to infinity, the following asymptotic result holds*

$$n\overline{W}_{p,\varepsilon}^p(\hat{a}_n, a) \xrightarrow{\mathcal{L}} \frac{1}{2}\sum_{i=1}^{N} \lambda_i \chi_i^2(1) \tag{13}$$

where $\lambda_1, \dots, \lambda_N$ are the non-negative eigenvalues of the matrix

$$\Sigma(a)^{1/2} \partial^2_{11} \overline{W}^p_{p,\varepsilon}(a,a) \Sigma(a)^{1/2},$$

and $\chi^2_1(1), \dots, \chi^2_N(1)$ are independent random variables with chi-squared distribution of degree 1.

Proof. From Corollary 2.4, we have that $\nabla \overline{W}^p_{p,\varepsilon}(a,a) = 0$. In order to apply a second order delta-method, the Hessian matrix $\nabla^2 \overline{W}^p_{p,\varepsilon}(a,b)$ of the Sinkhorn loss $\overline{W}^p_{p,\varepsilon}(a,b)$ needs to be non-singular in the neighborhood of $a = b$. Note that the Sinkhorn loss is at least C^3 (admitting a third continuous differential) on the interior of its domain, as proved in Theorem 2 by [23]. Moreover, since the function $a \mapsto W^p_{p,\varepsilon}(a,b)$ is ε-strongly convex (Theorem 3.4, [2]) and $a \mapsto -\frac{1}{2} W^p_{p,\varepsilon}(a,a)$ is (strictly) convex (Proposition 4, [14]), we have that the Hessian matrix of $a \mapsto \overline{W}^p_{p,\varepsilon}(a,b)$ is non-singular. We can thus apply Theorem 17 in [34] which states that from second order delta-method, the distributional limits of $n\overline{W}^p_{p,\varepsilon}(\hat{a}_n, a)$ is given by

$$\frac{1}{2} \mathcal{N}(0, \Sigma(a))^T \partial^2_{11} \overline{W}^p_{p,\varepsilon}(a,a) \mathcal{N}(0, \Sigma(a))$$

that can be rewritten as

$$\frac{1}{2} \sum_{i=1}^N \lambda_i \chi^2_i(1),$$

where $\lambda_1, \dots, \lambda_N$ are the eigenvalues of the matrix $\Sigma(a)^{1/2} \partial^2_{11} \overline{W}^p_{p,\varepsilon}(a,a) \Sigma(a)^{1/2}$. This concludes the distributional limit presented in relation (13). □

In the two samples case, the Hessian matrix is not guaranteed to be non-singular, in which case the asymptotic distribution is degenerated. Nevertheless, we have the following theorem.

Theorem 2.9. *Let $a = b$ be a probability distribution on Σ_N, and denote by $\hat{a}_n, \tilde{a}_m$ two empirical measures obtained by independent sampling data from a. Then, let us write the Hessian matrix*

$$\nabla^2 \overline{W}^p_{p,\varepsilon}(a,b) = \begin{pmatrix} A & B \\ B & C \end{pmatrix},$$

with $A = \partial^2_{11} \overline{W}^p_{p,\varepsilon}(a,b), C = \partial^2_{22} \overline{W}^p_{p,\varepsilon}(a,b)$ and $B = \partial^2_{12} W^p_{p,\varepsilon}(a,b)$. If its Schur complement $S = C - B^T A^{-1} B$ is non-singular in a neighborhood of $a = b$, then one has for $m/(n+m) \to \gamma \in (0,1)$ as $\min(n,m) \to +\infty$

$$\frac{nm}{n+m} \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \tilde{a}_m) \xrightarrow{\mathcal{L}} \frac{1}{2} \sum_{i=1}^N \tilde{\lambda}_i \chi^2_i(1), \tag{14}$$

where $\tilde{\lambda}_1, \dots, \tilde{\lambda}_N$ are the eigenvalues of the matrix of size $\mathbb{R}^{2N} \times \mathbb{R}^{2N}$ given by

$$(\sqrt{\gamma}\Sigma(a)^{1/2}, \sqrt{1-\gamma}\Sigma(a)^{1/2}) \nabla^2 \overline{W}^p_{p,\varepsilon}(a,a) (\sqrt{\gamma}\Sigma(a)^{1/2}, \sqrt{1-\gamma}\Sigma(a)^{1/2}),$$

and $\chi^2_1(1), \dots, \chi^2_N(1)$ are independent random variables with chi-squared distribution of degree 1*.*

Proof. As in the proof of Theorem 2.8, we have that both A and C are ε-strongly convex and therefore non-singular. The determinant $\det(\nabla^2 \overline{W}^p_{p,\varepsilon}(a,b)) = \det(A)$ $\det(S)$ is therefore non-zero in a neighborhood of $a = b$ if and only if the Schur complement S is invertible in a neighborhood of $a = b$. Therefore, applying Theorem 17 in [34] as in the one sample case, we obtain the distributional limit (14). This completes the proof of Theorem 2.9. □

Remark 4. *A sufficient condition to ensure the non-singularity of the Schur matrix S comes from the ε-strong convexity of A and C, implying that for any $x \in \mathbb{R}^N$*

$$x^T S x = x^T C x - x^T B^T A B x > \varepsilon \|x\|^2 - \varepsilon^{-1} \|Bx\|^2.$$

A sufficient condition for the non-singularity of S is therefore $\varepsilon > \sup_{x \in \mathbb{R}^N} \frac{\|Bx\|}{\|x\|}$ at the points $a = b$. Remark that since the global minimum is attained in the critical points $a = b$, we have that the Hessian $\overline{W}^p_{p,\varepsilon}$ is symmetric semi-definite positive at these points. Therefore its Schur complement $S = C - B^T A^{-1} B$ is also semi-definite positive (see e.g. Section A.5.5. in [4]).

2.4.2. Convergence in probability

Limits for empirical Sinkhorn loss can again be established from a corollary of the Delta-method as done in Theorem 2.6.

Theorem 2.10. *Using the same notations as introduced in the statement of Theorem 2.7, the following asymptotic results hold for all $a, b \in \Sigma_N$.*

1. *One sample. As $n \to +\infty$, one has that*

$$\sqrt{n}\left(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a, b) - \langle \hat{a}_n - a, u^{a,b}_\varepsilon - \frac{1}{2}(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle\right) \xrightarrow{\mathbb{P}} 0.$$

2. *Two samples – For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \gamma \in (0,1)$ as $\min(n,m) \to +\infty$, one has that*

$$\begin{aligned}\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) - &\overline{W}^p_{p,\varepsilon}(a, b) - (\sqrt{\gamma}\langle \hat{a}_n - a, u^{a,b}_\varepsilon - \frac{1}{2}(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle \\ &+ \sqrt{1-\gamma}\langle \hat{b}_m - b, v^{a,b}_\varepsilon - \frac{1}{2}(u^{b,b}_\varepsilon + v^{b,b}_\varepsilon)\rangle)) \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

Note that in the case $a = b$, this simplifies into

$$\begin{aligned}&\sqrt{n}\overline{W}^p_{p,\varepsilon}(\hat{a}_n, a) \xrightarrow{\mathbb{P}} 0 \\ &\rho_{n,m}\overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) \xrightarrow{\mathbb{P}} 0.\end{aligned}$$

2.5. Link with unregularized optimal transport

A natural question that arises is the behavior of distributional limits when we let ε tends to 0 at an appropriate rate depending on the sample size. Under such conditions, we recover the distributional limit given by Theorem 1 in Sommerfeld and Munk [33] in the setting of unregularized optimal transport.

