Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections

Anonymous Author(s)
Affiliation
Address
email

Abstract

Optimal transport maps between two probability distributions μ and ν on $\mathbb{R}^d$ have found extensive applications in both machine learning and statistics. In practice, these maps need to be estimated from data sampled according to μ and ν. Plug-in estimators are perhaps most popular in estimating transport maps in the field of computational optimal transport. In this paper, we provide a comprehensive analysis of the rates of convergences for general plug-in estimators defined via barycentric projections. Our main contribution is a new stability estimate for barycentric projections which proceeds under minimal smoothness assumptions and can be used to analyze general plug-in estimators. We illustrate the usefulness of this stability estimate by first providing rates of convergence for the natural discrete-discrete and semi-discrete estimators of optimal transport maps. We then use the same stability estimate to show that, under additional smoothness assumptions, kernel smoothed plug-in estimators speed up the rates of convergence and significantly mitigate the curse of dimensionality suffered by the natural discrete-discrete/semi-discrete estimators. As a by-product of our analysis, we also obtain faster rates of convergence for plug-in estimators of $W_2(\mu, \nu)$, the Wasserstein distance between μ and ν, under smoothness assumptions, thereby complementing recent results in Chizat et al. (2020). Finally, we illustrate the applicability of our results in obtaining rates of convergence for Wasserstein barycenters between two probability distributions and obtaining asymptotic detection thresholds for some recent optimal-transport based tests of independence.

1 Introduction

Given two random variables $X \sim \mu$ and $Y \sim \nu$, where μ, ν are probability measures on $\mathbb{R}^d$, $d \geq 1$, the problem of finding a "nice" map $T_0(\cdot)$ such that $T_0(X) \sim \nu$ has numerous applications in machine learning such as domain adaptation and data integration [34, 35, 38, 47, 58, 101], dimension reduction [11, 63, 85], generative models [57, 77, 83, 99], to name a few. Of particular interest is the case when $T_0(\cdot)$ is obtained by minimizing a cost function, a line of work initiated by Gaspard Monge [91] in 1781 (see (1.1) below), in which case $T_0(\cdot)$ is termed an *optimal transport* (OT) map and has applications in shape matching/transfer problems [30, 46, 96, 108], Bayesian statistics [45, 72, 76, 97], econometrics [14, 29, 44, 49, 53], nonparametric statistical inference [39–41, 102, 103]; also see [100, 114, 115] for book-length treatments on the subject. In this paper, we will focus on the OT map obtained using the standard Euclidean cost function, i.e.,

$$T_0 := \underset{T: T\#\mu = \nu}{\operatorname{argmin}} \ \mathbb{E}\|X - T(X)\|^2, \tag{1.1}$$

where $T\#\mu = \nu$ means $T(X) \sim \nu$ for $X \sim \mu$. The estimation of T_0 has attracted a lot of interest in recent years due to its myriad applications (as stated above) and interesting geometrical properties (see [19, 55, 86] and Definition 1.1 below). In practice, the main hurdle in constructing estimators for

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

T_0 is that the explicit forms of the measures μ, ν are unknown; instead only random samples

$$X_1, \ldots, X_m \sim \mu \qquad \text{and} \qquad Y_1, \ldots, Y_n \sim \nu$$

are available. A natural strategy in this scenario is to estimate T_0 using $\widetilde{T}_{m,n}$, where $\widetilde{T}_{m,n}$ is computed as in (1.1) with μ and ν replaced by $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ which are empirical approximations of μ and ν based on $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ respectively (see Definition 1.2). Such estimators are often called *plug-in estimators* and have been used extensively, see [6, 31, 64, 88, 89, 93, 104].

The main goal of this paper is to study the *rates of convergence* of general plug-in estimators of T_0 under a unified framework. We show that when $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ are chosen as $\widehat{\mu}_m$ and $\widehat{\nu}_n$ respectively, where $\widehat{\mu}_m$ and $\widehat{\nu}_n$ are the standard empirical distributions supported on m and n atoms, i.e.,

$$\widehat{\mu}_m := \frac{1}{m}\sum_{i=1}^{m} \delta_{X_i} \qquad \text{and} \qquad \widehat{\nu}_n := \frac{1}{n}\sum_{j=1}^{n} \delta_{Y_j}, \tag{1.2}$$

$\widetilde{T}_{m,n}$ (appropriately defined using Definition 1.2) converges at a rate of $m^{-2/d} + n^{-2/d}$ for $d \geq 4$. This rate happens to be minimax optimal under minimal smoothness assumptions (see [69, Theorem 6]) but suffers from the *curse of dimensionality*. We next show that, if μ and ν are known to admit sufficiently smooth densities, it is possible to apply a standard kernel smoothing technique on $\widehat{\mu}_m$ and $\widehat{\nu}_n$ to obtain plug-in estimators that mitigate the aforementioned curse of dimensionality.

