Central limit theorems for Sinkhorn divergence between probability distributions on finite spaces and statistical applications

Jérémie Bigot, Elsa Cazelles, Nicolas Papadakis

Central limit theorems for Sinkhorn divergence between probability distributions on finite spaces and statistical applications *

Jérémie Bigot†, Elsa Cazelles & Nicolas Papadakis

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

November 24, 2017

Abstract

The notion of 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 central limit theorems for empirical Sinkhorn divergences. We also propose a bootstrap procedure which allows to obtain new test statistics for measuring the discrepancies between multivariate probability distributions. The strategy of proof uses the notions of directional Hadamard differentiability and delta-method in this setting. It is inspired by the results in the work of Sommerfeld and Munk in [28] on the asymptotic distribution of empirical Wasserstein distance on finite space using un-regularized transportation costs. Simulated and real datasets are used to illustrate our approach. A comparison with existing methods to measure the discrepancy between multivariate distributions is also proposed.

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 entropically 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 [31] 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

*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).

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

relevant statistical measure in various fields such as clustering of discrete distributions [35], nonparametric Bayesian modelling [19], fingerprints comparison [28], unsupervised learning [2], and principal component analysis [4, 26].

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 [5] 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 as developed in [6] 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 [16] and more generally for high dimensional data analysis in multi-label learning [14], dictionnary learning [23] and image processing, see e.g. [7, 20] and references therein, text mining via bag-of-words comparison [15], averaging of neuroimaging data [18],

The goal of this paper is to analyze the potential benefits of Sinkhorn divergences for statistical inference from empirical probability measures, by deriving 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 Wasserstein distance is well understood for probability measures supported on $\mathbb{R}$ [13, 8, 9]. These results have then been extended for specific parametric distributions supported on $\mathbb{R}^d$ belonging to an elliptic class, see [22] and references therein. Recently, a central limit theorem has been established in [10] for empirical transportation cost for data sampled from absolutely continuous measures on $\mathbb{R}^d$ that holds for any $d \geq 1$. The case of discrete measures supported on a finite metric space has also been recently considered in [28] with the proof of the convergence (in the spirit of the central limit theorem) of empirical Wasserstein distances toward to the optimal value of a linear program. Ramdas and al. in [21] 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 ($d = 1$), these results leads to test statistics whose numerical implementation may become prohibitive for empirical measures supported on $\mathbb{R}^d$ with $d \geq 2$. This is due to the computational cost to evaluate a transport distance. Therefore, using test statistics based on Sinkhorn divergences may be of interest thanks to their fast computation through the iterative algorithm originally proposed in [5]. This paper is thus focused on the study of inference from discrete distributions in terms of entropic regularized transport costs. The results are inspired by the work in [28] on the asymptotic distribution of empirical Wasserstein distance on finite space using un-regularized transportation costs.

Our main contributions may be summarized as follows. First, for data sampled from one or two unknown discrete measures, we derive central limit theorems for empirical Sinkhorn

divergences. These results then lead to new test statistics for measuring the discrepancies between multivariate probability distributions. Finally, to illustrate the applicability of this approach to synthetic data, we propose a bootstrap procedure to estimate unknown quantities of interest in the computation of these test statistics (such as their non-asymptotic variance and quantiles). Simulated and real datasets are used to illustrate our approach. A comparison with existing methods to measure the discrepancy between multivariate distributions is also proposed.

1.3 Overview of the paper

In Section 2 we briefly recall the optimal transport problem between probability measures, and we introduce the Sinkhorn divergence following the presentation in [6]. Then, we discuss the notion of directional derivative of these divergences in order to obtain our main result on a central limit theorem for Sinkhorn divergence via an appropriate adaptation of the delta-method. A bootstrap procedure is discussed in Section 3. Numerical experiments are presented in Section 4 and Section 5 for synthetic data and real data, and we illustrate the benefits of a bootstrap procedure. Some perspectives are given in Section 6.

2 Distribution limits for empirical Sinkhorn divergences

2.1 Notation and definitions

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, 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 in x_i. As the space $\mathcal{X}$ is considered to be 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 $\lambda > 0$ be a regularization parameter. The Sinkhorn divergence [5] 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

$$p_\lambda(a, b) = \min_{T \in U(a,b)} \langle T, C \rangle - \lambda h(T), \text{ with } a \text{ and } b \text{ in } \Sigma_N, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ denotes the usual inner product between matrices, 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 $c_{i,j} = d(x_i,x_j)^p$,
- the regularization function $h(T) = -\sum_{i,j} t_{ij}\log t_{ij}$ is the negative entropy for a transport matrix $T \in U(a,b)$.

Remark 1. *The Sinkhorn divergence is not a metric on the space $\mathcal{P}_p(\mathcal{X})$ of discrete probability measures. In particular, $p_\lambda(a,b) \neq 0$ when $a = b$.*

Remark 2. *This entire section is also valid for more general cost matrices C.*

Computing the p-Wasserstein distance between discrete probability measures supported on $\mathcal{X}$ amounts to find a minimizer of $T \mapsto \langle T, C\rangle$ over $U(a,b)$. 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 in order to reduce its complexity that has been known since a long time [34]. This is the approach followed in [5] by adding an entropic regularization on the transport matrix. This yields the strictly convex (primal) problem (2) [5, 6].

Definition 2.2 (Dual problem)**.** Following [6], the dual version of the minimization problem (2) is given by

$$d_\lambda(a,b) = \max_{\alpha,\beta\in\mathbb{R}^N} \alpha^T a + \beta^T b - \sum_{i,j} \lambda e^{-\frac{1}{\lambda}(c_{ij}-\alpha_i-\beta_j)}, \tag{3}$$

in the sense that $d_\lambda(a,b) = p_\lambda(a,b)$.

There exists an explicit relation between the optimal solutions of the primal and dual problems above, and they can be computed through an iterative method called Sinkhorn's algorithm [6].

Proposition 2.1 (Sinkhorn's algorithm)**.** *Let $K = \exp(-C/\lambda)$ be the elementwise exponential of the matrix cost C divided by $-\lambda$. Then, there exists a pair a vectors $(u,v) \in \mathbb{R}_+^N \times \mathbb{R}_+^N$ such that the optimal solutions T_λ^* and $(\alpha_\lambda^*, \beta_\lambda^*)$ of problems (2) and (3) are respectively given by*

$$T_\lambda^* = \mathrm{diag}(u)K\,\mathrm{diag}(v), \text{ and } \alpha_\lambda^* = -\lambda\log(u),\ \beta_\lambda^* = -\lambda\log(v).$$

Moreover, such a pair (u,v) is unique up to scalar multiplication, and it can be recovered as a fixed point of the Sinkhorn map

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

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

Finally, the following notation will also be needed in the proofs. We define

$$f_\lambda : \begin{array}{ccc} \Sigma_N \times \Sigma_N \times \mathbb{R}^N \times \mathbb{R}^N & \longrightarrow & \mathbb{R} \\ (a,b,\alpha,\beta) & \longmapsto & \alpha^T a + \beta^T b - \sum_{i,j} \lambda e^{-\frac{1}{\lambda}(c_{ij}-\alpha_i-\beta_j)} \end{array} \tag{5}$$

We also denote by $\xrightarrow{\mathcal{L}}$ the convergence in distribution of a random variable and $\xrightarrow{\mathbb{P}}$ the convergence in propability. The notation $G \overset{\mathcal{L}}{\sim} a$ mean that G is a random variable taking its values in $\mathcal{X}$ with law $a = (a_1,\ldots,a_n) \in \Sigma_N$ (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.

2.2 Directional derivative of d_λ

We follow the presentation and notation in [28]. We recall that, if it exists, the Hadamard directional derivative of a function $g : D_g \subset \mathbb{R}^d$ at $z \in D_g$ in the direction h is defined as

$$g'_h(z) = \lim_{n\to\infty} \frac{g(z + t_n h_n) - g(z)}{t_n}$$

for any sequences $(t_n)_n$ such that $t_n \searrow 0$ and $h_n \to h$ with $z + t_n h_n \in D_g$ for all n. As explained in [28], the derivate $h \mapsto g'_h(z)$ is not necessarily a linear map contrary to the usual notion of Hadamard differentiability. A typical example being the function $g(z) = |z|$ (with $D_g = \mathbb{R}$) which is not Hadamard differentiable at $z = 0$ in the usual sense, but directionally differentiable with $g'_h(0) = |h|$.

Theorem 2.3. *The functional $(a, b) \mapsto d_\lambda(a, b)$ is directionally Hadamard differentiable at all $(a, b) \in int(\Sigma_N \times \Sigma_N)$ with derivative*

$$(h_1, h_2) \mapsto \max_{(\alpha,\beta)\in N_\lambda(a,b)} \langle \alpha, h_1\rangle + \langle \beta, h_2\rangle.$$

where

$$N_\lambda(a, b) = \{(\alpha, \beta) \in \mathbb{R}^N \times \mathbb{R}^N \text{ such that } f_\lambda(a, b, \alpha, \beta) = d_\lambda(a, b)\} \tag{6}$$

is the optimal set of solutions of the dual problem (3).

