*Submitted to the Annals of Statistics*
arXiv: arXiv:1812.09150

# ASYMPTOTIC DISTRIBUTION AND CONVERGENCE RATES OF STOCHASTIC ALGORITHMS FOR ENTROPIC OPTIMAL TRANSPORTATION BETWEEN PROBABILITY MEASURES

By Bernard Bercu and Jérémie Bigot

*Université de Bordeaux*

This paper is devoted to the stochastic approximation of entropically regularized Wasserstein distances between two probability measures, also known as Sinkhorn divergences. The semi-dual formulation of such regularized optimal transportation problems can be rewritten as a non-strongly concave optimisation problem. It allows to implement a Robbins-Monro stochastic algorithm to estimate the Sinkhorn divergence using a sequence of data sampled from one of the two distributions. Our main contribution is to establish the almost sure convergence and the asymptotic normality of a new recursive estimator of the Sinkhorn divergence between two probability measures in the discrete and semi-discrete settings. We also study the rate of convergence of the expected excess risk of this estimator in the absence of strong concavity of the objective function. Numerical experiments on synthetic and real datasets are also provided to illustrate the usefulness of our approach for data analysis.

## 1. Introduction.

1.1. *Optimal transport and regularized Wasserstein distances for data analysis.* The statistical analysis of high-dimensional data using tools from the theory of optimal transport [44] and the notion of Wasserstein distance between probability measures has recently gained increasing popularity. When elements in a dataset may be represented as probability distributions, the use of the Wasserstein distance leads to relevant statistics in various fields such as fingerprints comparison [43], clinical trials [33], metagenomics [20], clustering of discrete distributions [46], learning of generative models [4], or geodesic principal component analysis [8, 42]. Wasserstein distances are also of interest in the setting of statistical inference from empirical measures for hypothesis testing on discrepancies between multivariate distributions [43].

*MSC 2010 subject classifications:* Primary 62G05; secondary 62G20
*Keywords and phrases:* Stochastic optimisation, Convergence of random variables, Optimal transport, Entropic regularization, Sinkhorn divergence, Wasserstein distance

However, the cost to evaluate a Wasserstein distance between two discrete probability distributions with supports of equal size $K$ is generally of order $K^3 \log K$. Consequently, this evaluation represents a serious limitation for high-dimensional data analysis. To overcome this issue, Cuturi [14] has proposed to add an entropic regularization term to the linear program corresponding to a standard optimal transport problem. It leads to the notion of Sinkhorn divergence between two probability distributions which may be computed through an iterative algorithm where the cost of each iteration is of order $K^2$. This proposal makes feasible the use of regularized optimal transportation distance for data analysis, and it has found various applications in generative models [25], multi-label learning [22], dictionary learning [41], or image processing [15, 27, 38], to name but a few. For an overview of regularized optimal transport applied to machine learning, we refer the reader to the recent book of Cuturi and Peyré [16] as well as to [2, 14] for deterministic algorithms on the computation of Sinkhorn divergences.

1.2. *Main contributions and related works.* This paper is inspired by the recent work of Genevay, Cuturi, Peyré and Bach [24] on a very efficient statistical procedure to evaluate the possibly regularized Wasserstein distance $W_\varepsilon(\mu, \nu)$ between an arbitrary probability measure $\mu$ and a discrete distribution $\nu$ with finite support of size $J$. This statistical procedure is based on the well-known Robbins-Monro algorithm for stochastic optimisation [40] and its averaged version [36]. The keystone of their approach [24] is that $W_\varepsilon(\mu, \nu)$ can be rewritten in expectation form as

$$W_\varepsilon(\mu, \nu) = \max_{v \in \mathbb{R}^J} \mathbb{E}[h_\varepsilon(X, v)] \tag{1.1}$$

where $X$ is a random vector drawn from the unknown distribution $\mu$ and $h_\varepsilon(x, v)$ is a suitable function of the regularization parameter $\varepsilon \geqslant 0$. As it was shown in [24] on clouds of word embeddings, this statistical procedure is easy to implement with a low computational cost. For a sequence $(X_n)$ of independent and identically distributed random variables sharing the same distribution as $X$, we shall extend the statistical analysis of [24] by proving that, for $\varepsilon > 0$, the Robbins-Monro algorithm

$$\widehat{V}_{n+1} = \widehat{V}_n + \gamma_{n+1} \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \tag{1.2}$$

converges almost surely to a maximizer $v^*$ of $H_\varepsilon(v) = \mathbb{E}[h_\varepsilon(X, v)]$. We also investigate the asymptotic normality of $\widehat{V}_n$. It leads us to estimate the Sinkhorn divergence $W_\varepsilon(\mu, \nu)$ by the new recursive estimator

$$\widehat{W}_n = \frac{1}{n} \sum_{k=1}^{n} h_\varepsilon(X_k, \widehat{V}_{k-1}). \tag{1.3}$$

Under standard assumptions in stochastic optimisation, we shall prove that

$$\lim_{n\to\infty} \widehat{W}_n = W_\varepsilon(\mu,\nu) \qquad \text{a.s.} \tag{1.4}$$

as well as the asymptotic normality

$$\sqrt{n}\Big(\widehat{W}_n - W_\varepsilon(\mu,\nu)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0,\sigma_\varepsilon^2(\mu,\nu)\big) \tag{1.5}$$

where the asymptotic variance $\sigma_\varepsilon^2(\mu,\nu) = \mathbb{E}\big[h_\varepsilon^2(X,v^*)\big] - W_\epsilon^2(\mu,\nu)$ can also be estimated in a recursive manner. Finally, we analyze the rate of convergence of the expected excess risk $H_\varepsilon(v^*) - \mathbb{E}[\widehat{W}_n]$. We shall prove that the expected excess risk goes to zero faster than $1/\sqrt{n}$ in the absence of strong concavity of the objective function $H_\varepsilon$. These asymptotic results allow to better understand the convergence of stochastic algorithms for regularized optimal transport problems and to analyse the influence of their asymptotic variance on the quality of estimation. We shall also establish further results in the unregularized case where the regularization parameter $\varepsilon = 0$.

The asymptotic behavior of empirical Wasserstein distances when both $\mu$ and $\nu$ are absolutely continuous measures has been extensively studied over the last years [17, 18, 19, 21, 39]. For probability measures supported on finite spaces, limiting distributions for empirical Wasserstein distance have been obtained in [43], while the asymptotic distribution of empirical Sinkhorn divergence has been recently considered in [7, 30].

1.3. *Organisation of the paper.* Notation and formulation of the possibly regularized optimal transportation problem are presented in Section 2. Asymptotic properties of our stochastic algorithms for the regularized optimal transport are stated in Section 3 and further results for the unregularized optimal transport are provided in Section 4. Numerical experiments illustrating our theoretical results on simulated and real data are given in Section 5, where we consider the problem of estimating an optimal mapping between the distribution of spatial locations of reported incidents of crime in Chicago and the locations of Police stations. All the proofs are postponed to Appendices A and B that have been put in a supplementary file.

**2. Formulation of the optimal transportation problem.** Let $\mathcal{X}$ and $\mathcal{Y}$ be two metric spaces. Denote by $\mathcal{M}_+^1(\mathcal{X})$ and $\mathcal{M}_+^1(\mathcal{Y})$ the sets of probability measures on $\mathcal{X}$ and $\mathcal{Y}$, and by $\mathcal{C}_b(\mathcal{X})$ and $\mathcal{C}_b(\mathcal{Y})$ the spaces of bounded and continuous functions on $\mathcal{X}$ and $\mathcal{Y}$, respectively. When $\mathcal{X} = \{x_1,\dots,x_I\}$ and $\mathcal{Y} = \{y_1,\dots,y_J\}$ are finite sets, we identify the spaces $\mathcal{C}_b(\mathcal{X})$ and $\mathcal{C}_b(\mathcal{Y})$ by the Euclidean spaces $\mathbb{R}^I$ and $\mathbb{R}^J$. For $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and

$\nu \in \mathcal{M}_+^1(\mathcal{Y})$, let $\Pi(\mu,\nu)$ be the set of probability measures on $\mathcal{X}\times\mathcal{Y}$ with marginals $\mu$ and $\nu$. As formulated in Section 2 of [24], the problem of *entropic optimal transportation* between $\mu \in \mathcal{M}_+^1(\mathcal{X})$ and $\nu \in \mathcal{M}_+^1(\mathcal{Y})$ is as follows.

DEFINITION 2.1. *For any $(\mu,\nu) \in \mathcal{M}_+^1(\mathcal{X}) \times \mathcal{M}_+^1(\mathcal{Y})$, the Kantorovich formulation of the possibly regularized optimal transport between $\mu$ and $\nu$ is the following convex minimization problem*

$$W_\varepsilon(\mu,\nu) = \min_{\pi\in\Pi(\mu,\nu)} G_\varepsilon(\pi), \tag{2.1}$$

$$G_\varepsilon(\pi) = \int_{\mathcal{X}\times\mathcal{Y}} c(x,y)d\pi(x,y) + \varepsilon KL(\pi|\mu\otimes\nu), \tag{2.2}$$

*where $c : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}$ is a lower semi-continuous function referred to as the cost function of moving mass from location $x$ to $y$, $\varepsilon \geqslant 0$ is a regularization parameter, and KL stands for the Kullback-Leibler divergence, between $\pi$ and a positive measure $\xi$ on $\mathcal{X}\times\mathcal{Y}$, up to the additive term $\int_{\mathcal{X}\times\mathcal{Y}} d\xi(x,y)$, namely*

$$KL(\pi|\xi) = \int_{\mathcal{X}\times\mathcal{Y}} \Big(\log\Big(\frac{d\pi}{d\xi}(x,y)\Big) - 1\Big) d\pi(x,y).$$