Theorem 2.11. *Suppose that $\mathcal{X} \subset \mathbb{R}^q$, and consider the cost matrix C such that $c_{ij} = \|x_i - x_j\|^p$ where $\|\cdot\|$ stands for the Euclidean norm. We recall that $\mathcal{S}_0(a,b) \subset \mathbb{R}^N \times \mathbb{R}^N$ is the set of optimal solutions of the dual problem (3) for $\varepsilon = 0$.*

1. *One sample. Suppose that $(\varepsilon_n)_{n\geq 1}$ is a sequence of positive reals tending to zero such that*
$$\lim_{n\to+\infty} \sqrt{n}\varepsilon_n \log(1/\varepsilon_n) = 0. \tag{15}$$
Then, we have that
$$\sqrt{n}(\overline{W}^p_{p,\varepsilon_n}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon_n}(a,b)) \xrightarrow{\mathcal{L}} \max_{(u,v)\in\mathcal{S}_0(a,b)} \langle G, u\rangle. \tag{16}$$
2. *Two samples. Suppose that $(\varepsilon_{n,m})$ is a sequence of positive reals tending to zero as $\min(n,m) \to +\infty$ such that*
$$\lim_{\min(n,m)\to+\infty} \sqrt{\rho_{n,m}}\varepsilon_{n,m} \log(1/\varepsilon_{n,m}) = 0, \tag{17}$$
for $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \gamma \in (0,1)$. Then, one has that
$$\rho_{n,m}(\overline{W}^p_{p,\varepsilon_{n,m}}(\hat{a}_n, \hat{b}_m) - \overline{W}^p_{p,\varepsilon_{n,m}}(a,b)) \xrightarrow{\mathcal{L}} \max_{(u,v)\in\mathcal{S}_0(a,b)} \sqrt{\gamma}\langle G, u\rangle + \sqrt{1-\gamma}\langle H, v\rangle. \tag{18}$$

Proof. We will only prove the one sample case as both proofs work similarly. For that purpose, let us consider the decomposition

$$\begin{aligned}\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a,b)) &= \sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - W^p_p(\hat{a}_n, b)) \\ &+ \sqrt{n}(W^p_p(\hat{a}_n, b) - W^p_p(a,b)) + \sqrt{n}(W^p_p(a,b) - \overline{W}^p_{p,\varepsilon}(a,b)).\end{aligned} \tag{19}$$

From Theorem 1 in [33], we have that

$$\sqrt{n}(W^p_p(\hat{a}_n, b) - W^p_p(a,b)) \xrightarrow{\mathcal{L}} \max_{(u,v)\in\mathcal{S}_0(a,b)} \langle G, u\rangle. \tag{20}$$

Since $\mathcal{X}$ is a finite set, it follows that the cost c is a L-Lipschitz function separately in $x \in \mathcal{X}$ and $y \in \mathcal{X}$ with respect to the Euclidean distance. Therefore, it satisfies the assumptions of Theorem 1 in [17] that gives a bound on the error between the Sinkorn divergence and the unregularized transport for a given pair of distributions. It follows that for any $a, b \in \Sigma_N$ (possibly random),

$$0 \leq W^p_{p,\varepsilon}(a,b) - W^p_p(a,b) \leq 2\varepsilon q \log\left(\frac{e^2 L\,\mathrm{diam}(\mathcal{X})}{\varepsilon\sqrt{q}}\right)$$

where q is the dimension of the support space, and $\mathrm{diam}(\mathcal{X})$ is the diameter of $\mathcal{X}$ (i.e. $\mathrm{diam}(\mathcal{X}) = \sup_{x,y\in\mathcal{X}} \|x-y\|$) which is always finite in the discrete case. Then, as soon as the sequence $(\varepsilon_n)_{n\geq 1}$ satisfies (15), we obtain that

$$\sup_{(a,b)\in\Sigma_N\times\Sigma_N} \sqrt{n}(W^p_{p,\varepsilon_n}(a,b) - W^p_p(a,b)) \xrightarrow[n\to\infty]{} 0. \tag{21}$$

By definition of the Sinkhorn loss, one has that $W^p_{p,\varepsilon}(a,b) - \overline{W}^p_{p,\varepsilon}(a,b) = \frac{1}{2}\left(W^p_{p,\varepsilon}(a,a) + W^p_{p,\varepsilon}(b,b)\right)$. Therefore, using the upper bound (21), we get

$$\sqrt{n}(\overline{W}^p_{p,\varepsilon_n}(\hat{a}_n,b) - W^p_p(\hat{a}_n,b) \xrightarrow[n\to\infty]{\text{a.s.}} 0 \text{ and } \sqrt{n}(W^p_p(a,b) - \overline{W}^p_{p,\varepsilon_n}(a,b)) \xrightarrow[n\to\infty]{\text{a.s.}} 0. \tag{22}$$

Combining (19) with (20) and (22), and using Slutsky's theorem allow to complete the proof of Theorem 2.11. □

3. Use of the bootstrap for statistical inference

The results obtained in Section 2 on the distribution of the empirical Sinkhorn divergence and Sinkhorn loss are only asymptotic. It is thus of interest to estimate their non-asymptotic distribution using a bootstrap procedure. The bootstrap consists in drawing new samples from an empirical distribution $\hat{\mathbb{P}}_n$ that has been obtained from an unknown distribution $\mathbb{P}$. Therefore, conditionally on $\hat{\mathbb{P}}_n$, it allows to obtain new observations (considered as approximately sampled from $\mathbb{P}$) that can be used to approximate the distribution of a test statistics using Monte-Carlo experiments. We refer to [13] for a general introduction to the bootstrap procedure.

We can apply the delta-method to prove the consistency of the bootstrap in our setting using the bounded Lipschitz metric as defined below.

Definition 3.1. The Bounded Lipschitz (BL) metric between two probability measures μ, ν supported on Ω is defined by

$$d_{BL}(\mu,\nu) = \sup_{h\in BL_1(\Omega)} \int_\Omega h d(\mu - \nu)$$

where $BL_1(\Omega)$ is the set of real functions $\Omega \to \mathbb{R}$ such that $\|h\|_\infty + \|h\|_{\mathrm{Lip}} \leq 1$.

Our main result on the consistency of bootstrap samples can then be stated. Notice that similar results for the Sinkhorn divergence are obtained straightforward using the same arguments.

Theorem 3.2. *Let $a \neq b$ be in the simplex Σ_N. For $X_1, \ldots, X_n \overset{\mathcal{L}}{\sim} a$ and $Y_1, \ldots, Y_m \overset{\mathcal{L}}{\sim} b$, let $\hat{a}_n^*$ (resp. $\hat{b}_m^*$) be a bootstrap empirical distribution sampled from $\hat{a}_n$ (resp. $\hat{b}_m$) of size n (resp. m).*

1. *One sample case: $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n^*, b) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, b))$ converges in distribution (conditionally on $X_1, \ldots, X_n$) to $\langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a})\rangle$ for the BL metric, in the sense that*

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}[h(\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n^*, b) - W^p_{p,\varepsilon}(\hat{a}_n, b)))|X_1, \ldots, X_n] - \mathbb{E}[h\langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a}))\rangle]| \xrightarrow{\mathbb{P}} 0.$$

2. *Two samples case: $\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n^*, \hat{b}_m^*) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m))$ converges in distribution (conditionally on $X_1, \ldots, X_n$, $Y_1, \ldots, Y_m$) to*

$$\sqrt{\gamma}\langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a})\rangle + \sqrt{1-\gamma}\langle H, v_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{b,b} + v_\varepsilon^{b,b})\rangle$$

for the BL metric, in the sense that

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}[h(\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n^*, \hat{b}_m^*) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m)))|X_1, \ldots, X_n, Y_1, \ldots, Y_m] - \mathbb{E}[h(\sqrt{\gamma}\langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a})\rangle + \sqrt{1-\gamma}\langle H, v_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{b,b} + v_\varepsilon^{b,b})\rangle)]| \xrightarrow{\mathbb{P}} 0$$

Proof. We only prove the one sample case since the convergence for the two samples case can be shown with similar arguments. We know that $\sqrt{n}(\hat{a}_n - a)$ tends in distribution to $G \sim \mathcal{N}(0, \Sigma(a))$. Moreover $\sqrt{n}(\hat{a}_n^* - \hat{a}_n)$ converges (conditionally on $X_1, \ldots, X_n$) in distribution to G by Theorem 3.6.1 in [36]. Then, applying Theorem 3.9.11 in [36] on the consistency of the delta-method combined with the bootstrap allows us to obtain the statement of the present Theorem 3.2 in the case $a \neq b$. □

As explained in [6], the standard bootstrap fails under first order degeneracy, meaning for the null hypothesis case $a = b$. However, the authors propose a corrected version – called the Babu correction – of the bootstrap in their Theorem 3.2 given for the one sample case by

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}[h(n\{\overline{W}^p_{p,\varepsilon}(\hat{a}_n^*, a) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, a) - \partial_1 \overline{W}^p_{p,\varepsilon}(\hat{a}_n, a)(\hat{a}_n^* - \hat{a}_n, a)\}))|X_1, \ldots, X_n] - \mathbb{E}[h(\partial^2_{11} \overline{W}^p_{p,\varepsilon}(a, a)(G, a)]| \xrightarrow{\mathbb{P}} 0,$$

and for the two samples case by

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}[h(n\{\overline{W}^p_{p,\varepsilon}(\hat{a}_n^*, \hat{b}_m^*) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m)$$

$$- \nabla \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m)(\hat{a}^*_n - \hat{a}_n, \hat{b}^*_m - \hat{b}_m)\}))|X_1, \dots, X_n] \\ - \mathbb{E}[h(\nabla^2 \overline{W}^p_{p,\varepsilon}(a,a)(\sqrt{\gamma}G, \sqrt{1-\gamma}G)]| \xrightarrow{\mathbb{P}} 0.$$

Note that most of the requirements to apply Theorem 3.2 in [6] are trivial since the distributions are defined on a subset of $\mathbb{R}^N$ and the function $(a,b) \mapsto W^p_{p,\varepsilon}(a,b)$ is twice differentiable on all $\Sigma_N \times \Sigma_N$. However, the (*Assumption 3.3 in [6]*) on the second derivative requires a finer study that is left for future work. Hence, we stress that the Babu-bootstrap approach that we use in our numerical experiments is missing theoretical guarantees. Nevertheless, the results reported from our experiments on simulated and real data illustrate its correctness.