Proof. For $t > 0$ and $h_1, h_2, \alpha, \beta \in \mathbb{R}^N$, we define

$$\nabla^t_{h_1,h_2} f_\lambda(a, b, \alpha, \beta) = \frac{f_\lambda(a + th_1, b + th_2, \alpha, \beta) - f_\lambda(a, b, \alpha, \beta)}{t} = \alpha^T h_1 + \beta^T h_2.$$

Let $(a, b) \in \text{int}(\Sigma_N \times \Sigma_N)$ and $N_\lambda(a, b, r) = \{(\alpha, \beta) \in \mathbb{R}^N \times \mathbb{R}^N \text{ such that } f_\lambda(a, b, \alpha, \beta) > r\}$. By the existence of optimal solutions (see Proposition 2.1), there exists $(\alpha, \beta) \in N_\lambda(a, b, r)$ such that $\liminf_{t\to 0} \nabla^t_{h_1,h_2} f_\lambda(a, b, \alpha, \beta) = \liminf_{t\to 0} \alpha^T h_1 + \beta^T h_2 = \alpha^T h_1 + \beta^T h_2 > -\infty$. We can then apply Theorem 3.1 in [32] which gives the directional differentiability of d_λ. Remark that Proposition 2.1 in [32] provides sufficient conditions (by convexity of $(a, b) \mapsto f_\lambda(a, b, \alpha, \beta)$) to apply Theorem 3.1. Thus, the directional derivative $(d_\lambda)'_{h_1,h_2}(a, b)$ at (a, b) in the direction (h_1, h_2) exists, and it is given by

$$\begin{aligned}(d_\lambda)'_{h_1,h_2}(a, b) &= \lim_{r\nearrow d_\lambda(a,b)} \limsup_{t\to 0} \left(\sup_{(\alpha,\beta)\in N_\lambda(a,b,r)} \nabla^t_{h_1,h_2} f_\lambda(a, b, \alpha, \beta) \right) \\ &= \lim_{r\nearrow d_\lambda(a,b)} \left(\sup_{(\alpha,\beta)\in N_\lambda(a,b,r)} \alpha^T h_1 + \beta^T h_2 \right) = \sup_{(\alpha,\beta)\in N_\lambda(a,b)} \alpha^T h_1 + \beta^T h_2.\end{aligned}$$

Now, we argue as in the proof of Theorem 3 in [28]: to conclude that d_λ also admits a directional derivative in the Hadamard sense, it is sufficient to show that $(a, b) \mapsto d_\lambda(a, b)$ is locally Lipschitz (see Proposition 3.5 in [27]). The function $(a, b) \mapsto f_\lambda(a, b, \alpha, \beta)$ is linear for all α, β. Hence $(a, b) \mapsto d_\lambda(a, b)$ is convex. Since $\Sigma_N \times \Sigma_N$ is convex, we have that d_λ is locally Lipschitz on $int(\Sigma_N \times \Sigma_N)$, which completes the proof. □

Remark 3. *Since the boundary of the convex set $\Sigma_N \times \Sigma_N$ has zero d-Lebesgue measure, we have that $d_\lambda(a, b)$ is Hadamard directionally differentiable d-Lebesgue almost surely.*

2.3 Main result

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

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

We also define the multinomial covariance matrix

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

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

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

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

Theorem 2.4. *Recall that $K = \exp(-C/\lambda)$ is the matrix obtained by elementwise exponential of $-\frac{C}{\lambda}$. Then, the following central limit theorems holds for empirical Sinkhorn divergences.*

1. *Null hypothesis, i.e. $a = b$. Let $(u, v) \in \mathbb{R}_+^{N \times N}$ be a fixed point of the Sinkhorn map $S_{\{a,a\}}$ defined in (4)*

   (a) *H_0 - One sample.*

   $$\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a)) \xrightarrow{\mathcal{L}} \langle G, \lambda \log(u) \rangle. \tag{7}$$

   (b) *H_0 - Two samples. Let $\rho_{n,m} = \sqrt{(nm/(n+m))}$. If n and m tend to infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \gamma \in (0,1)$, then*

   $$\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, a)) \xrightarrow{\mathcal{L}} \langle G, \lambda \log(u) \rangle. \tag{8}$$

2. *Alternative case, i.e. $a \neq b$. Let $(u, v) \in \mathbb{R}_+^{N \times N}$ be a fixed point of the Sinkhorn map $S_{\{a,b\}}$*

   (a) *H_1 - One sample.*

   $$\sqrt{n}(d_\lambda(\hat{a}_n, b) - d_\lambda(a, b)) \xrightarrow{\mathcal{L}} \langle G, \lambda \log(u) \rangle. \tag{9}$$

   (b) *H_1 - Two samples. For $\rho_{n,m} = \sqrt{(nm/(n+m))}$ and $m/(n+m) \to \gamma \in (0,1)$,*

   $$\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b)) \xrightarrow{\mathcal{L}} \sqrt{\gamma} \langle G, \lambda \log(u) \rangle + \sqrt{1-\gamma} \langle H, \lambda \log(v) \rangle. \tag{10}$$

Proof. Following the proof of Theorem 1 in [28], we have that (e.g. thanks to Theorem 14.6 in [33])

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

since $n\hat{a}_n$ is a sample of a multinomial probability measure with probability a.

Therefore, for the one sample case, we apply the Delta-method for directionally differentiable functions in the sense of Hadamard (see Theorem 1 of Romisch in [24]). Thanks to Theorem 2.3, we directly get:

$$\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\lambda(a,a)} \langle G, \alpha\rangle \tag{11}$$

$$\sqrt{n}(d_\lambda(\hat{a}_n, b) - d_\lambda(a, b)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\lambda(a,b)} \langle G, \alpha\rangle \text{ for } a \neq b. \tag{12}$$

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 given in the statement of the Theorem. Then, applying again the delta-method for Hadamard directionally diffenrentiable functions, we obtain that for n and m tending to infinity such that $n \wedge m \to \infty$ and $m/(n+m) \to \gamma \in (0, 1)$,

$$\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\lambda(a,b)} \sqrt{\gamma}\langle G, \alpha\rangle + \sqrt{1-\gamma}\langle H, \beta\rangle. \tag{13}$$

In the null hypothesis case ($a = b$), this simplifies into

$$\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, a)) \xrightarrow{\mathcal{L}} \max_{(\alpha,\beta)\in N_\lambda(a,a)} \langle G, \alpha\rangle. \tag{14}$$

Now, thanks to Proposition 2.1, we know that there exists positive vectors $u \in \mathbb{R}_+^N$ and $v \in \mathbb{R}_+^N$ (unique up to scalar multiplication) such that an optimal solution in $N_\lambda(a, b)$ of d_λ is given by

$$\alpha^* = -\lambda \log(u), \quad \beta^* = -\lambda \log(v)$$

for a and b equal or not. From such results, for (u, v) obtained through Sinkhorn's algorithm (4), we can deduce that

$$\max_{(\alpha,\beta)\in N_\lambda(a,b)} \langle G, \alpha\rangle \overset{\mathcal{L}}{\sim} \max_{t\in\mathbb{R}} \langle G, -\lambda \log(u) + t\mathbb{1}_N\rangle \overset{\mathcal{L}}{\sim} \max_{t\in\mathbb{R}} \left(\langle G, -\lambda \log(u)\rangle + \langle G, t\mathbb{1}_N\rangle\right).$$

Moreover,

$$\langle G, t\mathbb{1}_N\rangle \overset{\mathcal{L}}{\sim} \mathcal{N}(t\mathbb{1}_N' \, \mathbb{E}(G), t\mathbb{1}_N'\Sigma(a)t\mathbb{1}_N) \overset{\mathcal{L}}{\sim} \mathcal{N}(0, t^2\mathbb{1}_N'\Sigma(a)\mathbb{1}_N) \overset{\mathcal{L}}{\sim} \mathcal{N}(0, 0) \overset{\mathcal{L}}{\sim} \delta_0$$

since G is centered in 0 and $\mathbb{1}_N'\Sigma(a)\mathbb{1}_N = 0$ for a in the simplex. Notice that $\langle G, -\lambda \log(u)\rangle \overset{\mathcal{L}}{\sim} \langle G, \lambda \log(u)\rangle$. Hence, let Y be a random variable of law δ_0. By independence, we have that $\langle G, -\lambda \log(u)\rangle + Y$ follows the same law as $\langle G, \lambda \log(u)\rangle$ since G is centered in 0. By the same process,

$$\max_{(\alpha,\beta)\in N_\lambda(a,b)} \sqrt{\gamma}\langle G, \alpha\rangle + \sqrt{1-\gamma}\langle H, \beta\rangle \overset{\mathcal{L}}{\sim} \sqrt{\gamma}\langle G, \lambda \log(u)\rangle + \sqrt{1-\gamma}\langle H, \lambda \log(v)\rangle.$$

Therefore we apply this result to the convergence in distribution obtained previously in (13) and (14), which concludes the proof. □

Distribution limits of empirical Sinkhorn divergences may also be characterized by the following result which follows from Theorem 1 of Romisch [24] using the property that $\Sigma_N \times \Sigma_N$ is a convex set.