For $\varepsilon = 0$, the quantity $W_0(\mu,\nu)$ is the *standard optimal transportation cost*, while for $\varepsilon > 0$, we shall refer to $W_\varepsilon(\mu,\nu)$ as the *Sinkhorn divergence* between the two probability measures $\mu$ and $\nu$. The choice of the cost function $c$ depends on the application, and it usually reflects some prior knowledge on the data or the problem at hand. Throughout the paper, following [45, Part I-4], we consider cost functions satisfying the following assumption which holds for many standard choices. The cost $c$ is a lower semi-continuous function satisfying, for all $(x,y) \in \mathcal{X}\times\mathcal{Y}$,

$$0 \leqslant c(x,y) \leqslant c_{\mathcal{X}}(x) + c_{\mathcal{Y}}(y) \tag{2.3}$$

where $c_{\mathcal{X}}$ and $c_{\mathcal{Y}}$ are real-valued functions such that $\int_{\mathcal{X}} c_{\mathcal{X}}(x)d\mu(x) < +\infty$ and $\int_{\mathcal{Y}} c_{\mathcal{Y}}(y)d\nu(y) < +\infty$. Under condition (2.3), $W_\varepsilon(\mu,\nu)$ is finite whatever is the value of the regularization parameter $\varepsilon \geqslant 0$. Moreover, we always have $W_0(\mu,\nu) \geqslant 0$, while $W_\varepsilon(\mu,\nu)$ can be negative for $\varepsilon > 0$ which comes from the expression of the KL divergence up to a constant additive term in Definition 2.1. We shall now define the dual and semi-dual formulations of the minimization problem (2.1) as introduced in [24]. For $\varepsilon = 0$, these formulations are classical results in optimal transport known as Kantorovich's duality. If

the regularization parameter $\varepsilon > 0$, it follows from [24, Proposition 2.1] that the dual expression of the minimization problem (2.1) is given by

$$W_\varepsilon(\mu,\nu) = \sup_{(u,v)\in\mathcal{C}_b(\mathcal{X})\times\mathcal{C}_b(\mathcal{Y})} \int_{\mathcal{X}\times\mathcal{Y}} \Big(u(x)+v(y)-w_c(x,y)\Big)d\mu(x)d\nu(y) \tag{2.4}$$

where

$$w_c(x,y) = \varepsilon \exp\Big(\frac{u(x)+v(y)-c(x,y)}{\varepsilon}\Big).$$

The proof in [24] to derive the strong duality statement (2.4) relies on a formal application of Fenchel-Rockafellar's theorem. Nevertheless, it appears that another proof can be found in the previous work [10] concerned by obtaining variational characterizations for the existence of a probability measure with given marginals, which is a problem closely related to regularized optimal transport, see also [31] for the connection between the Monge-Kantorovich and the Schrödinger problem. Indeed, it is known [16, Proposition 4.2] that the primal problem (2.1) can be refactored as the *I-projection* [12] problem

$$W_\varepsilon(\mu,\nu) = \varepsilon \min_{\pi\in\Pi(\mu,\nu)} \mathrm{KL}(\pi|\gamma) \tag{2.5}$$

with

$$d\gamma(x,y) = \exp\left(\frac{-c(x,y)}{\varepsilon}\right) d\mu(x)d\nu(x).$$

Hence, under condition (2.3), the duality result (2.4) can be derived from [10, Theorem 3.9]. The dual formulation (2.4) also follows as a particular instance of [11, Theorem 3.2] which considers unbalanced transport for marginals with different mass, but it is expressed as a maximization problem over the set of bounded functions $L^\infty(\mathcal{X})\times L^\infty(\mathcal{Y})$ rather than $\mathcal{C}_b(\mathcal{X})\times\mathcal{C}_b(\mathcal{Y})$. Then, arguing as in [24], the semi-dual of the convex minimization problem (2.1) is as follows. For any $\varepsilon \geqslant 0$, the optimal transportation between $\mu$ and $\nu$ is obtained by solving the concave maximization problem

$$W_\varepsilon(\mu,\nu) = \sup_{v\in\mathcal{C}_b(\mathcal{Y})} H_\varepsilon(v) \tag{2.6}$$

where

$$H_\varepsilon(v) = \int_{\mathcal{X}} v_{c,\varepsilon}(x)d\mu(x) + \int_{\mathcal{Y}} v(y)d\nu(y) - \varepsilon,$$

and for a given cost function satisfying (2.3), we define the $c$-transform of $v\in\mathcal{C}_b(\mathcal{Y})$ as

$$v_{c,\varepsilon}(x) = \begin{cases} \inf_{y\in\mathcal{Y}}\Big\{c(x,y)-v(y)\Big\} & \text{if} \quad \varepsilon = 0, \\ -\varepsilon\log\Big(\int_{\mathcal{Y}}\exp\Big(\frac{v(y)-c(x,y)}{\varepsilon}\Big)d\nu(y)\Big) & \text{if} \quad \varepsilon > 0. \end{cases}$$

Thanks to condition (2.3), the sup in (2.6) is a max meaning that it exists a dual variable $v^*$ such that $W_0(\mu,\nu) = H_0(v^*)$, see e.g. [45, Theorem 5.9]. In the regularized case $\varepsilon > 0$, the existence of maximizers of (2.6) appears to be a more delicate issue. From (2.3), the cost $c$ belongs to the set $L^1(\mu \otimes \nu)$ of integrable functions with respect to $\mu \otimes \nu$. Hence, combining [12, Corollary 3.2] with the characterization (2.5) of regularized optimal transport, there exist a pair of functions $(u^*, v^*)$ belonging to $L^1(\mu) \times L^1(\nu)$ and maximizing (2.4). However, an integrable function being not necessarily a bounded and continuous function, this result cannot be used to prove that $(u^*, v^*)$ is a pair of dual variables for the dual problem (2.4) when formulated as an optimisation over $\mathcal{C}_b(\mathcal{X}) \times \mathcal{C}_b(\mathcal{Y})$. The main difficulty seems to arise for unbounded costs. As a matter of fact, when the regularization parameter $\varepsilon > 0$, the existence of a function $v^* \in L^\infty(\nu)$ maximizing (2.6) is proved in [23, Theorem 7] under the additional assumption that the cost function $c$ is bounded.

In the rest of the paper, we shall now suppose that the cost $c$ is *not necessarily bounded*, but that, for any $\varepsilon \geqslant 0$, there exists $v^*$ such that $W_\varepsilon(\mu,\nu) = H_\varepsilon(v^*)$. This identity is the keystone result which allows us to formulate the problem of estimating $W_\varepsilon(\mu,\nu)$ in the setting of stochastic optimization. Indeed, the semi-dual problem (2.6) can be rewritten in expectation form as

$$W_\varepsilon(\mu,\nu) = \max_{v \in \mathcal{C}_b(\mathcal{Y})} \mathbb{E}[h_\varepsilon(X, v)] \tag{2.7}$$

where $X$ is a random vector drawn from the unknown distribution $\mu$, and for $x \in \mathcal{X}$ and $v \in \mathcal{C}_b(\mathcal{Y})$, $h_\varepsilon(x, v) = \int_{\mathcal{Y}} v(y) d\nu(y) + v_{c,\varepsilon}(x) - \varepsilon$. In all the sequel, we will further assume that $\nu$ is a discrete probability measure with finite support $\mathcal{Y} = \{y_1, \ldots, y_J\}$ in the sense that

$$\nu = \sum_{j=1}^{J} \nu_j \delta_{y_j} \tag{2.8}$$

where the locations $\{y_1, \ldots, y_J\}$ are a *known sequence* and $\delta$ stands for the standard Dirac measure. The weights $\{\nu_1, \ldots, \nu_J\}$ are a *known positive sequence* that sum up to one. By a slight abuse of notation, we identify $\nu$ to the vector of $\mathbb{R}^J$ with positive entries $(\nu_1, \ldots, \nu_J)$. We also denote by $\mathbf{0}_J$ and $\mathbf{1}_J$ the column vectors of $\mathbb{R}^J$ with all coordinates equal to zero and one respectively, and by $\langle\ ,\ \rangle$ the standard inner product in $\mathbb{R}^J$. Following the terminology from [24], the *discrete setting* corresponds to the supplementary assumption that $\mu$ is also a discrete probability measure with finite support while the *semi-discrete setting* is the general case where $\mu \in \mathcal{M}_+^1(\mathcal{X})$ is an

arbitrary probability measure that is absolutely continuous with respect to the Lebesgue measure and $\nu$ is a discrete measure with finite support (see e.g. Chapter 5 in [16] for an introduction to semi-discrete optimal transport problems and related references). When the size of the support of the measure $\nu$ is relatively small compared to the size of the support of $\mu$, we would recommend to use the stochastic approach proposed in this paper. This suggestion is supported by the numerical results reported in Section 5. Now, if $\nu$ is the discrete measure (2.8), the semi-dual problem (2.6) can be reformulated as

$$W_\varepsilon(\mu,\nu) = \max_{v\in\mathbb{R}^J} H_\varepsilon(v), \tag{2.9}$$

$$H_\varepsilon(v) = \mathbb{E}[h_\varepsilon(X,v)] \tag{2.10}$$

where

$$h_\varepsilon(x,v) = \begin{cases} \sum_{j=1}^J v_j\nu_j + \min\limits_{1\leqslant j\leqslant J}\Big\{c(x,y_j) - v_j\Big\} & \text{if } \ \varepsilon = 0, \\ \sum_{j=1}^J v_j\nu_j - \varepsilon\log\Big(\sum_{j=1}^J \exp\Big(\dfrac{v_j - c(x,y_j)}{\varepsilon}\Big)\nu_j\Big) - \varepsilon & \text{if } \ \varepsilon > 0. \end{cases} \tag{2.11}$$

**3. Asymptotic properties of stochastic algorithms for regularized optimal transport.** Throughout this section, we assume that $\varepsilon > 0$.

