Faster Wasserstein Distance Estimation with the Sinkhorn Divergence

Lénaïc Chizat[1*], **Pierre Roussillon**[2], **Flavien Léger**[2], **François-Xavier Vialard**[3], **Gabriel Peyré**[2]

1: Laboratoire de Mathématiques d'Orsay, CNRS, Université Paris-Saclay, Orsay, France
2: ENS, PSL University, Paris, France
3: Univ. Gustave Eiffel, CNRS, ESIEE Paris, Marne-la-Vallée, France

Abstract

The squared Wasserstein distance is a natural quantity to compare probability distributions in a non-parametric setting. This quantity is usually estimated with the plug-in estimator, defined via a discrete optimal transport problem which can be solved to ε-accuracy by adding an entropic regularization of order ε and using for instance Sinkhorn's algorithm. In this work, we propose instead to estimate it with the Sinkhorn divergence, which is also built on entropic regularization but includes debiasing terms. We show that, for smooth densities, this estimator has a comparable sample complexity but allows higher regularization levels, of order $\varepsilon^{1/2}$, which leads to improved computational complexity bounds and a strong speedup in practice. Our theoretical analysis covers the case of both randomly sampled densities and deterministic discretizations on uniform grids. We also propose and analyze an estimator based on Richardson extrapolation of the Sinkhorn divergence which enjoys improved statistical and computational efficiency guarantees, under a condition on the regularity of the approximation error, which is in particular satisfied for Gaussian densities. We finally demonstrate the efficiency of the proposed estimators with numerical experiments.

1 Introduction

Certain tasks in machine learning (implicit generative modeling [41], two-sample testing [50], structured prediction [25]) and imaging sciences (shape matching [31], computer graphics [9]) require to quantify how much two probability densities $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ differ. The squared Wasserstein distance $W_2^2(\mu, \nu)$ (defined below) is often well suited for this purpose because of its appealing geometrical properties [57, 52, 46] but it also raises important statistical and computational challenges. Indeed, in many practical settings, μ and ν are only accessed via empirical or discretized measures $\hat{\mu}_n, \hat{\nu}_n$ composed of n atoms. A standard workaround is to use the *plug-in estimator* $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$, but although it is efficient when μ and ν are discrete [55, 56], this estimator suffers from the curse of dimensionality when μ and ν have densities [59, Cor. 2], with an estimation error that scales as $n^{-2/d}$ as we show in Section 3. Moreover, solving the discrete optimal transport problem is computationally demanding when n is large, with a time complexity bound scaling as $n^2 \log(n)/\varepsilon^2$ to reach ε-accuracy with Sinkhorn's algorithm [20, 2]. These drawbacks give a strong motivation to define and study alternative estimators for $W_2^2(\mu, \nu)$ when μ and ν admit smooth densities.

Entropic regularization of optimal transport. In this paper, we consider instead estimators based on the idea of entropic regularization of optimal transport [61, 21, 36, 16]. When μ and ν have finite

[*] Corresponding author: `lenaic.chizat@universite-paris-saclay.fr`

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

second moments, the entropy regularized optimal transport cost is defined as

$$T_\lambda(\mu,\nu) \overset{\text{def.}}{=} \min_{\gamma\in\Pi(\mu,\nu)} \int_{(\mathbb{R}^d)^2} \|y-x\|_2^2 \,\mathrm{d}\gamma(x,y) + 2\lambda H(\gamma,\mu\otimes\nu) \tag{1}$$

where $\Pi(\mu,\nu)$ is the set of transport plans between μ and ν, $\lambda \geq 0$ is the regularization parameter, and $H(\gamma,\mu\otimes\nu)$ is the entropy of γ with respect to the product measure $\mu\otimes\nu$ (see details in the Notations paragraph). The squared Wasserstein distance is defined as $W_2^2(\mu,\nu) \overset{\text{def.}}{=} T_0(\mu,\nu)$. Entropic regularization has been popularized as a method to compute $W_2^2(\hat\mu_n,\hat\nu_n)$ efficiently or simply as a different notion of discrepancy between measures. In contrast, we use it as a tool to directly estimate $W_2^2(\mu,\nu)$. For this purpose, the choice $T_\lambda(\hat\mu_n,\hat\nu_n)$ is not ideal because its large bias requires to set λ to a small value, leading to computational difficulties.

The proposed estimators. The first estimator that we consider is $\hat S_{\lambda,n} = S_\lambda(\hat\mu_n,\hat\nu_n)$ where S_λ is the *Sinkhorn divergence* [50] defined as

$$S_\lambda(\mu,\nu) \overset{\text{def.}}{=} T_\lambda(\mu,\nu) - \frac{1}{2}\big(T_\lambda(\mu,\mu) + T_\lambda(\nu,\nu)\big). \tag{2}$$

In previous work [23], the debiasing terms have been theoretically justified as a mean to have $S_\lambda(\mu,\nu) \geq 0$ with equality when $\mu=\nu$, a property not satisfied by T_λ. In the present work, we show that they in fact allow, under regularity assumptions, to approximate $W_2^2(\mu,\nu)$ with an error of order λ^2, instead of $\lambda\log(1/\lambda)$ for the uncorrected quantity T_λ. We also consider the estimator $\hat R_{\lambda,n} = R_\lambda(\hat\mu_n,\hat\nu_n)$ where R_λ is built from S_λ via Richardson extrapolation as

$$R_\lambda(\mu,\nu) \overset{\text{def.}}{=} 2S_\lambda(\mu,\nu) - S_{\sqrt{2}\lambda}(\mu,\nu). \tag{3}$$

This estimator has a smaller approximation error in $o(\lambda^2)$ and potentially in $O(\lambda^4)$ under restrictive regularity assumptions.

Contributions. We make the following contributions:

- In Section 2, we exploit the dynamical formulation of (1) to show that $|S_\lambda(\mu,\nu) - W_2^2(\mu,\nu)| \leq \lambda^2 I$ where I depends on the Fisher information of μ, of ν and of the W_2-geodesic connecting them. We also give a second-order expansion of this approximation error and detail several situations where I admits a priori bounds.
- In Section 3.1, we prove a sample complexity bound for the plug-in estimator $W_2^2(\hat\mu_n,\hat\nu_n)$ of order $n^{-2/d}$ which has a tight exponent in contrast to the previously known rate $n^{-1/d}$. This is the baseline rate against which we compare the performance of $\hat S_{\lambda,n}$ and of $\hat R_{\lambda,n}$.
- In Section 3.2, we study the performance of the Sinkhorn divergence estimator $\hat S_{\lambda,n}$ given independent samples. We show that when λ is properly chosen, it enjoys comparable sample complexity bounds and improved computational guarantees in a certain sense. We also study the performance when the marginals are discretized on a uniform grid in Section 3.3.
- In Section 4, we study estimators based on Richardson extrapolation such as $\hat R_{\lambda,n}$. Under an abstract and stronger regularity assumption, this estimator enjoys better computational and sample complexity bounds than the plug-in estimator. We discuss this assumption and show that it is satisfied for Gaussian densities.
- In Section 5, we perform numerical experiments that confirm the benefits of the proposed estimators and suggest that our theoretical results could be extended in several ways.