As

$$\partial_1 \overline{W}^p_{p,\varepsilon}(\hat{a}_n, a)(\hat{a}^*_n - \hat{a}_n, a) = \langle u^{\hat{a}_n, a}, \hat{a}^*_n - \hat{a}_n \rangle$$
$$\nabla \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m)(\hat{a}^*_n - \hat{a}_n, \hat{b}^*_m - \hat{b}_m) = \langle u^{\hat{a}_n, \hat{b}_m}, \hat{a}^*_n - \hat{a}_n \rangle + \langle v^{\hat{a}_n, \hat{b}_m}, \hat{b}^*_m - \hat{b}_m \rangle$$

we can reformulate the Babu bootstrap as follows.

1. One sample case. For $(u^{\hat{a}_n,a}_\varepsilon, v^{\hat{a}_n,a}_\varepsilon) \in \mathcal{S}_\varepsilon(\hat{a}_n, a)$, we have that
$$n \left\{ \overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, a) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, a) - \langle u^{\hat{a}_n,a}, \hat{a}^*_n - \hat{a}_n \rangle \right\} \tag{23}$$
converges in distribution (conditionally on $X_1, \dots, X_n$,) to $\partial^2_{11} \overline{W}^p_{p,\varepsilon}(a,a)(G,a)$ for the BL metric.
2. Two samples case. For $(u^{\hat{a}_n,\hat{b}_m}_\varepsilon, v^{\hat{a}_n,\hat{b}_m}_\varepsilon) \in \mathcal{S}_\varepsilon(\hat{a}_n, \hat{b}_m)$ and $m/(n+m) \to \gamma \in (0,1)$, the quantity
$$\frac{nm}{n+m} \Big\{ \overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) \\ -(\langle u^{\hat{a}_n,\hat{b}_m}, \hat{a}^*_n - \hat{a}_n \rangle + \langle v^{\hat{a}_n,\hat{b}_m}, \hat{b}^*_m - \hat{b}_m \rangle) \Big\} \tag{24}$$
converges in distribution (conditionally on $X_1, \dots, X_n$, $Y_1, \dots, Y_m$) to
$$\nabla^2 \overline{W}^p_{p,\varepsilon}(a,a)(\sqrt{\gamma}G, \sqrt{1-\gamma}G)$$
for the BL metric.

4. Numerical experiments with synthetic data

We propose to illustrate Theorem 2.7, Theorem 2.8, Theorem 2.9 and Theorem 3.2 with simulated data consisting of random measures supported on a $l \times l$ square lattice (of regularly spaced points) $(x_i)_{i=1,\dots,N}$ in $\mathbb{R}^2$ (with $N = l^2$) for l ranging from 5 to 20. We use the squared Euclidean distance as the cost function C which therefore scales with the size of the grid. The range of interesting values for ε is thus closely linked to the size of the grid, as it can be seen in the expression of $K = \exp(-C/\varepsilon - \mathbb{1}_{N\times N})$. Hence, $\varepsilon = 100$ for a 5×5 grid corresponds to more regularization than $\varepsilon = 100$ for a 20×20 grid.

We ran our experiments on Matlab using the accelerated version [35][1] of the Sinkhorn transport algorithm [7]. Furthermore, we considered the numerical logarithmic stabilization described in [30] which allows to handle relatively small values of ε. Indeed, in small regularization regimes, the Sinkhorn algorithm quickly becomes unstable, even more for large grids with a small number of observations.

4.1. Convergence in distribution

We first illustrate the convergence in distribution of the empirical Sinkhorn loss (as stated in Theorem 2.7) for the hypothesis $a \neq b$ with either one sample or two samples.

4.1.1. Alternative $a \neq b$ – one sample

We consider the case where a is the uniform distribution on a square grid and

$$b \propto \mathbb{1}_N + \theta(1, 2, \dots, N)$$

is a distribution with linear trend depending on a slope parameter $\theta \geq 0$ that is fixed to 0.5, see Figure 1.

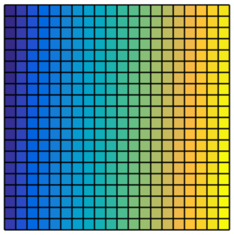


Fig 1. *Example of a distribution b with linear trend (with slope parameter $\theta = 0.5$ on a 20×20 grid).*

We generate $M = 10^3$ empirical distributions $\hat{a}_n$ (such that $n\hat{a}_n$ follows a multinomial distribution with parameter a) for different values of n and grid size. In this way, we obtain M realizations of $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a, b))$, and we use a kernel density estimate (with a data-driven bandwidth) to compare the distribution of these realizations to the density of the Gaussian distribution $\langle G, u^{a,b}_\varepsilon - 1/2(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle$. The results are reported in Figure 2 (grid 5×5) and Figure 3 (grid 20×20). It can be seen that the convergence of the empirical Sinkhorn loss to its asymptotic distribution ($n \to \infty$) is relatively fast.

Let us now shed some light on the bootstrap procedure. The results on bootstrap experiments are reported in Figure 4. From the uniform distribution

[1] http://www.math.u-bordeaux.fr/~npapadak/GOTMI/codes.html

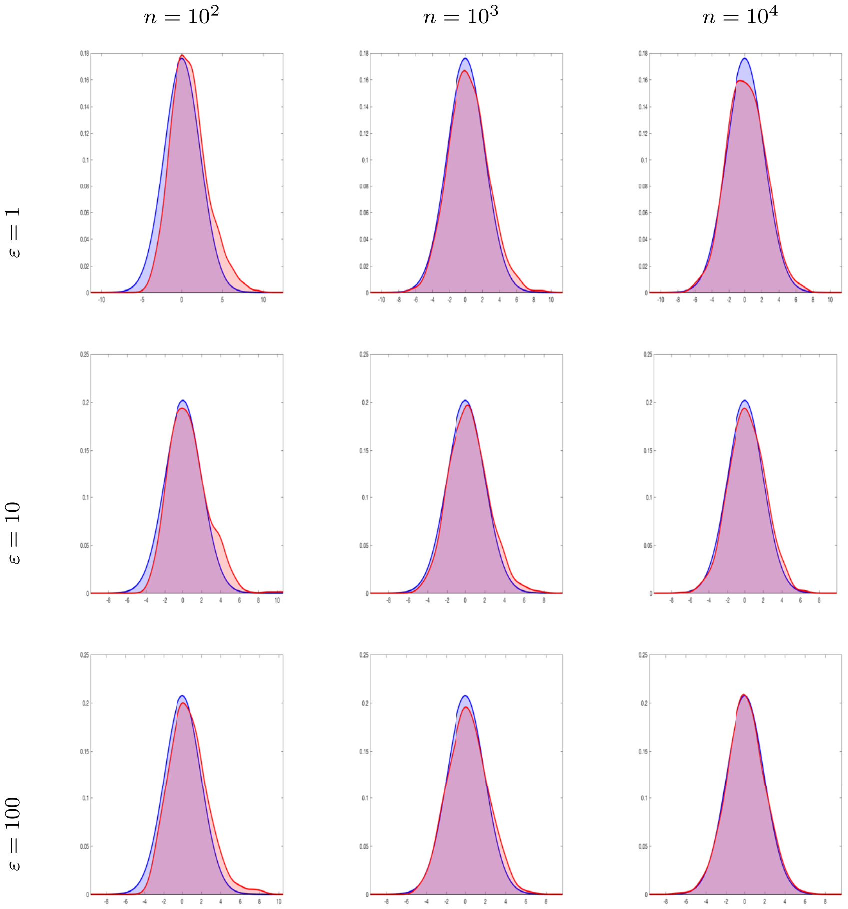


Fig 2. *Case $a \neq b$ with one sample. Illustration of the convergence in distribution of the empirical Sinkhorn loss for a 5×5 grid, $\varepsilon = 1, 10, 100$ and n ranging from 10^2 to 10^4. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a,b))$ (resp. $\langle G, u^{a,b}_\varepsilon - 1/2(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle$).*

a, we generate only one random distribution $\hat{a}_n$. The value of the realization $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a,b))$ is represented by the red vertical lines in Figure 4. Besides, we generate from $\hat{a}_n$, a sequence of $M = 10^3$ bootstrap samples of random measures denoted by $\hat{a}^*_n$ (such that $n\hat{a}^*_n$ follows a multinomial distribution with parameter $\hat{a}_n$). We use again a kernel density estimate (with a data-driven bandwidth) to compare the distribution of $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, b) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, b))$ to the distribution of $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a,b))$ displayed in Figure 2 and Figure 3. The green vertical lines in Figure 4 represent a confidence interval of level 95%. The observation represented by the red vertical line is mostly located within this confidence interval, and the density estimated by bootstrap decently captures the shape of the non-asymptotic distribution of Sinkhorn losses.