Our next contribution pertains to the estimation of $W_2^2(\mu, \nu)$ (the squared Wasserstein distance), see (1.3) below, a quantity of independent interest in statistics and machine learning with applications in structured prediction [50, 84], image analysis [17, 56], nonparametric testing [15, 95], generative modeling [9, 90], etc. In this paper, we also obtain rates of convergence for plug-in estimators $W_2^2(\widetilde{\mu}_m, \widetilde{\nu}_n)$ of $W_2^2(\mu, \nu)$. We show that kernel smoothing $\widehat{\mu}_m$ and $\widehat{\nu}_n$ can be used to obtain plug-in estimators of $W_2^2(\mu, \nu)$ that mitigate the curse of dimensionality as opposed to a direct plug-in approach using $\widehat{\mu}_m$ and $\widehat{\nu}_n$ (as used in [31, Theorem 2]). This provides an answer to the open question of estimating $W_2^2(\mu, \nu)$ when μ, ν admit smooth densities laid out in [31].

1.1 Background on optimal transport

In this section, we present some basic concepts and results associated with the OT problem that will play a crucial role in the sequel. Let $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$ denote the set of all Lebesgue absolutely continuous probability measures on $\mathbb{R}^d$ and $\mathcal{P}_2(\mathbb{R}^d)$ be the set of probability measures with finite second moments. Then the 2-*Wasserstein* distance (squared) between $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$ is defined as:

$$W_2^2(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int \|x - y\|^2 \, d\pi(x, y), \tag{1.3}$$

where $\Pi(\mu, \nu)$ is the set of probability measures on $\mathbb{R}^d \times \mathbb{R}^d$ with marginals μ and ν. The optimization problem in (1.3) is often called the *Kantorovich relaxation* (see [73, 74]) of the optimization problem in (1.1). The existence of a minimizer in (1.3) follows from [115, Theorem 4.1].

Proposition 1.1 (Brenier-McCann polar factorization theorem, see [86, 114])**.** *Suppose $\mu, \nu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$. Then there exists a μ-a.e. (almost everywhere) unique function $T_0(\cdot) : \mathbb{R}^d \to \mathbb{R}^d$, which is the gradient of a real-valued d-variate convex function, say $\varphi_0(\cdot) : \mathbb{R}^d \to \mathbb{R}$, such that $T_0 \# \mu = \nu$. Further, the distribution defined as $\pi(A \times B) = \mu(A \cap (T_0)^{-1}(B))$ for all Borel sets $A, B \subseteq \mathbb{R}^d$ is the unique minimizer in* (1.3) *provided $\mu, \nu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \cap \mathcal{P}_2(\mathbb{R}^d)$.*

Definition 1.1 (OT map and potential function)**.** *The function $T_0 : \mathbb{R}^d \to \mathbb{R}^d$ in Proposition 1.1 which satisfies $T_0 \# \mu = \nu$ will be called the OT map from μ to ν. The convex function $\varphi_0(\cdot)$ in Proposition 1.1 satisfying $\nabla \varphi_0 = T_0$ will be termed the OT potential.*

The next and final important ingredient is the alternate dual representation of (1.3) which gives:

$$\frac{1}{2} W_2^2(\mu, \nu) = \frac{1}{2}\int \|x\|^2 \, d\mu(x) + \frac{1}{2}\int \|y\|^2 \, d\nu(y) - \min_{f \in \mathcal{F}} \mathcal{S}_{\mu,\nu}(f), \qquad \text{where} \tag{1.4}$$

$$\mathcal{S}_{\mu,\nu}(f) = \int f \, d\mu + \int f^* \, d\nu. \tag{1.5}$$

Here $\mathcal{F}$ denotes the space of convex functions on $\mathbb{R}^d$ which are also elements of $L^1(\mu)$ and $f^*(\cdot)$ is the standard Legendre-Fenchel dual defined as:

$$f^*(x) := \sup_{y \in \mathbb{R}^d} [y^\top x - f(y)], \qquad \text{for } x \in \text{dom}(f). \tag{1.6}$$

1.2 Estimating OT map via barycentric projection

Recall the setting from the Introduction. Let $\widetilde{\mu}_m, \widetilde{\nu}_n \in \mathcal{P}_2(\mathbb{R}^d)$. Here $\widetilde{\mu}_m, \widetilde{\nu}_n$ need not be absolutely continuous and can be very general. Intuitively, $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ should be viewed as some empirical approximation of μ and ν respectively.

