Faster Wasserstein Distance Estimation with the Sinkhorn Divergence

Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, Gabriel Peyré

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 :

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

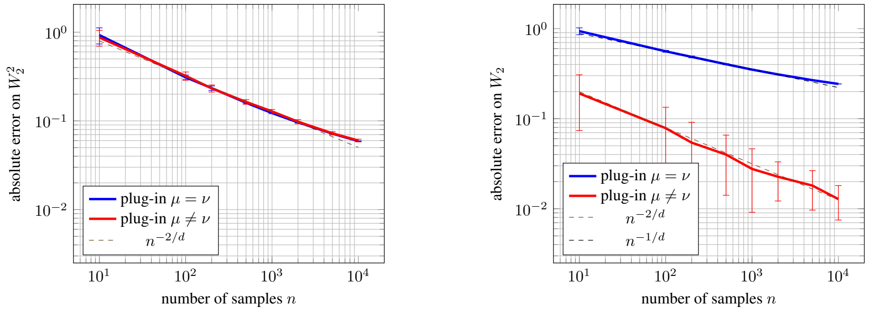

Figure 1: Estimation error of the plug-in estimator for μ, ν compactly supported with $d = 5$ (as detailed in Appendix G). Left: error on the cost W_2^2 has rate $n^{-2/d}$ (Theorem 2). Right: error on W_2 has rate $n^{-1/d}$ if $\mu = \nu$ and $n^{-2/d}$ if $\mu \neq \nu$ (Corollary 1) with $\mathbf{E}_\mu[x]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

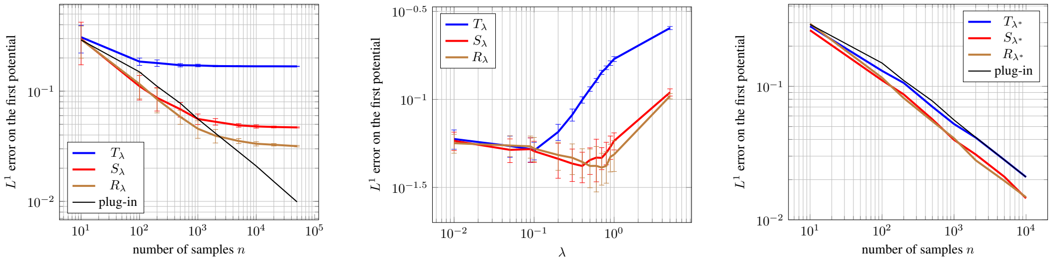

Figure 2: L^1 estimation error on the first potential for μ, ν smooth compactly supported distributions with $d = 5$. Left: as function of n for $\lambda = 1$. Middle: as a function of λ, for $n = 10000$. Right: as a function of n for the optimal $\lambda^*(n)$. Error bars show the standard deviation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

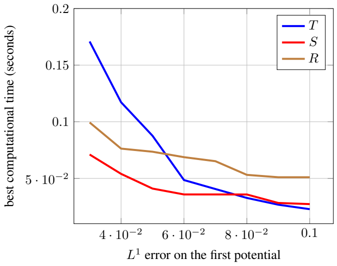

Figure 3: Best computational time achieved by the estimators to reach a given accuracy (after optimizing over n and λ), for μ, ν smooth compactly supported distributions with $d = 5$.

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.

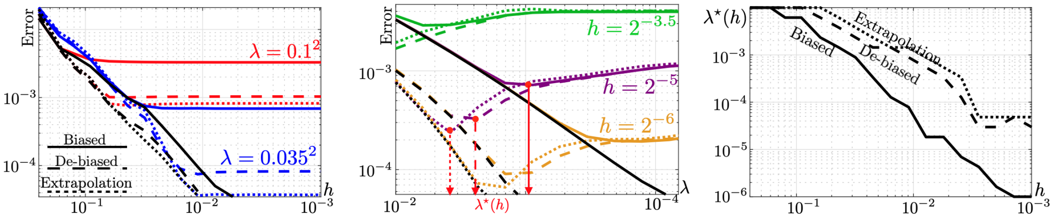

Figure 4: Left: L^1 error on the first potential φ as a function of the grid size h, for several value of λ. Middle: same error, displayed as a function of λ, for several grid sizes h. Right: evolution of the optimal regularization parameter $\lambda^\star(h)$ as a function of the grid size h.

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.

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.

Supplementary Material

Supplementary material for the paper: "Faster Wasserstein Distance Estimation with the Sinkhorn Divergence" authored by Lénaïc Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard and Gabriel Peyré (NeurIPS 2020). This supplementary material is organized as follows:

- Appendix A contains the proofs of Section 2,
- Appendix B recaps the convergence analysis of [20] to obtain Proposition 2,
- Appendix C contains the proofs of Section 3.1,
- Appendix D contains the proofs of Section 3.2,
- Appendix E contains the proofs of Section 3.3,
- in Appendix F, we derive the Taylor expansion for Gaussian distributions presented in Section 4,
- finally, Appendix G contains details on the settings of the numerical experiments and additional figures.

A Bounds on the approximation error

Dynamic entropy regularized optimal transport. Let us first justify how to obtain Eq. (4) since our conventions are slightly different than in [14]. In that reference, for μ and ν absolutely continuous with compact support, the authors define

$$\lambda C_\lambda(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \lambda H(\gamma, K)$$

where $K = (2\pi\lambda)^{-d/2}\exp(-\|y-x\|_2^2/(2\lambda))\mathrm{d}x\mathrm{d}y$ is the heat kernel at time $\lambda/2$. In contrast, we can see from Eq. (1) that

$$\frac{1}{2}T_\lambda(\mu,\nu) = \min_{\gamma\in\Pi(\mu,\nu)} \lambda H(\gamma, \tilde{K})$$

where $\tilde{K} = \exp(-\|y-x\|_2^2/(2\lambda))\mu(x)\nu(y)\mathrm{d}x\mathrm{d}y$. We directly deduce that $\frac{1}{2}T_\lambda(\mu,\nu) = \lambda C_\lambda(\mu,\nu) - \lambda H(\mu) - \lambda H(\nu) - \frac{d\lambda}{2}\log(2\pi\lambda)$. Thus Eq. (4) follows by the dynamic formulation of entropy regularized optimal transport in [14] which reads

$$\lambda C_\lambda(\mu,\nu) - \frac{\lambda}{2}H(\mu) - \frac{\lambda}{2}H(\nu) = \min_{\rho,v}\int_0^1\int_{\mathbb{R}^d}\big(\frac{1}{2}\|v(t,x)\|_2^2 + \frac{\lambda^2}{8}\|\nabla_x \log\rho(t,x)\|_2^2\big)\rho(t,x)\mathrm{d}x\mathrm{d}t$$

where the constraints on (ρ, v) are as in Eq. (4). Note that $\nabla_x\log\rho$ refers to the *weak logarithmic gradient* of ρ, which in particular does not requires $\rho > 0$ to be well defined, but only that for almost every $t\in[0,1]$, $\rho(t,\cdot)$ admits a distributional gradient which is an absolutely continuous measure with respect to $\rho(t,\cdot)$, and $\nabla_x\log\rho_t \overset{\text{def.}}{=} \frac{\mathrm{d}\nabla\rho_t}{\mathrm{d}\rho_t}$ refers to its density with respect to ρ_t (see e.g. [29]).

First order expansion. Let us state and prove a lemma that intervenes in the proof of Theorem 1. Arguments towards this expansion appeared in [14, Theorem 1.6] but under an abstract twice-differentiability assumption that is not needed in our statement.

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

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}J_\sigma|_{\sigma=0_+} = \frac{1}{4}I_0(\mu,\nu)$$

where, as in the proof of Theorem 1, $J_{\lambda^2}(\mu,\nu)$ refers to the right-hand side of (4). *More generally, the right derivative of $\sigma\mapsto J_\sigma$ exists for all $\sigma\geq 0$ and equals $\frac{1}{4}I_{\sqrt{\sigma}}(\mu,\nu)$.*

Proof. Since $\sigma\mapsto J_\sigma(\mu,\nu)$ is defined as an infimum of affine functions in σ, it is concave. Let $(\sigma_n)_{n\in\mathbb{N}}$ be a decreasing sequence of positive real numbers converging to 0 and let

$$\alpha_n = \frac{J_{\sigma_n} - J_0}{\sigma_n}.$$

By concavity, α_n is non-decreasing and admits a limit $J_0' = \frac{\mathrm{d}}{\mathrm{d}\sigma} J_\sigma|_{\sigma=0_+}$ that is the right derivative of J at 0. Our goal is to show that $J_0' = I_0(\mu,\nu)/4$. By the argument in the proof of Theorem (1), we have $\alpha_n \leq I_0(\mu,\nu)/4\ \forall n$ thus $J_0' \leq I_0(\mu,\nu)/4$, so we just have to prove the other inequality.

Let $(\rho_n, v_n)_{n\geq 0}$ be a sequence of minimizers for the r.h.s. of Eq. (4) (which is in fact unique although we do not use that fact here [30]) with $\lambda^2 = \sigma_n$ and let $V_n = \int_0^1 \int_{\mathbb{R}^d} \|v_n\|_2^2 \mathrm{d}\rho_n$ and $I_n = \int_0^1 \int_{\mathbb{R}^d} \|\nabla \log(\rho_n)\|_2^2 \mathrm{d}\rho_n$. Since V_n is uniformly bounded and converges to $V_0 = W_2^2(\mu,\nu)$, we have that ρ_n converges weakly (in duality with continuous functions with compact support) to ρ_0, the unique constant speed Wasserstein geodesic between μ and ν (see, e.g. [18, Cor. 4.10] or by an application of [29, Proposition 2.2] as below). Moreover, since $V_n \geq V_0$, it holds

$$\alpha_n = \frac{V_n - V_0}{\sigma_n} + \frac{1}{4} I_n \geq \frac{1}{4} I_n$$

and in particular we have the uniform bound $I_n \leq I_0$. It follows by [29, Proposition 2.2] applied to the quantity $I_n = \int_0^1 \int_{\mathbb{R}^d} \|\frac{\mathrm{d}(\nabla\rho_n)}{\mathrm{d}\rho_n}\|_2^2 \mathrm{d}\rho_n(x,t)$ that $\nabla\rho_n$, seen as a vector valued measure on $[0,1]\times\mathbb{R}^d$, admits a weak limit denoted ω which is absolutely continuous with respect to ρ_0 and that $\liminf I_n \geq \int_0^1 \int_{\mathbb{R}^d} \|\frac{\mathrm{d}\omega}{\mathrm{d}\rho_0}\|_2^2 \mathrm{d}\rho_0(t,x)$. Since for any compactly supported function $\varphi \in \mathcal{C}^1([0,1]\times\mathbb{R}^d;\mathbb{R}^d)$ it holds $\int \mathrm{div}_x(\varphi)\mathrm{d}\rho_n \to \int \mathrm{div}_x(\varphi)\mathrm{d}\rho_0$ and $\int \varphi\cdot \mathrm{d}(\nabla\rho_n) \to \int \varphi \cdot \mathrm{d}\omega$, we have that $\omega = \nabla_x \rho_0$ and thus the previous integral is precisely the Fisher information of ρ_0 integrated in time. It follows that $\liminf I_n \geq I_0$ hence $J_0' \geq \frac{1}{4} I_0$ which concludes the proof. Inspecting the above argument, we see that in fact it applies directly to the case $\sigma > 0$ (except that of course the trajectory recovered as $n\to\infty$ is $\rho_{\sqrt{\sigma}}$), hence our second claim. □

Bounds on the Fisher information of the geodesic. Let us now prove the bounds on the Fisher information of the Wasserstein geodesic that appear in Proposition 1. The main idea is to express $I_0(\mu,\nu)$ in terms of the initial and final densities and the Brenier potential.