3.1. *The stochastic Robbins-Monro algorithms.* Our goal is to estimate the Sinkhorn divergence $W_\varepsilon(\mu,\nu)$ using a stochastic Robbins-Monro algorithm [40]. For any $v \in \mathbb{R}^J$, the function $H_\varepsilon(v)$, given by (2.10), is the mean value of $h_\varepsilon(X,v)$ where $X$ is a random vector drawn from the unknown distribution $\mu$. For $\varepsilon > 0$, the function $h_\varepsilon$, defined by (2.11), is twice differentiable in the second variable. The gradient vector as well as the Hessian matrix of $h_\varepsilon$ can be easily calculated. More precisely, we have for any $x \in \mathcal{X}$,

$$\nabla_v h_\varepsilon(x,v) = \nu - \pi(x,v), \tag{3.1}$$

$$\nabla_v^2 h_\varepsilon(x,v) = \frac{1}{\varepsilon}\Big(\pi(x,v)\pi(x,v)^T - \mathrm{diag}(\pi(x,v))\Big) \tag{3.2}$$

where the $j^{th}$ component of the vector $\pi(x,v) \in \mathbb{R}^J$ is such that

$$\pi_j(x,v) = \Big(\sum_{k=1}^J \nu_k \exp\Big(\frac{v_k - c(x,y_k)}{\varepsilon}\Big)\Big)^{-1} \nu_j \exp\Big(\frac{v_j - c(x,y_j)}{\varepsilon}\Big). \tag{3.3}$$

Consequently, it follows from (2.10), (3.1) and (3.2) that the gradient vector and the Hessian matrix of $H_\varepsilon$ are given by

$$\nabla H_\varepsilon(v) = \mathbb{E}[\nabla_v h_\varepsilon(X, v)] = \nu - \mathbb{E}[\pi(X, v)], \tag{3.4}$$

$$\nabla^2 H_\varepsilon(v) = \mathbb{E}[\nabla_v^2 h_\varepsilon(X, v)] = \frac{1}{\varepsilon}\mathbb{E}\big[\pi(X, v)\pi(X, v)^T - \mathrm{diag}(\pi(X, v))\big]. \tag{3.5}$$

One can observe that for any $v \in \mathbb{R}^J$, $\nabla_v^2 H_\varepsilon(v)$ is a negative semi-definite matrix. Therefore, if $v^*$ is a maximizer of (2.9), we have $\nabla H_\varepsilon(v^*) = 0$ and for all $v \in \mathbb{R}^J$, $\langle v - v^*, \nabla H_\varepsilon(v)\rangle \leqslant 0$. It leads us to estimate the vector $v^*$ by the Robbins-Monro algorithm [40] given, for all $n \geqslant 0$, by

$$\widehat{V}_{n+1} = \widehat{V}_n + \gamma_{n+1}\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) \tag{3.6}$$

where the initial value $\widehat{V}_0$ is a square integrable random vector which can be arbitrarily chosen and $(\gamma_n)$ is a positive sequence of real numbers decreasing towards zero satisfying

$$\sum_{n=1}^{\infty} \gamma_n = +\infty \qquad \text{and} \qquad \sum_{n=1}^{\infty} \gamma_n^2 < +\infty. \tag{3.7}$$

Two main issues arise with this Robbins-Monro algorithm. First of all, we clearly have from (3.5) that for any $v \in \mathbb{R}^J$, $\nabla_v^2 H_\varepsilon(v)\mathbf{1}_J = \mathbf{0}_J$ which implies that zero is an eigenvalue of the Hessian matrix $\nabla_v^2 H_\varepsilon(v)$ associated with the eigenvector $\boldsymbol{v}_J = \frac{1}{\sqrt{J}}\mathbf{1}_J$. Next, it follows from [14] that the maximizer $v^*$ of (2.9) is unique up to a scalar translation of the form $v^* - t\boldsymbol{v}_J$ for any $t \in \mathbb{R}$. Throughout the paper, we shall denote by $v^*$ the maximizer of (2.9) satisfying $\langle v^*, \boldsymbol{v}_J\rangle = 0$ which means that $v^*$ belongs to $\langle \boldsymbol{v}_J\rangle^\perp$ where $\langle \boldsymbol{v}_J\rangle$ is the one-dimensional subspace of $\mathbb{R}^J$ spanned by $\boldsymbol{v}_J$. Therefore, to obtain a consistent estimator of $v^*$ it is necessary to slightly modify the Robbins-Monro algorithm (3.6).

*Algorithm 1.* A first strategy is as follows. It is easy to see from the expression (3.1) that the gradient $\nabla_v h_\varepsilon(x, v)$ is a vector of $\mathbb{R}^J$ belonging to the linear space $\langle \boldsymbol{v}_J\rangle^\perp$ for any vectors $x \in \mathcal{X}$ and $v \in \mathbb{R}^J$. Hence, if the initial value $\widehat{V}_0$ belongs to $\langle \boldsymbol{v}_J\rangle^\perp$, one has immediately that $(\widehat{V}_n)$ is a stochastic sequence with values in the subspace $\langle \boldsymbol{v}_J\rangle^\perp$. The analysis of its convergence to $v^*$ can thus be done by considering the restriction of the function $v \mapsto h_\varepsilon(x, v)$ to the linear subspace $\langle \boldsymbol{v}_J\rangle^\perp$.

*Algorithm 2.* A second strategy is to estimate $v^*$ by the modified stochastic Robbins-Monro algorithm given, for all $n \geqslant 0$, by

$$\widehat{V}_{n+1} = \widehat{V}_n + \gamma_{n+1}\big(\nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) - \alpha\langle \widehat{V}_n, \boldsymbol{v}_J\rangle \boldsymbol{v}_J\big) \tag{3.8}$$

where $\widehat{V}_0$ is a square integrable random vector which can be arbitrarily chosen, the sequence $(\gamma_n)$ satisfies (3.7), and where $\alpha$ is a typically small positive parameter. The role played by $\alpha$ is to overcome the fact that the Hessian matrix $\nabla_v^2 H_\varepsilon(v)$ is singular. One can observe that if $\widehat{V}_0 \in \langle \boldsymbol{v}_J\rangle^\perp$ then $\langle \widehat{V}_n, \boldsymbol{v}_J\rangle = 0$ for all $n \geqslant 0$, and thus Algorithm 2 is equivalent to Algorithm 1. By a slight abuse of notation, we shall also refer to Algorithm 1 as the case where $\alpha = 0$ and we refer to Algorithm 2 as the case where $\alpha > 0$, although it is clear that $\widehat{V}_n$ depends on $\alpha$ for Algorithm 2 when $\widehat{V}_0 \notin \langle \boldsymbol{v}_J\rangle^\perp$. One may also remark that Algorithm 2 corresponds to a stochastic ascent algorithm to compute a maximizer over $\mathbb{R}^J$ of the strictly concave function

$$H_{\varepsilon,\alpha}(v) = \mathbb{E}[h_{\varepsilon,\alpha}(X, v)] \tag{3.9}$$

where $h_{\varepsilon,\alpha}(x, v) = h_\varepsilon(x, v) - \frac{\alpha}{2}\big(\langle v, \boldsymbol{v}_J\rangle\big)^2$. An important role in the choice of $\alpha$ will be the control of the smallest eigenvalue of the Hessian matrix $\nabla_v^2 H_{\varepsilon,\alpha}(v^*)$. In the case $\alpha = 0$, the objective function $H_\varepsilon(v)$ has a bounded gradient. As a matter of fact, since $\|\nu\| \leqslant 1$ and $\|\pi(x, v)\| \leqslant 1$, it follows that for all $x \in \mathcal{X}$ and $v \in \mathbb{R}^J$,

$$\|\nabla_v h_\varepsilon(x, v)\| \leqslant \|\nu\| + \|\pi(x, v)\| \leqslant 2, \tag{3.10}$$

which ensures that $\|\nabla H_\varepsilon(v)\| \leqslant 2$. In the case $\alpha > 0$, we also have

$$\|\nabla_v h_{\varepsilon,\alpha}(x, v)\| \leqslant 2 + \alpha\|v\|, \tag{3.11}$$

which implies that $\|\nabla H_{\varepsilon,\alpha}(v)\| \leqslant 2 + \alpha\|v\|$. The gradient of the objective function $H_{\varepsilon,\alpha}(v)$ is clearly not bounded.

3.2. *Almost sure convergence and asymptotic normality.* Our first result concerns the almost sure convergence of the Robbins-Monro algorithms.

THEOREM 3.1. *For both algorithms, we have the almost sure convergence*

$$\lim_{n\to\infty} \widehat{V}_n = v^* \qquad a.s. \tag{3.12}$$

We now focus our attention on the asymptotic normality of the Robbins-Monro algorithms. For any $v \in \mathbb{R}^J$, let $\Gamma_\varepsilon(v)$ be the positive semidefinite matrix given by

$$\Gamma_\varepsilon(v) = \mathbb{E}\big[\pi(X, v)\pi(X, v)^T\big] - \mathbb{E}\big[\pi(X, v)\big]\mathbb{E}\big[\pi(X, v)^T\big]. \tag{3.13}$$

One can observe that for $v = v^*$, $\Gamma_\varepsilon(v^*) = \mathbb{E}\big[\pi(X, v^*)\pi(X, v^*)^T\big] - \nu\nu^T$, since $\nabla H_\varepsilon(v^*) = 0$ implies that $\nu = \mathbb{E}[\pi(X, v^*)]$. For any $v \in \mathbb{R}^J$, denote

$$A_\varepsilon(v) = \nabla^2 H_\varepsilon(v). \tag{3.14}$$

We shall see in Lemma A.1 that for any $v \in \mathbb{R}^J$, the matrix $A_\varepsilon(v)$ is negative semi-definite with $\text{rank}(A_\varepsilon(v)) = J - 1$. It means that the second smallest eigenvalue of the matrix $-A_\varepsilon(v)$ is always positive. By a slight abuse of notation, we shall denote by $\rho_{A_\varepsilon}(v)$ the second smallest eigenvalue of the matrix $-A_\varepsilon(v)$. Moreover, let

$$A_{\varepsilon,\alpha}(v) = \nabla_v^2 H_{\varepsilon,\alpha}(v) = A_\varepsilon(v) - \alpha \boldsymbol{v}_J \boldsymbol{v}_J^T. \tag{3.15}$$