Example 1.2 (Simple choices of $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$)**.** *Let $X_1, \ldots, X_m \overset{i.i.d.}{\sim} \mu$ and $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \nu$; in which case a natural choice would be to set $\widetilde{\mu}_m = \widehat{\mu}_m$ and $\widetilde{\nu}_n = \widehat{\nu}_n$ where $\widehat{\mu}_m$ and $\widehat{\nu}_n$ are the empirical distributions on $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ respectively, as defined in* (1.2)*. This is the standard choice adopted in the discrete-discrete Kantorovich relaxation; see [94, Section 2.3]. Another popular choice is $\widetilde{\mu}_m = \widehat{\mu}_m$, $\widetilde{\nu}_n = \nu$ or $\widetilde{\mu}_m = \mu$, $\widetilde{\nu}_n = \widehat{\nu}_n$. This is the semi-discrete Kantorovich problem and is popular when one of the measures is fully specified; see [27, 54].* ■

A natural way to estimate $T_0(\cdot)$, as defined in (1.1), would be to approximate it using the OT map from $\widetilde{\mu}_m$ to $\widetilde{\nu}_n$. However as $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ may not be elements of $\mathcal{P}_{\text{ac}}(\mathbb{R}^d)$, Proposition 1.1 does not apply and an OT map *may not exist* from $\widetilde{\mu}_m$ to $\widetilde{\nu}_n$. Such is the case in Example 1.2 in the *discrete-discrete* case when $m \neq n$. To circumvent this issue, we leverage the notion of barycentric projections (see [3, Definition 5.4.2]) defined below:

Definition 1.2 (Barycentric projection)**.** *Define the set*

$$\widetilde{\Gamma}_{\min} := \operatorname*{argmin}_{\pi \in \Pi(\widetilde{\mu}_m, \widetilde{\nu}_n)} \int \|x - y\|^2 \, d\pi(x, y).$$

The optimization problem above is the plug-in analog of the optimization problem on the right hand side of (1.3)*. Given any $\gamma \in \widetilde{\Gamma}_{\min}$, define the barycentric projection of γ as the conditional mean of y given x under γ, i.e.,*

$$\widetilde{T}_{m,n}(x) \equiv \widetilde{T}^{\gamma}_{m,n}(x) := \frac{\int_y y \, d\gamma(x, y)}{\int_y d\gamma(x, y)}, \qquad \textit{for } x \in \textit{supp}\,(\widetilde{\mu}_m)\,. \tag{1.7}$$

In general, $\widetilde{\Gamma}_{\min}$ need not be a singleton which is why we index the barycentric projection $\widetilde{T}^{\gamma}_{m,n}(\cdot)$ by $\gamma \in \widetilde{\Gamma}_{\min}$. Note that $\widetilde{T}^{\gamma}_{m,n}(\cdot)$ need not be a transport map; however, if an OT map exists then it must be equal to $\widetilde{T}^{\gamma}_{m,n}(\cdot)$ ($\widetilde{\mu}_m$-a.e.). Our goal is to obtain stochastic upper bounds for

$$\sup_{\gamma \in \widetilde{\Gamma}_{\min}} \int \left\| \widetilde{T}^{\gamma}_{m,n}(x) - T_0(x) \right\|^2 d\widetilde{\mu}_m(x). \tag{1.8}$$

In addition, our proof techniques also yield rates of convergence for

$$\left| W_2^2(\widetilde{\mu}_m, \widetilde{\nu}_n) - W_2^2(\mu, \nu) \right|. \tag{1.9}$$

In this paper, we will focus on $d \geq 2$. Due to the canonical ordering of $\mathbb{R}$, the case $d = 1$ can be handled easily using the classical Hungarian embedding theorem [78].