Proof of Proposition 1. Let us express $I_0(\mu,\nu)$ in terms of the densities ρ_0 and ρ_1 (of μ and ν respectively) and the Brenier potential φ which is the convex function such that $(\nabla\varphi)_\#\mu = \nu$, i.e. ν is the pushforward of μ by the map $\nabla\varphi$. Let $(\rho_t)_{t\in[0,1]}$ be the density of the W_2-geodesic between μ and ν. We start with the conservation of mass formula which holds under our regularity assumptions:

$$\rho_0(x) = \det(\nabla^2\varphi_t(x))\rho_t(\nabla\varphi_t(x)))$$

where $\varphi_t(x) \overset{\text{def.}}{=} (1-t)\|x\|_2^2/2 + t\varphi(x)$ is such that $(\nabla\varphi_t)_\#\rho_0 = \rho_t$. By taking the logarithm we get

$$\log\rho_0(x) = \log\rho_t(\nabla\varphi_t(x)) + \log\det(\nabla^2\varphi_t(x)).$$

Let us now take the gradient of this expression. We denote by $d^3\varphi(x) : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ the weak differential of $x \mapsto \nabla^2\varphi(x)$ (which exists for almost every x and is bounded since $\nabla^2\varphi$ is assumed Lipschitz) and by $[d^3\varphi(x)]^* : \mathbb{R}^{d\times d} \to \mathbb{R}^d$ its adjoint. Using the fact that the differential of $A \mapsto \log\det A$ at A is the scalar product with A^{-1} we get that for almost every $x\in\mathbb{R}^d$,

$$\nabla\log\rho_0(x) = \nabla^2\varphi_t(x)\nabla\log\rho_t(\nabla\varphi_t(x)) + [d^3\varphi_t(x)]^*[\nabla^2\varphi_t(x)]^{-1}. \tag{7}$$

It follows that

$$\begin{aligned}
I_0(\mu,\nu) &= \int_0^1 \int_{\mathbb{R}^d} \|\nabla\log\rho_t(x)\|_2^2 \rho_t(x)\mathrm{d}x\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d} \|\nabla\log\rho_t(\nabla\varphi_t(x))\|_2^2 \rho_0(x)\mathrm{d}x\mathrm{d}t \\
&= \int_0^1 \int_{\mathbb{R}^d} \|[\nabla^2\varphi_t(x)]^{-1}\nabla\log\rho_0(x) - t[\nabla^2\varphi_t(x)]^{-1}[d^3\varphi(x)]^*[\nabla^2\varphi_t(x)]^{-1}\|_2^2 \rho_0(x)\mathrm{d}x\mathrm{d}t
\end{aligned}$$

where we have used the fact that $d^3\varphi_t(x) = td^3\varphi(x)$.

General case. In the general case, we simply use the bounds $\nabla^2\varphi_t(x) \succeq ((1-t)+t\kappa)\mathrm{Id}$ and $\|d^3\varphi(x)\| \leq L$ almost everywhere in operator norm and the identity $|a+b|^2 \leq 2|a|^2 + 2|b|^2$ valid for any $a, b \in \mathbb{R}$ to get

$$\begin{aligned} I_0(\mu,\nu) &\leq 2\Big(\int_0^1 \frac{\mathrm{d}t}{(1+(\kappa-1))^2}\Big) I_0(\mu,\mu) + 2\Big(\int_0^1 \frac{t^2\mathrm{d}t}{(1+(\kappa-1))^4}\Big) L^2 \\ &= 2\kappa^{-1} I_0(\mu,\mu) + (2/3)\kappa^{-3}L^2. \end{aligned}$$

One dimensional case. When $d=1$, from Eq. (7) at time $t=1$, we get

$$\varphi'''(x) = \nabla \log \rho_0(x)\varphi''(x) - \nabla\rho_1(\nabla\varphi(x))\varphi''(x)^2.$$

Plugging this expression in the previous integral leads to:

$$I_0(\mu,\nu) = \int_{\mathbb{R}}\int_0^1 \left(\frac{(1-t)\nabla\log\rho_0 + t\nabla\log\rho_1(\nabla\varphi(x))\varphi''(x)^2}{\big((1-t)+t\varphi''(x)\big)^2}\right)^2 \mathrm{d}t\rho_0(x)\mathrm{d}x$$

With the valid change of variables $1-s = \frac{t\varphi''(x)}{(1-t)+t\varphi''(x)}$ (and thus $s = \frac{1-t}{(1-t)+t\varphi''(x)}$), we obtain:

$$\begin{aligned} I_0(\mu,\nu) &= \int_{\mathbb{R}}\int_0^1 \frac{1}{\varphi''(x)}\Big((1-s)\nabla\log\rho_0(x) + s\nabla\log\rho_1(\varphi'(x))\varphi''(x)\Big)^2 \mathrm{d}s\rho_0(x)\mathrm{d}x \\ &= \int_{\mathbb{R}}\int_0^1 \frac{1}{\varphi''(x)}\Big((1-s)\nabla\log\rho_0(x) + s\nabla\log(\rho_1\circ\varphi')(x)\Big)^2 \mathrm{d}s\rho_0(x)\mathrm{d}x \end{aligned}$$

This leads to the bound $I_0(\mu,\nu) \leq (2/3)\kappa^{-1}I_0(\mu,\mu) + (2/3)KI_0(\nu,\nu)$ since $(\varphi')_\#\mu = \nu$. □

Gaussian case. Let us now give the explicit expression of the Fisher information of the Wasserstein geodesic between Gaussian distributions, which is mentioned in Section 4. Whenever we deal with a positive semidefinite matrix A, the matrix $A^{1/2}$ refers to its unique positive semidefinite square root.

Proposition 9. *If $\mu = \mathcal{N}(0,A)$, $\nu = \mathcal{N}(0,B)$ then $I_0(\mu,\nu) = \operatorname{tr} S^{-1}$ with $S = (A^{1/2}BA^{1/2})^{1/2}$.*

Remark in particular that the expansion in Theorem 1 then gives

$$S_\lambda(\mu,\nu) - W_2^2(\mu,\nu) = \frac{1}{8}(2I_0(\mu,\nu) - I_0(\mu,\mu) - I_0(\nu,\nu)) = \frac{1}{8}(2\operatorname{tr}S^{-1} - \operatorname{tr}A^{-1} - \operatorname{tr}B^{-1})$$

which is consistent, as it must, with the expansion in Section 4.

Proof. When the Brenier potential $\varphi = \frac{1}{2}x^\top Hx$ is quadratic, we have by the proof of Proposition 1

$$I_0(\mu,\nu) = \int_{\mathbb{R}^d}\int_0^1 \|[\nabla^2\varphi_t(x)]^{-1}\nabla\log\rho_0(x)\|_2^2\rho_0(x)\mathrm{d}t\mathrm{d}x.$$

Putting ourselves in a basis diagonalizing H, the integration in time is explicit and we get

$$I_0(\mu,\nu) = \int_{\mathbb{R}^d} \|H^{-1/2}\nabla\log\rho_0(x)\|_2^2\rho_0(x)\mathrm{d}x.$$

It turns out that if $\mu = \mathcal{N}(0,A)$, $\nu = \mathcal{N}(0,B)$, then $\varphi(x) = \frac{1}{2}x^THx$ where [6]

$$H = A^{-1/2}\Big(A^{1/2}BA^{1/2}\Big)^{1/2}A^{-1/2}$$

and thus

$$\begin{aligned} I_0(\mu,\nu) &= \int_{\mathbb{R}^d} \|H^{-1/2}\nabla\log\rho_0(x)\|_2^2\rho_0(x)\mathrm{d}x. \\ &= \int_{\mathbb{R}^d} \|H^{-1/2}A^{-1}x\|_2^2\rho_0(x)\mathrm{d}x \\ &= \mathbf{E}_{X\sim\mathcal{N}(0,A)}\Big[X^T\big(A^{-1}H^{-1}A^{-1}\big)X\Big] \\ &= \operatorname{tr}\big(A^{-1}H^{-1}A^{-1}A\big) = \operatorname{tr}\big(A^{-1}H^{-1}\big) = \operatorname{tr}\big((A^{1/2}BA^{1/2})^{-1/2}\big) \end{aligned}$$

where the last row is obtained using [45, Eq. (378)]. □

B Computational complexity of Sinkhorn's algorithm

In this appendix, we recall the computational complexity analysis of Sinkhorn's algorithm from [20], in order to state Proposition 11 exactly as per our needs (while this result is implicit in [20]). There is nothing specific in this analysis about the squared-distance cost so we just assume that the cost $c : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is continuous, keeping in mind that in our case, $c(x,y) = \frac{1}{2}\|y-x\|_2^2$. We also consider a compact set $\mathcal{X} \subset \mathbb{R}^d$ and measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$ which are concentrated on this set. We consider the dual objective function of entropy regularized optimal transport [46]:

$$F_\lambda(u,v) = \int_{\mathbb{R}^d} u \mathrm{d}\mu + \int_{\mathbb{R}^d} v \mathrm{d}\nu + \lambda\Big(1 - \int_{(\mathbb{R}^d)^2} \exp((u(x)+v(y)-c(x,y))/\lambda)\mathrm{d}\mu(x)\mathrm{d}\nu(y)\Big). \tag{8}$$

By Fenchel duality, we have with $c(x,y) = \frac{1}{2}\|y-x\|_2^2$ that

$$\frac{1}{2}T_\lambda(\mu,\nu) = \max_{u,v} F_\lambda(u,v) \tag{9}$$

where the maximum is over pairs of continuous real-valued functions on $\mathbb{R}^d$, $(u,v) \in \mathcal{C}(\mathcal{X})^2$. Sinkhorn's algorithm is alternate maximization on u and v: it starts with $u_0, v_0 = 0$ and defines,

$$u_{k+1} = u_k - \lambda \log \int_{\mathbb{R}^d} \exp((u_k(\cdot) + v_k(y) - c(\cdot,y))/\lambda)\mathrm{d}\nu(y), \quad v_{k+1} = v_k \quad \text{if } k \text{ is odd}$$
$$v_{k+1} = v_k - \lambda \log \int_{\mathbb{R}^d} \exp((u_k(x) + v_k(\cdot) - c(x,\cdot))/\lambda)\mathrm{d}\mu(x), \quad u_{k+1} = u_k \quad \text{if } k \text{ is even.}$$

This form of the iterations that distinguishes between even and odd updates is convenient for the analysis, but beware that the index k here is twice the index appearing in Proposition 2, so the statements are adjusted consequently. We also introduce $\gamma_k = \exp((u_k(x)+v_k(y)-c(x,y))/\lambda)\mu\otimes\nu$, which is such that the update can be written: $u_{k+1} = u_k + \lambda \log(\mathrm{d}\mu/\mathrm{d}\pi^1_\#\gamma_k)$ if k odd and $v_{k+1} = v_k + \lambda\log(\mathrm{d}\nu/\mathrm{d}\pi^2_\#\gamma_k)$ if k even, where $\pi^1_\#\gamma$ is the marginal of γ on the first factor of $\mathbb{R}^d \times \mathbb{R}^d$ and $\pi^2_\#\gamma$ its marginal on the second. The following is a rearrangement of some intermediate results in [20] in a simplified form which is sufficient to our purpose.