It follows from Remark A.1 that for any $v \in \mathbb{R}^J$, the matrix $A_{\varepsilon,\alpha}(v)$ is a negative definite. We shall also denote $\rho_{A_{\varepsilon,\alpha}}(v) = -\lambda_{\max} A_{\varepsilon,\alpha}(v)$ where $\lambda_{\max} A_{\varepsilon,\alpha}(v)$ stands for the maximum eigenvalue of the matrix $A_{\varepsilon,\alpha}(v)$. It is not hard to see that $\rho_{A_{\varepsilon,\alpha}}(v) = \min(\rho_{A_\varepsilon}(v), \alpha)$. Hereafter, in order to unify the notation, we put

$$\rho^* = \begin{cases} \rho_{A_\varepsilon}(v^*) & \text{if} \quad \alpha = 0 \\ \rho_{A_{\varepsilon,\alpha}}(v^*) & \text{if} \quad \alpha > 0. \end{cases}$$

THEOREM 3.2. *Assume that the step $\gamma_n = \gamma/n$ where*

$$\gamma > \frac{1}{2\rho^*}. \tag{3.16}$$

*Then, for both algorithms, we have the asymptotic normality*

$$\sqrt{n}\big(\widehat{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}_J\big(0, \gamma\Sigma^*\big) \tag{3.17}$$

*where the asymptotic covariance matrix $\Sigma^*$ is the unique solutiuon of Lyapounov's equation with $\zeta = 1/2\gamma$*

$$\big(A_{\varepsilon,\alpha}(v^*) + \zeta I_J\big)\Sigma^* + \Sigma^*\big(A_{\varepsilon,\alpha}(v^*) + \zeta I_J\big) = -\Gamma_\varepsilon(v^*). \tag{3.18}$$

REMARK 3.1. *Theorem 3.2 is also true if $\gamma_n = \gamma/n^c$ with $\gamma > 0$ and $1/2 < c < 1$, see Pelletier [35], Theorem 1. To be more precise, we have the asymptotic normality*

$$\sqrt{n^c}\big(\widehat{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}_J\big(0, \gamma\Sigma^*\big).$$

*The convergence rate $n^c$ is clearly always slower than $n$, which means that the choice $\gamma_n = \gamma/n$ outperforms the choice $\gamma_n = \gamma/n^c$ in term of convergence rate. However, in the special case $\gamma_n = \gamma/n^c$, the restriction (3.16) involving the knownledge of $\rho^*$ is no longer needed.*

Some refinements on the asymptotic behavior of the Robbins-Monro algorithms are as follows.

THEOREM 3.3. *Assume that the step* $\gamma_n = \gamma/n$ *where* $\gamma > 0$ *satisfies* (3.16). *Then, for both algorithms, we have the quadratic strong law*

$$\lim_{n\to\infty} \frac{1}{\log n} \sum_{k=1}^{n} \big(\widehat{V}_k - v^*\big)\big(\widehat{V}_k - v^*\big)^T = \gamma\Sigma^* \qquad a.s. \tag{3.19}$$

*where* $\Sigma^*$ *is given by* (3.18). *Moreover, for any eigenvectors* $v \in \mathbb{R}^J$ *of the Hessian matrix* $A_{\varepsilon,\alpha}(v^*)$, *we have the law of iterated logarithm*

$$\begin{aligned} \limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} \langle v, \widehat{V}_n - v^*\rangle &= -\liminf_{n\to\infty} \left(\frac{n}{2\log\log n}\right)^{1/2} \langle v, \widehat{V}_n - v^*\rangle \\ &= \sqrt{\gamma v^T \Sigma^* v} \qquad a.s. \end{aligned} \tag{3.20}$$

*In particular,*

$$\limsup_{n\to\infty} \left(\frac{n}{2\log\log n}\right) \|\widehat{V}_n - v^*\|^2 \leqslant \gamma \mathrm{tr}(P^T\Sigma^* P) \qquad a.s. \tag{3.21}$$

*where* $P$ *is the matrix whose columns are the eigenvectors of* $A_{\varepsilon,\alpha}(v^*)$.

REMARK 3.2. *Theorem 3.3 also holds in the special case where* $\gamma_n = \gamma/n^c$ *with* $\gamma > 0$ *and* $1/2 < c < 1$, *see Pelletier [34] Theorems 1 and 3. For example, the quadratic strong law* (3.19) *has to be replaced by*

$$\lim_{n\to\infty} \frac{1}{n^{1-c}} \sum_{k=1}^{n} \big(\widehat{V}_k - v^*\big)\big(\widehat{V}_k - v^*\big)^T = \frac{\gamma}{1-c}\Sigma^* \qquad a.s. \tag{3.22}$$

*In the special case* $\gamma_n = \gamma/n^c$, *Theorem 3.3 is true without condition* (3.16).

The explicit calculation of the asymptotic covariance matrix $\Sigma^*$ in (3.17) is far from being simple since there is no closed-form solution of equation (3.18). To overcome this issue, one may use the averaged Robbins-Monro algorithm given by $\overline{V}_n = \frac{1}{n}\sum_{k=1}^{n} \widehat{V}_k$. which satisfies the second-order reccurence equation

$$\overline{V}_{n+1} = \Big(\frac{2n}{n+1}\Big)\overline{V}_n - \Big(\frac{n-1}{n+1}\Big)\overline{V}_{n-1} + \frac{\gamma_{n+1}}{n+1} Y_{n+1}$$

where the random vector $Y_{n+1}$ is given by

$$Y_{n+1} = \nabla_v h_\varepsilon(X_{n+1}, \widehat{V}_n) - \alpha\langle \widehat{V}_n, \boldsymbol{v}_J\rangle \boldsymbol{v}_J. \tag{3.23}$$

Our next result is devoted to the asymptotic behavior of the averaged Robbins-Monro algorithms.

THEOREM 3.4. *For both algorithms, we have the almost sure convergence*

$$\lim_{n\to\infty} \overline{V}_n = v^* \qquad a.s. \tag{3.24}$$

*Moreover, assume that the step* $\gamma_n = \gamma/n^c$ *where* $\gamma > 0$ *and* $1/2 < c < 1$. *Then, we have the asymptotic normality*

$$\sqrt{n}\big(\overline{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}_J\Big(0, (A_{\varepsilon,\alpha}(v^*))^{-1}\Gamma_\varepsilon(v^*)(A_{\varepsilon,\alpha}(v^*))^{-1}\Big). \tag{3.25}$$

*In particular, if the sequence* $(\overline{V}_n)$ *is associated with Algorithm 1, convergence* (3.25) *can be rewritten as*

$$\sqrt{n}\big(\overline{V}_n - v^*\big) \xrightarrow{\mathcal{L}} \mathcal{N}_J\Big(0, A_\varepsilon^\dagger(v^*)\Gamma_\varepsilon(v^*)A_\varepsilon^\dagger(v^*)\Big) \tag{3.26}$$

*where* $A_\varepsilon^\dagger(v^*)$ *stands for the Moore-Penrose inverse of* $A_\varepsilon(v^*)$.

REMARK 3.3. *We already saw that the Hessian matrix* $A_\varepsilon(v^*)$ *is negative semi-definite with* $\mathrm{rank}(A_\varepsilon(v^*)) = J-1$, *which implies that its Moore-Penrose inverse is given by* $A_\varepsilon^\dagger(v^*) = \sum_{j=1}^{J-1} \frac{1}{\lambda_j} v_j v_j^T$ *where* $\lambda_1, \ldots, \lambda_{J-1}$ *are the negative eigenvalues of* $A_\varepsilon(v_*)$ *and* $v_1, \ldots, v_{J-1}$ *are the associated orthonormal eigenvectors. Moreover,*

$$(A_{\varepsilon,\alpha}(v^*))^{-1} = A_\varepsilon^\dagger(v^*) - \frac{1}{\alpha}\boldsymbol{v}_J\boldsymbol{v}_J^T. \tag{3.27}$$

3.3. *Estimation of the Sinkhorn divergence.* Herafter, we focus our attention on the estimation of the Sinkhorn divergence $W_\epsilon(\mu,\nu)$. For that purpose, a natural recursive estimator of $W_\epsilon(\mu,\nu)$ is given by

$$\widehat{W}_n = \frac{1}{n}\sum_{k=1}^{n} h_\varepsilon(X_k, \widehat{V}_{k-1}). \tag{3.28}$$

Our first main result concerns the asymptotic behavior of the Sinkhorn divergence estimator $\widehat{W}_n$.

THEOREM 3.5. *Assume that the cost function* $c$ *satisfies for any* $y \in \mathcal{Y}$,

$$\int_{\mathcal{X}} c^4(x,y)d\mu(x) < +\infty. \tag{3.29}$$

*Then, for both algorithms, we have the almost sure convergence*

$$\lim_{n\to\infty} \widehat{W}_n = W_\epsilon(\mu,\nu) \qquad a.s. \tag{3.30}$$

*Moreover, assume that the step* $\gamma_n = \gamma/n$ *where* $\gamma > 0$ *satisfies* (3.16)*, or that* $\gamma_n = \gamma/n^c$ *where* $\gamma > 0$ *and* $1/2 < c < 1$*. Then, for both algorithms, we have the asymptotic normality*

$$\sqrt{n}\Big(\widehat{W}_n - W_\epsilon(\mu,\nu)\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, \sigma_\varepsilon^2(\mu,\nu)\big) \tag{3.31}$$

*where the asymptotic variance* $\sigma_\varepsilon^2(\mu,\nu) = \mathbb{E}\big[h_\varepsilon^2(X, v^*)\big] - W_\epsilon^2(\mu,\nu)$*.*

REMARK 3.4. *The asymptotic variance* $\sigma_\varepsilon^2(\mu,\nu)$ *can be estimated by*

$$\widehat{\sigma}_n^2 = \frac{1}{n}\sum_{k=1}^{n} h_\varepsilon^2(X_k, \widehat{V}_{k-1}) - \widehat{W}_n^2. \tag{3.32}$$

*Via the same lines as in the proof of* (3.30)*, we can show that* $\widehat{\sigma}_n^2 \to \sigma_\varepsilon^2(\mu,\nu)$ *a.s. Therefore, using Slutsky's Theorem, it follows from* (3.31) *that*

$$\sqrt{n}\Big(\frac{\widehat{W}_n - W_\epsilon(\mu,\nu)}{\widehat{\sigma}_n}\Big) \xrightarrow{\mathcal{L}} \mathcal{N}\big(0,1\big). \tag{3.33}$$