Theorem 2.5. *The following asymptotic result holds for empirical Sinkhorn divergences.*

1. *One sample*

$$\sqrt{n}\left(d_\lambda(\hat{a}_n, b) - d_\lambda(a, b) - \max_{(\alpha,\beta)\in N_\lambda(a,b)} \langle \hat{a}_n - a, \alpha\rangle\right) \xrightarrow{\mathbb{P}} 0.$$

2. *Two samples - For* $\rho_{n,m} = \sqrt{(nm/(n+m))}$ *and* $m/(n+m) \to \gamma \in (0,1)$,

$$\rho_{n,m}\left(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b) - \max_{(\alpha,\beta)\in N_\lambda(a,b)} (\langle \hat{a}_n - a, \alpha\rangle + \langle \hat{b}_m - b, \beta\rangle)\right) \xrightarrow{\mathbb{P}} 0.$$

3 Use of the bootstrap for statistical inference

The results obtained in Section 2 on the distribution of empirical Sinkhorn divergences are only asymptotic, and 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 [11] for a general introduction to the bootstrap procedure.

Nevertheless, as carefully explained in [28], for a test statistic based on functions that are only Hadamard directionally differentiability a classical bootstrap procedure is not consistent. To overcome this issue, we decide to choose α and β in $N_\lambda(a, b)$ (6) such that their components sum up to zero. In this way the optimal solution of the dual problem (3) becomes unique as initially remarked in [6]. We denote this solution by $(\alpha^0_\lambda, \beta^0_\lambda)$, and we let $N^0_\lambda(a, b) = \{(\alpha^0_\lambda, \beta^0_\lambda)\}$. Under this additional normalization, the previous results remain true. In particular, the directional derivative of d_λ at (a, b) becomes

$$d'_\lambda(a, b) : (h_1, h_2) \mapsto \langle \alpha^0_\lambda, h_1\rangle + \langle \beta^0_\lambda, h_2\rangle,$$

which is a linear map. Hence, by Proposition 2.1 in [12], the functional $(a, b) \mapsto d_\lambda(a, b)$ is Hadamard differentiable in the usual sense on $\text{int}(\Sigma_N \times \Sigma_N)$. We can thus apply the Delta-method to prove consistency of the bootstrap in our setting using the bounded Lipschitz metric defined below.

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

$$d_{BL}(\mu, \nu) = \sup_{h\in BL_1(\Omega)} \left| \int h d\mu - \int h d\nu \right|$$

where $BL_1(\Omega)$ is the set of real functions $\Omega \to \mathbb{R}$ with a Lipschitz norm bounded by 1.

Our main result adapted on the use of bootstrap samples can be stated as follows.

Theorem 3.2. *For $X_1, \dots, X_n \overset{\mathcal{L}}{\sim} a$ and $Y_1, \dots, Y_m \overset{\mathcal{L}}{\sim} b$, let $\hat{a}_n^*$ and $\hat{b}_m^*$ be boostrap versions of $\hat{a}_n$ and $\hat{b}_m$ respectively.*

1. *One sample case: $\sqrt{n}(d_\lambda(\hat{a}_n^*, b) - d_\lambda(\hat{a}_n, b))$ converges in distribution (conditionally on $X_1, \dots, X_n$) to $\langle G, \alpha_\lambda^0 \rangle$ for the BL metric, in the sense that*

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}(h(\sqrt{n}(d_\lambda(\hat{a}_n, b) - d_\lambda(\hat{a}_n^*, b)))|X_1, \dots, X_n] - \mathbb{E}[h\langle G, \alpha_\lambda^0 \rangle]| \xrightarrow{\mathbb{P}} 0$$

2. *Two samples case: $\rho_{n,m}(d_\lambda(\hat{a}_n^*, \hat{b}_m^*) - d_\lambda(\hat{a}_n, \hat{b}_m))$ converges in distribution (conditionally on $X_1, \dots, X_n$, $Y_1, \dots, Y_m$) to $\sqrt{\gamma}\langle G, \alpha_\lambda^0 \rangle + \sqrt{1-\gamma}\langle H, \beta_\lambda^0 \rangle$ for the BL metric, in the sense that*

$$\sup_{h \in BL_1(\mathbb{R})} |\mathbb{E}(h(\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(\hat{a}_n^*, \hat{b}_m^*)))|X_1, \dots, X_n, Y_1, \dots, Y_m] - \mathbb{E}[h(\sqrt{\gamma}\langle G, \alpha_\lambda^0 \rangle + \sqrt{1-\gamma}\langle H, \beta_\lambda^0 \rangle)]| \xrightarrow{\mathbb{P}} 0$$

Proof. We only prove the one sample case since both convergence can be shown by 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 $X_1, \dots, X_n$) in distribution to G by Theorem 3.6.1 in [30]. Theorem 3.9.11 in the same book, on the consistency of the Delta-method combined with bootstrap, allows us to conclude. □

4 Numerical experiments with synthetic data

We propose to illustrate Theorem 2.4 and Theorem 3.2 with simulated data consisting of random measures supported on a $p \times p$ square grid of regularly spaced points $(x_i)_{i=1,\dots,N}$ in $\mathbb{R}^2$ (with $N = p^2$) for p ranging from 5 to 20. We use the squared Euclidean distance. Therefore, the cost C scales with the size of the grid. The range of interesting values for λ is thus closely linked to the size of the grid (as it can be seen in the expression of $K = \exp(-C/\lambda)$. Hence, $\lambda = 100$ for a 5×5 grid corresponds to more regularization than $\lambda = 100$ for a 20×20 grid.

We ran our experiments on Matlab using the accelerate version [29][1] of the Sinkhorn transport algorithm [5]. Furthermore, we considered the numerical logarithmic stabilization described in [25] which allows to handle small values of λ.

4.1 Convergence in distribution

We first illustrate the convergence in distribution of empirical Sinkhorn divergences (as stated in Theorem 2.4) for either the hypothesis H_0 with one sample, or the hypothesis H_1 with two samples.

Hypothesis H_0 - One sample. We consider the case where a is the uniform distribution on a square 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 $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$, and we use a kernel density estimate

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

(with a data-driven bandwith) to compare the distribution of these realizations to the density of the Gaussian distribution $\langle G, \lambda \log(u) \rangle$. The results are reported in Figure 1.

It can be seen that the convergence of empirical Sinkhorn divergences to its asymptotic distribution ($n \to \infty$) is relatively slow. Moreover, for a fixed number n of observations, the convergence becomes slower as λ increases. We can also notice that for various values of (n, λ), the non-asymptotic distribution of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ seems to be non-Gaussian. This justifies the use of the bootstrap procedure described in Section 3.

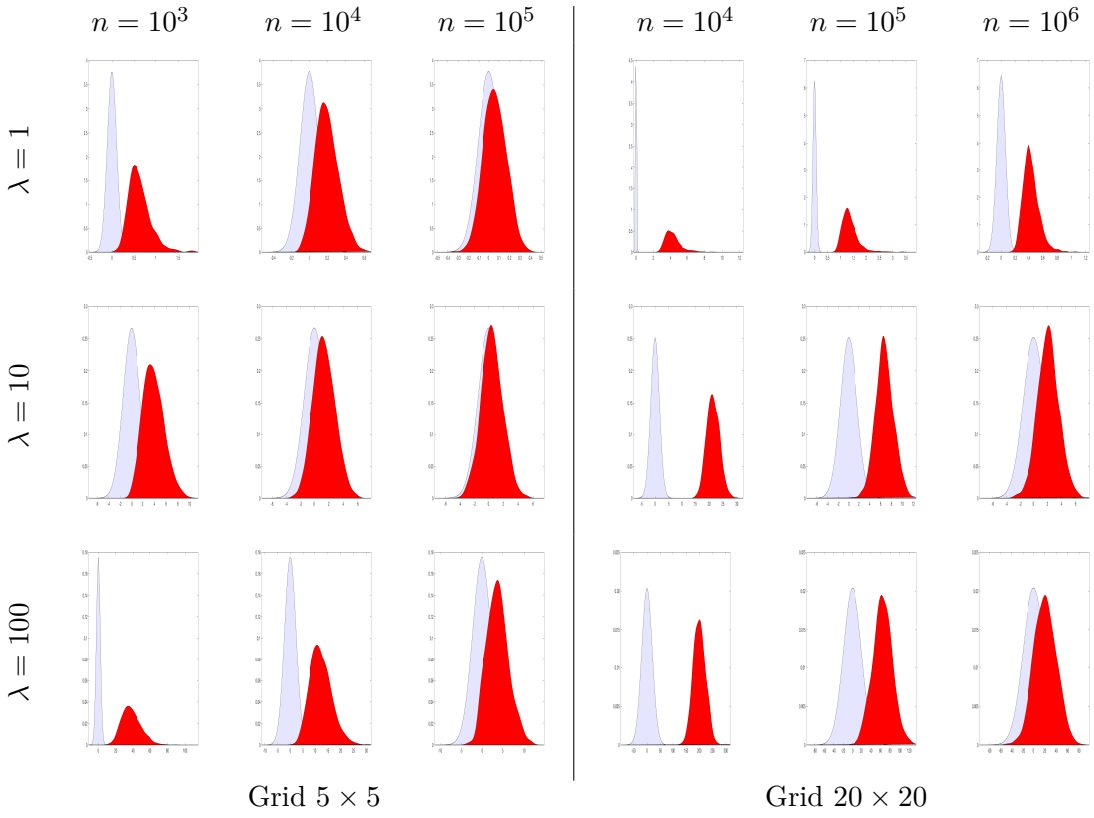


Figure 1: Hypothesis H_0 with one sample. Illustration of the convergence in distribution of empirical Sinkhorn divergences for a 5×5 grid (left) and a 20×20 grid (right), for $\lambda = 1, 10, 100$ and n ranging from 10^3 to 10^6. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ (resp. $\langle G, \lambda \log(u) \rangle$).