Proposition 10. *Assume $c \geq 0$ and let $\|c\|_\infty = \sup_{(x,y)\in\mathcal{X}^2} c(x,y)$. Sinkhorn's iterates satisfy, for $k \geq 1$,*

$$0 \leq \max_{u,v} F_\lambda(u,v) - F_\lambda(u_k,v_k) \leq \frac{2\|c\|_\infty^2}{\lambda k}$$

Proof. First, remark that the iterations are such that $\int \mathrm{d}\gamma_k = 1$ for $k \geq 1$, so it holds $F_\lambda(u,v) = \int u\mathrm{d}\mu + \int v \mathrm{d}\nu$ for $(u,v) = (u_k,v_k)$ and also for any maximizer $(u,v) = (u^*,v^*)$. The key of the proof is the following equality first noticed by [2]. If k is odd, then

$$\begin{aligned} F_\lambda(u_{k+1},v_{k+1}) - F_\lambda(u_k,v_k) &= -\lambda \int \log\Big(\int \exp((u_k(x)+v_k(y)-c(x,y))/\lambda)\mathrm{d}\nu(y)\Big)\mathrm{d}\mu(x) \\ &= \lambda \int \log(\mathrm{d}\mu/\mathrm{d}\pi^1_\#\gamma_k)\mathrm{d}\mu = \lambda H(\mu, \pi^1_\#\gamma_k). \end{aligned}$$

Let us define $\Delta_k = F_\lambda(u^*,v^*) - F_\lambda(u_k,v_k) \geq 0$. Using Pinsker's inequality and Lemma 2 it follows

$$\Delta_k - \Delta_{k+1} \geq \frac{\lambda}{2}\|\mu - \pi^1_\#\gamma_k\|_1^2 \geq \frac{\lambda}{2\|c\|_\infty^2}\Delta_k^2.$$

We can similarly prove the same inequality for k even. We conclude as in the usual proof of gradient descent for smooth functions [43, Thm. 2.1.14]: by dividing by $\Delta_k\Delta_{k+1}$ we have

$$\frac{1}{\Delta_{k+1}} - \frac{1}{\Delta_k} \geq \frac{\lambda}{2\|c\|_\infty^2}\frac{\Delta_k}{\Delta_{k+1}} \geq \frac{\lambda}{2\|c\|_\infty^2}.$$

Summing these inequalities yields a telescopic sum and we get $1/\Delta_k \geq \lambda k/(2\|c\|_\infty^2)$ which allows to conclude. □

From this analysis, we deduce the following complexity to approximate T_λ and T_0 using Sinkhorn's iterations, adapted from [20].

Proposition 11. *Assume that $\mu_n = \sum_{i=1}^n p_i\delta_{x_i}$ and $\nu_n = \sum_{j=1}^n q_j\delta_{y_j}$ are discrete measures with n atoms such that $p_i, q_j \geq \alpha/n$ for some $\alpha > 0$. Then Sinkhorn's algorithm returns an ε-accurate estimation of $T_\lambda(\mu,\nu)$ in time $O(n^2\|c\|_\infty^2/(\lambda\varepsilon))$. Moreover, fixing $\lambda = \varepsilon/4(\log(n)+\log(1/\alpha))$, it returns an ε-accurate estimation of $T_0(\mu,\nu)$ in $O(n^2\log(n)\|c\|_\infty^2/\varepsilon^2)$ operations.*

Proof. The first claim is a direct consequence of Proposition 10 since when μ and ν have a finite support of size n, an iteration of Sinkhorn can be performed with $O(n^2)$ operations. The second claim follows from the bound

$$0 \leq T_\lambda(\mu,\nu) - T_0(\mu,\nu) \leq 2\lambda H(\gamma^*, \mu\otimes\nu) \leq 4\lambda(\log n + \log(1/\alpha))$$

where γ^* is the optimal transport plan for T_0. □

Lemma 2. *Under the assumptions and notations of Proposition 10 it holds*

$$\Delta_k \leq \|c\|_\infty\Big(\|\mu - \pi^1_\#\gamma_k\|_1 + \|\nu - \pi^2_\#\gamma_k\|_1\Big)$$

where $\|\mu\|_1 \overset{\text{def.}}{=} \sup_{\|u\|_\infty\leq 1}\int u(x)\mathrm{d}\mu(x)$ denotes the total variation norm in the space of measures.

Proof. Remark that F_λ is differentiable in (u,v) with gradient $(\mu - \pi^1_\#\gamma_k, \nu - \pi^2_\#\gamma_k)$ at (u_k, v_k). The concavity inequality then gives

$$\Delta_k \leq \int (u^* - u_k)\mathrm{d}(\mu - \pi^1_\#\gamma_k) + \int (v^* - v_k)\mathrm{d}(\nu - \pi^2_\#\gamma_k).$$

Also, for any $u \in \mathcal{C}(\mathcal{X})$ and $\alpha = (\max u + \min u)/2$, using the fact that $\int \mu = \int \pi^1_\#\gamma_k$, we have

$$\int u\mathrm{d}(\mu - \pi^1_\#\gamma_k) = \int (u-\alpha)\mathrm{d}(\mu - \pi^1_\#\gamma_k) \leq \frac{1}{2}(\max u - \min u)\|\mu - \pi^1_\#\gamma_k\|_1.$$

Finally, for $u = u^*$ or $u = u_k$ for $k \geq 1$, we have, for some $v \in \mathcal{C}(\mathcal{X})$, and for all $x, x' \in \mathcal{X}$

$$u(x) = -\lambda\log\int \exp((v(y) - c(x,y))/\lambda)\mathrm{d}\nu(y) \leq \|c\|_\infty - u(x')$$

because $c(x,y) \leq c(x',y) + \|c\|_\infty$. Thus $(\max u - \min u)/2 \leq \|c\|_\infty/2$. The conclusion follows by bounding all terms this way. □

C Properties of the plug-in estimator

In this section we prove Theorem 2 about the rate of convergence of $T_0(\hat{\mu}_n, \hat{\nu}_n)$ to $T_0(\mu,\nu)$ (we recall that, by definition $W_2^2(\mu,\nu) = T_0(\mu,\nu)$). We start with the following lemma which bounds the estimation error by simpler quantities. Note that we consider measures on the centered ball of radius R in $\mathbb{R}^d$, for some $R > 0$, which is without loss of generality compared to other bounded sets since $T_\lambda(\mu,\nu)$ is invariant by translation of both measures. In the following lemma $\mu_n, \nu_n \in \mathcal{P}(\mathbb{R}^d)$ can be unrelated to μ, ν but this lemma will later be applied to the case where μ_n, ν_n are empirical distributions of n samples, hence our choice of notation.

Lemma 3. *Let $\mu, \nu, \mu_n, \nu_n \in \mathcal{P}(\mathbb{R}^d)$ be concentrated on the centered ball of radius R. Then it holds*

$$\begin{aligned}\Big|\frac{1}{2}T_0(\mu,\nu) - \frac{1}{2}T_0(\mu_n,\nu_n)\Big| &\leq \Big|\frac{1}{2}\int\|x\|_2^2\mathrm{d}(\mu-\mu_n)(x)\Big| + \Big|\frac{1}{2}\int\|x\|_2^2\mathrm{d}(\nu-\nu_n)(x)\Big| \\ &\quad + \sup_{\varphi\in\mathcal{F}_R}\Big|\int\varphi\mathrm{d}(\mu_n-\mu)\Big| + \sup_{\varphi\in\mathcal{F}_R}\Big|\int\varphi\mathrm{d}(\nu_n-\nu)\Big|\end{aligned}$$

where $\mathcal{F}_R$ is the set of convex and R-Lipschitz functions on the ball of radius R.

Proof. The first part of the proof is fairly classical. By Kantorovich duality, we have

$$\frac{1}{2}T_0(\mu,\nu) = \max_{u,v\in\mathcal{C}(\mathcal{X})} \int u \mathrm{d}\mu + \int v \mathrm{d}\nu$$

where $\mathcal{X}$ is the closed ball of radius R and under the constraint that $u(x)+v(y) \leq \frac{1}{2}\|y-x\|_2^2$ for all $(x,y)\in\mathcal{X}^2$ and there exists a maximizer [52]. By expanding the square and changing the unknown $(\varphi,\psi) = (\frac{1}{2}\|\cdot\|_2^2 - u, \frac{1}{2}\|\cdot\|_2^2 - v)$, we can equivalently write

$$\frac{1}{2}T_0(\mu,\nu) = \frac{1}{2}\int \|x\|_2^2 \mathrm{d}\mu(x) + \frac{1}{2}\int \|x\|_2^2 \mathrm{d}\nu(x) - \min_{\varphi,\psi\in\mathcal{C}(\mathcal{X})} \Big(\int \varphi \mathrm{d}\mu + \int \psi \mathrm{d}\nu\Big)$$

under the constraint that $\varphi(x)+\psi(y)\geq \langle x,y\rangle$ for all $(x,y)\in\mathcal{X}^2$. In the minimization problem, fix an arbitrary $\psi\in\mathcal{C}(\mathcal{X})$ and notice that the value of the objective cannot increase if we replace φ by ψ^* defined by $\psi^*(x) = \max_{y\in\mathcal{X}}\langle x,y\rangle - \psi(y)$ and the couple (ψ^*,ψ) still satisfies the constraint. Repeating this process by now fixing ψ^*, we find that the couple (ψ^*,ψ^{**}) satisfies the constraint and has a smaller objective value. Now, as a supremum of affine functions, ψ^* is convex. For any $y_0\in\mathcal{X}$, let x_0 be such that $\psi^*(y_0) = \langle x_0,y_0\rangle - \psi(x_0)$, and observe that for all $y\in\mathcal{X}$

$$\begin{cases}\psi^*(y_0) = \langle x_0,y_0\rangle - \psi(y_0)\\ \psi^*(y) \geq \langle x_0,y\rangle - \psi(y)\end{cases} \quad\Rightarrow\quad \psi^*(y_0)-\psi^*(y) \leq \langle x_0, y_0-y\rangle \leq R\|y_0-y\|_2.$$

Since y_0 and y are arbitrary, this shows that ψ^* is R-Lipschitz, i.e., $|\psi^*(y)-\psi^*(y')|\leq R\|y-y'\|_2$ for all $(y,y')\in\mathcal{X}^2$. We thus have

$$\frac{1}{2}T_0(\mu,\nu) = \frac{1}{2}\int \|x\|_2^2 \mathrm{d}\mu(x) + \frac{1}{2}\int \|x\|_2^2 \mathrm{d}\nu(x) - \min_{\varphi\in\mathcal{F}_\mathcal{R}} \Big(\int \varphi \mathrm{d}\mu + \int \varphi^* \mathrm{d}\nu\Big).$$

The rest of the proof is inspired by [40, Prop. 2] (which analyzes the sample complexity of T_λ for $\lambda>0$). Let us denote $\mathcal{S}_{\mu,\nu}(\varphi) \overset{\text{def.}}{=} \int\varphi\mathrm{d}\mu + \int\varphi^*\mathrm{d}\nu$ and $\varphi_{\mu,\nu}$ the minimizer of $\mathcal{S}_{\mu,\nu}$ over $\mathcal{F}_R$. By optimality, we have

$$\mathcal{S}_{\mu,\nu}(\varphi_{\mu_n,\nu}) - \mathcal{S}_{\mu_n,\nu}(\varphi_{\mu_n,\nu}) \leq \mathcal{S}_{\mu,\nu}(\varphi_{\mu,\nu}) - \mathcal{S}_{\mu_n,\nu}(\varphi_{\mu_n,\nu}) \leq \mathcal{S}_{\mu,\nu}(\varphi_{\mu,\nu}) - \mathcal{S}_{\mu_n,\nu}(\varphi_{\mu,\nu}).$$

It follows that

$$\begin{aligned}|\mathcal{S}_{\mu,\nu}(\varphi_{\mu,\nu}) - \mathcal{S}_{\mu_n,\nu}(\varphi_{\mu_n,\nu})| &\leq \max\Big\{|\mathcal{S}_{\mu,\nu}(\varphi_{\mu_n,\nu}) - \mathcal{S}_{\mu_n,\nu}(\varphi_{\mu_n,\nu})|, |\mathcal{S}_{\mu,\nu}(\varphi_{\mu,\nu}) - \mathcal{S}_{\mu_n,\nu}(\varphi_{\mu,\nu})|\Big\}\\ &\leq \sup_{\varphi\in\mathcal{F}_R} |\mathcal{S}_{\mu,\nu}(\varphi) - \mathcal{S}_{\mu_n,\nu}(\varphi)| = \sup_{\varphi\in\mathcal{F}_R}\Big|\int\varphi\mathrm{d}(\mu_n-\mu)\Big|.\end{aligned}$$

As a consequence, we have

$$\Big|\frac{1}{2}T_0(\mu,\nu) - \frac{1}{2}T_0(\mu_n,\nu)\Big| \leq \Big|\frac{1}{2}\int\|x\|_2^2\mathrm{d}(\mu-\mu_n)(x)\Big| + \sup_{\varphi\in\mathcal{F}_R}\Big|\int\varphi\mathrm{d}(\mu_n-\mu)\Big|.$$

We finally conclude with the triangle inequality

$$|T_0(\mu,\nu) - T_0(\mu_n,\nu_n)| \leq |T_0(\mu,\nu) - T_0(\mu_n,\nu)| + |T_0(\mu_n,\nu) - T_0(\mu_n,\nu_n)|$$

and by bounding the second term in the same fashion. □

The next technical step is to bound the supremum of an empirical process that appears in the bound of Lemma 3.