*Convergence* (3.33) *allows us to construct confidence intervals for the Sinkhorn divergence* $W_\epsilon(\mu,\nu)$ *as illustrated in the numerical experiments of Section 5.*

Our second main result is devoted to the expected excess risk of the Sinkhorn divergence estimator $\widehat{W}_n$. It follows from (3.28) that

$$\mathbb{E}[\widehat{W}_n] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}\Big[\mathbb{E}[h_\varepsilon(X_k, \widehat{V}_{k-1})|\mathcal{F}_{k-1}]\Big] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[H_\varepsilon(\widehat{V}_{k-1})].$$

Hence, the *expected excess risk* of $\widehat{W}_n$ is defined as the non-negative quantity

$$\widehat{R}_n = H_\varepsilon(v^*) - \mathbb{E}[\widehat{W}_n] = \frac{1}{n}\sum_{k=1}^{n}\Big(H_\varepsilon(v^*) - \mathbb{E}[H_\varepsilon(\widehat{V}_{k-1})]\Big). \tag{3.34}$$

It is well known that only assuming concavity of the objective function leads to convergence rates for the expected excess risk $\widehat{R}_n$ of the order $1/\sqrt{n}$ for the Robbins-Monro algorithm. This rate of convergence cannot be improved without supplementary assumptions such as the strong concavity of the objective function $H_\varepsilon$, which leads to faster rates of convergence of the order $1/n^d$ for some $1/2 < d \leqslant 1$ which depends on the decay of the step $\gamma_n = \gamma/n^c$ where $1/2 < c < 1$. We refer the reader to [5], [6], [26] for a recent overview on the convergence rates of first-order stochastic algorithms.

However, as it was already shown in [24], the objective function $H_\varepsilon$ in the semi-dual problem (2.6) cannot be strongly concave, even by restricting the maximization to the subset $\langle \boldsymbol{v}_J \rangle^\perp$. Indeed, the gradient $v \mapsto \nabla H_\varepsilon(v)$ being bounded on $\mathbb{R}^J$, it follows that $H_\varepsilon$ is a Lipschitz function, and thus it cannot be strongly concave on $\langle \boldsymbol{v}_J \rangle^\perp$, see e.g. Section 3.2 in [6]. Nevertheless, for the stochastic optimization problem (2.6), it is possible to derive rates of convergence faster than $1/\sqrt{n}$ for the expected excess risk $\widehat{R}_n$. To this end, we borrow some ideas related to the so-called notion of generalized self-concordance coming from the recent contribution of Bach [5] and leading to faster rates of convergence for stochastic algorithms with non-strongly concave objective functions.

THEOREM 3.6. *Assume that $\widehat{V}_0$ is a random vector such that, for any integer $p \geqslant 1$, $\mathbb{E}[\|\widehat{V}_0\|^p]$ is finite. Moreover, assume that the step $\gamma_n = \gamma/n^c$ where $\gamma > 0$ and $2/3 < c < 1$. In addition, suppose that $0 < \varepsilon \leqslant 1$ and*

$$\theta_\varepsilon \leqslant \gamma\rho^* \leqslant 1 \quad \text{where} \quad \theta_\varepsilon = \frac{\sqrt{2}}{4}\Big(1 - \exp\Big(-\frac{\sqrt{2}}{\varepsilon}\Big)\Big)^{-1}. \tag{3.35}$$

*Then, there exists a positive constant $C$ such that for any $n \geqslant 1$,*

$$H_\varepsilon(v^*) - \mathbb{E}[H_\varepsilon(\widehat{V}_n)] \leqslant \frac{C}{n^{2c-1}}. \tag{3.36}$$

REMARK 3.5. *It is easy to see that the assumption $\varepsilon \leqslant 1$ implies that $\theta_\varepsilon < 1/2$. Consequently, the condition (3.35) is not really restrictive and it is fulfilled by a suitable choice of $\gamma$ depending on $\rho^*$. By inequality (A.4) and Remark A.2, one has the following bounds $\min_{1 \leqslant j \leqslant J} \nu_j \leqslant \varepsilon\rho^* \leqslant 1$, which are used to choose the parameter $\gamma$ in the numerical experiments carried out in Section 5. Finally, it follows from (3.34) together with inequality (3.36) that*

$$\widehat{R}_n \leqslant \frac{C}{n}\sum_{k=1}^{n} \frac{1}{k^{2c-1}} \leqslant C\Big(\frac{1}{n} + \frac{1}{2(1-c)}\frac{1}{n^{2c-1}}\Big). \tag{3.37}$$

*Consequently, if $c > 3/4$, inequality (3.37) shows that the expected excess risk of $\widehat{R}_n$ may converge to zero faster than $1/\sqrt{n}$ when the sequence $(\widehat{V}_n)$ is given by Algorithm 1.*

**4. Further results on the unregularized case.** We now focus our attention on the unregularized case where $\varepsilon = 0$. The function $h_0$, defined by (2.11), is not differentiable. Nevertheless, as remarked in [24], it follows from (2.11) that for any $x \in \mathcal{X}$, a supergradient $\partial_v h_0(x, v)$ of $h_0(x, \cdot)$ at $v$ is

$$\partial_v h_0(x, v) = \nu - \pi^0(x, v) \tag{4.1}$$

where $\pi^0(x,v)$ denotes the vector of $\mathbb{R}^J$ with all entries equal to zero expect the $j^*$-th which is equal to one, that is $\pi_j^0(x,v) = \mathbb{1}_{j=j^*}$, where

$$j^* \in \operatorname*{argmin}_{1\leqslant j\leqslant J}\Big\{c(x,y_j) - v_j\Big\}.$$

One can note that for $x \in \mathcal{X}$ and $v \in \mathbb{R}^J$, the integer $j^*$ may be not unique. In this case, the set of supergradients $\partial_v h_0(x,v)$ is not a singleton. In contrast with the regularized case where $\varepsilon > 0$, the function $H_0$ does not necessarily have a unique maximizer $v^*$ over $\langle \boldsymbol{v}_J\rangle^\perp$. To estimate such a maximizer $v^*$, which is assumed to belong to $\langle \boldsymbol{v}_J\rangle^\perp$, we shall consider a standard stochastic supergradient ascent given, for all $n \geqslant 0$, by

$$\widehat{V}_{n+1}^0 = \widehat{V}_n^0 + \gamma_{n+1}\big(\partial_v h_0(X_{n+1}, \widehat{V}_n^0) - \alpha\langle \widehat{V}_n^0, \boldsymbol{v}_J\rangle \boldsymbol{v}_J\big) \tag{4.2}$$

where the sequence $(\gamma_n)$ satisfies (3.7), $\alpha$ is a typically small positive parameter and $\partial_v h_0(X_{n+1}, \widehat{V}_n^0)$ is any supergradient of $h_0(X_{n+1},\cdot)$ at $\widehat{V}_n^0$ of the form (4.1). Therefore, a recursive estimator of $W_0(\mu,\nu)$ is given by

$$\widehat{W}_n^0 = \frac{1}{n}\sum_{k=1}^n h_0(X_k, \widehat{V}_{k-1}^0). \tag{4.3}$$

In order to investigate the asymptotic properties of the random sequences $(\widehat{V}_n^0)$ and $(\widehat{W}_n^0)$, two issues need to be addressed: the regularity of the function $H_0$ and the uniqueness of its maximizer over $\langle \boldsymbol{v}_J\rangle^\perp$. In the discrete setting, the function $H_0$ is clearly not differentiable. However, in the semi-discrete setting where $\mu$ is absolutely continuous, the differentiability of $H_0$ has been proved in [29] under appropriate regularity assumptions on the cost function and the measure $\mu$. More precisely, by [29, Theorem 2.1], we obtain the following conditions ensuring the $\mathcal{C}^1$-smoothness of $H_0$.

PROPOSITION 4.1. *Assume that that $\mathcal{X} = \mathbb{R}^d$. Moreover, suppose that*

**(i)** *For all $1 \leqslant j \leqslant J$, the function $x \mapsto c(x,y_j)$ is continuous.*
**(ii)** *The measure $\mu$ is absolutely continuous with bounded and compactly supported probability density function.*
**(iii)** *For any $j \neq k$ and for all $t \in \mathbb{R}$, the set $\{x \in \mathbb{R}^d, c(x,y_j) - c(x,y_k) = t\}$ has zero Lebesgue measure.*

*Then, the function $H_0$ is continuously differentiable on $\mathbb{R}^J$,*

$$\nabla H_0(v) = \int_{\mathbb{R}^d} \partial_v h_0(x,v) d\mu(x)$$

*where $\partial_v h_0(x,v)$ is any supergradient of $h_0(x,\cdot)$ at $v$ of the form (4.1).*

Guaranteeing the uniqueness of the maximizer $v^*$ of $H_0$ over $\langle \boldsymbol{v}_J \rangle^\perp$ is more involved. In particular, in the semi-discrete setting, we are not aware of any sufficient conditions ensuring such a property. Nevertheless, if one assumes the uniqueness of $v^*$ up to scalar translations, then under the assumptions of Proposition 4.1, it follows that $\nabla H_0$ is continuous. Therefore, arguing as in the proofs of Theorem 3.1 and Theorem 3.5, one obtains under the assumptions of Proposition 4.1 that

$$\lim_{n\to\infty} \widehat{V}_n^0 = v^* \qquad \text{a.s.}$$

and

$$\lim_{n\to\infty} \widehat{W}_n^0 = W_0(\mu,\nu) \qquad \text{a.s.}$$

Furthermore, under additional assumptions it follows from [32, Theorem 6] that $H_0$ is twice continuously differentiable. Nevertheless, in contrast to the regularized case, the second smallest eigenvalue of $A_0(v^*) = -\nabla^2 H_0(v^*)$ may be zero. Therefore, the proof of the asymptotic normality for $(\widehat{V}_n^0)$ and $(\widehat{W}_n^0)$ is much more tricky and left open for future work.