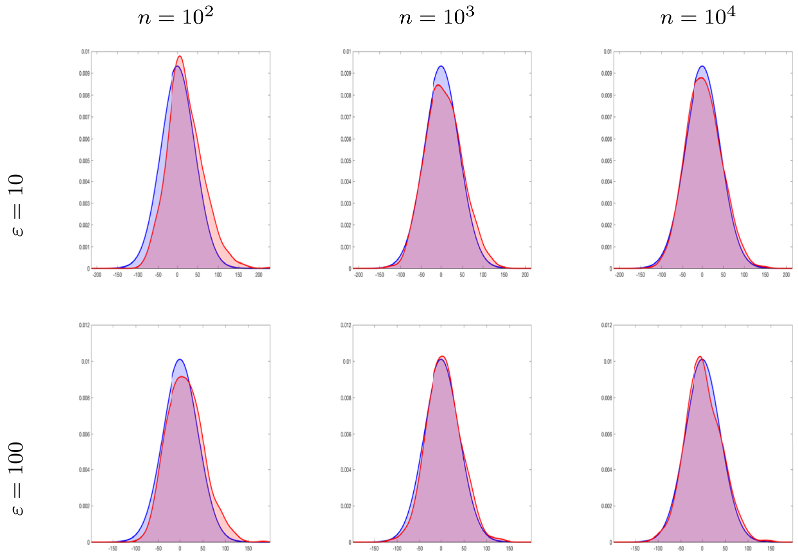


Fig 3. *Case $a \neq b$ with one sample. Illustration of the convergence in distribution of empirical Sinkhorn loss for a 20×20 grid, $\varepsilon = 10, 100$ and n ranging from 10^2 to 10^4. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a, b))$ (resp. $\langle G, u^{a,b}_\varepsilon - 1/2(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle$).*

4.1.2. *Alternative $a \neq b$ – two samples*

We consider the same setting as before, excepting that data are now both sampled from distributions a and b. Hence, we run $M = 10^3$ experiments to obtain a kernel density estimation of the distribution of

$$\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) - \overline{W}^p_{p,\varepsilon}(a, b)),$$

that is compared to the density of the Gaussian variable

$$\sqrt{\gamma}\langle G, u^{a,b}_\varepsilon - \frac{1}{2}(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle + \sqrt{1-\gamma}\langle H, v^{a,b}_\varepsilon - \frac{1}{2}(u^{b,b}_\varepsilon + v^{b,b}_\varepsilon)\rangle,$$

for different values of n and m. The results are reported in Figure 5. The convergence does not seem as good as in the one sample case, this must be due to the randomness coming from both $\hat{a}_n$ and $\hat{b}_m$.

We also report in Figure 6 results on the consistency of the bootstrap procedure under the hypothesis H_1 with two samples. From the distributions a and b, we generate two random distributions $\hat{a}_n$ and $\hat{b}_m$. The value of the realization $\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) - \overline{W}^p_{p,\varepsilon}(a, b))$ is represented by the red vertical lines in Figure 6. Then, we generate from $\hat{a}_n$ and $\hat{b}_m$, two sequences of $M = 10^3$ bootstrap samples of random measures denoted by $\hat{a}^*_n$ and $\hat{b}^*_m$. We use again a kernel density estimate (with a data-driven bandwith) to compare the green

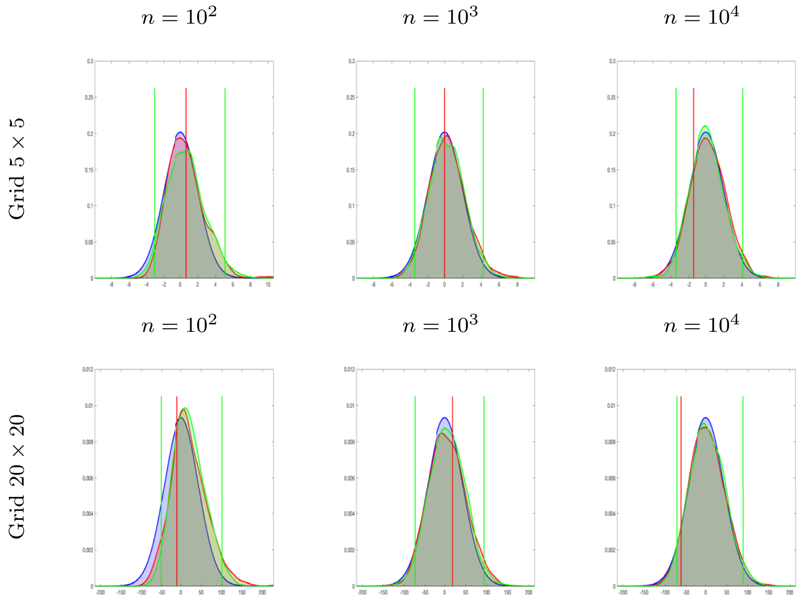


FIG 4. *Case $a \neq b$ with one sample. Illustration of the bootstrap with $\varepsilon = 10$, grids of size 5×5 and 20×20 to approximate the non-asymptotic distribution of empirical Sinkhorn losses. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, b) - \overline{W}^p_{p,\varepsilon}(a, b))$ (resp. $\langle G, u^{a,b}_\varepsilon - 1/2(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle$). The green density represents the distribution of the random variable $\sqrt{n}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, b) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, b))$ in Theorem 3.2.*

distribution of $\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m))$ to the red distribution of $\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) - \overline{W}^p_{p,\varepsilon}(a, b))$ displayed in Figure 6. The green vertical lines in Figure 6 represent a confidence interval of level 95%. We can draw the same conclusion as in the one sample case. All these experiments thus perfectly illustrate the Theorem 2.7.

4.1.3. Hypothesis $a = b$ – one sample

As in the previous cases, we consider a to be the uniform distribution on a square grid. We recall that the distributional limit in the right hand side of (13) is the following mixture of random variables with chi-squared distribution of degree 1

$$\frac{1}{2}\sum_{i=1}^{N} \lambda_i \chi^2_i(1) \text{ for } \lambda_1, \ldots, \lambda_N \text{ the eigenvalues of } \Sigma(a)^{1/2} \partial^2_{11} \overline{W}^p_{p,\varepsilon}(a, a) \Sigma(a)^{1/2}.$$

It appears to be difficult to compute the density of this distributional limit or to draw samples from it, since computing the Hessian matrix $\partial^2_{11} \overline{W}^p_{p,\varepsilon}(a, a)$ is a delicate task. We thus leave this problem open for future work, and only rely

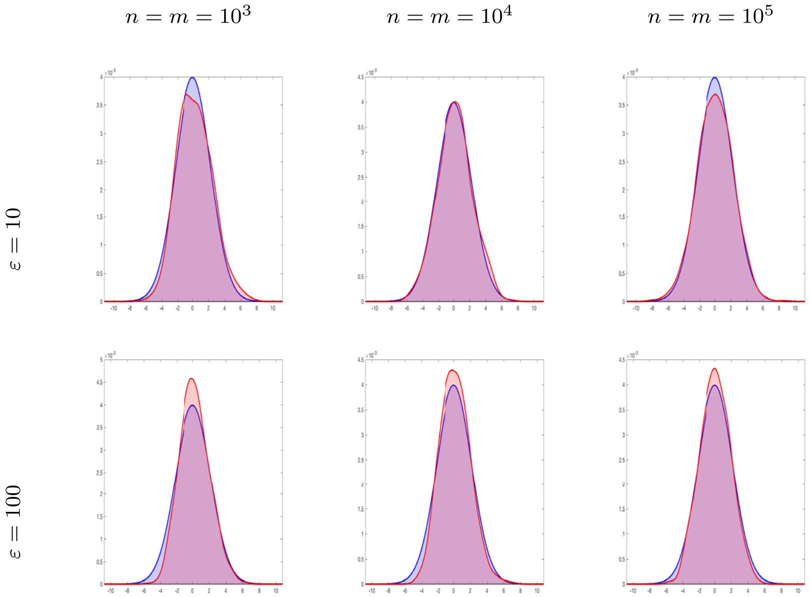


FIG 5. *Case $a \neq b$ with two samples. Illustration of the convergence in distribution of empirical Sinkhorn loss for a 5×5 grid, for $\varepsilon = 10, 100$, $n = m$ and n ranging from 10^3 to 10^5. Densities in red (resp. blue) represent the distribution of $\rho_{n,m}(\overline{W}_{p,\varepsilon}^p(\hat{a}_n, \hat{b}_m) - \overline{W}_{p,\varepsilon}^p(a,b))$ (resp. $\sqrt{\gamma}\langle G, u_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{a,a} + v_\varepsilon^{a,a})\rangle + \sqrt{1-\gamma}\langle H, v_\varepsilon^{a,b} - \frac{1}{2}(u_\varepsilon^{b,b} + v_\varepsilon^{b,b})\rangle$ with $\gamma = 1/2$).*

on the non-asymptotic distribution of $n\overline{W}_{p,\varepsilon}^p(\hat{a}_n, a)$. This justifies the use of the bootstrap procedure described in Section 3. We display the bootstrap statistic in Figure 7. The shape of the non-asymptotic density of $n\overline{W}_{p,\varepsilon}^p(\hat{a}_n, a)$ (red curves in Figure 7) looks chi-squared distributed. In particular, it only takes positive values. The bootstrap distribution in green also recovers the most significant mass location of the red density.