Lemma 4. *Let $\mu\in\mathcal{P}(\mathbb{R}^d)$ be concentrated on the ball of radius R and $\hat{\mu}_n$ an empirical distribution of n independent samples. Then it holds*

$$\mathbf{E}\Big[\sup_{\varphi\in\mathcal{F}_R}\Big|\int\varphi\mathrm{d}(\hat{\mu}_n-\mu)\Big|\Big] \lesssim \begin{cases} R^2 n^{-1/2} & \text{if } d<4,\\ R^2 n^{-1/2}\log(n) & \text{if } d=4,\\ R^2 n^{-2/d} & \text{if } d>4\end{cases}$$

where the notation $\lesssim$ hides a constant depending only on d and $\mathcal{F}_R$ is defined in Lemma 3.

Proof. First notice that we can include in the definition of $\mathcal{F}_R$ the property that $\varphi(0) = 0$ without changing the supremum. With this additional property, we in particular have that $\|\varphi\|_\infty \leq R^2$ for all $\varphi \in \mathcal{F}_R$. By a classical symmetrization argument [58, Thm. 4.10], we have

$$\mathbf{E}\Big[\sup_{\varphi\in\mathcal{F}_R}\Big|\int \varphi \mathrm{d}(\hat{\mu}_n-\mu)\Big|\Big] \leq 2\underbrace{\mathbf{E}_{\sigma,X}\Big[\sup_{\varphi\in\mathcal{F}_R}\Big|\frac{1}{n}\sum_{i=1}^n \sigma_i\varphi(X_i)\Big|\Big]}_{\mathcal{R}_n(\mathcal{F}_R,\mu)}$$

where $\sigma_1,\dots,\sigma_n$ are independent Rademacher random variables taking the values $\{-1,+1\}$ with equal probability and $X_1,\dots,X_n$ are independent random variables with law μ. This quantity $\mathcal{R}_n(\mathcal{F}_R,\mu)$ is the Rademacher complexity of $\mathcal{F}_R$ under the distribution μ. It can be bounded by Dudley's chaining technique (see [58, Thm. 5.22] and the associated discussion): it holds, for some universal constant $C > 0$,

$$\mathcal{R}_n(\mathcal{F}_R,\mu) \leq C\inf_{\delta>0}\Big(\delta + n^{-1/2}\int_\delta^{R^2}\sqrt{\log N_\infty(\mathcal{F}_R,u)}\mathrm{d}u\Big)$$

where $N_\infty(\mathcal{F}_R,u)$ is the covering number of the set $\mathcal{F}_R$ for the metric $\|\cdot\|_\infty$ at scale u. Then we use the covering number bound of Bronshtein [10], as reported in [32, Thm. 1] which states that there exists constants $C_1, C_2 > 0$ depending only on d such that if $u/R^2 \leq C_1$ then

$$\log N_\infty(\mathcal{F}_R,u) \leq C_2(u/R^2)^{-d/2}.$$

After a change of variable we thus have that

$$\mathcal{R}_n(\mathcal{F}_R,\mu) \lesssim R^2\Big(\inf_{\delta>0}\delta + n^{-1/2}\int_\delta^1 u^{-d/4}\mathrm{d}u\Big).$$

The claim follows by optimizing over δ which gives $\delta = 0$ for $d < 4$, $\delta = n^{-1/2}$ for $d = 4$ and $\delta = n^{-2/d}$ for $d > 4$. □

We are now in position to conclude the proof of Theorem 2.

Proof of Theorem 2. Let us assume without loss of generality that μ,ν are concentrated on the centered closed ball of radius R in $\mathbb{R}^d$ (which can be taken as $R = 1/2$ under our assumptions, but let us continue with an arbitrary R for explicitness of the proof). Given Lemma 3 and Lemma 4, it only remains to bound the quantity

$$A \overset{\text{def.}}{=} \mathbf{E}\Big|\frac{1}{2}\int \|x\|_2^2\mathrm{d}(\mu-\hat{\mu}_n)(x)\Big|$$

and the corresponding quantity for ν. Considering independent samples of the random variable $Y = \frac{1}{2}\|X\|_2^2$ where the law of X is μ, our goal is to bound $A = \mathbf{E}|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbf{E}Y|$. By Chebyshev's inequality and the fact that the variance of Y is bounded by R^4, we have for all $t \geq 0$,

$$\mathbf{P}\Big[\big|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbf{E}Y\big| \geq t\Big] \leq \min\{1, R^4/(nt^2)\}.$$

Finally, by the integral representation of the expectation of a nonnegative random variable we have

$$A = \int_0^\infty \mathbf{P}\Big[\big|\frac{1}{n}\sum_{i=1}^n Y_i - \mathbf{E}Y\big| \geq t\Big]\mathrm{d}t \leq \frac{R^2}{\sqrt{n}} + \int_{R^2n^{-1/2}}^\infty \frac{R^4}{nt^2}\mathrm{d}t = 2R^2n^{-1/2}$$

which is sufficient to conclude. The concentration bound is proved separately in Proposition 12. □

Let us now prove the concentration bound, which is a consequence of the bounded difference inequality. We give a unified proof for T_λ and T_0 since the argument is similar. The result for T_0 was known [59] but we are not aware of a similar result for $\lambda > 0$ (note that the concentration bound in [26] has an undesirable exponential dependency in λ and the central limit theorem in [40] does not a priori gives the dependency in λ).

Proposition 12. *Assume that $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ are concentrated on a set of diameter D. It holds for all $t \geq 0$, $\lambda \geq 0$ and $n \geq 1$,*

$$\mathbf{P}\Big[\big|T_\lambda(\mu_n, \nu_n) - \mathbf{E}[T_\lambda(\mu_n, \nu_n)]\big| \geq t\Big] \leq 2\exp(-nt^2/D^4).$$

Proof. As in [59], we want to apply the bounded difference inequality but we study the stability of the primal problem (instead of the dual) in order to cover the regularized case painlessly. The empirical measures are of the form $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\nu_n = \frac{1}{n}\sum_{j=1}^n \delta_{y_j}$. Let $c \in \mathbb{R}^{n\times n}$ be the cost matrix with entries $c_{i,j} = \frac{1}{2}\|x_i - y_j\|_2^2$. With those notations, it holds

$$\frac{1}{2}T_\lambda(\mu_n, \nu_n) = \min \sum_{i,j} c_{i,j}P_{i,j} + \lambda \sum_{i,j} P_{i,j}\log(n^2 P_{i,j})$$

where the minimum is over matrices $P \in \mathbb{R}_+^{n\times n}$ such that $P\mathbf{1} = \mathbf{1}/n$ and $P^\top \mathbf{1} = \mathbf{1}/n$ (i.e. nP is bistochastic). Let P^* be a minimizer. Now let $\tilde{\mu}_n = \frac{1}{n}(\sum_{i=1}^{n-1}\delta_{x_i} + \delta_{\tilde{x}_n})$ for some $\tilde{x}_i$ in the same set of diameter D. This changes one row in the cost matrix, each entry in this row being changed by at most $D^2/2$. Thus using P^* as a candidate in the minimization problem defining $T_\lambda(\tilde{\mu}_n, \nu_n)$ we get $T_\lambda(\tilde{\mu}_n, \nu_n) \leq T_\lambda(\mu_n, \nu_n) + \frac{D^2}{n}$. Interchanging the role of μ_n and $\tilde{\mu}_n$, we get the reverse inequality and thus

$$|T_\lambda(\tilde{\mu}_n, \nu_n) - T_\lambda(\mu_n, \nu_n| \leq \frac{D^2}{n}.$$

The same stability can be shown about perturbing ν_n by one sample. The proposition follows by applying the bounded difference inequality [58, Cor. 2.21], paying attention to the fact that the total number of samples is $2n$. □

Finally, let us give the details of the proof of Proposition 3.

Proof of Proposition 3. By the concentration result we have that with probability $1 - \delta$, $|W_2^2(\mu_n, \nu_n) - \mathbf{E}W_2^2(\mu_n, \nu_n)| \lesssim n^{-1/2}\sqrt{\log(2/\delta)}$. Let us break down the proof into three cases depending on the dimension d.

If $d < 4$, then by choosing $n \gtrsim \log(2/\delta)\varepsilon^{-2}$, the quantity $W_2^2(\mu_n, \nu_n)$ has the desired accuracy with probability $1 - \delta$. Also choosing $\lambda \lesssim \varepsilon/(2\log n)$ guarantees that $|\hat{T}_{\lambda,n} - \hat{W}_2^2| \lesssim \varepsilon$. Thus, the computational complexity is $O(n^2/(\lambda\varepsilon)) = \tilde{O}(\varepsilon^{-6})$.

If $d > 4$, we can choose $n \gtrsim \log(2/\delta)^{d/4}\varepsilon^{-d/2}$ to reach the desired accuracy, which leads to a computational complexity in $\tilde{O}(\varepsilon^{-d-2})$.

Finally if $d = 4$, we can choose n such that $\varepsilon \asymp n^{-1/2}(\log(n) + \sqrt{\log(2/\delta)})$ which leads to a computational complexity in $O(n^2\log(n)\varepsilon^{-2}) = O(\varepsilon^{-6}(\log n + \sqrt{\log(2/\delta)})^4) = \tilde{O}(\varepsilon^{-6})$. □

D Analysis of the Sinkhorn divergence estimator given samples

Let us first state a result on the sample complexity to estimate S_λ with $\hat{S}_{\lambda,n}$ which is defined, given $x_1, \dots, x_n$ i.i.d. samples from μ and $y_1, \dots, y_n$ i.i.d. samples from ν, as $\hat{S}_{\lambda,n} = S_\lambda(\hat{\mu}_n, \hat{\nu}_n)$ as in Eq. (2) where $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{\nu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$. Since the following result has not yet been stated in the precise form that we use, we give a short proof below. It essentially just requires to combine the results from [40] and [26].