Previous Works. Without additional assumptions, no estimator achieves better statistical rates than the plug-in estimator [44, Thm. 3]. Recent breakthroughs in statistical optimal transport [60, 33] have shown that other estimators can exploit smoothness assumptions to attain faster and nearly minimax estimation rates for W_2 or the dual potentials, but they are *a priori* not computationally efficient. In contrast, our goal in this paper is to improve the computational efficiency of estimating $W_2^2(\mu,\nu)$ and we are not aiming at statistical optimality.

The idea of entropic regularization has a long history in computational optimal transport. It has been shown in [2, 20] that solving $T_\lambda(\hat\mu_n,\hat\nu_n)$ to ε-accuracy requires $O(n^2/(\lambda\varepsilon))$ arithmetic operations using Sinkhorn's algorithm if the domain is bounded (see Appendix B). We use this bound in our discussions on computational complexity because it cleanly quantifies how harder the problem becomes as λ becomes smaller and also because Sinkhorn's algorithm is simple to implement and widely used in practice. Choosing $\lambda \asymp \varepsilon/\log(n)$ allows in turn to estimate $W_2^2(\hat{\mu}_n, \hat{\nu}_n)$ to ε-accuracy in $O(n^2\log(n)/\varepsilon^2)$ operations [20]. There are however various algorithms with better guarantees both for the regularized [20, 1, 13] and the unregularized problem [37, 47, 7]. In our numerical experiments, we use Sinkhorn's iterations combined with Anderson's acceleration [3, 53], which in practice strongly speeds up convergence.

In front of the difficulty to estimate $W_2^2(\mu,\nu)$, researchers have also turned their attention to similar but more tractable discrepancy measures such as the sliced Wasserstein distance [49] or the Sinkhorn divergence [50], which can be both estimated at the parametric rate [26, 40, 39, 42]. However, there is "no free lunch" and unconditional statistical efficiency comes at the price of lack of adaptivity and discriminative power. In particular, it is known that when $\lambda \to \infty$, $S_\lambda(\mu,\nu)$ converges to the squared distance between the expectations of μ and ν, which is a degenerate form of Kernel Mean Discrepancy [27, 23]. This shows that the discriminative power of S_λ decreases as λ increases, but this phenomenon is not yet well understood nor quantified. From a theoretical viewpoint, we thus believe that seeing S_λ as an estimator for W_2^2 allows to clarify the trade-offs at play in the choice of λ between the statistical, approximation and computational errors.

Notations. For two probability measures $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$, we denote by $\Pi(\mu,\nu)$ the set of transport plans between μ and ν, which is the set of measures $\gamma \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ with marginal μ (resp. ν) on the first (resp. second) factor of $\mathbb{R}^d \times \mathbb{R}^d$. The quantity $H(\mu,\nu)$ is the entropy of μ relative to ν, defined as $H(\mu,\nu) \overset{\text{def.}}{=} \int \log(\mathrm{d}\mu/\mathrm{d}\nu)\mathrm{d}\mu$ when μ is absolutely continuous with respect to ν, and $+\infty$ otherwise. When μ has a density with respect to the Lebesgue measure, written $\mu(x)$, we define $H(\mu) \overset{\text{def.}}{=} \int \log(\mu(x))\mu(x)\mathrm{d}x$ its entropy relative to the Lebesgue measure. Finally, $\mu \otimes \nu \in \mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$ is the product measure characterized by $(\mu\otimes\nu)(A\times B) = \mu(A)\nu(B)$ for any pair of Borel sets $A, B \subset \mathbb{R}^d$.

2 Refined approximation bound for the Sinkhorn divergence

In this section, we study the approximation error of S_λ. To this goal, we leverage the dynamical formulation of entropic optimal transport [12, 28, 30, 14] which states that, for $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ absolutely continuous probability measures with compact support,

$$T_\lambda(\mu,\nu) + d\lambda\log(2\pi\lambda) + \lambda(H(\mu)+H(\nu)) = \\ \inf_{\rho,v} \int_0^1 \int_{\mathbb{R}^d} \Big(\|v(t,x)\|_2^2 + \frac{\lambda^2}{4}\|\nabla_x \log(\rho(t,x))\|_2^2\Big)\rho(t,x)\,\mathrm{d}x\,\mathrm{d}t\,, \quad (4)$$

where the infimum is taken over time-dependent probability measures $\rho(t,x)$ that interpolate between μ at $t=0$ and ν at $t=1$, and time-dependent vector fields $v(t,x)$ under the continuity equation constraint $\partial_t\rho(t,x) + \mathrm{div}(\rho(t,x)v(t,x)) = 0$ where div is the usual divergence operator. The first term in the r.h.s. of Eq. (4) is the kinetic energy and the second is the Fisher information integrated in time. For $\lambda \geq 0$, there exists a unique minimizer of the r.h.s. [30] denoted by ρ_λ and we define

$$I_\lambda(\mu,\nu) \overset{\text{def.}}{=} \int_0^1 \int_{\mathbb{R}^d} \|\nabla_x \log(\rho_\lambda(t,x))\|_2^2\,\rho_\lambda(t,x)\,\mathrm{d}x\,\mathrm{d}t\,. \quad (5)$$

Remark that $I_0(\mu,\mu)$ is the Fisher information of μ and $I_0(\mu,\nu)$ is the Fisher information of the Wasserstein geodesic between μ and ν. Building on [14], we next show that the Sinkhorn divergence approximates $W_2^2(\mu,\nu)$ with an error in $O(\lambda^2)$, as suggested by Eq. (4).

Theorem 1. *Assume that $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ have bounded densities and supports. It holds*

$$\big|S_\lambda(\mu,\nu) - W_2^2(\mu,\nu)\big| \leq \frac{\lambda^2}{4}\max\big\{I_0(\mu,\nu), (I_0(\mu,\mu)+I_0(\nu,\nu))/2\big\}\,.$$

If moreover $I_0(\mu,\nu), I_0(\mu,\mu), I_0(\nu,\nu) < \infty$ then

$$S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = \frac{\lambda^2}{4}\big(I_0(\mu,\nu) - (I_0(\mu,\mu)+I_0(\nu,\nu))/2\big) + o(\lambda^2).$$

Proof. Denote the right-hand side of (4) by $J_{\lambda^2}(\mu,\nu)$ and note that $S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = (J_{\lambda^2}(\mu,\nu) - J_0(\mu,\nu)) - (J_{\lambda^2}(\mu,\mu) + J_{\lambda^2}(\nu,\nu))/2$ and $J_0(\mu,\mu) = J_0(\nu,\nu) = 0$. Since ρ_0 is feasible in Eq. (4), we have $J_0(\mu,\nu) \leq J_{\lambda^2}(\mu,\nu) \leq J_0(\mu,\nu) + (\lambda^2/4)I_0(\mu,\nu)$, hence the bound. For the second claim, we prove in Appendix A (Lemma 1) that the right derivative at 0 of $\sigma \mapsto J_\sigma$ is $\frac{1}{4}I_0(\mu,\nu)$, which justifies the Taylor expansion. □

The Fisher information of μ or ν can be bounded by assuming regularity of the densities, but bounding $I_0(\mu,\nu)$ is more subtle. Next, we bound $I_0(\mu,\nu)$ assuming regularity on the *Brenier potential* φ, which is the convex function such that $\nabla\varphi$ is the optimal transport map from μ to ν [52].