1.3 Contributions

1. We provide a new and flexible stability estimate Theorem 2.1 which yields a unified approach to obtaining *rates* of convergence for general plug-in estimators of the OT map $T_0(\cdot)$. Unlike existing stability estimates, Theorem 2.1 holds for the barycentric projection (which is the same as the OT map when it exists) and does not require any smoothness assumptions on $\widetilde{\mu}_m$, $\widetilde{\nu}_n$ or $\widetilde{T}^{\gamma}_{m,n}(\cdot)$; also see Remark 2.1 for a comparison with the existing literature.
2. in Sections 2.1 and 2.2, we use Theorem 2.1 to bound (1.8) and (1.9):

- In Section 2.1, we show that in both the discrete-discrete and semi-discrete Kantorovich relaxation problems (see Example 1.2), the rate of convergence of (1.8) is $m^{-2/d} + n^{-2/d}$ for $d \geq 4$ when T_0 is assumed to be Lipschitz (see Theorem 2.2), which is the minimax rate (see [69, Theorem 6]). To the best of our knowledge, rates of convergence for these natural estimators weren't previously established in the literature.

- In Section 2.2, we show that the curse of dimensionality in the above rates can be mitigated provided μ and ν admit Hölder smooth (see Definition B.4) densities. Our plug-in estimator in this case is obtained by choosing $\widetilde{\mu}_m$ (and $\widetilde{\nu}_n$) as the convolution of $\widehat{\mu}_m$ (and $\widehat{\nu}_n$) and a smooth kernel with an appropriate bandwidth. Under this choice, the rate of convergence in (1.8) is $m^{-\left(\frac{s+2}{d} \wedge \frac{1}{2}\right)} + n^{-\left(\frac{s+2}{d} \wedge \frac{1}{2}\right)}$, where s denotes the degree of Hölder smoothness (see Theorem 2.4). Clearly, if $2(s+2) \geq d$, the rate of convergence becomes dimension-free and mitigates the curse of dimensionality. We also show the same rate of convergence also holds for (1.9) (see Proposition 2.5) which makes a strong case in favor of incorporating smoothness in the construction of plug-in estimators as was conjectured in [31].

3. In Section 2.3, we use a discretization technique from [117] to construct discrete approximations to the smoothed $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ from the previous paragraph that in turn yield exactly computable plug-in estimators for T_0 that also achieve the same statistical guarantees as the smoothed plug-in estimator from Section 2.2 (see Theorem 2.6). However the number of atoms required in the discretizations and correspondingly the computational complexity increases with the degree of Hölder smoothness; this highlights a statistical and computational tradeoff.

4. We provide implications of our results in popular applications of OT such as estimating the barycenter of two multivariate probability distributions (see Theorem 3.1 in Section 3.1) and in nonparametric independence testing (see Theorem 3.3 in Section 3.2).

1.4 Related work

Many recent works have focused on obtaining consistent estimators of T_0 using the plug-in principle, see [27, 54] (in the semi-discrete problem) and [41, 66, 118] (in the discrete-discrete problem). In [54], the authors have studied the rate of convergence of the semi-discrete optimal transport map from ν (absolutely continuous) to $\widehat{\mu}_m$. This paper complements the aforementioned papers by studying the rates of convergence for general plug-in estimators in a unified fashion. In two other papers [8, Theorem 1.1] and [82, Section 4], the authors use a "Voronoi tessellation" approach to estimate T_0, however the rates obtained in this paper, even in the absence of smoothness, are strictly better than those in [8, 82]. Perhaps the most closely related paper to ours would be [64]. In [64], the author uses variational techniques to arrive at stability estimates while we exploit the Lipschitz nature of the OT map (see Definition 1.1). Further the rates in this paper have exponents $\frac{s+2}{d} \wedge \frac{1}{2}$ which are *strictly better* than the exponents $\frac{s+2}{2(s+2)+d}$ obtained in [64, Proposition 1], where s is the degree of Hölder smoothness of the densities of μ, ν (see Definition B.4). In another line of work [69], the authors use theoretical wavelet based estimators (not of the plug-in type) of T_0 to obtain nearly minimax optimal rates of convergence. However these estimators are computationally infeasible and hard to approximate. In contrast, our goal is not to obtain incomputable minimax optimal estimators but instead, we focus on obtaining rates of convergence for plug-in estimators that are a lot more popular in computational OT [6, 31, 64, 88, 89, 93, 104]. Some of our estimators (see Section 2.3) can, for example, be provably constructed in polynomial time unlike the wavelet estimator in [69].