4.1.4. *Hypothesis $a = b$ – two samples*

We still consider $a = b$ to be the uniform distribution on a square grid and we sample two measures from a denoted $\hat{a}_n, \hat{b}_m$. We compute the non-asymptotic distribution of $(nm/(m+n))\overline{W}_{p,\varepsilon}^p(\hat{a}_n, \hat{b}_m)$ which, from Theorem 2.9, must converge to $\frac{1}{2}\sum_{i=1}^N \tilde{\lambda}_i \chi_i^2(1)$ with $\{\tilde{\lambda}_i\}_i$ the eigenvalues of

$$\operatorname{diag}(\sqrt{\gamma}\Sigma(a)^{1/2}, \sqrt{1-\gamma}\Sigma(a)^{1/2})\nabla^2 \overline{W}_{p,\varepsilon}^p(a,a)\operatorname{diag}(\sqrt{\gamma}\Sigma(a)^{1/2}, \sqrt{1-\gamma}\Sigma(a)^{1/2}).$$

The results are displayed in red in Figure 8, together with the bootstrap distribution (in green) $\rho_{n,m}^2(\overline{W}_{p,\varepsilon}^p(\hat{a}_n^*, \hat{b}_m^*) - \overline{W}_{p,\varepsilon}^p(\hat{a}_n, \hat{b}_m) - \langle u^{\hat{a}_n, \hat{b}_m}, \hat{a}_n^* - \hat{a}_n\rangle - \langle v^{\hat{a}_n, \hat{b}_m}, \hat{b}_m^* - \hat{b}_m\rangle)$. We obtain similar results to the one sample case.

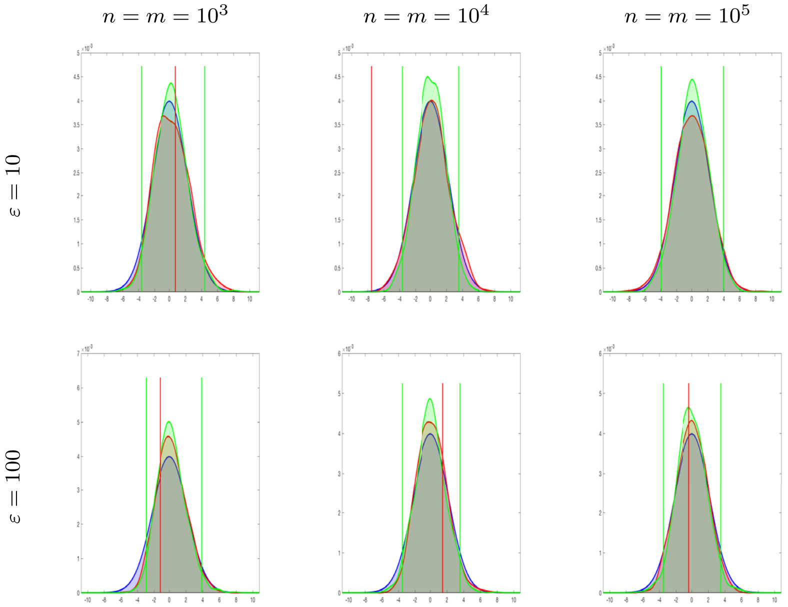


FIG 6. *Case $a \neq b$ with two samples. Illustration of the bootstrap with $\varepsilon = 10$ for the grid of size 5×5 and $\varepsilon = 100$ for the grid 20×20 to approximate the non-asymptotic distribution of empirical Sinkhorn divergences. Densities in red (resp. blue) represent the distribution of $\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) - \overline{W}^p_{p,\varepsilon}(a,b))$ (resp. $\sqrt{\gamma}\langle G, u^{a,b}_\varepsilon - \frac{1}{2}(u^{a,a}_\varepsilon + v^{a,a}_\varepsilon)\rangle + \sqrt{1-\gamma}\langle H, v^{a,b}_\varepsilon - \frac{1}{2}(u^{b,b}_\varepsilon + v^{b,b}_\varepsilon)\rangle$). The green density is the distribution of the random variable $\rho_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m))$ in Theorem 3.2.*

4.2. Estimation of test power using the bootstrap

One sample – distribution with linear trend and varying slope parameter The consistency and usefulness of the bootstrap procedure is illustrated by studying the statistical power (that is $\mathbb{P}(\text{Reject } H_0|H_1 \text{ is true})$) of statistical tests (at level 5%) based on the empirical Sinkhorn loss. For this purpose, we choose a to be uniform and b to be a distribution with linear trend whose slope parameter θ is ranging from 0 to 0.1 on a 5×5 grid. We assume that we observe a single realization of an empirical measure $\hat{b}_m$ sampled from b with $m = 10^3$. Then, we generate $M = 10^3$ bootstrap samples of random measures $\hat{b}^*_{m,j}$ from $\hat{b}_m$ (with $1 \leq j \leq M$), which allows the computation of the p-value

$$\begin{aligned} p\text{-value} = \#\{j \text{ such that } & n|\overline{W}^p_{p,\varepsilon}(a, \hat{b}^*_{m,j}) - \overline{W}^p_{p,\varepsilon}(a, \hat{b}_m) - \langle v^{a,\hat{b}_m}, \hat{b}^*_{m,j} - \hat{b}_m\rangle| \\ & \geq n\overline{W}^p_{p,\varepsilon}(a, \hat{b}_m)\}/M. \end{aligned}$$

This experiments is repeated 100 times, in order to estimate the power (at level u) of a test based on $n\overline{W}^p_{p,\varepsilon}(a, \hat{b}_m)$ by comparing the resulting sequence of p-values to the value u. The results are reported in Figure 9 (left). It can

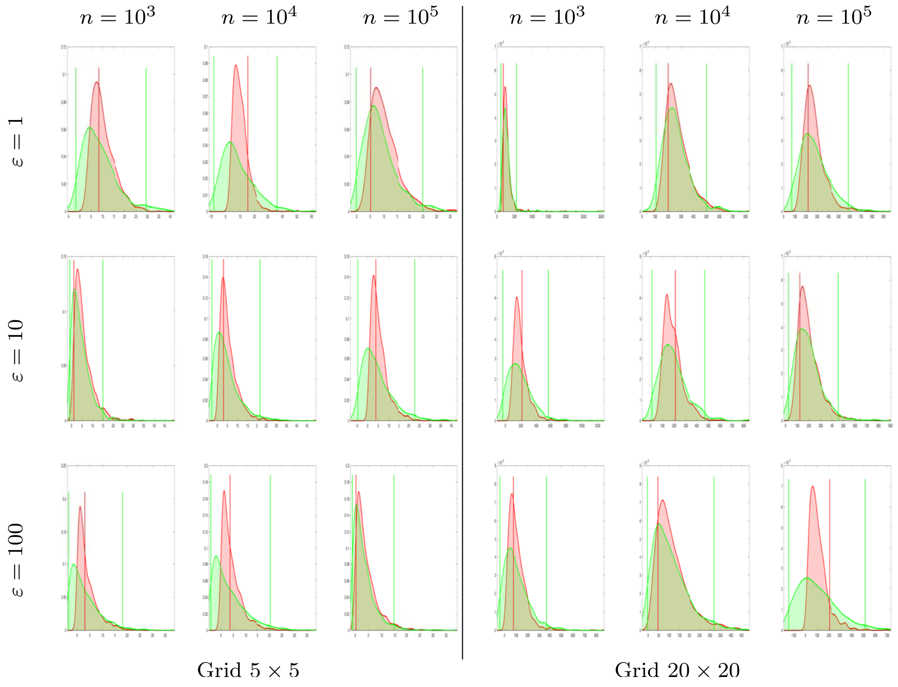


FIG 7. *Case $a = b$ with one sample. Illustration of the bootstrap with $\varepsilon = 1, 10, 100$ and two grids of size 5×5 (left) and 20×20 (right) to approximate the non-asymptotic distribution of the empirical Sinkhorn loss. Densities in red represent the distribution of $n\overline{W}_{p,\varepsilon}^p(\hat{a}_n, a)$. The green density represents the distribution of the random variable $n(\overline{W}_{p,\varepsilon}^p(\hat{a}_n, a) - \overline{W}_{p,\varepsilon}^p(\hat{a}_n^*, a) - \langle u^{\hat{a}_n, a}, \hat{a}_n^* - \hat{a}_n \rangle)$ in (23).*