Let us now shed some light on the bootstrap procedure described in Section 3. The results on bootstrap experiments are reported in Figure 2. From the uniform distribution a, we generate one random distribution $\hat{a}_n$. The value of the realization $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ is represented by the red vertical lines in Figure 2.

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 bandwith) to compare the distribution of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ to the distribution of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ displayed in Figure 1. The green vertical lines in Figure 2 represent a confidence interval of level 95%. The observation represented by the red vetical line is consistently located with respect to this confidence interval, and the density estimated by bootstrap decently captures the shape of

the non-asymptotic distribution of Sinkhorn divergences.

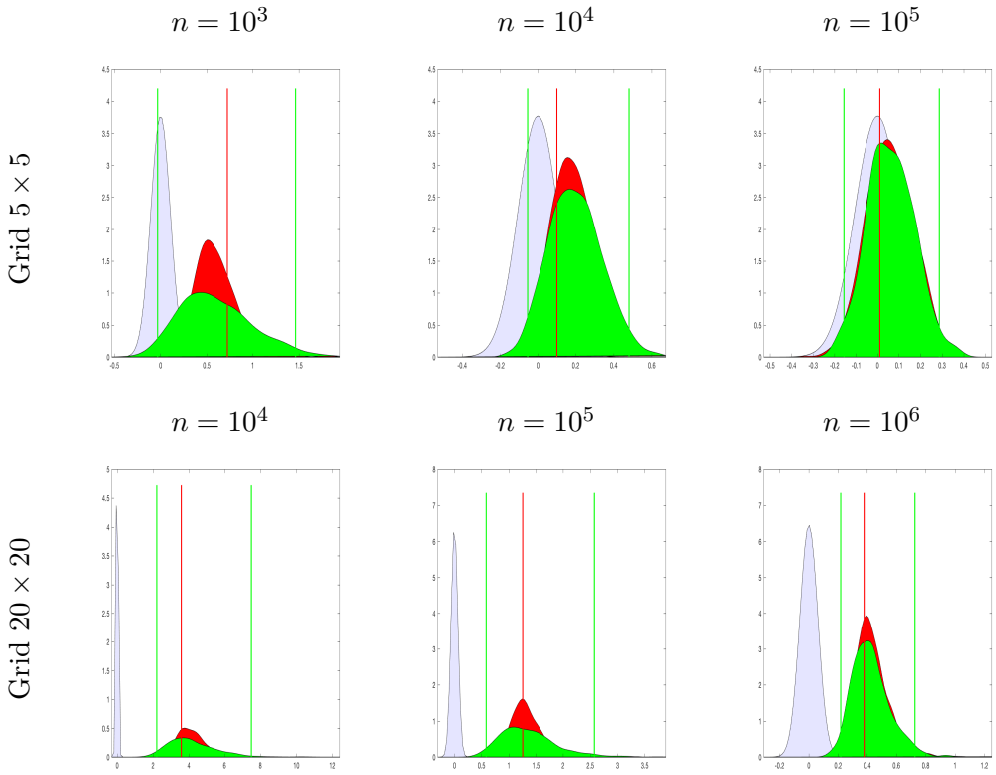


Figure 2: Hypothesis H_0 with one sample. Illustration of the bootstrap with $\lambda = 1$ and two grids of size 5×5 and 20×20 to approximate the non-asymptotic distribution of empirical Sinkhorn divergences. Densities in red (resp. light blue) represent the distribution of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ (resp. $\langle G, \lambda \log(u) \rangle$). The green density represents the distribution of the random variable $\sqrt{n}(d_\lambda(\hat{a}_n^*, a) - d_\lambda(\hat{a}_n, a))$ in Theorem 3.2.

Hypothesis H_1 - Two samples We consider now the setting where a is still a uniform distribution, and

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

is a distribution with linear trend depending on a slope parameter $\theta \geq 0$ that is fixed to 0.5, see Figure 3.

As previously, we run $M = 10^3$ experiments to obtain a kernel density estimation of the distribution of

$$\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b)),$$

that we compare to the density of the Gaussian variable with mean 0 and variance

$$\lambda \sqrt{\gamma \log(u)^t \Sigma(a) \log(u) (1 - \gamma) \log(v)^t \Sigma(b) \log(v)}.$$

The results are reported in Figure 4. The convergence of empirical Sinkhorn divergences to their asymptotic distribution seems to be much faster under the hypothesis H_1, but increasing the regularization parameter still makes this convergence slower.

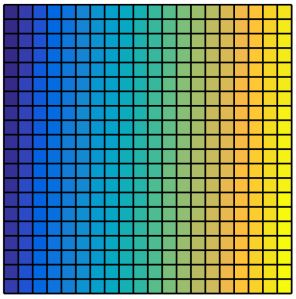


Figure 3: Example of a distribution b with linear trend (with slope parameter $\theta = 0.5$ on a 20×20 grid).

Remark 4. *A possible explanation for the slow convergence under the hypothesis H_0 is that, in this setting, the Sinkhorn divergence $d_\lambda(a, a)$ is very close to 0, but as soon as we generate an empirical measure $\hat{a}_n$, the value of $d_\lambda(\hat{a}_n, a)$ seems to explode in comparison to the divergence between a and itself.*

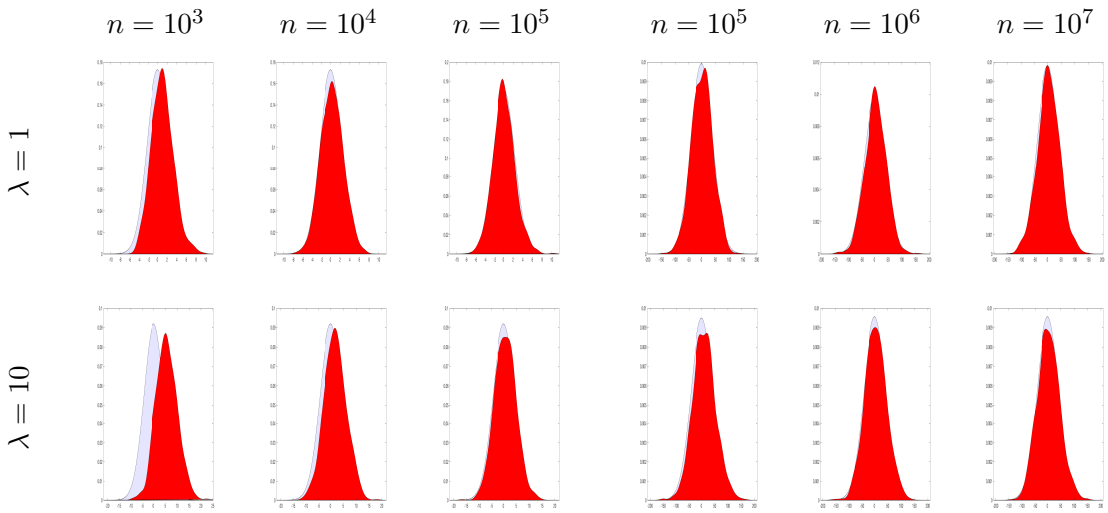


Figure 4: Hypothesis H_1 - two samples. Illustration of the convergence in distribution of empirical Sinkhorn divergences for a 5×5 grid (left) and a 20×20 grid (right), for $\lambda = 1, 10$, $n = m$ and n ranging from 10^3 to 10^7. Densities in red (resp. blue) represent the distribution of $\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b))$ (resp. $\sqrt{\gamma}\langle G, \lambda \log(u)\rangle + \sqrt{1-\gamma}\langle H, \lambda \log(v)\rangle$ with $\gamma = 1/2$).

We also report in Figure 5 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 $\sqrt{n}(d_\lambda(\hat{a}_n, \hat{b}_n) - d_\lambda(a, b))$ is represented by the red vertical lines in Figure 5. 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 distribution of $\rho_{n,m}(d_\lambda(\hat{a}_n^*, \hat{b}_m^*) - d_\lambda(\hat{a}_n, \hat{b}_m))$ to the red distribution of $\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b))$ displayed

in Figure 5. The green vertical lines in Figure 5 represent a confidence interval of level 95%. The observation represented by the red vertical line is consistently located with respect to this confidence interval, and the green density estimated by bootstrap captures very well the shape and location of the non-asymptotic distribution of Sinkhorn divergences.