Lemma 5. *Let $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ be concentrated on a set of diameter 1, let $\hat{\mu}_n, \hat{\nu}_n$ be empirical distributions with n independent samples and let $d' = 2\lfloor d/2 \rfloor$. Then*

$$\mathbf{E}\Big[|\hat{S}_{\lambda,n} - S_\lambda(\mu, \nu)|\Big] \lesssim (1 + \lambda^{-d'/2})n^{-1/2}$$

where $\lesssim$ hides a constant that only depends on d.

Proof. It has been shown in [40, Cor. 2], with a strategy similar to that employed in the end of the proof of Lemma 3, that

$$\Big|\frac{1}{2}T_\lambda(\hat{\mu}_n, \hat{\nu}_n) - \frac{1}{2}T_\lambda(\mu, \nu)\Big| \leq \sup_{f\in\mathcal{F}}\Big|\int f\mathrm{d}(\hat{\mu}_n - \mu)\Big| + \sup_{f\in\mathcal{F}}\Big|\int f\mathrm{d}(\hat{\nu}_n - \nu)\Big|$$

where $\mathcal{F}$ is any class of functions that is large enough to contain all the solutions to Eq. (9) for all pairs of measures $\mu, \nu \in \mathcal{P}(\mathbb{R}^d)$ concentrated on a set of diameter 1. It was shown in [26, Thm. 2] that $\mathcal{F}$ can be chosen as a ball in the Sobolev space H^s, $s \geq 1$ with diameter $C(1 + \lambda^{1-s})$ for some $C > 0$ that only depends on d and s. In particular, for $s = d'/2 + 1$, H^s is a reproducible kernel Hilbert space. Thus, using the notion of Rademacher complexity introduced in the proof of Lemma 4 and its bound for balls in reproducible kernel Hilbert spaces (as in [26, Prop. 2]), it follows

$$\mathbf{E}\Big[\sup_{f\in\mathcal{F}}\Big|\int f\mathrm{d}(\hat{\mu}_n - \mu)\Big|\Big] \leq 2\mathcal{R}_n(\mathcal{F}, \mu) \lesssim (1 + \lambda^{-d'/2})n^{-1/2}.$$

This is sufficient to bound the expected estimation error of T_λ. Let us now turn our attention to $\hat{S}_{\lambda,n}$. It holds

$$\begin{aligned}|\hat{S}_{\lambda,n} - S_\lambda(\mu, \nu)| \leq |T_\lambda(\hat{\mu}_n, \hat{\nu}_n) - T_\lambda(\mu, \nu)| \\ + \frac{1}{2}|T_\lambda(\hat{\mu}_n, \hat{\mu}_n) - T_\lambda(\mu, \mu)| + \frac{1}{2}|T_\lambda(\hat{\nu}_n, \hat{\nu}_n) - T_\lambda(\nu, \nu)|.\end{aligned}$$

The argument in [40] goes through for each term and it follows that $\hat{S}_{\lambda,n}$ admits the same statistical bound (up to a constant) than $\hat{T}_{\lambda,n}$. □

Proof of Proposition 4. Let $W_2^2 = W_2^2(\mu, \nu)$ and $S_\lambda = S_\lambda(\mu, \nu)$. We consider the following error decomposition:

$$\mathbf{E}\big[|\hat{S}_{\lambda,n} - W_2^2|\big] \leq \mathbf{E}\big[|\hat{S}_{\lambda,n} - S_\lambda|\big] + |S_\lambda - W_2^2| \lesssim (1 + \lambda^{-d'/2})n^{-1/2} + \lambda^2$$

where the first bound is from Lemma 5 and the second bound is an assumption. We then optimize the bound in λ which gives $\lambda \asymp n^{-1/(d'+4)}$ and an error in $n^{-2/(d'+4)}$. For the concentration bound, we use the argument in the proof of Proposition 12 in Appendix C. Observe that if only one of the samples drawn from μ is changed, the resulting change in $\hat{S}_{\lambda,n}$ is at most $2/n$ which leads to, by the bounded difference inequality,

$$\mathbf{P}\Big[\big|S_\lambda(\mu_n, \nu_n) - \mathbf{E}[S_\lambda(\mu_n, \nu_n)]\big| \geq t\Big] \leq 2\exp(-nt^2/4).$$

□

Proof of Proposition 5. For $\hat{S}^{(k)}_{\lambda,n}$ we consider the error decomposition

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

Let us choose $\lambda \asymp n^{-1/(d'+4)}$ as in Proposition 4. By the concentration result of Proposition 4, we have that with probability $1 - \delta$, $|\hat{S}_{\lambda,n} - \mathbf{E}\hat{S}_{\lambda,n}| \lesssim n^{-1/2}\sqrt{\log(2/\delta)}$ and thus $|\hat{S}_{\lambda,n} - W_2^2| \lesssim n^{-2/(d'+4)} + n^{-1/2}\sqrt{\log(2/\delta)}$. Thus by choosing $n \gtrsim \log(2/\delta)\varepsilon^{-(d'+4)/2}$ the quantity $\hat{S}_{\lambda,n}$ has the desired accuracy with probability $1 - \delta$. It follows that the computational complexity is $O(n^2/(\lambda\varepsilon)) = \tilde{O}(\varepsilon^{-d'-5.5})$.

For the second claim, we just remark that $n^{-2/(d'+4)}$ dominates the rate of the plug-in estimator given in Theorem 2 for all d, so both estimators can achieve an error of this order. However the difference is that with $\hat{S}_{\lambda,n}$ a regularization level $\lambda \asymp \varepsilon^{-1/2}$ is sufficient while $\lambda \lesssim \varepsilon/\log(n)$ is required for $\hat{T}_{\lambda,n}$ to achieve this error ε. The time complexity bounds then follows by Proposition 2. □

E Analysis of deterministic discretization

In this section, we consider probability distributions on the torus $\mu, \nu \in \mathcal{P}(\mathbb{T}^d)$ with densities with respect to the Lebesgue measure (also denoted μ, ν) and $c(x,y) = \frac{1}{2}\|[y-x]\|_2^2$ which is half the squared distance on the torus. We denote $[x] = x + k_0$ where $k_0 \in \mathbb{Z}^d$ is such that $\|x+k\|_2$ is minimal (k_0 is unique Lebesgue almost everywhere). We denote by (u_λ, v_λ) the couple of minimizers of Eq. (8) that are fixed points of Sinkhorn's iterations

$$u_\lambda(x) = -\lambda \log \int e^{(v_\lambda(y)-c(x,y))/\lambda} \mathrm{d}\nu(y), \quad v_\lambda(y) = -\lambda \log \int e^{(u_\lambda(x)-c(x,y))/\lambda} \mathrm{d}\mu(x) \tag{10}$$

and such that $u_\lambda(0) = 0$. These properties uniquely define (u_λ, v_λ) and we consider $p_\lambda(x,y) = \exp\big((u_\lambda(x) + v_\lambda(y) - c(x,y))/\lambda\big)\mu(x)\nu(y)$ which is the unique solution to (1). The following lemma gives some regularity estimates on p_λ. What is required in its proof is regularity of the marginals and of the cost function (which we fix to be the half squared-norm cost for consistency).

Lemma 6 (Regularity of p_λ)**.** *Assume that $\mu, \nu \in \mathcal{P}(\mathbb{T}^d)$ admit M-Lipschitz continuous log-densities. Then for almost every $z \in (\mathbb{T}^d)^2$ it holds*

$$\|\nabla \log p_\lambda(z)\|_2 \leq 4\sqrt{d}\lambda^{-1} + 2M.$$

Moreover, it holds for all $z, z' \in (\mathbb{T}^d)^2$

$$|p_\lambda(z) - p_\lambda(z')| \leq (e^{(4\sqrt{d}\lambda^{-1}+M)\|[z-z']\|_2} - 1)p_\lambda(z).$$

Proof. By differentiating the definition of p_λ, we have for almost every $(x,y) \in (\mathbb{T}^d)^2$

$$\nabla_x \log p_\lambda(x,y) = \frac{1}{\lambda}(\nabla u_\lambda(x) - [x-y]) + \nabla m(x).$$

where m is the log-density of μ. By differentiating Eq. (10), we also have

$$\nabla u_\lambda(x) = \int [x-y] e^{(u_\lambda(x)+v_\lambda(y)-c(x,y))/\lambda} \mathrm{d}\nu(y)$$

and thus $\|\nabla u_\lambda(x)\|_2 \leq \sup_{y\in\mathbb{T}^d} \|[y-x]\|_2 = \sqrt{d}$. It follows that $\sup_{x,y\in\mathbb{T}^d} \|\nabla_x \log p_\lambda(x,y)\|_2 \leq \frac{2\sqrt{d}}{\lambda} + M$, from which we deduce the first bound by also taking into account the ∇_y component. Now let $\alpha = 4\sqrt{d}\lambda^{-1} + 2M$. By Grönwall's inequality, we have $e^{-\alpha\|[z-z']\|_2}p_\lambda(z) \leq p_\lambda(z') \leq e^{\alpha\|[z-z']\|_2}p_\lambda(z)$ for all $z, z' \in (\mathbb{T}^d)^2$. It follows that $|p_\lambda(z') - p_\lambda(z)| \leq \max\{e^{\alpha\|[z'-z]\|_2} - 1, 1 - e^{-\alpha\|[z'-z]\|_2}\}p_\lambda(z)$ which implies our claim. □

For a measure $\mu \in \mathcal{P}(\mathbb{T}^d)$ we call μ_h its finite volume discretization at resolution $h = 1/m$ for $m \in \mathcal{N}$ on the grid $(\mathbb{Z}/m\mathbb{Z})^d$. It is built via the following process: let $q_h : \mathbb{T}^d \to \mathbb{T}^d$ be defined by $q_h(x_1, \dots, x_d) = (\frac{1}{m}\lfloor mx_1 + 1/2\rfloor, \dots, \frac{1}{m}\lfloor mx_d + 1/2\rfloor)$. It maps each point $x \in \mathbb{T}^d$ to its closest point on the grid $(\mathbb{Z}/m\mathbb{Z})^d$ (with some arbitrary rule for ties). Then let $\mu_h \overset{\text{def.}}{=} (q_h)_\# \mu$ which gives to each point in the grid the mass that μ gives to its surrounding cell. Also, let us label the points in $(\mathbb{Z}/m\mathbb{Z})^d$ from 1 to $n = m^d$ as $(x_i)_{i=1}^n$ (we also use the notation $y_i = x_i$) and let us call $Q_j \subset \mathbb{T}^d$ the set of points which are mapped to the point labeled by $j \in \{1, \dots, n\}$. We also call $Q_{i,j} = Q_i \times Q_j \subset (\mathbb{T}^d)^2$. We now state and prove a result that is slightly more precise than Proposition 6 were we control the error made by replacing measures by their discretization in the estimation of T_λ.

Proposition 13 (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*

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

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

Proof. The principle of the proof is to build admissible transport plans for the continuous (resp. discrete) problem from an admissible transport plan for the discrete (resp. continuous) problem and to bound the associated primal objectives functions.