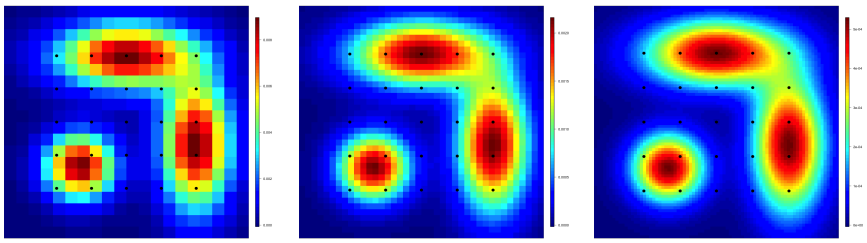
**5. Statistical applications and numerical experiments.** In this section, we report results on numerical experiments for probability measures $\mu$ and $\nu$ with supports included in $\mathbb{R}^d$ for $d \geqslant 2$ using synthetic and real data sets. All numerical experiments are carried out with the statistical computing environment R [37], and they are based on iid samples from $\mu$. We mainly investigate the numerical behavior of the two recursive estimators $\widehat{W}_n$ and $\widehat{\sigma}_n^2$. The reader has to keep in mind that the estimators $\widehat{W}_n$ and $\widehat{\sigma}_n^2$ depend on the positive value of the regularization parameter $\varepsilon$ as well as on the positive value of $\alpha$ and the statistical characteristics of $\mu$ and $\nu$. However, for the sake of simplicity, we have chosen to denote them as $\widehat{W}_n$ and $\widehat{\sigma}_n^2$. We carry out our numerical experiments for different values of $\varepsilon$ to illustrate the convergence of the recursive algorithms proposed in this paper as $n$ increases. Following the discussion in Section 3 on the calibration of the step size for $\varepsilon > 0$, we took $\gamma_n = \gamma/n^c$ with $c = 0.51$ and $\gamma = \varepsilon/(2\nu_{\min})$ where $\nu_{\min}$ stands for $\nu_{\min} = \min_{1\leqslant j\leqslant J} \nu_j$. In the unregularized case $\varepsilon = 0$ of Section 4, we took $\gamma = \varepsilon_{\min}/(4\nu_{\min})$ where $\varepsilon_{\min} = 0.01$ is the smallest value of regularization used in these numerical experiments. The cost function is chosen as the standard Euclidean distance,

$$c(x,y) = \Big(\sum_{k=1}^{d} \big|x^{(k)} - y^{(k)}\big|^2\Big)^{1/2}.$$

In our numerical experiments, we have found that Algorithm 1 with $\alpha = 0$ and $\widehat{V}_0 = 0$, and Algorithm 2 with $\alpha = \nu_{\min}/\varepsilon$ and $\widehat{V}_0 = \boldsymbol{v}_J$ share the same

numerical behavior for all sufficiently large values of $n$, that is $n \geqslant 10^2$. Consequently, we only report here the results obtained with Algorithm 1.

5.1. *Discrete setting in dimension two.* We first consider a setting in dimension $d = 2$ with discrete probability measures, and we investigate the regularized case where $\varepsilon > 0$. The measure $\nu$ is assumed to be the uniform measure on a grid $\mathcal{Y} \subset [0,1]^2$ made of $J = 25$ regularly spaced points. The measure $\mu$ is obtained by projecting a mixture of Gaussian densities on an $N \times N$ regular grid of $[0,1]^2$. The cardinality $I = N^2$ of its support $\mathcal{X}$ varies in the numerical experiments. The two measures are displayed in Figure 1 for different values of $N$.



(a) $N = 20$ (b) $N = 40$ (c) $N = 80$

FIG 1. *Discrete measure $\mu$ supported on a grid of $I = N^2$ regularly spaced points $[0,1]^2$ with colors indicating the intensity of the weights $(\mu_i)$. Discrete measure $\nu$ supported on $J = 25$ black dots points with uniform weights $\nu_j = 1/J$.*

The computation of the Sinkhorn divergence $W_\varepsilon(\mu, \nu)$ is done via the package `Barycenter`[1]. It allows us to solve the semi-dual maximization problem (2.6) using the Sinkhorn algorithm [14] which is a fixed point iteration algorithm for obtaining a solution of the primal problem (2.1) of the form

$$\pi^* = \mathrm{diag}(u^*) S_\varepsilon \, \mathrm{diag}(v^*) \qquad \text{with} \qquad S_\varepsilon = \exp(-\varepsilon^{-1} \boldsymbol{C}) \tag{5.1}$$

where $\boldsymbol{C} \in \mathbb{R}^{I \times J}$ is the cost matrix, $(u^*, v^*) \in \mathbb{R}_+^I \times \mathbb{R}_+^J$ is a pair of optimal dual variables and $\exp(\cdot)$ denotes the entrywise exponential. To obtain such a pair of dual variables in an iterative way, the Sinkhorn algorithm alternately scales the row and column sums of matrices written in the form (5.1) to match the marginals $\mu \in \mathbb{R}_+^J$ and $\nu \in \mathbb{R}_+^I$, that is $u^{n+1} = \mu ./ (S_\varepsilon v^n)$ and $v^{n+1} = \nu ./ (S_\varepsilon^T u^{n+1})$, where $./$ denotes the elementwise ratio between vectors. Hence, at each iteration, the computational cost of the Sinkhorn

[1] https://cran.r-project.org/package=Barycenter

algorithm is of the order $I^2 + J^2$. An advantage of stochastic algorithms for optimal transport is that the computational cost of the recursive estimators $\widehat{W}_n$ and $\widehat{\sigma}_n^2$ at each iteration of (3.6) and (3.32) is only of order $J$ as discussed in details in [24]. Moreover, the computation of these estimators do not require the full knowledge of the measure $\mu$, and the storage of the full cost matrix $\boldsymbol{C}$. The computational cost at each iteration of the Sinkhorn algorithm can be reduced by using a greedy coordinate descent algorithm referred to as the Greenkhorn algorithm [3] which consists in only updating one row or column of a matrix written in the form (5.1) by selecting the one that most violates the constraint that its row and columns sums should match the desired marginal $\mu$ and $\nu$. As described in [3], it is possible to implement this algorithm in such a way that the computational cost at each iteration is linear in the dimension of the input measures that is of order $I + J$. A stochastic version of the Greenkhorn algorithm has also been proposed in [1], where, instead of selecting the column or row which most violates the constraint, one row or column is randomly selected according to probability chosen in such a way that the columns and rows with highest violation are updated more frequently. Note that the stochastic Greenkhorn algorithm makes use of the full knowledge of $\mu$, and it is thus a stochastic algorithm of a different nature than the Robbins-Monro algorithm investigated in this paper. In particular, our approach does not use the knowledge of $\mu$, and the recursive estimators estimators $\widehat{W}_n$ and $\widehat{\sigma}_n^2$ have not been considered so far in the literature. In the discrete setting, it is proposed in [24] to use a stochastic averaged gradient algorithm (which uses the knowledge of $\mu$) to estimate $v^*$, and we refer to [24, Section 3] for detailed experiments on the comparison of this approach to the Sinkhorn algorithm.

In Figure 2, we report numerical results on the comparison between the recursive estimator $\widehat{W}_n$ and the numerical approximation of $W_\varepsilon(\mu, \nu)$ using either the Greenkhorn algorithm or its stochastic version as a function of the iterations whose computational costs are linear in the dimension of the input measures. The output of the Sinkhorn algorithm is used as the ground truth for $W_\varepsilon(\mu, \nu)$. Using the results from Section 3, one can construct confidence intervals for the Sinkhorn divergence between $\mu$ and $\nu$ by considering

$$\sqrt{n}\Big(\frac{\widehat{W}_n - W_\epsilon(\mu, \nu)}{\widehat{\sigma}_n}\Big) \tag{5.2}$$

to be approximately normally distributed. One can see in Figure 2 that the 95% confidence intervals always contain the value $W_\varepsilon(\mu, \nu)$ for $n \geqslant 3.10^5$ and all values of $\varepsilon$ and cardinality $I = N^2$ of the support of $\mu$. The Greenkhorn algorithm and its stochastic version perform similarly. For small values of

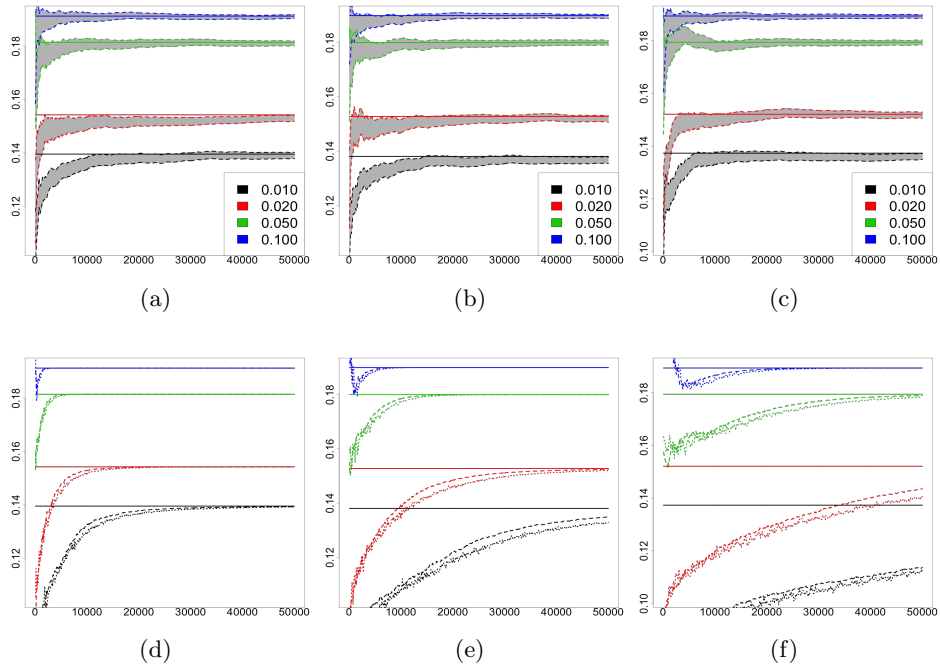



FIG 2. *First row - 95% confidence interval for* $\widehat{W}_n \pm 1.96\widehat{\sigma}_n/\sqrt{n}$ *where* $10^3 \leqslant n \leqslant 5.10^5$. *Second row - Approximation the Sinkhorn divergence using either the Greenkhorn algorithm (dashed lines) or its stochastic version (dotted lines) as the iterations n grows from* $10^3$ *to* $5.10^5$. *In all the figures, the solid and horizontal lines correspond to the value* $W_\epsilon(\mu,\nu)$ *obtained with the Sinkhorn algorithm.*

$N \leqslant 20$ and large values of $\varepsilon$, we observe that these algorithms converge in very few iterations. However, for larger values of $N$, that is larger sizes $I$ of the support of $\mu$, and for small values of $\varepsilon$, the recursive estimator $\widehat{W}_n$ clearly outperforms such Greenkhorn algorithms in the number of required iterations to obtain a satisfactory approximation of $W_\varepsilon(\mu,\nu)$.