Proposition 1. *Let $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ be absolutely continuous with compact support. Assume that the Brenier potential φ has a Hessian satisfying $0 \prec \kappa\mathrm{Id} \preceq \nabla^2\varphi \preceq K\mathrm{Id}$ and that $\nabla^2\varphi$ is L-Lipschitz continuous, then $I_0(\mu,\nu) \leq 2\kappa^{-1}(I_0(\mu,\mu) + \kappa^{-2}L^2/3)$. In particular, if φ is quadratic then $I_0(\mu,\nu) \leq 2\kappa^{-1}I_0(\mu,\mu)$. If $d = 1$, then $I_0(\mu,\nu) \leq \frac{2}{3}(\kappa^{-1}I_0(\mu,\mu) + KI_0(\nu,\nu))$.*

Sufficient conditions on the densities of μ and ν to guarantee bounds on $\nabla^2\varphi$ are known (e.g. bounds on their first derivative and on their log-densities over their convex support [17, Thm 3.3]). However, the assumption that $\nabla^2\varphi$ is Lipschitz continuous is more demanding and potentially not sharp as it can be avoided when $d = 1$. Note that the Brenier potential φ is quadratic whenever the densities are in the same family of elliptically contoured distributions [6]. For Gaussian densities, we show in Appendix A that $I_0(\mu,\nu)$ admits an explicit expression, given in Section 4.

3 Performance analysis of the Sinkhorn divergence estimator

In this section, we discuss the performance of the Sinkhorn divergence estimator in two situations: when we observe independent samples or when we have access to discretized densities. But first, we study the plug-in estimator, which is the baseline against which our estimators are compared.

3.1 Analysis of the plug-in estimator

A tighter statistical bound for the plug-in estimator. Let us first study the rate of convergence of $W_2^2(\hat\mu_n,\hat\nu_n)$ towards $W_2^2(\mu,\nu)$ where $\hat\mu_n$ and $\hat\nu_n$ are empirical distributions of n independent samples. This is well-studied in the case $\mu = \nu$, but the case $\mu \neq \nu$ was not specifically covered in the literature except for discrete measures [55].

Theorem 2. *If $\mu,\nu \in \mathcal{P}(\mathbb{R}^d)$ are supported on a set of diameter 1 then it holds*

$$\mathbf{E}\big[|W_2^2(\hat\mu_n,\hat\nu_n) - W_2^2(\mu,\nu)|\big] \lesssim \begin{cases} n^{-2/d} & \text{if } d > 4, \\ n^{-1/2}\log(n) & \text{if } d = 4, \\ n^{-1/2} & \text{if } d < 4, \end{cases}$$

where the notation $\lesssim$ hides constants that only depend on the dimension d. Also, this estimator concentrates well around its expectation, in the sense that for all $t \geq 0$,

$$\mathbf{P}\Big[|W_2^2(\hat\mu_n,\hat\nu_n) - \mathbf{E}[W_2^2(\hat\mu_n,\hat\nu_n)]| \geq t\Big] \leq 2\exp(-nt^2).$$

To prove this result in Appendix C, we first upper bound the expected error by the Rademacher complexity of a certain set of convex and Lipschitz functions. We use Dudley's chaining and a bound on the covering number of this set of functions due to Bronshtein [10] to conclude. The concentration bound is already present in a similar form in [59, Prop. 20]. When $\mu = \nu$, this bound is well-known and has a sharp exponent [59, 54, 8, 19, 24]. However, perhaps surprisingly, this result implies that the plug-in estimator $W_2(\hat\mu_n,\hat\nu_n)$ (without the square) converges at the rate $n^{-2/d}$ when $\mu \neq \nu$, while only a bound in $n^{-1/d}$ (the rate when $\mu = \nu$) was known. This is the content of the following corollary. See Figure 1 for a numerical illustration of these rates.

Corollary 1. *Assume that μ,ν are supported on a set of diameter 1 and satisfy $W_2(\mu,\nu) \geq \alpha > 0$. Then $\mathbf{E}\big[|W_2(\mu_n,\nu_n) - W_2(\mu,\nu)|\big]$ enjoys the bound given in Theorem 2 multiplied by $1/\alpha$.*

Proof. It is sufficient to take expectations in the following inequality :

$$|W_2(\hat\mu_n,\hat\nu_n) - W_2(\mu,\nu)| = \frac{|W_2^2(\hat\mu_n,\hat\nu_n) - W_2^2(\mu,\nu)|}{W_2(\mu,\nu) + W_2(\hat\mu_n,\hat\nu_n)} \leq \frac{1}{\alpha}|W_2^2(\hat\mu_n,\hat\nu_n) - W_2^2(\mu,\nu)|. \quad \square$$

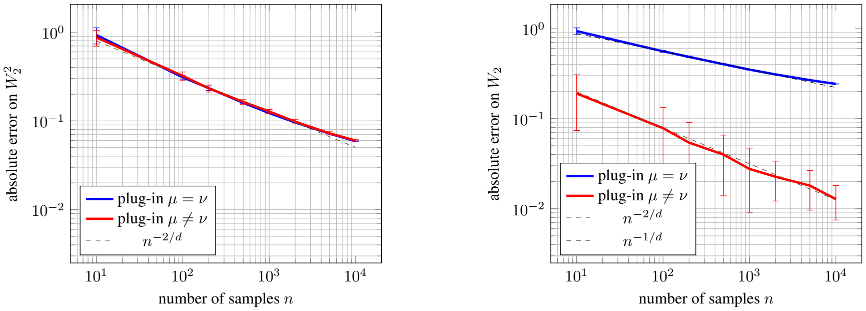

Figure 1: Estimation error of the plug-in estimator for μ, ν compactly supported with $d = 5$ (as detailed in Appendix G). Left: error on the cost W_2^2 has rate $n^{-2/d}$ (Theorem 2). Right: error on W_2 has rate $n^{-1/d}$ if $\mu = \nu$ and $n^{-2/d}$ if $\mu \neq \nu$ (Corollary 1) with $\mathbf{E}_\mu[x] = 0$ and $\mathbf{E}_\nu[x] = (1, \ldots, 1)$.