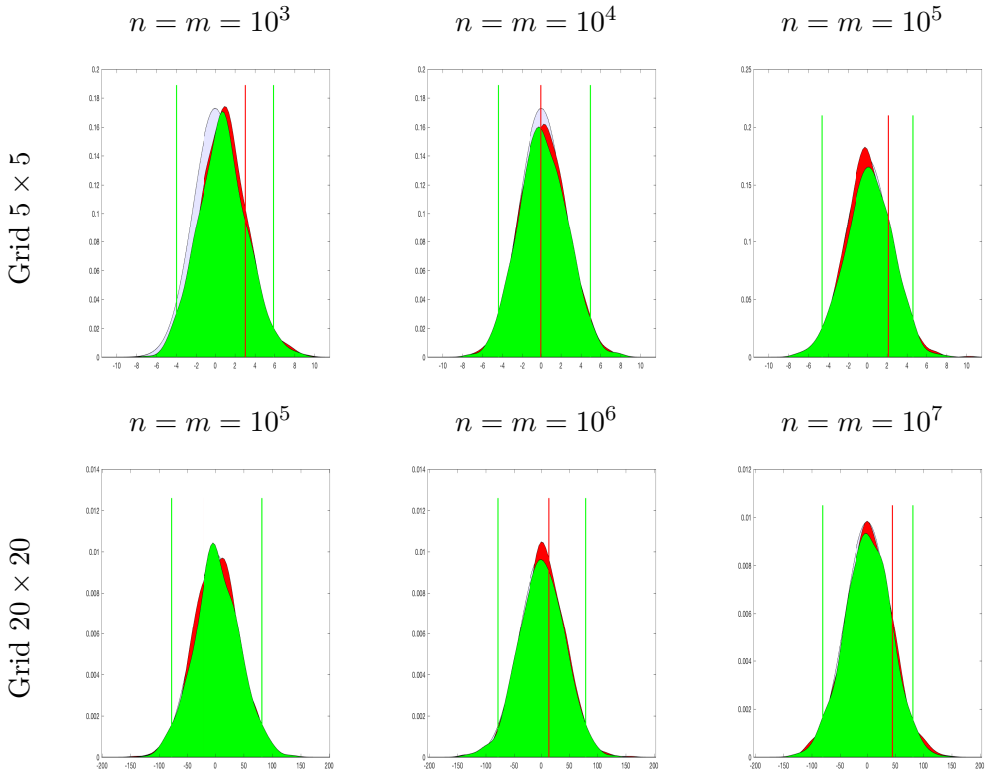


Figure 5: Hypothesis H_1 - two samples. Illustration of the bootstrap with $\lambda = 1$ and two grids of size 5×5 and 20×20 to approximate the non-asymptotic distribution of empirical Sinkhorn divergences. Densities in red (resp. blue) represent the distribution of $\rho_{n,m}(d_\lambda(\hat{a}_n, \hat{b}_m) - d_\lambda(a, b))$ (resp. $\sqrt{\gamma}\langle G, \lambda \log(u)\rangle + \sqrt{1-\gamma}\langle H, \lambda \log(v)\rangle$). The green density is the distribution of the random variable $\rho_{n,m}(d_\lambda(\hat{a}^*_n, \hat{b}^*_m) - d_\lambda(\hat{a}_n, \hat{b}_m))$ in Theorem 3.2.

4.2 Estimation of test power using the bootstrap

One sample - distribution with linear trend and varying slope parameter. We illustrate the consistency and usefulness of the bootstrap procedure 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 empirical Sinkhorn divergences. For this purpose, we choose a to be uniform on a 5×5 grid, and b to be a distribution with linear trend whose slope parameter θ is ranging from 0 to 0.15. We assume that we observe a single realization of an empirical measure $\hat{a}_n$ sampled from a with $n = 10^3$. Then, we generate $M = 10^3$ bootstrap samples of random measures $\hat{a}^*_{n,j}$ from $\hat{a}_n$ (with $1 \leq j \leq M$), which allows the computation of the p-value

$$p\text{-value} = \#\{j \text{ such that } \sqrt{n}|d_\lambda(\hat{a}^*_{n,j}, b) - d_\lambda(\hat{a}_n, b)| \geq \sqrt{n}|d_\lambda(\hat{a}_n, b) - d_\lambda(a, b)|\}/M.$$

This experiments is repeated 100 times, in order to estimate the power (at level α) of a test based on $\sqrt{n}(d_\lambda(\hat{a}_n, b) - d_\lambda(a, b))$ by comparing the resulting sequence of p-values to the value α. The results are reported in Figure 6.

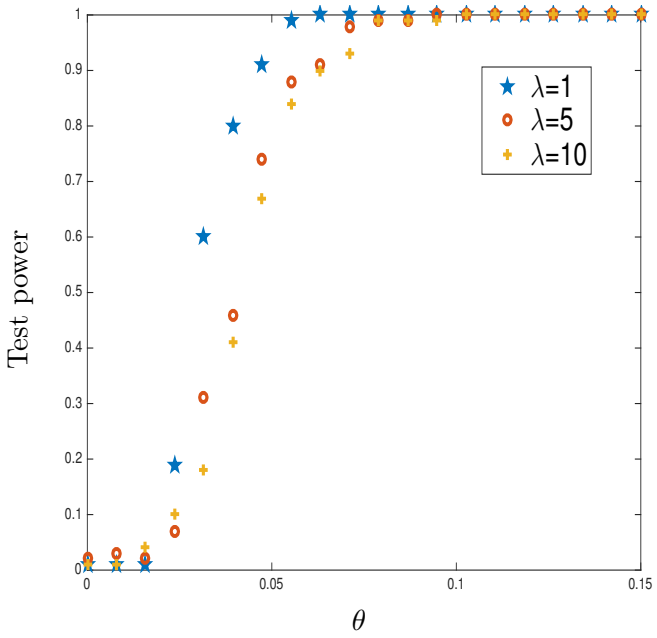


Figure 6: Test power (probability of accepting H_0 knowing that H_1 is true) on a 5×5 grid in the one sample case, as a function of the slope parameter θ ranging from 0 to 0.15 for $\lambda = 1$ (blue), $\lambda = 5$ (orange) and $\lambda = 10$ (yellow), with $n = 10^3$.

It can be seen that this test is a good discriminant, especially when λ is small. As soon as the slope θ increases and b sufficiently differs from a, then the probability of rejecting H_0 increases. Moreover, for a fixed value of the slope parameter θ of distribution b, the test power becomes larger as λ gets smaller. This suggests the use of a small regularization parameter λ to be more accurate for discriminating two measures using statistical testing based on empirical Sinkhorn divergences.

5 Analysis of real data

We consider a dataset containing the locations of reported incidents of crime (with the exception of murders) in Chicago in 2014 which is publicly available[2], and that has been recently studied in [3] and [17]. Victims' addresses are shown at the block level only (specific locations are not identified) in order to (i) protect the privacy of victims and (ii) have a sufficient amount of data for the statistical analysis. The city of Chicago is represented as a two-dimensional grid $\mathcal{X} = \{x_1, \ldots, x_N\}$ of size $N = 27 \times 18 = 486$ of equi-spaced points $x_i = (x_i^{(1)}, x_i^{(2)}) \in [1, 27] \times [1, 18] \subset \mathbb{R}^2$. For each month $1 \leq k \leq 12$ of the year 2014, the spatial locations of reported incidents of crime in Chicago are available. This yields to a

[2] https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2/data

dataset made of 12 empirical measures

$$\hat{\mu}_k = \sum_{i=1}^{N} \hat{a}_i^{(k)} \delta_{x_i} \text{ for } 1 \leq k \leq 12,$$

where $\hat{a}_i^{(k)}$ is the relative frequency of reported crimes for month k at location x_i. We denote by $n = n_k$ the number of reported crimes for month k. This dataset is displayed in Figure 7 and 8. To compute the cost matrix C, we use the squared Euclidean distance between the spatial locations $x_i \in \mathbb{R}^2$.