5.2. *Semi-discrete setting in dimension* $d \geqslant 2$.

**Simulated data.** We now consider a synthetic example where $\mu$ is an absolutely continuous measure obtained as a mixture of three Gaussian densities with support truncated to $[0,1]^d$ for $d \geqslant 2$. We shall let the dimension $d$ growing as well as the size $J$ of the support of $\nu$ when $d$ increases. For each $d \geqslant 2$, $\nu$ is chosen as the uniform discrete probability measure supported on $J = 5^d$ points drawn uniformly on the hypercube $[0,1]^d$. We report results for $d \in \{3,4,5\}$. There exist various algorithms for semi-discrete optimal

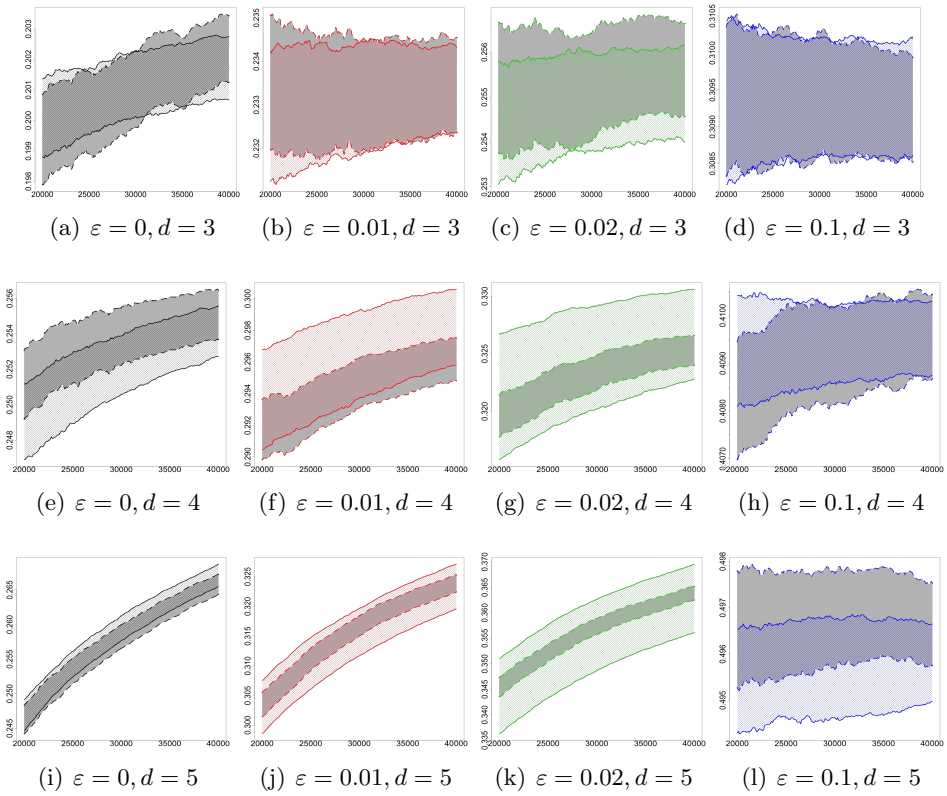
transport in the unregularized case to evaluate $W_0(\mu,\nu)$. We refer to [32, Section 1.2] for an overview and a discussion of their computational cost. These approaches are based on the knowledge of the measure $\mu$ that is generally projected over a partition of $[0,1]^d$. However, available implementations[2] based on the works in [28, 29] are typically restricted to the dimension $d = 2$. For larger values of $d$, projecting $\mu$ on a sufficiently finite partition becomes computationally prohibitive, and storing the resulting cost matrix $\boldsymbol{C}$ becomes too memory demanding which makes a direct use of Sinkhorn or Greenkhorn algorithms not feasible. Moreover, to the best of our knowledge, apart from stochastic approaches as in [24], there is no other method to evaluate $W_\varepsilon(\mu,\nu)$ in the semi-discrete setting.

In the following numerical experiments, we briefly study how the recursive estimators $\widehat{W}_n$ and $\widehat{\sigma}_n$ scales with increasing dimension $d$ and support size $J = 5^d$, for $d \in \{3,4,5\}$, in both unregularized and regularized cases. First, we observe that for various values of the regularization parameter $\varepsilon$ and the dimension $d$, the confidence intervals obtained via the Gaussian approximation (5.2) give an accurate estimation of the range of variation of $\widehat{W}_n$ calculated by Monte Carlo simulations as shown in Figure 3. Note that, in these numerical experiments, we conjecture that the Gaussian approximation (5.2) also holds for $\varepsilon = 0$.

Finally, in Figure 4, we display the evolution of the size $2 \times 1.96\widehat{\sigma}_n/\sqrt{n}$ of the confidence intervals for $W_\varepsilon(\mu,\nu)$ (after $n = 4.10^4$ iterations) based on the Gaussian approximation (5.2) as the dimension $d$ increases and $J = 5^{\lceil\sqrt{d}\rceil}$ for $2 \leqslant d \leqslant 20$. The size of these confidence intervals is clearly an increasing function of $d$. This suggests that the number $n$ of iterations should increase with $d$ to keep the same level of accuracy when estimating $W_\varepsilon(\mu,\nu)$.

**Real data.** We consider a dataset containing spatial locations $X_1,\ldots,X_N$ of reported incidents of crime with the exception of murders in Chicago in 2014, publicly available at `https://data.cityofchicago.org`. These $N$ data points are ordered in a chronological manner from January to December. Victims' addresses are shown at the block level only. Specific locations are not identified in order to protect the privacy of victims and to have a sufficient amount of data for the statistical analysis. For simplicity, spatial locations of the city of Chicago are represented on the unit square $[0,1]^2$. For the year 2014, $N = 16104$ spatial locations of reported incidents of crime in Chicago are available. They are displayed in Figure 5(a). Chicago has $J = 23$ Police stations whose locations are shown in Figure 5(b) with a kernel density estimation of the unknown distribution $\mu$ of crime locations.

[2] `https://github.com/mrgt/PyMongeAmpere` and `https://cran.r-project.org/package=transport`



(a) $\varepsilon = 0, d = 3$ (b) $\varepsilon = 0.01, d = 3$ (c) $\varepsilon = 0.02, d = 3$ (d) $\varepsilon = 0.1, d = 3$

(e) $\varepsilon = 0, d = 4$ (f) $\varepsilon = 0.01, d = 4$ (g) $\varepsilon = 0.02, d = 4$ (h) $\varepsilon = 0.1, d = 4$

(i) $\varepsilon = 0, d = 5$ (j) $\varepsilon = 0.01, d = 5$ (k) $\varepsilon = 0.02, d = 5$ (l) $\varepsilon = 0.1, d = 5$

FIG 3. *Comparison of confidence intervals for $W_\varepsilon(\mu, \nu)$ based on Gaussian approximation and Monte Carlo simulations along the iterations for various values of $\varepsilon$, $d$ and $J = 5^d$. The grey areas represent the 95% confidence intervals $\widehat{W}_n \pm 1.96\widehat{\sigma}_n/\sqrt{n}$ for $10^4 \leqslant n \leqslant 4.10^5$, while the colored areas are Monte Carlo confidence intervals obtained from $N = 100$ repetitions.*

We assume that Police stations have the same capacity, and they are thus modeled by the uniform discrete measure $\nu$ on these locations. We first report the evolution of the recursive confidence intervals $\widehat{W}_n \pm 1.96\widehat{\sigma}_n/\sqrt{n}$ for various values of $\varepsilon$ in the unregularized and regularized cases. To evaluate the convergence of our stochastic algorithm, we have also computed the values of $W_\varepsilon(\widehat{\mu}_N, \nu)$ where $\widehat{\mu}_N$ is the standard empirical measure approximating $\mu$.

For the regularized case $\varepsilon > 0$, we used the Sinkhorn algorithm. For $\varepsilon = 0$, we followed the method proposed in [28] that is specific to the Euclidean cost $c(x, y) = \|x - y\|_2$ and implemented in the package `Transport`. One can observe in Figure 5(c) a very good convergence of the algorithm for different values of $\varepsilon$.

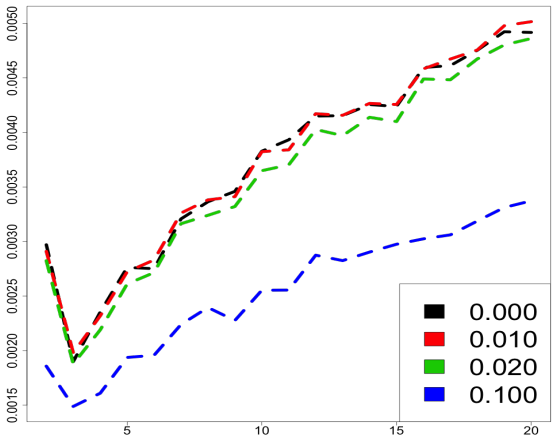



FIG 4. *Evolution of the size of the confidence intervals for* $W_\varepsilon(\mu,\nu)$ *as a function of the dimension* $d$ *with* $J = 5^{\lceil\sqrt{d}\rceil}$.