Computational complexity via Sinkhorn's algorithm. In previous work [2, 20], solving $T_\lambda(\hat\mu_n, \hat\nu_n)$ with $\lambda > 0$ has been studied as a computationally efficient way to compute $T_0(\hat\mu_n, \hat\nu_n)$ and related quantities. One standard algorithm to compute T_λ is Sinkhorn's algorithm, which can be interpreted as alternate block maximization on the dual of Eq. (1), see Appendix B. Given two discrete marginals $\hat\mu_n = \sum_{i=1}^n p_i \delta_{x_i}$ and $\hat\nu_n = \sum_{i=1}^n q_j \delta_{y_j}$, let us define the cost matrix with entries $c_{i,j} = \frac{1}{2}\|x_i - y_j\|_2^2$. The iterates $u^{(k)}, v^{(k)} \in \mathbb{R}^n$, $k \geq 1$ of Sinkhorn's algorithm are defined as follows: let $v^{(0)} = 0 \in \mathbb{R}^n$ and let

$$u_i^{(k)} = -\lambda \log\Big(\sum_{j=1}^n e^{(v_j^{(k-1)} - c_{i,j})/\lambda} q_j\Big) \quad \text{and} \quad v_j^{(k)} = -\lambda \log\Big(\sum_{i=1}^n e^{(u_i^{(k)} - c_{i,j})/\lambda} p_i\Big). \tag{6}$$

An estimate for $\hat T_{\lambda,n} \overset{\text{def.}}{=} T_\lambda(\hat\mu_n, \hat\nu_n)$ is then given by $\hat T_{\lambda,n}^{(k)} = 2\sum_{i=1}^n u_i^{(k)} p_i + v_i^{(k)} q_i$. These iterations enjoy the following guarantee, proved in [20] (see details in Appendix B).

Proposition 2. *It holds* $|\hat T_{\lambda,n}^{(k)} - \hat T_{\lambda,n}| \leq 2\|c\|_\infty^2/(\lambda k)$ *where* $\|c\|_\infty = \max_{i,j} \|x_i - y_j\|_2^2/2$.

In particular, taking into account the fact that each iteration requires $O(n^2)$ arithmetic operations, Sinkhorn's algorithm returns an ε-accurate estimation of $\hat T_{\lambda,n}$ in time $O(n^2\|c\|_\infty^2/(\lambda\varepsilon))$. Moreover, if $\alpha > 0$ is such that $p_i, q_j \geq \alpha/n$, we have the approximation bound $|\hat T_{\lambda,n} - \hat T_{0,n}| \leq 4\lambda\log(n/\alpha)$ which follows by bounding the relative entropy of admissible transport plans [2]. By fixing $\lambda = \varepsilon/4(\log(n/\alpha))$, we thus obtain an ε-accurate estimation of $\hat T_{0,n}$ in $O(n^2\log(n/\alpha)\|c\|_\infty^2/\varepsilon^2)$ operations. As a consequence, by combining Theorem 2 and Proposition 2, we can thus give the following computational complexity bound to estimate $W_2^2(\mu, \nu)$ given random samples that takes into account the number of samples and the regularization level required to reach a certain accuracy.

Proposition 3. *Assume that μ, ν are supported on a set of diameter* 1. *Using $\hat T_{\lambda,n}^{(k)}$, an ε-accurate estimation of $W_2^2(\mu, \nu)$ is achieved with probability $1 - \delta$ in $\tilde O(\varepsilon^{-\max\{6, d+2\}})$ operations, where $\tilde O$ hides poly-log factors in $1/\varepsilon$ and $1/\delta$.*

Proof idea. We write $W_2^2 \overset{\text{def.}}{=} W_2^2(\mu, \nu)$, $\hat W_2^2 \overset{\text{def.}}{=} W_2^2(\hat\mu_n, \hat\nu_n)$ and consider the error decomposition

$$|\hat T_{\lambda,n}^{(k)} - W_2^2| \leq |\hat T_{\lambda,n}^{(k)} - \hat T_{\lambda,n}| + |\hat T_{\lambda,n} - \hat W_2^2| + |\hat W_2^2 - \mathbf{E}[\hat W_2^2]| + \mathbf{E}|\hat W_2^2 - W_2^2||$$

where each term has been bounded in the previous discussion, see details in Appendix C. □

3.2 Performance of the Sinkhorn divergence estimator given random samples

Statistical performance. Let us now turn to our object of interest which is the Sinkhorn divergence estimator $\hat S_{\lambda,n} \overset{\text{def.}}{=} S_\lambda(\hat\mu_n, \hat\nu_n)$, defined from n independent samples from μ and ν. We note that all the results in this section also apply to the estimator $T_\lambda(\hat\mu_n, \hat\nu_n) - (T_\lambda(\hat\mu_{n/2}, \hat\mu'_{n/2}) + T_\lambda(\hat\nu_{n/2}, \hat\nu'_{n/2}))/2$ where $\hat\mu_{n/2}$ (resp. $\hat\mu'_{n/2}$) is the empirical distribution of the first (resp. second) half samples from μ (assuming n even for conciseness), which is a natural alternative definition. The following result gives the expected error of the estimator $\hat S_{\lambda,n}$.

Proposition 4. *Let μ, ν be supported on a set of diameter 1 and assume that $|S_\lambda(\mu,\nu) - W_2^2(\mu,\nu)| \leq \lambda^2 I$ for some $I > 0$ (see guarantees in Section 2). Then, with the choice $\lambda = n^{\frac{-1}{d'+4}}$, it holds*

$$\mathbf{E}\big[|\hat{S}_{\lambda,n} - W_2^2(\mu,\nu)|\big] \lesssim n^{\frac{-2}{d'+4}}.$$

where $d' = 2\lfloor d/2 \rfloor$ and $\lesssim$ hides a constant depending only on I and d. Also, this estimator concentrates well around its expectation: for all $t, \lambda \geq 0$, $\mathbf{P}\Big[|\hat{S}_{\lambda,n} - \mathbf{E}[\hat{S}_{\lambda,n}]| \geq t\Big] \leq 2\exp(-nt^2/4)$.

Observe that when d is large, the exponent $-2/(d'+4)$ is equivalent to $-2/d$ which is the rate of the plug-in estimator as shown in Theorem 2. However, except for $d = 1$, this exponent is slightly worse and we believe that this is due to a weakness in our bound. In fact, in our numerical experiments we observe that $\hat{S}_{\lambda,n}$ is in fact more statistically efficient than the plug-in estimator (cf. Figure 2).

Computational performance. An ideal theoretical goal would be to exhibit a computational advantage for using $\hat{S}_{\lambda,n}$ in the sense of Proposition 3, but unfortunately the statistical bound in Proposition 4 is not strong enough to allow for such a result. Still, there is a clear computational advantage in using $\hat{S}_{\lambda,n}$ which is that to attain an accuracy ε, it requires a regularization level λ of order $\varepsilon^{1/2}$ instead of ε for the plug-in estimator. This advantage can be formalized as follows, where $\hat{S}^{(k)}_{\lambda,n}$ is the estimation of $\hat{S}_{\lambda,n}$ obtained after k Sinkhorn's iterations.