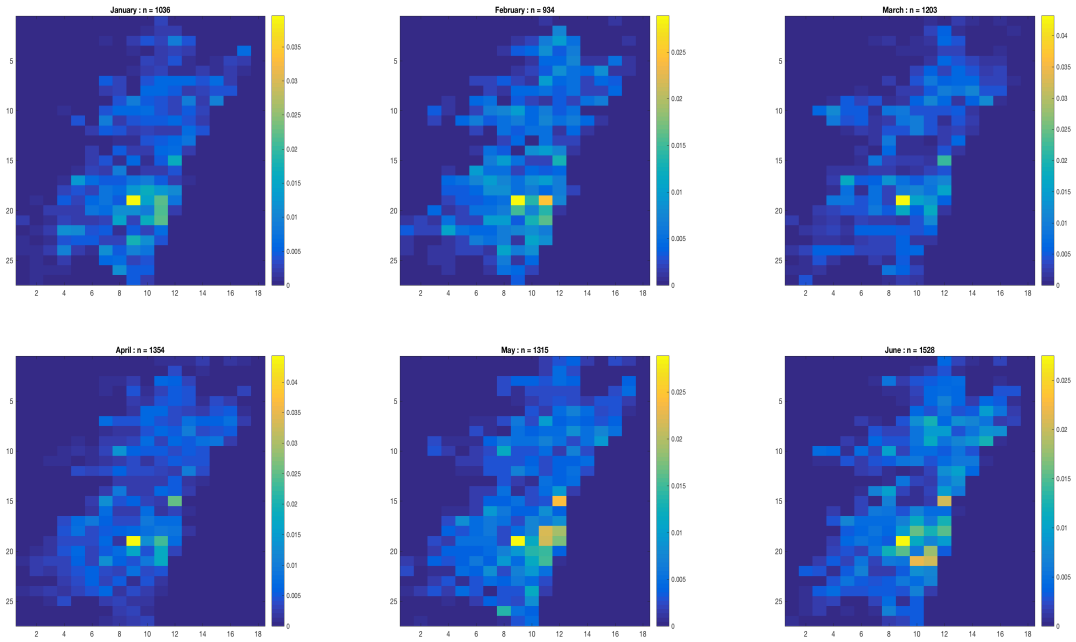


Figure 7: Spatial locations of reported incidents (relative frequencies) of crime in Chicago for the first 6 months of 2014 over a two-dimensional grid of size 27×18.

5.1 Testing the hypothesis of uniform distribution of crimes locations

We first test the null hypothesis that the distribution of crimes locations over the whole year 2014 is uniform. To this end, we consider the Euclidean barycenter of the dataset $(\hat{\mu}_k)_{1 \leq k \leq n}$ defined as

$$\bar{\mu}_{12} = \frac{1}{12} \sum_{k=1}^{12} \hat{\mu}_k = \sum_{i=1}^{N} \bar{a}_i \delta_{x_i}$$

which represents the locations of crime in 2014. This discrete measure is displayed in Figure 9(a). It can be seen that $\bar{\mu}_{12}$ is a discrete empirical measure consisting of $n = 16104$ observations such that $\bar{a}_i = 0$ for many locations x_i. We use the one sample testing procedure described previously, and a bootstrap approach to estimate the distribution of the test statistics

$$\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$$

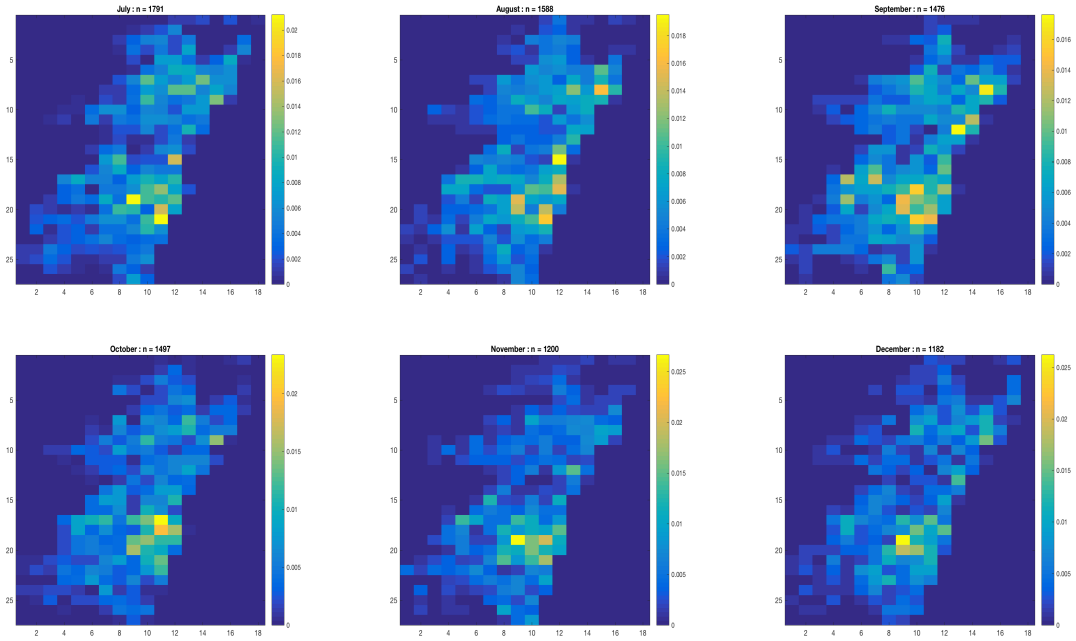


Figure 8: Spatial locations of reported incidents (relative frequencies) of crime in Chicago for the last 6 months of 2014 over a two-dimensional grid of size 27×18.

with $\hat{a}_n = \bar{\mu}_{12}$ and a the uniform distribution over the support of $\bar{\mu}_{12}$ defined as $\{x_i \ : \ \bar{a}_i \neq 0, 1 \leq i \leq N\}$, see Figure 9(b). We report results for $\lambda = 1$ and $\lambda = 5$ by displaying in Figure 9(cd) an estimation of the density of the bootstrap statistics $\sqrt{n}(d_\lambda(\hat{a}_n^*, a) - d_\lambda(\hat{a}_n, a))$. For both values of λ, the value of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ is outside the support of this density, and the null hypothesis that crimes are uniformly distributed (over the support of $\bar{\mu}_{12}$ is thus rejected.

5.2 Testing the hypothesis of equal distributions between months

We propose now to investigate the possibility of equal distributions of crime locations between different months. To this end, we first compute a reference measure using data from the first 6 months. Under the assumption that the distribution of crime locations does not change from one month to another, it is natural to consider the Euclidean barycenter

$$\bar{\mu}_6 = \frac{1}{6} \sum_{k=1}^{6} \hat{\mu}_k,$$

as a reference measure to which the data from the last 6 months of 2014 can be compared. The measure $\bar{\mu}_6$ is displayed in Figure 10(a) and Figure 11(a).

One sample testing. We use the one sample testing procedure described previously, and a bootstrap approach to estimate the distribution of the test statistics

$$\sqrt{n_k}(d_\lambda(\hat{a}_{n_k}, a) - d_\lambda(a, a))$$

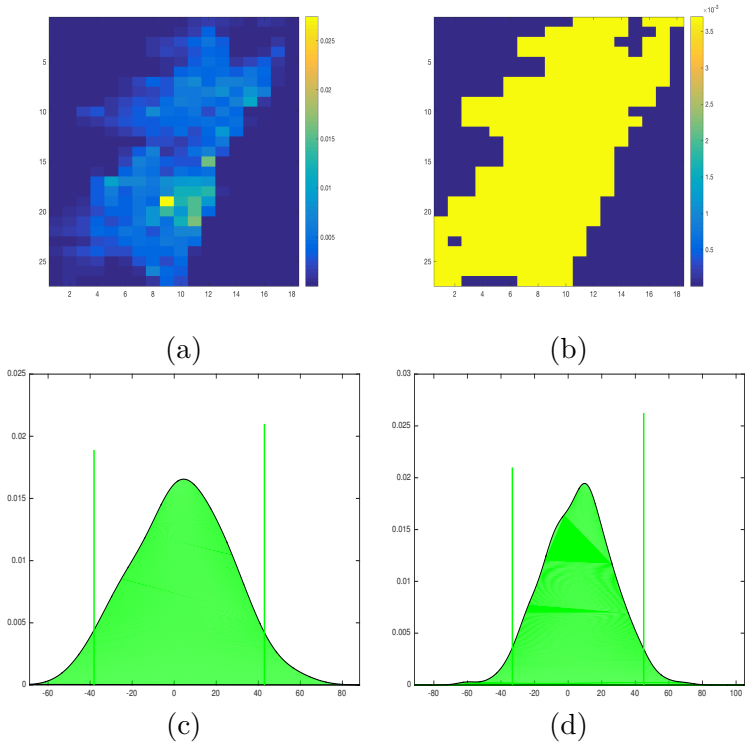


Figure 9: Testing uniform distribution of crimes locations. (a) Euclidean barycenter $\bar{\mu}_{12}$ (empirical measure corresponding to locations of crime in Chicago for the whole year 2014 over a two-dimensional grid of size 27×18), (b) Uniform distribution a over the support of $\bar{\mu}_{12}$. Green densities represent the distribution of the bootstrap statistics $\sqrt{n}(d_\lambda(\hat{a}_n^*, a) - d_\lambda(\hat{a}_n, a))$ (vertical bars represent a confidence interval of level 95%) for (c) $\lambda = 1$ and (d) $\lambda = 5$. The value of $\sqrt{n}(d_\lambda(\hat{a}_n, a) - d_\lambda(a, a))$ (with $\hat{a}_n = \bar{\mu}_{12}$) is outside the support $[-100, 100]$ for each value of λ, and it is thus not represented.

.

with $a = \bar{\mu}_6$ and $\hat{a}_{n_k} = \hat{\mu}_k$, for $7 \leq k \leq 12$. We report results for $\lambda = 1$ by displaying in Figure 10 an estimation of the density of the bootstrap statistics $\sqrt{n_k}(d_\lambda(\hat{a}_{n_k}^*, a) - d_\lambda(\hat{a}_{n_k}, a))$, and the values of the observations $\sqrt{n_k}(d_\lambda(\hat{a}_{n_k}, a) - d_\lambda(a, a))$ for the last 6 months of 2014. It can be seen that, at level 5%, the null hypothesis that the distribution of crime locations is equal to the reference measure $\bar{\mu}_6$ is accepted for the months of September, October, November and December, but that it is rejected for the months of July and August.