In terms of obtaining rates of convergence for (1.9), some attempts include [98, 104] where parametric rates are obtained when μ, ν are known to be finitely supported or are both Gaussian. In a related problem, bounds for $W_2^2(\widehat{\mu}_m, \mu)$ were obtained in [5, 42, 48, 110]. Using these bounds, it is easy to get a $n^{-1/d}$ rate of convergence for (1.9). This rate was recently improved to $n^{-2/d}$ in [31] under no smoothness assumptions. Our rates coincide with the $n^{-2/d}$ rate from [31] under no smoothness assumptions. But further, we show in this paper that the curse of dimensionality in the above rate can be mitigated by incorporating smoothness into the plug-in procedure. In another line of research, theoretical wavelet based estimators have been used in [92, 117] to estimate $W_2^2(\mu, \nu)$ but these estimators are difficult to construct or approximate in practice.

2 Main results

Recall the definition of $\varphi_0(\cdot)$ from Definition 1.1. The following is our main result.

Theorem 2.1 (Stability estimate)**.** *Suppose that $\mu, \nu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d) \cap \mathcal{P}_2(\mathbb{R}^d)$ and $\widetilde{\mu}_m, \widetilde{\nu}_n \in \mathcal{P}_2(\mathbb{R}^d)$. Assume that $T_0(\cdot)$ (as defined in* (1.1)*) is L-Lipschitz ($L > 0$). Then,*

$$\sup_{\gamma \in \widetilde{\Gamma}_{\min}} \int \|\widetilde{T}^{\gamma}_{m,n}(x) - T_0(x)\|^2 \, d\widetilde{\mu}_m(x) \leq L \max\left\{ \left| \int \Psi^*_{\widetilde{\mu}_m, \widetilde{\nu}_n} \, d(\widetilde{\nu}_n - \overline{\nu}_m) \right|, \left| \int \Psi^*_{\widetilde{\mu}_m, \overline{\nu}_m} \, d(\widetilde{\nu}_n - \overline{\nu}_m) \right| \right\}$$
$$+ 2L \int \varphi^*_0(y) \, d(\widetilde{\nu}_n - \overline{\nu}_m)(y), \tag{2.1}$$

where $\overline{\nu}_m = T_0 \# \widetilde{\mu}_m$, $\varphi^*_0(\cdot)$ is defined as in (1.6)*, and with $\mathcal{S}_{\cdot,\cdot}(\cdot)$ defined as in* (1.5)*, $\Psi_{\widetilde{\mu}_m, \widetilde{\nu}_n}(\cdot) := \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{S}_{\widetilde{\mu}_m, \widetilde{\nu}_n}(f)$, $\quad \Psi_{\widetilde{\mu}_m, \overline{\nu}_m}(\cdot) := \operatorname{argmin}_{f \in \mathcal{F}} \mathcal{S}_{\widetilde{\mu}_m, \overline{\nu}_m}(f)$.*

The proof of Theorem 2.1 (see Appendix A.1) starts along the same lines as the proof of the curvature estimate in [55, Proposition 3.3]. This is followed by some careful manipulations of $W_2^2(\cdot, \cdot)$ (as in (1.3)) and an application of the conditional version of Jensen's inequality, see (A.3). The final step of the proof uses the dual representation in (1.4) with techniques similar to some intermediate steps in the proof of [87, Proposition 2] and [31, Lemma 3].

Remark 2.1 (Comparison with other stability estimates)**.** *Theorem 2.1 provides some important advantages to existing stability estimates in the literature. One of the earliest results in this direction can be found in [55, Proposition 3.3] but their bound involves a push-forward constraint which makes it hard to use for rate of convergence analysis. A bound similar to Theorem 2.1 is presented in [54, Lemma 5.1] but there the authors assume the existence of an OT map from $\widetilde{\mu}_m$ to $\widetilde{\nu}_n$. Therefore, it does not apply to the discrete-discrete problem where $\widetilde{\mu}_m = \widehat{\mu}_m$ and $\widetilde{\nu}_n = \widehat{\nu}_n$ with $m \neq n$. Overcoming all these limitations is an important contribution of Theorem 2.1 and allows us to deal with popular plug-in estimators all in one go. The stability estimate in [69, Proposition 10] on the other hand requires $\widetilde{\mu}_m$, $\widetilde{\nu}_n$ to be sufficiently smooth and hence it does not hold for discrete-discrete or semi-discrete plug-in estimators (see Example 1.2). Further their result requires all the measures involved to be compactly supported unlike the much milder requirements of Theorem 2.1. However, a shortcoming of Theorem 2.1 is that it is hard to obtain rates faster than $n^{-1/2}$ using it directly, whereas [69] can obtain rates arbitrarly close to n^{-1}. This is a price we pay for analyzing natural and popular plug-in estimators as opposed to the (incomputable) wavelet based estimators in [69].*