Proposition 5. *Under the assumptions of Proposition 4, an ε-accurate estimation of $W_2^2(\mu,\nu)$ can be obtained with probability $1-\delta$ in $\tilde{O}(\varepsilon^{-(d'+5.5)})$ computations via $\hat{S}^{(k)}_{\lambda,n}$ where $d' = 2\lfloor d/2 \rfloor$ and $\tilde{O}$ hides a poly-log factor in $1/\delta$. Given n samples, both estimators can achieve with probability $1-\delta$ an accuracy $\varepsilon \asymp n^{-2/(d'+4)}$, but in time $\tilde{O}(n^2\varepsilon^{-1.5})$ via $\hat{S}^{(k)}_{\lambda,n}$ and in time $\tilde{O}(n^2\varepsilon^{-2})$ via $T^{(k)}_{\lambda,n}$.*

Proof idea. For $\hat{T}^{(k)}_{\lambda,n}$, we consider the error decomposition of Proposition 3, while for $\hat{S}^{(k)}_{\lambda,n}$, we write

$$|\hat{S}^{(k)}_{\lambda,n} - W_2^2| \leq |\hat{S}^{(k)}_{\lambda,n} - \hat{S}_{\lambda,n}| + |\hat{S}_{\lambda,n} - \mathbf{E}[\hat{S}_{\lambda,n}]| + \mathbf{E}|\hat{S}_{\lambda,n} - S_\lambda| + |S_\lambda - W_2^2|.$$

The key difference with the decomposition in the proof of Proposition 3 is that the error induced by the entropic regularization is bounded on the population quantities instead of the empirical ones. These terms have been bounded in the previous discussion, see details in Appendix D. □

3.3 Performance of the Sinkhorn divergence estimator given densities discretized on grids

In this section, we consider the case where the marginals μ and ν are not randomly sampled, but instead are accessed via their discretized densities which is the common situation in imaging sciences. We show a stability property of the entropy regularized optimal transport which leads to improved error bounds compared to the plug-in estimator.

For simplicity, we consider measures on the d dimensional torus $\mathbb{T}^d = (\mathbb{R}/\mathbb{Z})^d$ with its usual distance denoted by $\|[x-y]\|_2$. For a measure $\mu \in \mathcal{P}(\mathbb{T}^d)$ its discretization μ_h at resolution $h = 1/m$ for an integer m is the discrete measure with $n = m^d$ atoms supported on the regular grid $(\mathbb{Z}/m\mathbb{Z})^d$ which gives to each point the mass of μ on its surrounding cell. The following approximation result suggests that regularizing the optimal transport problem increases the stability under such a discretization.

Proposition 6 (Stability under discretization)**.** *Assume that $\mu, \nu \in \mathcal{P}(\mathbb{T}^d)$ admit M-Lipschitz continuous log-densities and let $C > 0$ be any constant. If $h(M + \lambda^{-1}) \leq C$ then*

$$|T_\lambda(\mu_h, \nu_h) - T_\lambda(\mu,\nu)| \lesssim \min\{h, h^2(\lambda^{-1} + M + 1)\}$$

where $\lesssim$ hides constants that only depend on d and C.

This bound implies an error of order h^2 for the entropy regularized problem while it is not known whether such a bound is possible for $\lambda = 0$, where a naive analysis suggests a bound of order h. When combined with the approximation error and the analysis of Sinkhorn's iterations, this yields the following performance guarantees for $S_\lambda(\mu_h, \nu_h)$ as defined in Eq. (2).

Proposition 7. *Assume that $\mu, \nu \in \mathcal{P}(\mathbb{T}^d)$ admit Lipschitz continuous log-densities and that $I_0(\mu,\nu)$ is finite. We can estimate $W_2^2(\mu,\nu)$ to ε-accuracy:*

- *with $T_\lambda(\mu_h, \nu_h)$ in time $\tilde{O}(\varepsilon^{-(2d+2)})$ by setting $h \asymp \varepsilon$ and $\lambda \asymp \varepsilon/\log(1/\varepsilon)$,*
- *with $S_\lambda(\mu_h, \nu_h)$ in time $O(\varepsilon^{-(3d/2+3/2)})$ by setting $h \asymp \varepsilon^{3/4}$ and $\lambda \asymp \varepsilon^{1/2}$.*

This result suggests that $S_\lambda(\mu_h, \nu_h)$ estimates $W_2^2(\mu, \nu)$ both faster and more accurately than $T_\lambda(\mu_h, \nu_h)$ for their respective optimal λ, and this behavior is observed in numerical experiments (cf. Figure 4). Our aim with Proposition 7 is to illustrate the potential usefulness of the debiasing terms beyond the random sampling setting, but we stress that we are just comparing simple upper bounds which are not intended to be the best possible (in particular, we are not exploiting the fact that the computational cost of each Sinkhorn iteration could be reduced from $O(n^2)$ to $O(n\log(n))$ using discrete convolutions [5, Sec. 6.3.1]). In fact, in a similar setting, a completely different analysis of Sinkhorn's iterations is carried in [5, Cor.1.4], where a time complexity in $\tilde{O}(\varepsilon^{-(2d+1)})$ is derived for $T_\lambda(\mu_h, \nu_h)$.

4 Towards faster estimation with Richardson extrapolation

The systematic bias induced by the Fisher information terms in Theorem 1 can be removed using Richardson extrapolation [35, 51], which usefulness in machine learning was recently pointed out in [4]. This technique consists in taking linear combinations of S_λ for various values of $\lambda > 0$ in order to estimate S_0, by cancelling the successive terms of the Taylor expansion of S_λ at 0. Since in our context the first term of $S_\lambda - S_0$ is of order λ^2, this suggests to define (among other possible choices) $R_\lambda \overset{\text{def.}}{=} 2S_\lambda - S_{\sqrt{2}\lambda}$. Indeed, whenever $S_\lambda = S_0 + \lambda^2 I + o(\lambda^2)$ for some $I \in \mathbb{R}$, such as under the assumptions of Theorem 1, this quantity satisfies $R_\lambda = S_0 + o(\lambda^2)$.

Efficiency of R_λ under an abstract assumption. A difficulty with R_λ, or other extrapolated estimators, is that understanding their performance requires a fine understanding of the regularization path $\lambda \mapsto S_\lambda$. By remarking that in Eq. (4), λ appears only via its square after debiasing, we might conjecture that if S_λ admits a 4th order Taylor expansion at $\lambda = 0$, then the third term vanish. Before giving some arguments in favor of this property, let us state what it implies in terms of the performance of $\hat{R}_{\lambda,n} = \hat{S}_{\lambda,n} - \hat{S}_{\sqrt{2}\lambda,n}$, the extrapolation of the estimator $\hat{S}_{\lambda,n}$.