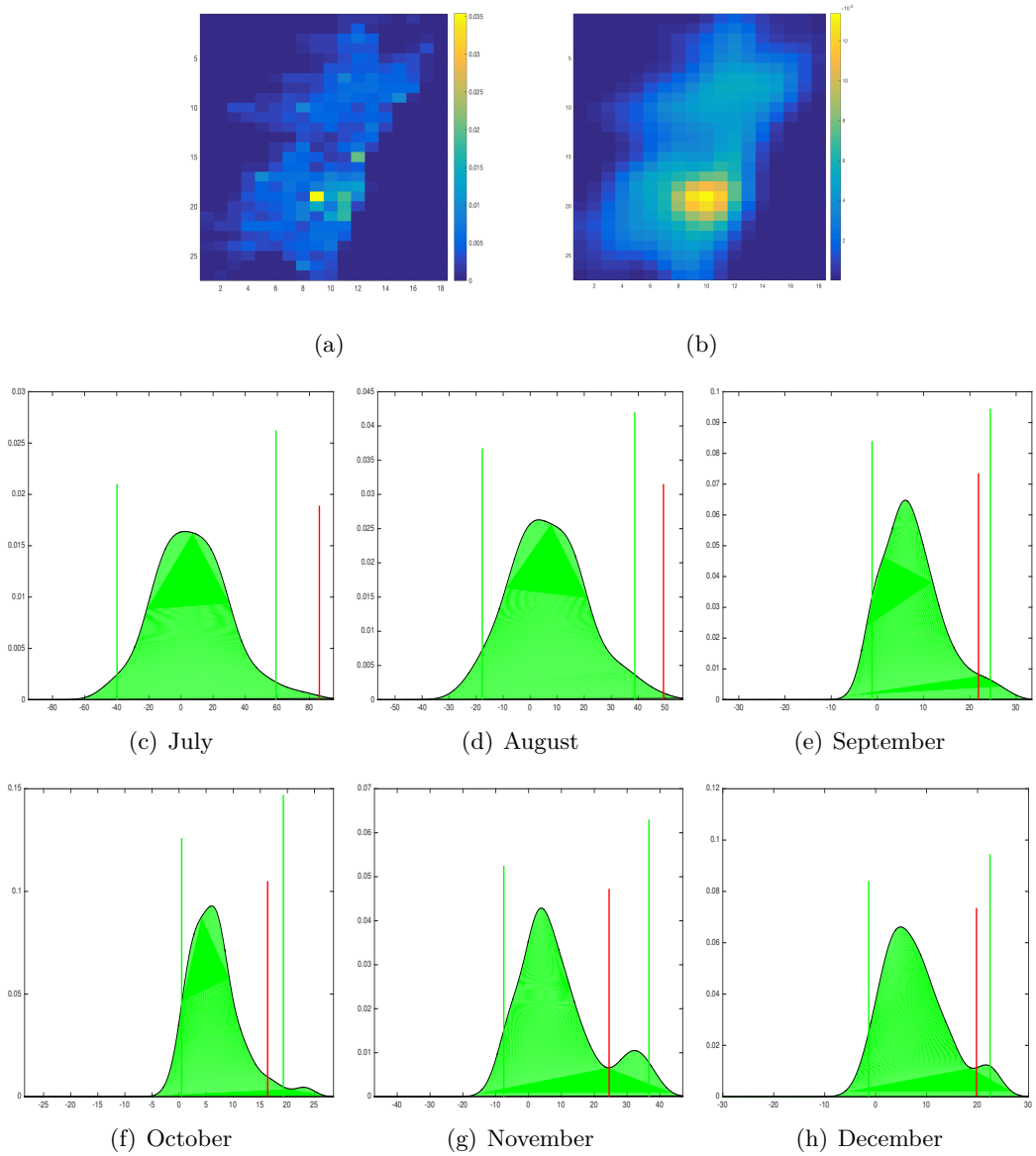
Alternatively, one may think of using a smoothed Wasserstein barycenter $\bar{\mu}_6^\lambda$ of the data $(\hat{\mu}_k)_{1 \leq k \leq 6}$ as a reference measure that is defined as

$$\bar{\mu}_6^\lambda = \operatorname*{arg\,min}_{\mu \in \mathcal{P}_p(\mathcal{X})} \frac{1}{6} \sum_{k=1}^{6} p_\lambda(\hat{\mu}_k, \mu).$$

To compute such a smoothed Wasserstein barycenter, we use the algorithmic approach proposed in [7], and we display $\bar{\mu}_6^\lambda$ for $\lambda = 1$ in Figure 10(b) and $\lambda = 0.3$ in Figure 11(b).

For $\lambda = 1$, this smoothed Wasserstein barycenters is visually quite different from the measures $(\hat{\mu}_k)_{7 \leq k \leq 12}$ that are displayed in Figure 8. For $\lambda = 1$, we found that using $\bar{\mu}_6^\lambda$ as a

reference measure in one sample testing (with $\hat{a}_{n_k} = \hat{\mu}_k$ and $a = \bar{\mu}_6^\lambda$) leads to reject the null hypothesis that the distribution of crime locations is equal to $\bar{\mu}_6^\lambda$ for all $7 \leq k \leq 12$ (last 6 months of 2014). As a consequence we do not display the corresponding results.



(a) (b)

(c) July (d) August (e) September

(f) October (g) November (h) December

Figure 10: Testing equality of distributions over months for $\lambda = 1$ with the Euclidean barycenter as a reference measure. (a) Euclidean barycenter $\bar{\mu}_6$ (empirical measure corresponding to locations of crime in Chicago for the first 6 months of 2014). (b) Smoothed Wassertein barycenter $\bar{\mu}_6^\lambda$ of the measures $(\hat{\mu}_k)_{1 \leq k \leq 6}$ for $\lambda = 1$. (c)-(h) Green densities represent the distribution of the bootstrap statistics $\sqrt{n_k}(d_\lambda(\hat{a}^*_{n_k}, a) - d_\lambda(\hat{a}_{n_k}, a))$ for the last 6 months of 2014, with $a = \bar{\mu}_6$ and $\hat{a}_{n_k} = \hat{\mu}_k$, for $7 \leq k \leq 12$. The green vertical bars represent a confidence interval of level 95% for each density. The red vertical bars represent the value of $\sqrt{n_k}(d_\lambda(\hat{a}_{n_k}, a) - d_\lambda(a, a))$.

For $\lambda = 0.3$, the Wasserstein barycenter $\bar{\mu}_6^\lambda$ is a slightly smoothed version of the Euclidean one $\bar{\mu}_6$. We display in Figure 11 an estimation of the density of the bootstrap statistics $\sqrt{n_k}(d_\lambda(\hat{a}^*_{n_k}, a) - d_\lambda(\hat{a}_{n_k}, a))$, and the values of the observations $\sqrt{n_k}(d_\lambda(\hat{a}_{n_k}, a) - d_\lambda(a, a))$ for the last 6 months of 2014, with $a = \bar{\mu}_6^\lambda$ and $\lambda = 0.3$. At level 5%, the null hypothesis that the distribution of crime locations is equal to the reference measure $\bar{\mu}_6^\lambda$ is accepted for the months of November and December, just as in the case where the Euclidean barycenter $\bar{\mu}_6$ is the reference measure. However, the null hypothesis is rejected for the four others months July, August, September and October.