From discrete to continuous plans. Consider any $\gamma_h \in \Pi(\mu_h, \nu_h)$ and consider $\gamma \in \Pi(\mu, \nu)$ the (unique) measure with a constant density with respect to $\mu \otimes \nu$ on each cell $Q_{i,j}$ and such that $(q_h \otimes q_h)_\# \gamma = \gamma_h$ (see [26, Def. 1] for a detailed construction in $\mathbb{R}^d$). By construction, it holds $H(\gamma, \mu \otimes \nu) = H(\gamma_h, \mu_h \otimes \nu_h)$. Let us bound the difference $\Delta_{i,j} = \int_{Q_{i,j}} (\frac{1}{2}\|[y-x]\|_2^2 - \frac{1}{2}\|[x_i - y_j]\|_2^2)\mathrm{d}\gamma(x,y)$. For clarity, let us assume that $[x-y] = x-y$ for all $(x,y) \in Q_{i,j}$, the argument being the same in each cell. We start with a second order Taylor expansion of the cost (which is exact with our quadratic cost):

$$\begin{aligned}\frac{1}{2}\|y-x\|_2^2 - \frac{1}{2}\|x_i - y_j\|_2^2 = (x_i - y_j)^\top (x - x_i) + (y_j - x_i)^\top (y - y_i)\\ + \frac{1}{2}\|x - x_i\|_2^2 + \frac{1}{2}\|y - y_i\|_2^2 - (x-x_i)^\top (y - y_i).\end{aligned}$$

Integrating the terms in the second row over $Q_{i,j}$, we get a quantity bounded by $dh^2/2$. For the terms in the first row, we see that we have to bound integrals of the form $|\sum_j \int_{Q_{i,j}} (x_i - y_j)^\top (x - x_i)\mathrm{d}\gamma(x,y)| \leq \sqrt{d}|\int_{Q_i} (x - x_i)\mu(x)\mathrm{d}x|$. So let us consider specifically the following integral:

$$\begin{aligned}\Delta_i &= \Big|\int_{Q_i} (x - x_i)\mu(x)\mathrm{d}x\Big|\\ &= \Big|\int_{Q_i} (x - x_i)(\mu(x) - |Q_i|^{-1}\int_{Q_i} \mu(x')\mathrm{d}x')\mathrm{d}x\Big|\\ &\leq \sqrt{d}h|Q_i|^{-1}\int_{Q_i^2} |\mu(x) - \mu(x')|\mathrm{d}x\mathrm{d}x'\end{aligned}$$

where we used the fact that x_i is the center of mass of Q_i for the Lebesgue measure and we denoted $|Q_i|$ the Lebesgue measure of Q_i. Now, since $\log \mu$ is M-Lipschitz an application of Grönwall's inequality as in the proof of Lemma 6 shows that $|\mu(x) - \mu(x')| \leq (e^{M\|x-x'\|_2} - 1)\mu(x)$. It thus follows that

$$\Delta_i \leq \sqrt{d}h(e^{Mh\sqrt{d}} - 1)\mu(Q_i) \lesssim Mh^2\mu(Q_i).$$

Putting all the bounds together and summing over all cells $Q_{i,j}$ we get

$$\int_{(\mathbb{T}^d)^2} \Big(\frac{1}{2}\|[y-x]\|_2^2 - \frac{1}{2}\|[x_i - y_j]\|_2^2\Big)\mathrm{d}\gamma(x,y) \lesssim h^2(1+M).$$

From this it follows that for $\lambda \geq 0$, we have $T_\lambda(\mu,\nu) - T_\lambda(\mu_h,\nu_h) \lesssim h^2(1+M)$.

From continuous to discrete plans. Consider any $\gamma \in \Pi(\mu,\nu)$ and consider its discretization $\gamma_h = (q_h \otimes q_h)_\# \gamma$. By the "information processing inequality", it holds $H(\gamma_h, \mu_h \otimes \nu_h) \leq H(\gamma, \mu \otimes \nu)$. Also, since the cost function is $\sqrt{d}$-Lipschitz on $\mathbb{T}^d$, we have the naive discretization bound

$$\Big|\int_{(\mathbb{T}^d)^2} \frac{1}{2}\|x-y\|_2^2 \mathrm{d}(\gamma - \gamma_h)(x,y)\Big| \lesssim h.$$

This is sufficient to deduce that $T_\lambda(\mu_h,\nu_h) - T_\lambda(\mu,\nu) \lesssim h$ for all $\lambda \geq 0$. Let us see however that a finer discretization bound can be given when γ is the optimal solution of the entropy regularized problem using the regularity shown in Lemma 6. We denote $z = (x,y) \in (\mathbb{T}^d)^2$ and $z_{i,j} = (x_i, y_i)$ and we have, by decomposing the error into a first and second order term as in the first part of the proof,

$$\begin{aligned}\Big|\int_{(\mathbb{T}^d)^2} \frac{1}{2}\|y-x\|_2^2 \mathrm{d}(\gamma(x,y) - \gamma_h(x,y))\Big| &= \Big|\sum_{i,j}\int_{Q_{i,j}} (\frac{1}{2}\|y-x\|_2^2 - \frac{1}{2}\|x_i - y_j\|_2^2)\mathrm{d}\gamma(x,y)\Big|\\ &\lesssim \sum_{i,j}\Big|\int_{Q_{i,j}} (z - z_{i,j})p_\lambda(z)\mathrm{d}z\Big| + h^2.\end{aligned}$$

It remains to estimate the integral terms as can be done as in the first part of the proof by using the regularity of $\log p_\lambda$ given by Lemma 6

$$\begin{aligned}\Big|\int_{Q_{i,j}} (z - z_{i,j})p_\lambda(z)\mathrm{d}z\Big| &\lesssim h|Q_{i,j}|^{-1}\int_{Q_{i,j}}\int_{Q_{i,j}} |p_\lambda(z) - p_\lambda(z')|\mathrm{d}z\mathrm{d}z'\\ &\leq h(e^{(4\sqrt{d}\lambda^{-1}+M)\sqrt{d}h} - 1)p_\lambda(Q_{i,j})\\ &\lesssim h^2(\lambda^{-1}+M)p_\lambda(Q_{i,j}).\end{aligned}$$

The conclusion follows by summing over all cells $Q_{i,j}$. □

We now proceed to the proof of Proposition 7. This proof would be immediate if we were working on $\mathbb{R}^d$ by combining the stability of Proposition 6 with the approximation error of Theorem 1. However, our framework in this section is that of the torus, and has to be so because there is no compactly supported measures with continuous log-densities on $\mathbb{R}^d$. In the setting of the torus, the equivalence from Eq. (4) holds for a slightly different cost function built from the heat kernel on the torus, as proved in [30] for general manifolds. This cost function is

$$\tilde{c}_\lambda(x,y) = -\lambda \log\Big(\sum_{k\in\mathbb{Z}^d} \exp\Big(-\frac{1}{2\lambda}\|x-y-k\|_2^2\Big)\Big).$$

Let $\tilde{T}_\lambda(\mu,\nu)$ be the entropy regularized optimal transport cost as defined in Eq. (1) where the cost function $c(x,y)=\frac{1}{2}\|[x-y]\|_2^2$ is replaced by $\tilde{c}_\lambda$, and let $\tilde{S}_\lambda$ be the corresponding Sinkhorn divergence, as defined in Eq. (2). A direct extension of Theorem 1 then gives that if $\mu,\nu\in\mathcal{P}(\mathbb{R}^d)$ have bounded densities and supports then

$$|\tilde{S}_\lambda(\mu,\nu) - W_2^2(\mu,\nu)| \le \frac{\lambda^2}{4}\max\{2I_0(\mu,\nu), I_0(\mu,\mu)+I_0(\nu,\nu)\}. \tag{11}$$

In the next lemma, we control the error that is made when replacing $\tilde{S}_\lambda$ by S_λ, which is asymptotically exponentially small.

Lemma 7. *Assume that $\mu,\nu\in\mathcal{P}(\mathbb{T}^d)$ admit log-densities which are Lipschitz continuous. Then there exists $c_1,c_1',c_2>0$ such that*

$$0 \le T_\lambda(\mu,\nu) - \tilde{T}_\lambda(\mu,\nu) \le c_1 e^{-c_2/\lambda}.$$

In particular, we have $|\tilde{S}_\lambda(\mu,\nu) - S_\lambda(\mu,\nu)| \le c_1' e^{-c_2/\lambda}$.

In contrast to the other statements in this paper, this one is purely asymptotic in the sense that the constants may depend on μ and ν. This is due to a technical difficulty near the cut-locus where the convergence of $\tilde{c}_\lambda$ towards c is only in $O(\lambda)$ which is too slow for our purposes. We can avoid this difficulty by exploiting the fact that the optimal transport map stays away from the cut locus and using the uniform convergence of the dual potentials (u_λ,v_λ) towards (u_0,v_0) but we are not aware of quantitative versions of these results.

Proof. The inequality $\tilde{T}_\lambda(\mu,\nu)\le T_\lambda(\mu,\nu)$ is immediate since $\tilde{c}_\lambda\le c$. The main difficulty is thus to prove the other bound. For this, let (u_λ,v_λ) be the unique pair of maximizers of Eq. (9) such that $u_\lambda(0)=0$. As $\lambda\to 0$, this pair converges uniformly to a couple of functions (u_0,v_0) which is the unique solution to the unregularized dual problem such that $u_0(0)=0$, see e.g. [5]. Letting $\tilde{F}_\lambda$ be the dual of the regularized problem Eq. (8) where c is replaced by $\tilde{c}_\lambda$, we have $\frac{1}{2}\tilde{T}_\lambda(\mu,\nu)=\sup \tilde{F}_\lambda(u,v)$ where the supremum is over pairs of continuous functions on the torus. Thus we have

$$\begin{aligned}\frac{1}{2}T_\lambda(\mu,\nu) - \frac{1}{2}\tilde{T}_\lambda(\mu,\nu) &\le F_\lambda(u_\lambda,v_\lambda) - \tilde{F}_\lambda(u_\lambda,v_\lambda)\\ &= \lambda\int_{(\mathbb{T}^d)^2} e^{(u_\lambda(x)+v_\lambda(y)-c(x,y))/\lambda}\big(e^{(c-\tilde{c})/\lambda}-1\big)\mathrm{d}\mu(x)\mathrm{d}\nu(y).\end{aligned}$$

It remains to bound this integral and we will do so by dividing the domain $(\mathbb{T}^d)^2$ into two sets.

By the regularity theory of optimal transport on the torus [15], we know that u_0 is continuously differentiable (note that our assumption on the regularity of μ and ν is indeed stronger than Hölder continuity). It follows by [5, Lem. 2.4] that the optimal transport map T is continuous and its graph $G=\{(x,T(x))\ ;\ x\in\mathbb{T}^d\}$ does not intersect the singular set S of $(x,y)\mapsto\|[y-x]\|_2^2$, i.e. the set where this function is not differentiable. As both sets are compact, they are thus at a positive distance $2\delta>0$ from each other. Let G_δ be the closed set of points that are at a distance less than or equal to δ from G (which is itself at a distance δ from S). Since in our context G is precisely the set of points (x,y) where $u_0(x)+v_0(y)=c(x,y)$ (see again [5, Lem. 2.4]), there exists $\alpha>0$ such that $u_0(x)+v_0(x)-c(x,y)\le -2\alpha$ for all $(x,y)\in G_\delta^c=(\mathbb{T}^d)^2\setminus G_\delta$.