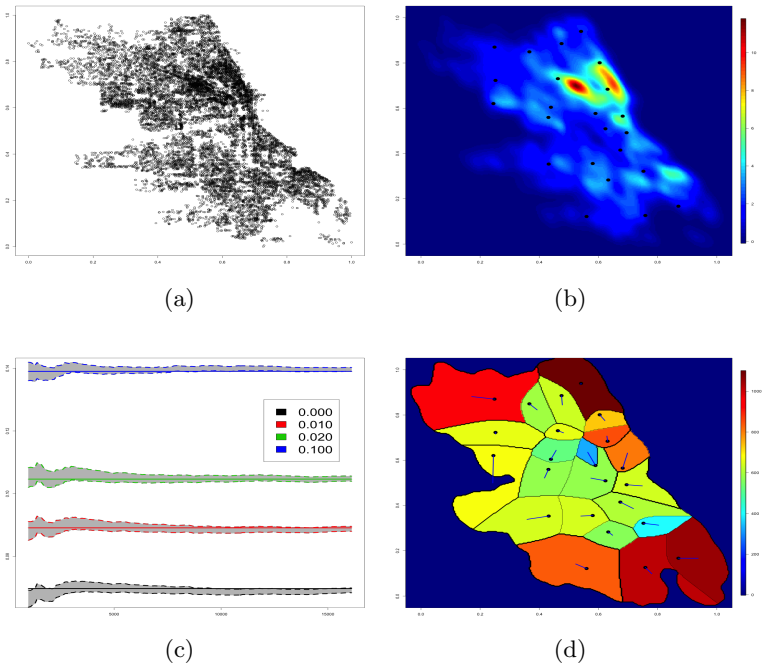
Finally, we consider the problem of estimating an optimal partition of the city of Chicago into 23 districts matching expected locations of crimes with the capacity of Police stations so that the expected cost of travelling from a station to a crime's location is minimal. This can be done by estimating, in the unregularized case, an optimal map $T$ which pushes forward $\mu$ onto $\nu$. Since $\mu$ is absolutely continuous, it is well-known [9, 13] that there exists a unique optimal mapping $T : \mathrm{supp}(\mu) \to \{y_1, \ldots, y_J\}$ which pushes forward $\mu$ onto $\nu$. This mapping is clearly piecewise constant. It follows from Corollary 1.2 in [29] that for all $1 \leqslant j \leqslant J$,

$$T^{-1}(y_j) = \Big\{ x \in \mathrm{supp}(\mu), c(x,y_j) - v^*_{j,0} \leqslant c(x,y_k) - v^*_{k,0} \text{ for all } 1 \leqslant k \leqslant J \Big\}$$

where $v^* \in \mathbb{R}^J$ is any maximiser of the semi-dual problem (2.6). The sets $\{T^{-1}(y_j)\}$ are the so-called Laguerre cells that correspond to an important concept from computational geometry (see e.g. [28, 29] and Chapter 5 in [16]). Then, based on a sample $X_1, \ldots, X_N$ from $\mu$, it is natural to estimate the Laguerre cells by

$$\widehat{T}_N^{-1}(y_j) = \Big\{ x \in \mathrm{supp}(\mu), c(x,y_j) - \widehat{V}^0_{N,j} \leqslant c(x,y_k) - \widehat{V}^0_{N,k} \text{ for all } 1 \leqslant k \leqslant J \Big\}$$

where $\widehat{V}^0_{N,j}$ stands for $j$-entry of the vector $\widehat{V}^0_N$ obtained from (4.2). An example of estimated Laguerre cells is given in Figure 5(d). We observe that cells of small size are located near the modes of the estimated distribution of crime locations.




(a) (b)

(c) (d)

FIG 5. *First row: (a)* $N = 16104$ *spatial locations of reported incidents crime in Chicago. (b) Police stations locations in black dots and kernel density estimation of the distribution of crime locations. Second row: (c) The solid horizontal lines correspond to* $W_\varepsilon(\widehat{\mu}_N, \nu)$ *while the grey areas represent the* 95% *confidence intervals* $\widehat{W}_n \pm 1.96\widehat{\sigma}_n/\sqrt{n}$ *for* $1 \leqslant n \leqslant N$*. (d) Estimated Laguerre cells associated to* $\widehat{V}_N^0$*. The color in each cells corresponds to the number of reported incidents of crime it contains and the blue segments allow to link cells to the Police stations.*

**Acknowledgements.** The authors would like to thank the associate editor and the three referees for their suggestions and constructive comments which helped to improve the paper substantially, in particular on the investigation of the classical transportation problem without regularization. Jérémie Bigot is a member of Institut Universitaire de France (IUF), and this work has been carried out with financial support from the IUF.

## SUPPLEMENTARY MATERIAL

**Supplementary material: Proofs of the main results**
(). The supplement consists of Appendix A and Appendix B that contain the proofs of the main results.

**References.**

[1] Abid, B. K., and Gower, R. M. Greedy stochastic algorithms for entropy-regularized optimal transport problems. In *Proceedings of the 21th International Conference on Artificial Intelligence and Statistics* (Lanzarote, Spain, Apr. 2018).

[2] Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA* (2017), pp. 1961–1971.

[3] Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (USA, 2017), NIPS'17, Curran Associates Inc., pp. 1961–1971.

[4] Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017* (2017), pp. 214–223.

[5] Bach, F. R. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *Journal of Machine Learning Research 15*, 1 (2014), 595–627.

[6] Bach, F. R., and Moulines, E. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011. Proceedings of a meeting held 12-14 December 2011, Granada, Spain.* (2011), pp. 451–459.

[7] Bigot, J., Cazelles, E., and Papadakis, N. Central limit theorems for Sinkhorn divergence between probability distributions on finite spaces and statistical applications. Preprint - arXiv:1711.08947, Nov. 2017.

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

[9] Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math. 44*, 4 (1991), 375–417.

[10] Cattiaux, P., and Gamboa, F. Large deviations and variational theorems for marginal problems. *Bernoulli 5*, 1 (1999), 81–108.

[11] Chizat, L., Peyré, G., Schmitzer, B., and Vialard, F. Scaling algorithms for unbalanced optimal transport problems. *Math. Comput. 87*, 314 (2018), 2563–2609.

[12] Csiszar, I. $i$-divergence geometry of probability distributions and minimization problems. *Ann. Probab. 3*, 1 (1975), 146–158.

[13] Cuesta, J. A., and Matran, C. Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab. 17*, 3 (1989), 1264–1276.

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

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

[16] Cuturi, M., and Peyré, G. *Computational Optimal Transport.* Book available at https://optimaltransport.github.io/book/, 2017.

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

[18] del Barrio, E., Giné, E., and Utzet, F. Asymptotics for $L_2$ functionals of

the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli 11*, 1 (2005), 131–189.

[19] del Barrio, E., and Loubes, J.-M. Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab. 47*, 2 (03 2019), 926–951.

[20] Evans, S. N., and Matsen, F. A. The phylogenetic Kantorovich-Rubinstein metric for environmental sequence samples. *J. R. Stat. Soc. Ser. B. Stat. Methodol. 74*, 3 (2012), 569–592.

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

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

[23] Genevay, A. *Entropy-regularized optimal transport for machine learning*. PhD thesis, Université Paris-Dauphine, 2019.

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

[25] Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics* (2018), A. Storkey and F. Perez-Cruz, Eds., vol. 84 of *Proceedings of Machine Learning Research*, PMLR, pp. 1608–1617.

[26] Godichon-Baggioni, A. Lp and almost sure rates of convergence of averaged stochastic gradient algorithms: locally strongly convex objective. *ESAIM: Probability and Statistics To be published* (2019).

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

[28] Hartmann, V., and Schuhmacher, D. Semi-discrete optimal transport - the case p = 1. *Preprint, arXiv:1706.07650* (2018).

[29] Kitagawa, J., Mérigot, Q., and B., T. Convergence of a newton algorithm for semi-discrete optimal transport. *Journal of the European Math Society To be published* (2018).

[30] Klatt, M., Tameling, C., and Munk, A. Empirical regularized optimal transport: Statistical theory and applications. *Preprint, arXiv:1810.09880* (2018).

[31] Léonard, C. A survey of the Schrödinger problem and some of its connections with optimal transport. *Discrete Contin. Dyn. Syst. 34*, 4 (2014), 1533–1574.

[32] Mérigot, Q., Meyron, J., and Thibert, B. An algorithm for optimal transport between a simplex soup and a point cloud. *SIAM Journal on Imaging Sciences 11*, 2 (2018), 1363–1389.

[33] Munk, A., and Czado, C. Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B Stat. Methodol. 60*, 1 (1998), 223–241.

[34] Pelletier, M. On the almost sure asymptotic behaviour of stochastic algorithms. *Stochastic Processes and their Applications 78*, 2 (1998), 217–244.

[35] Pelletier, M. Weak convergence rates for stochastic approximation with application to multiple targets and simulated annealing. *Ann. Appl. Probab. 8*, 1 (1998), 10–44.

[36] Polyak, B. T., and Juditsky, A. Acceleration of stochastic approximation by

averaging. *SIAM J. Control Optim. 30*, 4 (1992), 838–855.

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

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

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

[40] Robbins, H., and Monro, S. A stochastic approximation method. *The Annals of Mathematical Statistics 22*, 3 (September 1951), 400–407.

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

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

[43] Sommerfeld, M., and Munk, A. Inference for empirical Wasserstein distances on finite spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology). 80*, 1 (2018), 219–238.

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

[45] Villani, C. *Optimal transport: old and new*, vol. 338. Springer Science & Business Media, 2008.

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

Université de Bordeaux
Institut de Mathématiques de Bordeaux et CNRS (UMR 5251)
351, cours de la Libération
33405 Talence
E-mail: bernard.bercu@u-bordeaux.fr
jeremie.bigot@u-bordeaux.fr