Proposition 8. *Assume that μ, ν are compactly supported, that $S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = \lambda^2 I + O(\lambda^4)$ for some $I \in \mathbb{R}$ and let $d' = 2\lfloor d/2 \rfloor$. Then with $\lambda \asymp n^{-1/(d'+8)}$ it holds*

$$\mathbf{E}\big[|\hat{R}_{\lambda,n} - W_2^2(\mu,\nu)|\big] \lesssim n^{-4/(d'+8)}.$$

Moreover, with probability $1-\delta$, this estimator returns an ε-accurate estimation of $W_2^2(\mu,\nu)$ with $\tilde{O}(\varepsilon^{-(d'+11)/2})$ computations via Sinkhorn's algorithm where $\tilde{O}$ hides poly-log factors in $1/\delta$.

Proof. We use Lemma 5 to get

$$\mathbf{E}[|\hat{R}_{\lambda,n} - W_2^2(\mu,\nu)|] \leq \mathbf{E}[|\hat{R}_{\lambda,n} - R_\lambda(\mu,\nu)|] + |R_\lambda(\mu,\nu) - W_2^2(\mu,\nu)| \lesssim (1+\lambda^{-d'/2})n^{-1/2} + \lambda^4$$

and optimize the bound in λ. For the last claim we proceed as in the proof of Proposition 3. □

Under this abstract assumption, there is thus a clear statistical improvement over the plug-in estimator for $d > 8$ and a computational improvement for $d > 6$. Notice that a similar performance analysis could be done in the deterministic setting of Section 3.3. In the rest of this section we discuss the assumption of Proposition 8. First we show that it is satisfied in the Gaussian case and second we propose formal calculations towards a 4th order Taylor expansion of T_λ.

Gaussian case. Let $\mu = \mathcal{N}(a, A)$ and $\nu = \mathcal{N}(b, B)$ be Gaussian probability distributions with means $a, b \in \mathbb{R}^d$ and positive definite covariances $A, B \in \mathbb{R}^{d\times d}$. In this case, it is well known that $W_2^2(\mu,\nu) = \|a-b\|_2^2 + \mathrm{d}_\mathrm{b}^2(A,B)$ where $\mathrm{d}_\mathrm{b}^2(A,B) \overset{\text{def.}}{=} \mathrm{tr}(A) + \mathrm{tr}(B) - 2\,\mathrm{tr}(S)$ with $S = (A^{1/2}BA^{1/2})^{1/2}$ is the squared Bures distance [6]. More recently, an explicit expression for $T_\lambda(\mu,\nu)$ was derived in [34, 11, 38]. By a Taylor expansion of this expression (see Appendix F), we find that

$$S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = -\frac{\lambda^2}{8}\mathrm{d}_\mathrm{b}^2(A^{-1}, B^{-1}) + \frac{\lambda^4}{384}\mathrm{d}_\mathrm{b}^2(A^{-3}, B^{-3}) + O(\lambda^5).$$

This expansion shows that the hypotheses of Proposition 8 are satisfied (to the exception of the compactness assumption, but note that sample complexity bounds for S_λ are also known in this case [40]). Also we can explicitly compute the Fisher information $I_0(\mu,\nu) = \mathrm{tr}(S^{-1})$ (Appendix A) which shows that the second order term is consistent, as it must, with the expansion in Theorem 1.

Formal fourth order expansion. Denoting $J_{\lambda^2}(\mu,\nu)$ the r.h.s. of Eq. (4), we show in Lemma 1 that $\sigma \mapsto J_\sigma$ admits a right derivative at all $\sigma \geq 0$ which is the Fisher information $\frac{1}{4} I_{\sqrt{\sigma}}(\mu,\nu)$ defined in Eq. (5). Thus, if we assume that $\sigma \mapsto I_{\sqrt{\sigma}}(\mu,\nu)$ admits a right derivative I_0' at 0, then it holds

$$T_\lambda(\mu,\nu) \;=\; T_0(\mu,\nu) - d\lambda\log(2\pi\lambda) - \lambda(H(\mu) + H(\nu)) + \frac{\lambda^2}{4} I_0(\mu,\nu) + \frac{\lambda^4}{8} I_0' + o(\lambda^4)\,,$$

where $I_0' = \frac{\mathrm{d}}{\mathrm{d}(\lambda^2)} I_\lambda(\mu,\nu)|_{\lambda=0} = \int_0^1 \int_{\mathbb{R}^d} (\|\nabla \log \rho_0\|^2 - 2\Delta\rho_0/\rho_0)\delta_{\lambda^2}\rho_\lambda|_{\lambda=0}\, dx$ is the variation of Fisher information in the direction of $\delta_{\lambda^2}\rho_\lambda|_{\lambda=0_+}$, the first variation of ρ_λ w.r.t. λ^2. Hence under this abstract regularity assumption on $I_{\sqrt{\sigma}}(\mu,\nu)$, the result of Proposition 8 holds true.

5 Numerical experiments

In this section, we assess the statistical and computational efficiency of the proposed estimators on synthetic problems[2]. While this is what our theory controls, the error on the scalar $W_2^2(\mu,\nu)$ is not a suitable quantity to plot as it might vanish spuriously as we vary other parameters (such as n or λ), which hinders interpretation of the plots (see Appendix G). Instead, we propose to observe a more stringent and stable quantity, namely the L_1 error on the estimated dual potential φ, which is the Lagrange multiplier associated to the first marginal constraint in Eq. (1). This dual potential is the gradient of $W_2^2(\mu,\nu)$ with respect to μ [52, Prop. 7.17], a quantity of high interest when training machine learning models with W_2^2 as a loss function.

Specifically, given $v^{(k)} \in \mathbb{R}^n$ obtained after k Sinkhorn's iterations with discrete marginals μ_n, ν_n as in Eq. (6), we define the function $\hat{u}_{\mu,\nu}(x) = -\lambda \log(\sum_{j=1}^n e^{(v_j^{(k)} - \frac{1}{2}\|x-y_j\|_2^2)/\lambda} q_j)$. The quantity we plot is $\int |\hat{\varphi}_{\lambda,n}(x) - \varphi(x)| \mathrm{d}\mu(x)$ estimated via Monte Carlo integration or on a fine grid, where $\hat{\varphi}_{\lambda,n}$ is defined as follows: (i) $\hat{\varphi}_{\lambda,n} = 2\hat{u}_{\mu,\nu}$ for the biased estimator $\hat{T}_{\lambda,n}$, (ii) $\hat{\varphi}_{\lambda,n} = 2\hat{u}_{\mu,\nu} - (\hat{u}_{\mu,\mu'} + \hat{v}_{\mu,\mu'})$ for the debiased estimator $\hat{S}_{\lambda,n}$ and (iii) $2\hat{\varphi}_{\lambda,n} - \hat{\varphi}_{\sqrt{2}\lambda,n}$ for the extrapolated estimator $\hat{R}_{\lambda,n}$.