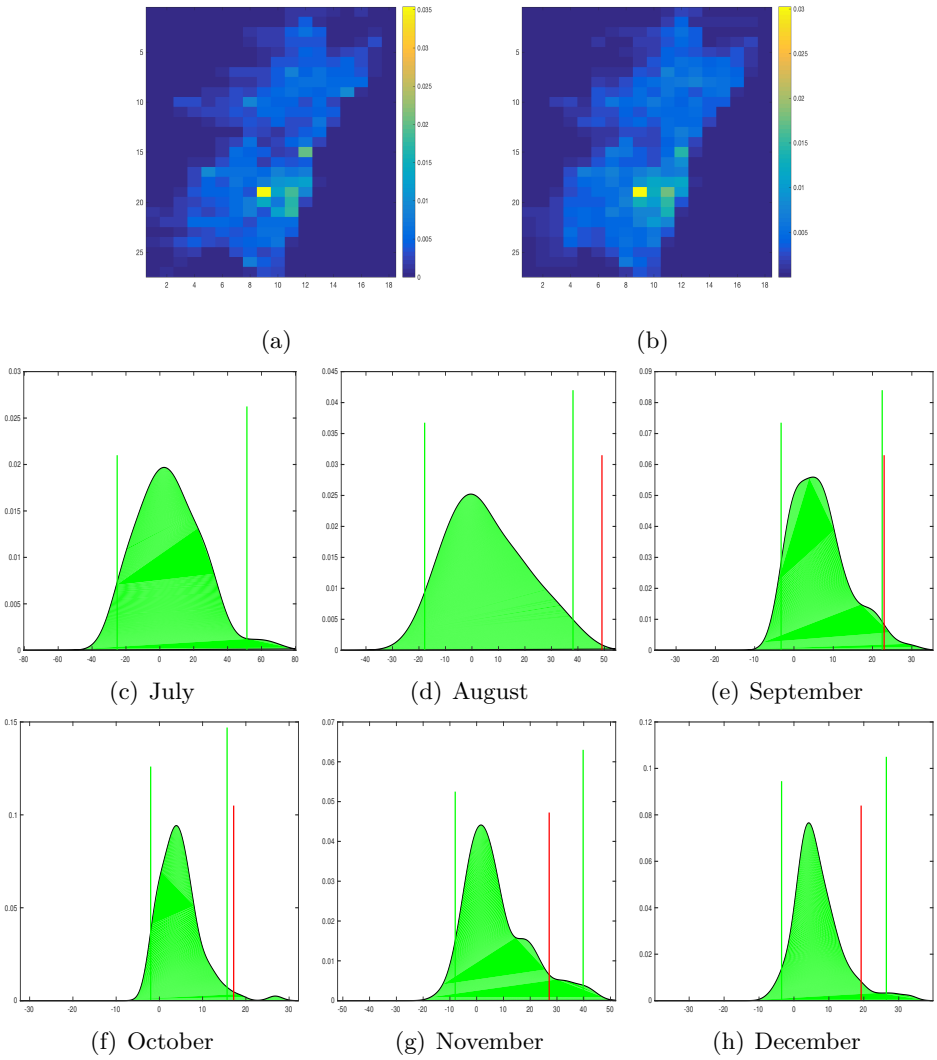
Two samples testing. We finally consider the problem of testing the hypothesis that the distributions of crime locations between two months (from July to December) are equal to the reference measure $a = \bar{\mu}_6$ (Euclidean barycenter over the first 6 months of 2014) using the two samples test statistic based on Sinkhorn divergence for $\lambda = 1$ and $\lambda = 5$ combined with a bootstrap procedure. We report in Table 1 and Table 2 the estimated p-values corresponding to such tests for all pairs of different months from July to December 2014. For both values of λ the interpretation of the results is similar. They tend to support the hypothesis that the distribution of crime locations is the same when comparing two months among September, October, November and December, and that this distribution is different when the comparison is done with the month of July. The results for August are more difficult to interpret, as it can be concluded that the distribution of crime locations for this month is equal to that of July, September, October and December.

As remarked in [28], 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 compare the results that have been obtained using our approach with those given by a kernel based test proposed in [1] that is implemented in the R package `ks`. The test statistics in [1] uses the integrated square distance between two kernel-based density estimates computed from two empirical measures with a data-based choice of bandwidth. We report in Table 3 the p-values corresponding to this test for all pairs of different months from July to December 2014. It can be seen that the p-values obtained with this test are larger than those obtained with our testing procedure. Nevertheless, the conclusions on the equality of distributions of crime locations between different months are roughly the same than previously.

6 Future works

We intend to further investigate the benefits of the use of Sinkhorn divergences to propose novel testing procedure 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. [7] and references therein) for the analysis of variance of multiple and multivariate random measures (MANOVA). However, as pointed out in [7], a critical issue in this setting will be the choice of the regularization parameter λ, as it has a large influence on the shape of the estimated Wasserstein barycenter. Our simulations in Section 5 show that using a smoothed Wasserstein barycenter as a reference measure may lead to different results from the use of an Euclidean barycenter for hypothesis testing of the equality of distributions.

Another issue is that, for one or two samples testing, the use of Sinkhorn divergences leads



(a) (b)

(c) July (d) August (e) September

(f) October (g) November (h) December

Figure 11: Testing equality of distributions over months for $\lambda = 0.3$ with the smoothed Wasserstein barycenter as a reference measure. (a) Euclidean barycenter $\bar{\mu}_6$ (empirical measure corresponding to locations of crime in Chicago for the first 6 months of 2014). (b) Smoothed Wassertein barycenter $\bar{\mu}_6^\lambda$ of the measures $(\hat{\mu}_k)_{1\leq k\leq 6}$ for $\lambda = 0.3$. (c)-(h) Green densities represent the distribution of the bootstrap statistics $\sqrt{n_k}(d_\lambda(\hat{a}^*_{n_k}, a) - d_\lambda(\hat{a}_{n_k}, a))$ for the last 6 months of 2014, with $a = \bar{\mu}_6^\lambda$ and $\hat{a}_{n_k} = \hat{\mu}_k$, for $7 \leq k \leq 12$. The green vertical bars represent a confidence interval of level 95% for each density. The red vertical bars represent the value of $\sqrt{n_k}(d_\lambda(\hat{a}_{n_k}, a) - d_\lambda(a, a))$.

to a biased statistics in the sense that its expectation $d_\lambda(a, b)$ is not equal to zero under the hypothesis that $a = b$. A possible alternative to avoid this issue would be to use the so-called notion of Sinkhorn loss defined as

$$\bar{d}_\lambda(a, b) := 2d_\lambda(a, b) - d_\lambda(a, a) - d_\lambda(b, b),$$

that has been recently introduced in [16], and which satisfies the property that $\bar{d}_\lambda(a, b) = 0$ when $a = b$. An interesting extension of the results in this paper would thus be to develop

	July	August	September	October	November	December
July	1	0.07	0.04	0.01	$< 10^{-2}$	0.08
August		1	0.16	0.14	0.01	0.12
September			1	0.18	0.07	0.20
October				1	0.06	0.05
November					1	0.10
December						1

Table 1: Two samples testing of equal distributions between pairs of different months from July to December using a test statistics based on Sinkhorn divergence for $\lambda = 1$ with reference measure $a = \bar{\mu}_6$ (Euclidean barycenter over the first 6 months of 2014). The table reports estimated p-values using a bootstrap procedure for the test statistics $\rho_{n_k,n_\ell}(d_\lambda(\hat{a}_{n_k}, \hat{b}_{n_\ell}) - d_\lambda(a,a))$ (with $\hat{a}_{n_k} = \hat{\mu}_k$ and $\hat{b}_{n_\ell} = \hat{\mu}_\ell$) for $7 \leq k \leq \ell \leq 12$.

	July	August	September	October	November	December
July	1	0.12	0.04	0.01	$< 10^{-2}$	0.05
August		1	0.25	0.11	0.01	0.10
September			1	0.40	0.06	0.20
October				1	0.06	0.05
November					1	0.06
December						1

Table 2: Two samples testing of equal distributions between pairs of different months from July to December using a test statistics based on Sinkhorn divergence for $\lambda = 5$ with reference measure $a = \bar{\mu}_6$ (Euclidean barycenter over the first 6 months of 2014). The table reports estimated p-values using a bootstrap procedure for the test statistics $\rho_{n_k,n_\ell}(d_\lambda(\hat{a}_{n_k}, \hat{b}_{n_\ell}) - d_\lambda(a,a))$ (with $\hat{a}_{n_k} = \hat{\mu}_k$ and $\hat{b}_{n_\ell} = \hat{\mu}_\ell$) for $7 \leq k \leq \ell \leq 12$.

	July	August	September	October	November	December
July	1	0.14	0.04	0.04	0.10	0.09
August		1	0.25	0.06	0.12	0.16
September			1	0.11	0.30	0.30
October				1	0.16	0.14
November					1	0.43
December						1

Table 3: Two samples testing with kernel smoothing. The table reports p-values using the kernel based test proposed in [1] for testing equality of distributions between different pairs of months from July to December 2014.

test statistics based on the Sinkhorn loss for the comparison of multivariate distributions. We believe this can be done using tools in this paper on the delta method for differentiable functions in the sense of Hadamard.

References

[1] N. H. Anderson, P. Hall, and D. M. Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

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

[3] J. Bigot, E. Cazelles, and N. Papadakis. Penalized Barycenters in the Wasserstein Space. *ArXiv e-prints*, June 2016.

[4] 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.

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

[6] 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.

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

[8] 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.

[9] 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.

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

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

[12] Z. Fang and A. Santos. Inference on directionally differentiable functions. *arXiv preprint arXiv:1404.3763*, 2014.

[13] 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.

[14] 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.

[15] 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.

[16] A. Genevay, G. Peyré, and M. Cuturi. Sinkhorn-autodiff: Tractable Wasserstein learning of generative models. Preprint 1706.00292, Arxiv, 2017.

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

[18] 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.

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

[20] 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.

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

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

[23] 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.

[24] W. Römisch. Delta method, infinite dimensional. *Encyclopedia of statistical sciences*, 2005.

[25] 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.

[26] 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.

[27] A. Shapiro. On concepts of directional differentiability. *Journal of optimization theory and applications*, 66(3):477–487, 1990.

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

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

[30] A. W. Van Der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[31] C. Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, 2003.

[32] C. Wang and F. Zhao. Directional derivatives of optimal value functions in mathematical programming. *Journal of optimization theory and applications*, 82(2):397–404, 1994.

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

[34] 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.

[35] 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.