be seen that the resulting testing procedures are good discriminants for the three values of the regularization parameters ε that we considered. As soon as the slope θ increases then b sufficiently differs from a, and the probability of rejecting H_0 thus increases. We have also chosen to report results obtained with the Sinkhorn loss corresponding to optimal transport regularized by the entropy $H(T) = \sum_{ij} t_{ij} \log(t_{ij})$ instead of the relative entropy $H(T|a \otimes b) = \sum_{i,j} \log\left(\frac{t_{ij}}{a_i b_j}\right) t_{ij}$ (see Figure 9 (right)). Indeed, we remark that in the case of the relative entropy, the power of the test seems to highly depend on the value of ε. More precisely, for a fixed value of the slope parameter θ (or distribution b), the test power is larger as ε increases. On the other hand, when using the Sinkhorn loss computed with the entropy, the power of the test seems to be the same for any value of ε.

Remark 5. *The truly interesting property of the Sinkhorn loss over the Sinkhorn divergence is that in theory, for any $\varepsilon > 0$, we will obtain a steady ε-dependent asymptotic distribution, and that any regularization allows us to perform test statistics. In practice, more regularization leads to a blending of in-*

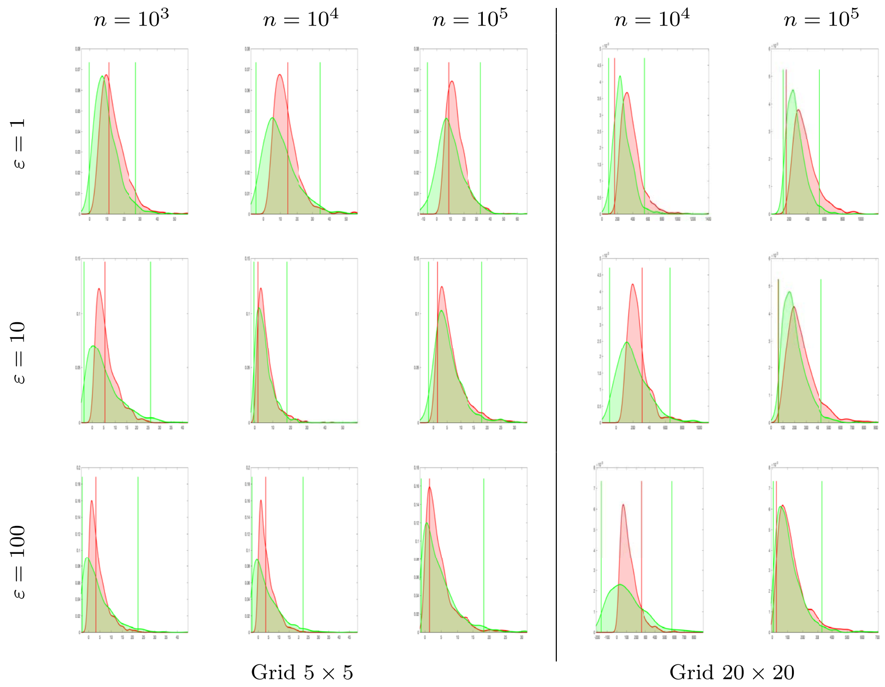


FIG 8. *Case $a = b$ with two samples. Illustration of the bootstrap with $\varepsilon = 1, 10, 100$ and two grids of size 5×5 (left) and 20×20 (right) to approximate the non-asymptotic distribution of the empirical Sinkhorn loss. Densities in red represent the distribution of $\rho^2_{n,m} \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m)$. The green density represents the distribution of the random variable $\rho^2_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - \overline{W}^p_{p,\varepsilon}(\hat{a}_n, \hat{b}_m) - \langle u^{\hat{a}_n, \hat{b}_m}, \hat{a}^*_n - \hat{a}_n \rangle - \langle v^{\hat{a}_n, \hat{b}_m}, \hat{b}^*_m - \hat{b}_m \rangle)$ in (24).*

formation. More precisely, the entropy will spread the mass of the distributions, and in some points of the grid, the differences of masses between the two distributions can be the result of regularization. On the other hand, when very few observations are available, and that measures are sparsely distributed on the grid, a large ε will still allow to perform a statistical study.

5. Analysis of real data

We consider a dataset of colored images representing landscapes and foliage taken during Autumn (20 images) and Winter (17 images), see Figure 10 for examples. These images, provided by [25], are available at `http://tabby.vision.mcgill.ca/html/welcome.html`.

Each image is transformed into a color histogram on a three-dimensional grid (RGB colors) of size $N^3 = 16^3 = 4096$ of equi-spaced points. We will denote by $a_1, \dots, a_{20}$ the autumn histograms and $w_1, \dots, w_{17}$ the winter histograms. To compute the cost matrix C, we again use the squared Euclidean distance between the spatial integer locations $x_i \in [0; 255]^3$.

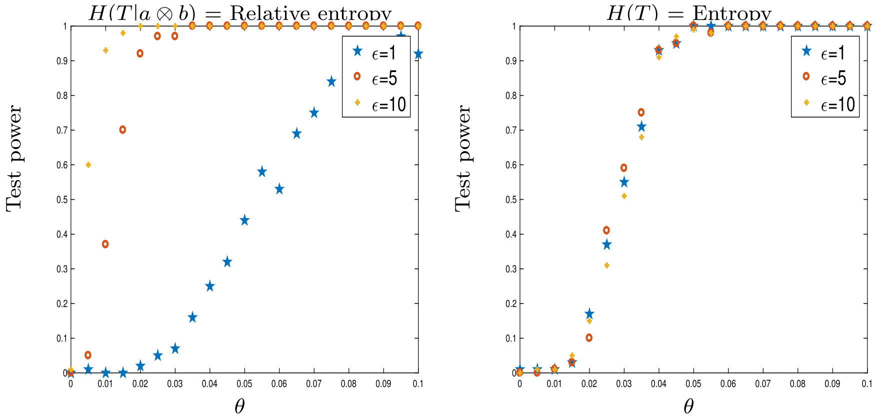


FIG 9. *Test power (probability of rejecting H_0 knowing that H_1 is true) on a 5×5 grid in the one sample case, as a function of the slope parameter θ ranging from 0 to 0.15 for $\varepsilon = 1$ (blue), $\varepsilon = 5$ (orange) and $\varepsilon = 10$ (yellow), with $n = 10^3$. (left) $H(T|a \otimes b)$ = Relative entropy, (right) $H(T)$ = Entropy.*

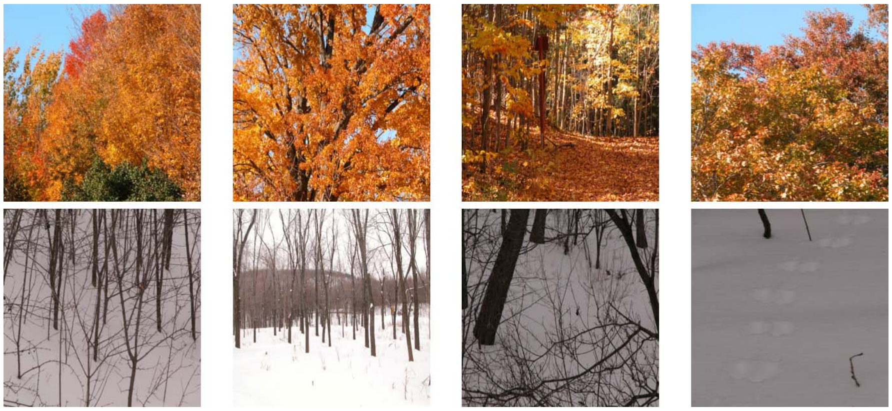
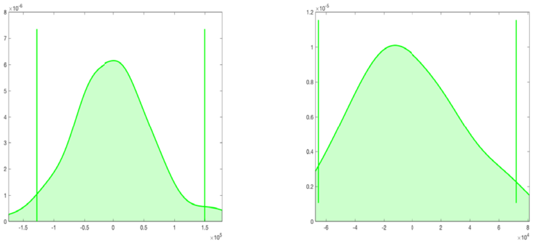


FIG 10. *Samples of 768×576 colored images from autumn (first row) and winter (second row).*