Random sampling. Figure 2 shows the approximation error for the estimators T_λ, S_λ and R_λ in the random sampling setting. Here, $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ with $d = 5$ are smooth elliptically contoured distributions with compact support and are such that the optimal potential φ is quadratic and admits a closed-form, as well as the transport cost (see Appendix G). These properties guarantee that the conclusions of Proposition 4 apply. As expected, for a given λ, S_λ and R_λ have a much smaller bias than T_λ (left plot). Looking at the performance as a function of λ (middle plot), we see that the error is minimal for some λ^* that is much larger than what is needed for T_λ to achieve a comparable accuracy. Also, choosing the best λ^* for each n (right panel), we see that S_{λ^*} has the same rate as the plug-in estimator (estimated with T_λ with a small λ), with a better constant. We remark that R_λ does not converge faster, which does not contradict ours results since we have no guarantee on the specific quantity plotted here.

Overall, these estimators require less samples and a larger λ to achieve a given accuracy compared to T_λ, which leads to substantial computational gains. This is illustrated on Figure 3 where for a target L^1 error on the potential, we chose the largest λ and smallest n that achieve this error, with $\lambda \in [0.1, 1]$ and $n \in [10, 100000]$. We report the computational time using the Sinkhorn's iterations of Eq. (6) stopped when the ℓ_1-error on the marginals is below 10^{-5}. We observe that for small target accuracies, the estimators S_λ and R_λ compare favorably to T_λ. In practical settings, one does not know *a priori* the best choice for λ, but many machine learning tasks involving W_2^2 come with a performance criterion, in which case cross-validation can be used to select this parameter.

[2] The code to reproduce these experiments is available at this webpage `https://gitlab.com/proussillon/wasserstein-estimation-sinkhorn-divergence`.

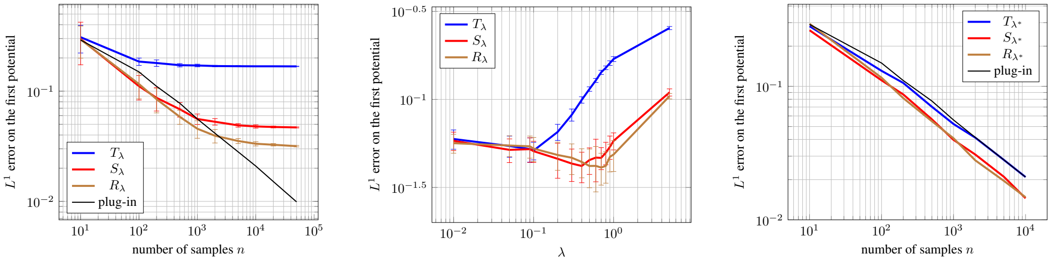

Figure 2: L^1 estimation error on the first potential for μ, ν smooth compactly supported distributions with $d = 5$. Left: as function of n for $\lambda = 1$. Middle: as a function of λ, for $n = 10000$. Right: as a function of n for the optimal $\lambda^*(n)$. Error bars show the standard deviation on 30 realizations.

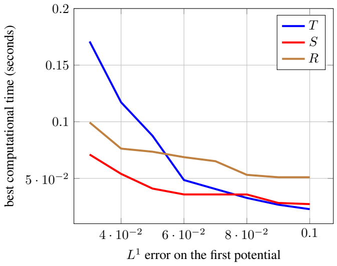

Figure 3: Best computational time achieved by the estimators to reach a given accuracy (after optimizing over n and λ), for μ, ν smooth compactly supported distributions with $d = 5$.

Discretization on grids. Figure 4 shows the evolution of the errors for densities (μ, ν) on the 1-D torus, the setting of Proposition 7. In this case, one can compute efficiently the dual potentials φ using cumulative functions [48]. This figure shows that, as expected, for a fixed (h, λ) the error of S_λ and R_λ is systematically lower than that of T_λ. Even when selecting the optimal regularization $\lambda^\star(h)$ for each h and for each method (which is a fair comparison), the error of S_λ and R_λ is still lower. Furthermore, the optimal parameter $\lambda^\star(h)$ is systematically larger for S_λ and R_λ. Additional figures showing visual comparisons of the potentials and their approximations are provided in the appendix.

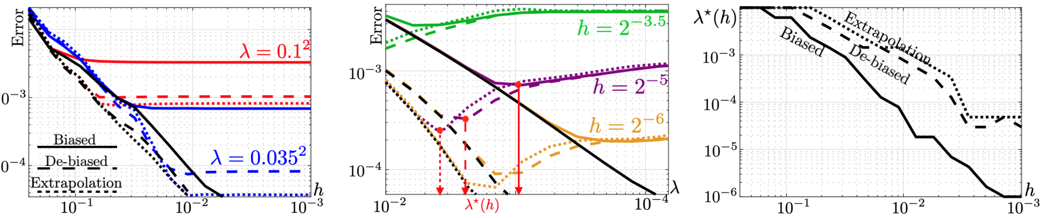

Figure 4: Left: L^1 error on the first potential φ as a function of the grid size h, for several value of λ. Middle: same error, displayed as a function of λ, for several grid sizes h. Right: evolution of the optimal regularization parameter $\lambda^\star(h)$ as a function of the grid size h.

6 Conclusion and open questions

In this paper we have exhibited the usefulness of entropic regularization with debiasing for the estimation of the squared Wasserstein distance: it may increase both accuracy and efficiency when the problem has a smooth nature. Numerical experiments suggest that the theory could be extended in several directions. First, the Sinkhorn divergence estimator appears at least as statistical efficient as the plug-in estimator, while our bound is slightly weaker. Also, the estimation of Kantorovich potentials seems to enjoy similar guarantees, but this is not covered by our theory.

Broader Impact

Broader impact statement does not apply for this paper, which is of theoretical nature.

Acknowledgments

The works of Pierre Roussillon, Flavien Léger and Gabriel Peyré is supported by the ERC grant NORIA and by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, Rafael Oliveira, and Avi Wigderson. Much faster algorithms for matrix scaling. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 890–901. IEEE, 2017.

[2] Jason Altschuler, Jonathan Niles-Weed, and Philippe Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974, 2017.

[3] Donald G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM (JACM)*, 12(4):547–560, 1965.

[4] Francis Bach. On the effectiveness of Richardson extrapolation in machine learning. *arXiv preprint arXiv:2002.02835*, 2020.

[5] Robert J. Berman. The Sinkhorn algorithm, parabolic optimal transport and geometric Monge–Ampère equations. *Numerische Mathematik*, 145(4):771–836, 2020.

[6] Rajendra Bhatia, Tanvi Jain, and Yongdo Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[7] Jose Blanchet, Arun Jambulapati, Carson Kent, and Aaron Sidford. Towards optimal running times for optimal transport. *arXiv preprint arXiv:1810.07717*, 2018.

[8] Emmanuel Boissard and Thibaut Le Gouic. On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'IHP Probabilités et statistiques*, volume 50, pages 539–563, 2014.