Let (I) and (II) be the value of the integral above on G_δ^c and G_δ respectively, so that $T_\lambda(\mu,\nu) - \tilde{T}_\lambda(\mu,\nu) \leq 2(I) + 2(II)$. On the one hand, by uniform convergence of the potentials, there exists $\lambda_0 > 0$ such that $\forall \lambda < \lambda_0$, $\|u_\lambda - u_0\|_\infty + \|v_\lambda - v_0\|_\infty \leq \alpha$ and thus $\forall \lambda \leq \lambda_0$,

$$(I) \leq \lambda e^{-\alpha/\lambda} |e^{\|c-\tilde{c}_\lambda\|_\infty/\lambda} + 1| = o(e^{-\alpha/(2\lambda)})$$

because $\tilde{c}_\lambda$ converges uniformly to c as $\lambda \to 0$. On the other hand

$$(II) \leq \lambda \sup_{z\in G_\delta} (e^{(c(z)-\tilde{c}(z))/\lambda} - 1) = \lambda \sup_{z\in G_\delta} \sum_{k\in\mathbb{Z}^d\setminus\{k_0(z)\}} e^{(\|z-k_0(z)\|_2^2 - \|z-k\|_2^2)/(2\lambda)}$$

where $k_0(z)$ is such that $\|[z]\|_2 = \|z - k_0\|_2$ and is unique for $z \in G_\delta$. Letting $\beta = \inf_{z\in G_\delta, k\neq k_0(z)} \|z-k\|_2^2 - \|z-k_0(z)\|_2^2$, we have $\beta > 0$ since G_δ is at a positive distance from the singular set S and we have $(II) \lesssim \lambda e^{-\beta/(2\lambda)}$ because the series $\sum_{k\neq k_0} e^{(\beta + \|z-k_0\|_2^2 - \|z-k\|_2^2)/(2\lambda)}$ is nonincreasing in λ (notice that the exponent is nonpositive). Summing (I) and (II) leads to the result. □

We are finally in a position to prove Proposition 7.

Proof of Proposition 7. We decompose the error as

$$|S_\lambda(\mu_h,\nu_h) - W_2^2(\mu,\nu)| \leq |S_\lambda(\mu_h,\nu_h) - S_\lambda(\mu,\nu)| + |S_\lambda(\mu,\nu) - \tilde{S}_\lambda(\mu,\nu)| + |\tilde{S}_\lambda(\mu,\nu) - W_2^2(\mu,\nu)|.$$

The first term is in $O(h^2(\lambda^{-1} + M + 1))$ by Proposition 6. The second term is bounded by $c_1 e^{-c_2/\lambda}$ by Lemma 7. The third term is bounded by $(\lambda^2/4)\max\{I_0(\mu,\nu), I_0(\mu,\mu) + I_0(\nu,\nu)\}$ as seen in Eq. (11), which is a variation of Theorem 1. Moreover, the assumption that μ and ν have M-Lipschitz continuous log-densities leads to the bound $I_0(\mu,\mu), I_0(\nu,\nu) \leq M^2$, which justifies why the statement of Proposition 7 does not requires specifically that these quantities be finite. Thus, we have

$$|S_\lambda(\mu_h,\nu_h) - W_2^2(\mu,\nu)| \lesssim h^2\lambda^{-1} + \lambda^2.$$

Minimizing in λ suggests to take $\lambda = h^{2/3}$ and leads to an error bound in $O(h^{4/3})$. In terms of the accuracy ε, we thus have $h \asymp \varepsilon^{3/4}$ and $\lambda \asymp \varepsilon^{1/2}$. The computational complexity bound follows by Proposition 2 which gives a bound in $O(n^2\lambda^{-1}\varepsilon^{-1})$ and the fact that $n = h^{-d} \asymp \varepsilon^{-3d/4}$, hence a bound in $O(\varepsilon^{-3d/2-3/2})$.

For the computational complexity bound via T_λ, we use the error decomposition

$$|T_\lambda(\mu_h,\nu_h) - W_2^2(\mu,\nu)| \leq |T_\lambda(\mu_h,\nu_h) - T_0(\mu_h,\nu_h)| + |T_0(\mu_h,\nu_h) - T_0(\mu,\nu)|$$

where the first term is in $O(\lambda\log(n))$ and the second term is in $O(h)$ by Proposition 6. Thus to reach an accuracy $\varepsilon > 0$, we may choose $h \asymp \varepsilon$ and $\lambda \asymp \varepsilon/\log(n)$ which leads to a time complexity in $\tilde{O}(\varepsilon^{-2d-2})$. □

F Analysis of the 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 $A, B \in \mathbb{R}^{d\times d}$. The following explicit formula for T_λ is proven in [34]:

$$T_\lambda(\mu,\nu) = \|a-b\|_2^2 + \operatorname{tr}(A) + \operatorname{tr}(B) - 2\operatorname{tr}(D_\lambda^{AB}) + d\lambda(1 - \log(2\lambda)) + \lambda \log\det(2D_\lambda^{AB} + \lambda I)$$

where $A^{1/2}$ denotes the unique positive definite square root of a positive definite matrix A and $D_\lambda^{AB} = (A^{1/2}BA^{1/2} + \lambda^2 I/4)^{1/2}$ (notice that $A^{1/2}BA^{1/2} = M^\top M$ for $M = B^{1/2}A^{1/2}$ is positive definite). When $\lambda = 0$, we recover the well known explicit formula (see e.g. [6]):

$$W_2^2(\mu,\nu) = \|a-b\|_2^2 + \operatorname{tr}(A) + \operatorname{tr}(B) - 2\operatorname{tr}(S).$$

where $S = (A^{1/2}BA^{1/2})^{1/2}$. Notice that this expression involves the squared Bures distance [6] between positive definite matrices defined as $\mathrm{d}_\mathrm{b}^2(A,B) \overset{\text{def.}}{=} \operatorname{tr}(A) + \operatorname{tr}(B) - 2\operatorname{tr}(S)$.

The expression above leads to the following formula for $\Delta = S_\lambda(\mu,\nu) - W_2^2(\mu,\nu)$:

$$\begin{aligned}\Delta = (\operatorname{tr}(D_\lambda^{AA}) - \operatorname{tr}(D_0^{AA})) + (\operatorname{tr}(D_\lambda^{BB}) - \operatorname{tr}(D_0^{BB})) - 2(\operatorname{tr}(D_\lambda^{AB}) - \operatorname{tr}(D_0^{AB}))\\ + \frac{\lambda}{2}\big(2\log\det(2D_\lambda^{AB} + \lambda I) - \log\det(2D_\lambda^{AA} + \lambda I) - \log\det(2D_\lambda^{BB} + \lambda I)\big).\end{aligned}$$

Fourth-order expansion of Δ. Let us first expand individual terms using the fact that all the matrices involved are positive definite. We have

$$\begin{aligned} D_\lambda^{AA} &= A(I + (\lambda^2/4)A^{-2})^{1/2} \\ &= A + \frac{\lambda^2}{8}A^{-1} - \frac{\lambda^4}{128}A^{-3} + O(\lambda^5). \end{aligned}$$

Also, since $\log\det(I + \lambda A) = \lambda\operatorname{tr}(A) - (\lambda^2/2)\operatorname{tr}(A^2) + (\lambda^3/3)\operatorname{tr}(A^3) + O(\lambda^4)$, we obtain the expansion

$$\begin{aligned} \frac{\lambda}{2}\log\det(2D_\lambda^{AA} + \lambda I) &= \frac{\lambda}{2}\log\det(2A + (\lambda^2/4)A^{-1} + \lambda I + O(\lambda^4)) \\ &= \frac{\lambda}{2}\log\det(2A) + \frac{\lambda}{2}\log\det(I + (\lambda/2)A^{-1} + (\lambda^2/8)A^{-2} + O(\lambda^4)) \\ &= \frac{\lambda}{2}\log\det(2A) + \frac{\lambda^2}{4}\operatorname{tr}(A^{-1}) - \frac{\lambda^4}{96}\operatorname{tr}(A^{-3}) + O(\lambda^5). \end{aligned}$$

Putting all pieces together with the notation $S = (A^{1/2}BA^{1/2})^{1/2}$ leads to

$$\begin{aligned} \Delta = \frac{\lambda^2}{8}\operatorname{tr}(A^{-1}) - \frac{\lambda^4}{128}\operatorname{tr}(A^{-3}) + \frac{\lambda^2}{8}\operatorname{tr}(B^{-1}) - \frac{\lambda^4}{128}\operatorname{tr}(B^{-3}) - \frac{\lambda^2}{4}\operatorname{tr}(S^{-1}) + \frac{\lambda^4}{64}\operatorname{tr}(S^{-3}) \\ + \lambda\log\det(2S) - \frac{\lambda}{2}\log\det(2A) - \frac{\lambda}{2}\log\det(2B) \\ + \frac{\lambda^2}{2}\operatorname{tr}(S^{-1}) - \frac{\lambda^2}{4}\operatorname{tr}(A^{-1}) - \frac{\lambda^2}{4}\operatorname{tr}(B^{-1}) - \frac{\lambda^4}{48}\operatorname{tr}(S^{-3}) + \frac{\lambda^4}{96}\operatorname{tr}(A^{-3}) + \frac{\lambda^4}{96}\operatorname{tr}(B^{-3}) + O(\lambda^5). \end{aligned}$$

The $\log\det$ terms cancel each other and some simplifications in the other terms lead to

$$\Delta = \frac{\lambda^2}{8}\big(2\operatorname{tr}(S^{-1}) - \operatorname{tr}(A^{-1}) - \operatorname{tr}(B^{-1})\big) - \frac{\lambda^4}{384}\big(2\operatorname{tr}(S^{-3}) - \operatorname{tr}(A^{-3}) - \operatorname{tr}(B^{-3})\big) + O(\lambda^5).$$

Interestingly, this expression can be expressed purely in terms of Bures distances:

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

This shows that the terms in this expansion are non-zero unless $A = B$ and also determines their sign.

G Numerical settings and additional experiments

G.1 Sampling method

In this paragraph, we detail the setting of the *random sampling* experiments (Figure 2 and Figure 6). In those experiments, the distributions μ and ν are elliptically contoured and centered, which allows to have a closed form expression for the optimal transport cost T_0 and the dual potential φ (the Lagrange multiplier associated to the first marginal constraint in the computation of $T_0(\mu,\nu)$ in Eq. (1)), which only depends on the two covariances [6]. Specifically, given two measures μ, ν that belong to the same family of elliptically contoured distributions, with respective covariances A and B and with 0 means, we have

$$T_0(\mu,\nu) = \mathrm{d}_\mathrm{b}^2(A,B) \qquad \text{and} \qquad \varphi(x) = x^\top(\mathrm{Id} - M)x$$

where $\mathrm{d}_\mathrm{b}^2(A,B) = \operatorname{tr}(A) + \operatorname{tr}(B) - 2\operatorname{tr}(S)$ and $M = A^{1/2}SA^{1/2}$ where S is as defined in Appendix F. Let us detail how we have chosen the covariances and our choice of elliptically contoured distribution.

Choice of the covariances. The covariances $A, B \in \mathbb{R}^{d\times d}$ are generated randomly, independently and identically according to the following process, that we detail for A. Let $M \in \mathbb{R}^{d\times k}$ be a random matrix with i.i.d. entries following a standard normal distribution $\mathcal{N}(0,1)$, with $k = d/\alpha$ for some $\alpha \in (0,1)$. We then define $\tilde{A} = MM^\top$, which is a random positive semidefinite matrix. By non-asymptotic versions of the Marčenko-Pastur Theorem (e.g. [58, Eq.(1.11)]), the eigenvalues of $\tilde{A}$ are contained within a small enlargement of the interval $[(1-\sqrt{\alpha})^2, (1+\sqrt{\alpha})^2]$ with a high probability that increases with d. We then define $A = \tilde{A}/\operatorname{tr}\tilde{A}$. With our choice $\alpha = 1/3$, this allows to define generic covariance matrices of trace 1 with a controlled anisotropy: the ratio between the largest and smallest eigenvalue is with high probability of order 0.07 for large d (but note that since we work with relatively small values of d, this ratio is subject to fluctuations).

Choice of the distributions. Given a covariance A we generate a sample X as follows:

1. $U \sim \mathcal{U}(\mathbb{S}^{d-1})$ (U is uniformly distributed on the sphere in $\mathbb{R}^d$)
2. $Z \sim \mathcal{N}(0,1)$
3. $R = \alpha|\arctan(Z/\beta)|^{1/d}$ where $\alpha > 0$ is such that $\mathbf{E}[R^2] = d$
4. $X = R \cdot A^{1/2}U$

Here $\beta > 0$ is a free parameter that determines the shape of the distribution and we have chosen $\beta = 2$ because it tends to yield nice bell shaped densities (see Figure 5). Also, α is a quantity that only depends on d and β that we estimate via Monte-Carlo integration. Let us describe the distribution of X.