5.1. Testing the hypothesis of equal color distribution between seasons

We first test the null hypothesis that the color distribution of the images in Autumn is the same as the color distribution of the images in Winter. To this end, we consider the mean histogram of the dataset for each season, that we denote

$$\bar{a}_{20} = \frac{1}{20}\sum_{k=1}^{20} a_k \qquad \text{and} \qquad \bar{w}_{17} = \frac{1}{17}\sum_{k=1}^{17} w_k.$$



(a) (b)

FIG 11. *Testing equality of color distributions between Autumn and Winter for a grid of size* $16^3 = 4096$. *Green densities represent the distribution of the bootstrap statistics* $\rho_{n,m}^2(\overline{W}_{p,\varepsilon}^p(\hat{a}_n^*, \hat{w}_m^*) - \overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17}) - (\langle u^{\bar{a}_{20},\bar{w}_{17}}, \hat{a}_n^* - \bar{a}_{20}\rangle + \langle v^{\bar{a}_{20},\bar{w}_{17}}, \hat{w}_m^* - \bar{w}_{17}\rangle))$ *(vertical bars represent a confidence interval of level* 95%*) for (a)* $\varepsilon = 10$ *and (b)* $\varepsilon = 100$. *The value of* $\rho_{n,m}^2 \overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17})$ *is outside the support of the green density for each value of* ε, *and it is thus not represented.*

Notice that both $\bar{a}_{20}$ and $\bar{w}_{17}$ are discrete empirical measures admitting a zero mass for many locations x_i.

We use the two samples testing procedure described previously, and a bootstrap approach to estimate the distribution of the test statistics

$$\rho_{n,m}^2 \overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17}).$$

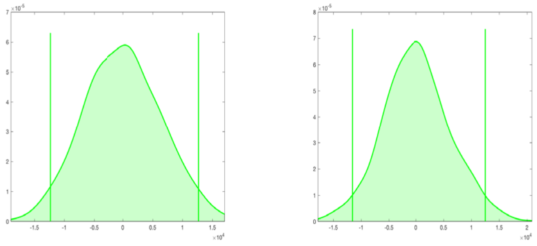
Notice also that n and m respectively correspond to the number of observations for the empirical Autumn distribution $\bar{a}_{20}$ and the empirical Winter distribution $\bar{w}_{17}$, which is the total number of pixels times the number of images. Therefore, $n = 20 * 768 * 576 = 8847360$ and $m = 17 * 768 * 576 = 7520256$. We report the results of the testing procedure for $\varepsilon = 10, 100$ by displaying in Figure 11 an estimation of $M = 100$ observations of the bootstrap statistic's density

$$\rho_{n,m}^2(\overline{W}_{p,\varepsilon}^p(\hat{a}_n^*, \hat{w}_m^*) - \overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17}) \\ - (\langle u^{\bar{a}_{20},\bar{w}_{17}}, \hat{a}_n^* - \bar{a}_{20}\rangle + \langle v^{\bar{a}_{20},\bar{w}_{17}}, \hat{w}_m^* - \bar{w}_{17}\rangle)),$$

where $\hat{a}_n^*$ and $\hat{w}_m^*$ are respectively bootstrap samples of $\bar{a}_{20}$ and $\bar{w}_{17}$, and $(u^{\bar{a}_{20},\bar{w}_{17}}, v^{\bar{a}_{20},\bar{w}_{17}})$ are the optimal dual variables associated to $(\bar{a}_{20}, \bar{w}_{17})$ in problem (3).

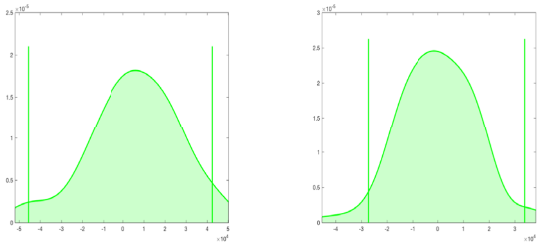
For $\varepsilon = 10, 100$, the value of $\rho_{n,m}^2(\overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17}))$ is outside the support of this density, and the null hypothesis that the color distributions of images taken during Autumn and Winter are the same is thus rejected. In particular, the test statistic $\rho_{n,m}^2(\overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17}))$ is equal to 6.07×10^7 for $\varepsilon = 10$ and to 5.03×10^7 for $\varepsilon = 100$.

We also run the exact same experiments for a smaller grid (size $8^3 = 512$) and a higher number of observations ($M = 1000$). The results are displayed in Figure 12. The distributions $\rho_{n,m}^2(\overline{W}_{p,\varepsilon}^p(\hat{a}_n^*, \hat{w}_m^*) - \overline{W}_{p,\varepsilon}^p(\bar{a}_{20}, \bar{w}_{17}) - (\langle u^{\bar{a}_{20},\bar{w}_{17}}, \hat{a}_n^* - \bar{a}_{20}\rangle + \langle v^{\bar{a}_{20},\bar{w}_{17}}, \hat{w}_m^* - \bar{w}_{17}\rangle))$ are much more centered around 0 (we gain a factor 10). However, we obtain the same conclusion as before, with a test statistic equal to 9.39×10^6 for $\varepsilon = 10$ and 8.50×10^6 for $\varepsilon = 100$.



(a) (b)

Fig 12. *Testing equality of color distributions between Autumn and Winter for a grid of size* $8^3 = 512$. *Green densities represent the distribution of the bootstrap statistics* $\rho^2_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{w}^*_m) - \overline{W}^p_{p,\varepsilon}(\bar{a}_{20}, \bar{w}_{17}) - (\langle u^{\bar{a}_{20}, \bar{w}_{17}}, \hat{a}^*_n - \bar{a}_{20}\rangle + \langle v^{\bar{a}_{20}, \bar{w}_{17}}, \hat{w}^*_m - \bar{w}_{17}\rangle))$ *(vertical bars represent a confidence interval of level* 95%*) for (a)* $\varepsilon = 10$ *and (b)* $\varepsilon = 100$. *The value of* $\rho^2_{n,m}\overline{W}^p_{p,\varepsilon}(\bar{a}_{20}, \bar{w}_{17})$ *is outside the support of the green density for each value of* ε, *and it is thus not represented.*



(a) (b)

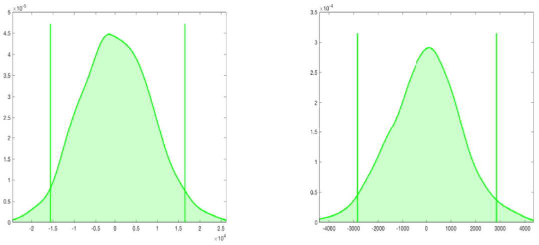
Fig 13. *Testing equality of color distributions when splitting the autumn dataset into two for a grid of size* $16^3 = 512$. *Green densities represent the distribution of the bootstrap statistics* $\rho^2_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - \overline{W}^p_{p,\varepsilon}(\bar{a}_{1\to 10}, \bar{a}_{11\to 20}) - (\langle u^{\bar{a}_{1\to 11}, \bar{a}_{11\to 20}}, \hat{a}^*_n - \bar{a}_{1\to 11}\rangle + \langle v^{\bar{a}_{1\to 11}, \bar{a}_{11\to 20}}, \hat{b}^*_m - \bar{a}_{11\to 20}\rangle))$ *(vertical bars represent a confidence interval of level* 95%*) for (a)* $\varepsilon = 10$ *and (b)* $\varepsilon = 100$. *The value of* $\rho^2_{n,m}\overline{W}^p_{p,\varepsilon}(\bar{a}_{1\to 10}, \bar{a}_{11\to 20})$ *is outside the support of the green density for each value of* ε, *and it is thus not represented.*

5.2. Testing the hypothesis of equal distribution when splitting the Autumn dataset

We propose now to investigate the equality of distributions within the same dataset of Autumn histograms. To this end, we arbitrarily split the Autumn dataset into two subsets of 10 images and we compute their mean distribution

$$\bar{a}_{1\to 10} = \frac{1}{10}\sum_{k=1}^{10} a_k \qquad \text{and} \qquad \bar{a}_{11\to 20} = \frac{1}{10}\sum_{k=11}^{20} a_k,$$

for which $n = m = 10 * 768 * 576 = 4423680$. The procedure is then similar to the Autumn versus Winter case in Subsection 5.1, meaning that we sample $M = 100$ bootstrap distributions $\hat{a}^*_n$ and $\hat{b}^*_m$ from respectively $\bar{a}_{1\to 10}$ and $\bar{a}_{11\to 20}$.