[9] Nicolas Bonneel, Michiel Van De Panne, Sylvain Paris, and Wolfgang Heidrich. Displacement interpolation using Lagrangian mass transport. In *Proceedings of the 2011 SIGGRAPH Asia Conference*, pages 1–12, 2011.

[10] E.M. Bronshtein. ε-entropy of convex sets and functions. *Siberian Mathematical Journal*, 17(3):393–398, 1976.

[11] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. Optimal steering of a linear stochastic system to a final probability distribution, part i. *IEEE Transactions on Automatic Control*, 61(5):1158–1169, 2015.

[12] Yongxin Chen, Tryphon T. Georgiou, and Michele Pavon. On the relation between optimal transport and Schrödinger bridges: A stochastic control viewpoint. *Journal of Optimization Theory and Applications*, 169(2):671–691, 2016.

[13] Michael B. Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *2017 IEEE 58th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 902–913. IEEE, 2017.

[14] Giovanni Conforti and Luca Tamanini. A formula for the time derivative of the entropic cost and applications. *arXiv preprint arXiv:1912.10555*, 2019.

[15] Dario Cordero-Erausquin. Sur le transport de mesures périodiques. *Comptes Rendus de l'Académie des Sciences-Series I-Mathematics*, 329(3):199–202, 1999.

[16] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, 2013.

[17] Guido De Philippis and Alessio Figalli. The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51(4):527–580, 2014.

[18] Jean Dolbeault, Bruno Nazaret, and Giuseppe Savaré. A new class of transport distances between measures. *Calculus of Variations and Partial Differential Equations*, 34(2):193–231, 2009.

[19] Richard Mansfield Dudley. The speed of mean Glivenko–Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

[20] Pavel Dvurechensky, Alexander Gasnikov, and Alexey Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. In *International Conference on Machine Learning*, pages 1367–1376, 2018.

[21] Sven Erlander and Neil F. Stewart. *The gravity model in transportation analysis: theory and extensions*, volume 3. Vsp, 1990.

[22] Kai-Tai Fang, Samuel Kotz, and Kai Wang Ng. *Symmetric Multivariate and Related Distributions*. London: Chapman and Hall, 2018.

[23] Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun-ichi Amari, Alain Trouvé, and Gabriel Peyré. Interpolating between optimal transport and MMD using Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2681–2690, 2019.

[24] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[25] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya, and Tomaso A. Poggio. Learning with a Wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015.

[26] Aude Genevay, Lénaïc Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583, 2019.

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

[28] Ivan Gentil, Christian Léonard, and Luigia Ripani. About the analogy between optimal transport and minimal entropy. In *Annales de la Faculté des Sciences de Toulouse: Mathématiques*, volume 26, pages 569–600, 2017.

[29] Ugo Gianazza, Giuseppe Savaré, and Giuseppe Toscani. The Wasserstein gradient flow of the Fisher information and the quantum drift-diffusion equation. *Archive for Rational Mechanics and Analysis*, 194(1):133–220, 2009.

[30] Nicola Gigli and Luca Tamanini. Benamou–Brenier and duality formulas for the entropic cost on $RCD^*(K, N)$ spaces. *Probability Theory and Related Fields*, 176(1-2):1–34, 2020.

[31] Joan Glaunes, Alain Trouvé, and Laurent Younes. Diffeomorphic matching of distributions: A new approach for unlabelled point-sets and sub-manifolds matching. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 2, pages II–II. IEEE, 2004.

[32] Adityanand Guntuboyina and Bodhisattva Sen. L_1 covering numbers for uniformly bounded convex functions. In *Conference on Learning Theory*, pages 12–1, 2012.

[33] Jan-Christian Hütter and Philippe Rigollet. Minimax rates of estimation for smooth optimal transport maps. *arXiv preprint arXiv:1905.05828*, 2019.

[34] Hicham Janati, Boris Muzellec, Gabriel Peyré, and Marco Cuturi. Entropic optimal transport between (unbalanced) Gaussian measures has a closed form. *arXiv preprint arXiv:2006.02572*, 2020.

[35] D.C. Joyce. Survey of extrapolation processes in numerical analysis. *Siam Review*, 13(4):435–490, 1971.

[36] Jeffrey J. Kosowsky and Alan L. Yuille. The invisible hand algorithm: Solving the assignment problem with statistical physics. *Neural Networks*, 7(3):477–490, 1994.

[37] Nathaniel Lahn, Deepika Mulchandani, and Sharath Raghvendra. A graph theoretic additive approximation of optimal transport. In *Advances in Neural Information Processing Systems*, pages 13813–13823, 2019.

[38] Anton Mallasto, Augusto Gerolin, and Hà Quang Minh. Entropy-regularized 2-Wasserstein distance between Gaussian measures. *arXiv preprint arXiv:2006.03416*, 2020.

[39] Tudor Manole, Sivaraman Balakrishnan, and Larry Wasserman. Minimax confidence intervals for the sliced Wasserstein distance. *arXiv preprint arXiv:1909.07862*, 2019.

[40] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. In *Advances in Neural Information Processing Systems*, pages 4543–4553, 2019.

[41] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[42] Kimia Nadjahi, Alain Durmus, Umut Simsekli, and Roland Badeau. Asymptotic guarantees for learning generative models with the Sliced-Wasserstein distance. In *Advances in Neural Information Processing Systems*, pages 250–260, 2019.

[43] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[44] Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

[45] Kaare Brandt Petersen and Michael Syskind Pedersen. The matrix cookbook, nov 2012. *URL http://www2. imm. dtu. dk/pubdb/p. php*, 3274, 2012.

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

[47] Kent Quanrud. Approximating optimal transport with linear programs. In *2nd Symposium on Simplicity in Algorithms (SOSA 2019)*, volume 69, pages 6:1–6:9, 2018.

[48] Julien Rabin, Julie Delon, and Yann Gousseau. Transportation distances on the circle. *Journal of Mathematical Imaging and Vision*, 41(1-2):147, 2011.

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

[50] Aaditya Ramdas, Nicolás García Trillos, and Marco Cuturi. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47, 2017.

[51] Lewis Fry Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A*, 210(459-470):307–357, 1911.

[52] Filippo Santambrogio. *Optimal Transport for Applied Mathematicians*. Springer, 2015.

[53] Damien Scieur, Alexandre d'Aspremont, and Francis Bach. Regularized nonlinear acceleration. In *Advances In Neural Information Processing Systems*, pages 712–720, 2016.

[54] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

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

[56] Carla Tameling, Max Sommerfeld, and Axel Munk. Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29(5):2744–2781, 2019.

[57] Cédric Villani. *Optimal transport: Old and New*, volume 338. Springer Science & Business Media, 2008.

[58] Martin J. Wainwright. *High-Dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

[59] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A):2620–2648, 2019.

[60] Jonathan Weed and Quentin Berthet. Estimation of smooth densities in Wasserstein distance. In *Conference on Learning Theory*, pages 3118–3119, 2019.

[61] Alan Geoffrey 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.