Proposition 14. *The law of X is elliptically contoured, centered, and has a compact support. Its covariance is A and its density with respect to the Lebesgue measure (denoted by $\mu(x)$) is given by*

$$\mu(x) \propto (1+\tan(y)^2)\exp(-\beta^2\tan(y)^2/2) \tag{12}$$

where $y = (\|x\|_{A^{-1}}/\alpha)^d$ and $\|x\|^2_{A^{-1}} = x^\top A^{-1}x$. In particular, if A is nonsingular then its Fisher information is finite: $I_0(\mu,\mu) < \infty$.

It follows that if μ and ν are the densities of random variables generated via this procedure, with respective covariances A and B, then Theorem 1 together with Proposition 1 guarantee that Proposition 4 applies. We illustrate the results of Proposition 14 in Figure 5.

Proof. By construction μ is elliptically contoured and centered [22, Chap. 2]. It is compactly supported because the range of $z \mapsto |\arctan(z/\beta)|$ is $[0,\pi/2)$. Also the covariance of X is

$$\mathbf{E}\big[XX^\top\big] = \frac{1}{d}\mathbf{E}\big[R^2\big]A = A.$$

Let $Y = \arctan(Z/\beta)$ and let F_Y (resp. f_Y) be the cumulative (resp. probability) distribution function of Y. We have for $x \in \mathbb{R}$,

$$F_R(x) = \mathbf{P}\big[R \le x\big] = \mathbf{P}\big[\alpha|\arctan(Z/\beta)|^{1/d} \le x\big] = \mathbf{P}\big[|Y| \le (x/\alpha)^d\big] = F_{|Y|}((x/\alpha)^d).$$

Differentiating this relation, it follows that $f_R(x) \propto x^{d-1}f_{|Y|}((x/\alpha)^d)$. Then by [22, Thm. 2.9 & Eq. (2.43)], we have

$$\mu(x) \propto \|x\|^{1-d}_{A^{-1}} f_R(\|x\|_{A^{-1}}) \propto f_{|Y|}((\|x\|_{A^{-1}}/\alpha)^d).$$

It thus remains to compute the density $f_{|Y|}$ which, by symmetry of Y around 0, is precisely twice the density f_Y for nonnegative arguments. Denoting $g(z) = \arctan(z/\beta)$, by the change of variable formula, we have

$$f_Y(y) = \frac{f_Z(g^{-1}(y))}{g'(g^{-1}(y))} \propto (1+\tan(y)^2)\cdot\exp(-\beta^2\tan(y)^2/2)$$

which gives the density of μ, up to a multiplicative constant. Let us now show that the Fisher information $I_0(\mu,\mu) = \int_{\mathbb{R}^d}\|\frac{\nabla\mu(x)}{\mu(x)}\|_2^2\mu(x)\mathrm{d}x$ is finite, with the assumption that $A = \mathrm{Id}$ for simplicity (the general case can be treated similarly). We have $\mu(x) = f_Y(h(\|x\|_2))$ with $h(r) = (r/\alpha)^d$ and by direct computations:

$$I_0(\mu,\mu) \propto \int_{\mathbb{R}^d}\Big(\frac{f'(h(\|x\|_2))}{f(h(\|x\|_2))}\Big)^2\|x\|_2^{2d-2}f(h(\|x\|_2))\mathrm{d}x \propto \int_0^{\pi/2}\Big(\frac{f'(h(r))}{f(h(r))}\Big)^2 r^{3d-3}f(h(r))\mathrm{d}r$$
$$f'_Y(y) \propto \exp(-\beta^2\tan(y)^2/2)\big(\beta^2\tan(y)(1+\tan(y)^2)(1-\beta^2\tan(y)^2)\big).$$

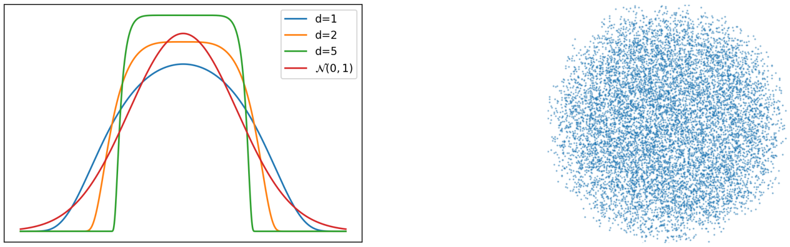

Figure 5: Density used for the random sampling experiments, when $A = \mathrm{Id}/d$. Left: radial profile of the density as given by Eq. (12), i.e. $t \mapsto \mu(t\vec{u})$ for some $\vec{u} \in \mathbb{S}^{d-1}$. Right: 10^4 samples for $d = 2$.

Then by posing $z = \tan h(r)$, we get

$$I_0(\mu, \mu) \propto \int_{\mathbb{R}_+} (\beta^4 z^2 (1 - \beta^2 z^2)^2 \arctan(z)^{2-2/d}) \exp(-\beta^2 z^2/2) \mathrm{d}z$$

where $\propto$ in those computations just means that the right-hand side is finite if and only if the left-hand side is finite. Since the right-hand side is finite, this shows that $I_0(\mu, \mu) < \infty$. □

G.2 Additional random sampling experiment

On Figure 6, we show the same experiment as in Section 5 but in dimension $d = 10$ and moreover we report the error on the transport cost $T_0(\mu, \nu)$ and the rate of Theorem 2, which were not shown on Figure 2. The plot on the right shows the estimation error on $T_0(\mu, \nu)$, which is the quantity that we control in our theoretical analysis. This plot confirms several of our results: (i) the convergence rate in $n^{-2/d}$ of the plug-in estimator proved in Theorem 2 (note that we compute it with a small entropic regularization, which might explain the slight deviation from the rate $n^{-2/d}$ that we observe for n large), and (ii) the fact that T_λ has a much larger bias than S_λ and R_λ. Even more interestingly, S_λ and R_λ have a smaller error than the plug-in estimator. However, we should also be cautious when interpreting such a plot because $T_0(\mu, \nu)$ is a scalar, and it is easy to make the error vanish when varying a parameter, such as n or λ. In particular, the local minimum observed for S_λ and R_λ is simply due to the fact that the error changes its sign as n grows.

This phenomenon led us to report the error on a different quantity, the L^1 error on the potential, which is not subject to this phenomenon and which also raises interesting open questions. Notice however that this quantity may behave quite differently than the estimation error on $T_0(\mu, \nu)$. In particular, we see on Figure 5-(left), that the rate of convergence of the plug-in estimator is in fact faster than $n^{-2/d}$ in this experiment.

G.3 Additional figures for the discretization experiment

Figure 7 shows the same setting as on Figure 4 and gives more details. The densities of μ and ν on the 1-dimensional torus $\mathbb{T}$ are shown on the top row at several levels of discretization. The two other rows show the evolution of the estimated potentials as n varies for the optimal λ (middle row) or as λ varies for n large (bottom row) towards the true potentials (u_0, v_0) (shown in dark color). Here u_0 is the Lagrange multiplier associated to the first marginal constraint in the computation of $T_0(\mu, \nu)$ in Eq. (1) and v_0 is the one associated to the second marginal constraint. On Figure 7, we denote by (u_h, v_h) the potentials associated to the estimator T_λ and by $(\bar{u}_h, \bar{v}_h)$ those associated to the estimator S_λ, as defined in Section 5. This figure illustrates that for λ large, the error is systematically smaller with the debiasing terms.

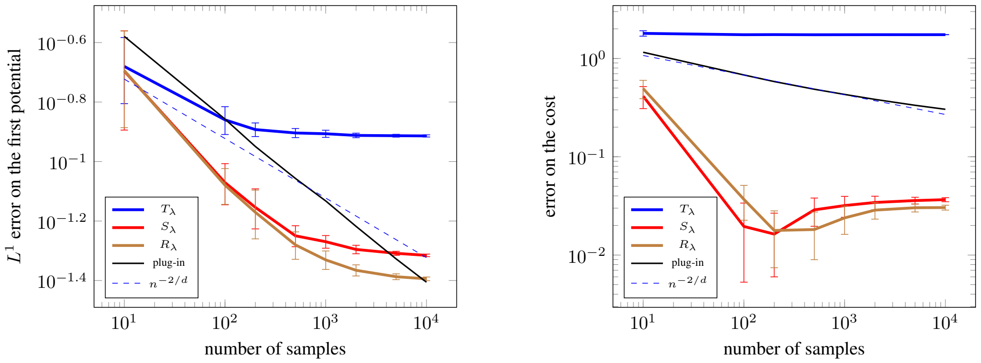

Figure 6: L^1 error on the first potential (left) and error on the estimated cost (right) for different estimators, for μ, ν smooth compactly supported distributions with $d = 10$, as a function of n for $\lambda = 1$. Error bars show the standard deviation on 30 realizations

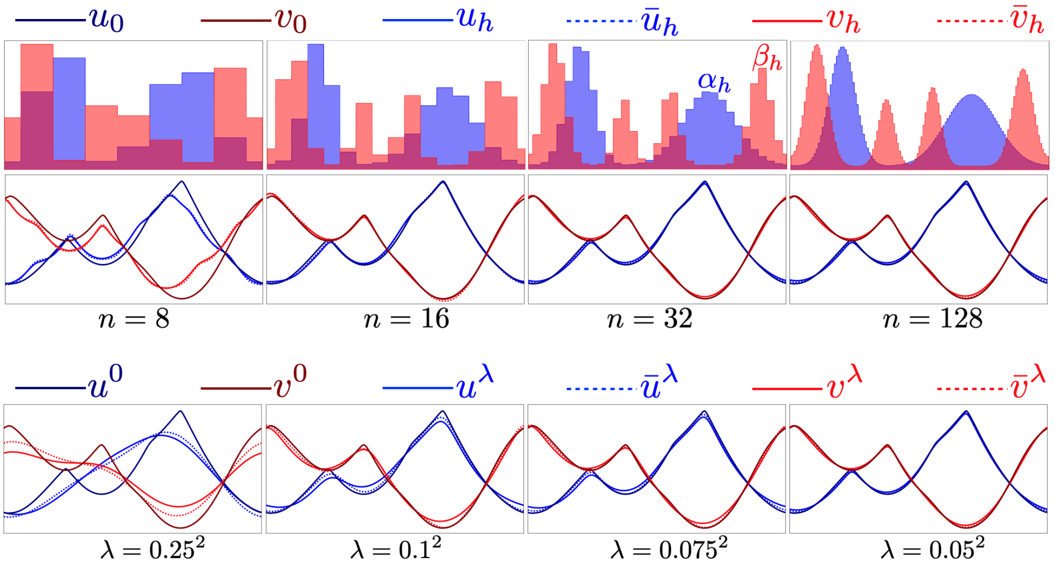

Figure 7: Rows 1 and 2: convergence of the dual potentials $(u_{0,h}, v_{0,h})$ and $(\bar{u}_{0,h}, \bar{v}_{0,h})$ towards (u_0, v_0) for decreasing sampling step h. The top row shows the discretized measures (μ_h, ν_h) (the measure is a sum of Dirac masses, which is vizualized as a piecewise constant function to indicate the cells over which the densities have been integrated). Last row: same but for the convergence of $(u_{\lambda,0}, v_{\lambda,0})$ and $(\bar{u}_{\lambda,0}, \bar{v}_{\lambda,0})$ as λ gets smaller.