(a) (b)

FIG 14. *Testing equality of color distributions when splitting the autumn dataset into two for a grid of size* $8^3 = 512$. *Green densities represent the distribution of the bootstrap statistics* $\rho^2_{n,m}(\overline{W}^p_{p,\varepsilon}(\hat{a}^*_n, \hat{b}^*_m) - \overline{W}^p_{p,\varepsilon}(\bar{a}_{1\to 10}, \bar{a}_{11\to 20}) - (\langle u^{\bar{a}_{1\to 11}, \bar{a}_{11\to 20}}, \hat{a}^*_n - \bar{a}_{1\to 11}\rangle + \langle v^{\bar{a}_{1\to 11}, \bar{a}_{11\to 20}}, \hat{b}^*_m - \bar{a}_{11\to 20}\rangle))$ *(vertical bars represent a confidence interval of level* 95%*) for (a)* $\varepsilon = 10$ *and (b)* $\varepsilon = 100$. *The value of* $\rho^2_{n,m}\overline{W}^p_{p,\varepsilon}(\bar{a}_{1\to 10}, \bar{a}_{11\to 20})$ *is outside the support of the green density for each value of* ε*, and it is thus not represented.*

The results are displayed in Figure 13. We obtain similar results than in the two seasons case, the null hypothesis that both Autumn distributions follow the same law is thus rejected. On the other hand, the test statistics are smaller in this case as the histogram of color seems to be closer. Indeed, the quantity $\rho^2_{n,m}\overline{W}^p_{p,\varepsilon}(\bar{a}_{1\to 10}, \bar{a}_{11\to 20})$ is equal to 11.02×10^6 for $\varepsilon = 10$ and 5.50×10^6 for $\varepsilon = 100$.

Similarly to the Winter VS Autumn case, we also run the same Autumn VS Autumn experiments for a grid of size $8^3 = 512$ and $M = 1000$ observations. The results are displayed in Figure 14 for test statistics equal to 14.07×10^5 for $\varepsilon = 10$ and 3.41×10^5 for $\varepsilon = 100$.

Remark 6. *For comparison purpose, we ran a* χ^2 *test of homogeneity for testing the hypothesis of equal distributions of colors. The obtained test statistic in the Autumn vs Winter case is equal to* $\chi^2_{AW} = 6.96 \times 10^4$, *and in the Autumn splitting case to* $\chi^2_{AA} = 4.06 \times 10^4$. *Even if* χ^2_{AA} *is indeed smaller than* χ^2_{AW}*, the contrast between these two is weaker than with the Sinkhorn loss test.*

6. Future works

As remarked in [33], there exists a vast literature for two-sample testing using univariate data. However, in a multivariate setting, it is difficult to consider that there exist standard methods to test the equality of two distributions. We thus intend to further investigate the benefits of the use of the empirical Sinkhorn loss to propose novel testing procedures able to compare multivariate distributions for real data analysis. A first perspective is to apply the methodology developed in this paper to more than two samples using the notion of smoothed Wasserstein barycenters (see e.g. [9] and references therein) for the analysis of variance of multiple and multivariate random measures (MANOVA). However, as pointed out in [9], a critical issue in this setting will be the choice of the regularization parameter ε, as it has a large influence on the shape of the estimated Wasserstein barycenter. Another interesting extension of the results presented in this paper would be to obtain the eigenvalues of the Hessian matrix of the Sinkhorn loss, in order to compute the distributional limit under the null hypothesis of equality of distributions.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein GAN. *arXiv preprint* arXiv:1701.07875, 2017.

[2] J. Bigot, E. Cazelles, and N. Papadakis. Data-driven regularization of wasserstein barycenters with an application to multivariate density registration. *ArXiv e-prints*, 1804.08962, 2018.

[3] J. Bigot, R. Gouet, T. Klein, A. López, et al. Geodesic pca in the Wasserstein space by convex pca. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1–26, 2017. MR3606732

[4] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. MR2061575

[5] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. Geodesic pca versus log-pca of histograms in the wasserstein space. *SIAM Journal on Scientific Computing*, 40(2):B429–B456, 2018. MR3780753

[6] Q. Chen and Z. Fang. Inference on functionals under first order degeneracy. *SSRN*, 2018. MR3958414

[7] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pages 2292–2300. 2013.

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

[9] M. Cuturi and G. Peyré. A smoothed dual approach for variational Wasserstein problems. *SIAM Journal on Imaging Sciences*, 9(1):320–343, 2016. MR3466197

[10] E. del Barrio, J. A. Cuesta-Albertos, C. Matrán, and J. M. Rodriguez-Rodriguez. Tests of goodness of fit based on the L_2-Wasserstein distance. *Ann. Statist.*, 27(4):1230–1239, 1999. MR1740113

[11] E. del Barrio, E. Giné, and F. Utzet. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189, 2005. MR2121458

[12] E. del Barrio and J.-M. Loubes. Central limit theorems for empirical transportation cost in general dimension. arXiv:1705.01299v1, 2017. MR3916938

[13] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall, New York, 1993. MR1270903

[14] J. Feydy, T. Séjourné, F.-X. Vialard, S.-I. Amari, A. Trouvé, and G. Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. *arXiv preprint* arXiv:1810.08278, 2018.

[15] G. Freitag and A. Munk. On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships. *J. Multivariate Anal.*, 94(1):123–158, 2005. MR2161214

[16] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[17] A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. *arXiv preprint* arXiv:1810.02733, 2018.

[18] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In D. D. Lee, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Proc. NIPS'16*, pages 3432–3440. Curran Associates, Inc., 2016.

[19] A. Genevay, G. Peyré, and M. Cuturi. Sinkhorn-autodiff: Tractable Wasserstein learning of generative models. *arXiv preprint* 1706.00292, 2017.

[20] A. Gramfort, G. Peyré, and M. Cuturi. Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging*, pages 261–272. Springer, 2015.

[21] M. Klatt, C. Tameling, and A. Munk. Empirical regularized optimal transport: Statistical theory and applications. *arXiv preprint* arXiv:1810.09880, 2018.

[22] T.-T. Lu and S.-H. Shiou. Inverses of 2 × 2 block matrices. *Computers & Mathematics with Applications*, 43(1-2):119–129, 2002. MR1873248

[23] G. Luise, A. Rudi, M. Pontil, and C. Ciliberto. Differential properties of Sinkhorn approximation for learning with Wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 5859–5870, 2018.

[24] X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.*, 41(1):370–400, 02 2013. MR3059422

[25] A. Olmos and F. A. Kingdom. A biologically inspired algorithm for the recovery of shading and reflectance images. *Perception*, 33(12):1463–1473, 2004.

[26] J. Rabin and N. Papadakis. Convex color image segmentation with optimal transport distances. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 256–269. Springer, 2015. MR3394935

[27] A. Ramdas, N. G. Trillos, and M. Cuturi. On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017. MR3608466

[28] T. Rippl, A. Munk, and A. Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.*, 151:90–109, 2016. MR3545279

[29] A. Rolet, M. Cuturi, and G. Peyré. Fast dictionary learning with a smoothed Wasserstein loss. In *Proc. International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2016.

[30] M. A. Schmitz, M. Heitz, N. Bonneel, F. M. N. Mboula, D. Coeurjolly, M. Cuturi, G. Peyré, and J.-L. Starck. Wasserstein dictionary learning: Optimal transport-based unsupervised non-linear dictionary learning. *arXiv*

preprint arXiv:1708.01955, 2017. MR3593193

[31] V. Seguy and M. Cuturi. Principal geodesic analysis for probability measures under the optimal transport metric. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 3294–3302. Curran Associates, Inc., 2015.

[32] J. Solomon, F. de Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. In *ACM Transactions on Graphics (SIGGRAPH'15)*, 2015.

[33] M. Sommerfeld and A. Munk. Inference for empirical wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016. MR3744719

[34] J. Sourati, M. Akcakaya, T. K. Leen, D. Erdogmus, and J. G. Dy. Asymptotic analysis of objectives based on Fisher information in active learning. *Journal of Machine Learning Research*, 18(34):1–41, 2017. MR3646629

[35] A. Thibault, L. Chizat, C. Dossal, and N. Papadakis. Overrelaxed Sinkhorn-Knopp algorithm for regularized optimal transport. *arXiv preprint* arXiv:1711.01851, 2017.

[36] A. W. Van Der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996. MR1385671

[37] C. Villani. *Topics in Optimal Transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003. MR1964483

[38] L. Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2011. MR2055670

[39] A. G. Wilson. The use of entropy maximising models, in the theory of trip distribution, mode split and route split. *Journal of Transport Economics and Policy*, pages 108–126, 1969.

[40] J. Ye, P. Wu, J. Z. Wang, and J. Li. Fast discrete distribution clustering using Wasserstein barycenter with sparse support. *IEEE Trans. Signal Processing*, 65(9):2317–2332, 2017. MR3620352

[41] C. Zalinescu. *Convex analysis in general vector spaces*. World Scientific, 2002. MR1921556