Remark 2.2 (*How to use Theorem 2.1 to obtain rates of convergence?*)**.** *Note that the second term on the right hand side of* (2.1)*, under appropriate moment assumptions, is $O_p(m^{-1/2} + n^{-1/2})$ (free of dimension) by a direct application of Markov's inequality. We therefore focus on the first term. By* (1.5)*, $\Psi^*_{\widetilde{\mu}_m, \widetilde{\nu}_n}(\cdot)$, $\Psi^*_{\widetilde{\mu}_m, \overline{\nu}_m}(\cdot) \in \mathcal{F}$. Further, by Caffarelli's regularity theory [21–23], depending on the "smoothness" of $\widetilde{\mu}_m$, $\widetilde{\nu}_n$, it can be shown that there exists a further class of functions $\mathcal{F}_s$ (see Remarks B.3 and B.5) such that $\Psi^*_{\widetilde{\mu}_m, \widetilde{\nu}_n}(\cdot)$, $\Psi^*_{\widetilde{\mu}_m, \overline{\nu}_m}(\cdot) \in \mathcal{F} \cap \mathcal{F}_s$. Thus, we can bound the first term on the right hand side of* (2.1) *as:*

$$\max\left\{ \left| \int \Psi^*_{\widetilde{\mu}_m, \widetilde{\nu}_n} \, d(\widetilde{\nu}_n - \overline{\nu}_m) \right|, \left| \int \Psi^*_{\widetilde{\mu}_m, \overline{\nu}_m} \, d(\widetilde{\nu}_n - \overline{\nu}_m) \right| \right\} \leq \sup_{f \in \mathcal{F} \cap \mathcal{F}_s} \left| \int f \, d(\widetilde{\nu}_n - \overline{\nu}_m) \right|. \tag{2.2}$$

The right hand side of (2.2) *can now be bounded using the corresponding Dudley's entropy integral bounds using empirical process techniques, see [112, Lemmas 19.35-19.37].*

To conclude, the two main steps in our strategy are identifying the family of functions $\mathcal{F}_s$ and computing Dudley's entropy integral. Further, the more the smoothness of $\widetilde{\mu}_m$, $\widetilde{\nu}_n$, the smaller is the class of functions $\mathcal{F}_s$ and smaller the supremum on the right hand side of (2.2). This shows why better rates can be expected under smoothness assumptions.

2.1 Natural non-smooth plug-in estimator

In this case, we discuss the rates of convergence for the discrete-discrete problem and the semi-discrete problem, where *no smoothness* is available on $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$.

Theorem 2.2. *Suppose that $T_0(\cdot)$ is L-Lipschitz, ν is compactly supported and $\mathbb{E}\exp(t|X_1|^\alpha) < \infty$ for some $t > 0$, $\alpha > 0$.*

(Discrete-discrete): *Set $\widetilde{\mu}_m = \widehat{\mu}_m$ and $\widetilde{\nu}_n = \widehat{\nu}_n$. Then the following holds:*

$$\sup_{\gamma \in \widetilde{\Gamma}_{\min}} \int \|\widetilde{T}^{\gamma}_{m,n}(x) - T_0(x)\|^2 \, d\widetilde{\mu}_m(x) = O_p\left(r_d^{(m,n)} \times (\log(1 + \max\{m,n\}))^{(d+2)/\alpha}\right), \quad (2.3)$$

$$\textit{where} \quad r_d^{(m,n)} := \begin{cases} m^{-1/2} + n^{-1/2} & \textit{for } d = 2,3, \\ m^{-1/2}\log(1+m) + n^{-1/2}\log(1+n) & \textit{for } d = 4, \\ m^{-2/d} + n^{-2/d} & \textit{for } d \geq 5. \end{cases} \quad (2.4)$$

The same bound holds for $|W_2^2(\widetilde{\mu}_m, \widetilde{\nu}_n) - W_2^2(\mu,\nu)|$ without assuming $T_0(\cdot)$ is Lipschitz.

(Semi-discrete): *Set $\widetilde{\mu}_m = \mu$, $\widetilde{\nu}_n = \widehat{\nu}_n$ or $\widetilde{\mu}_m = \widehat{\mu}_m$, $\widetilde{\nu}_n = \nu$. Then the left hand side of* (2.3) *is $O_p(r_d^{(n,n)} \times (\log(1+n))^{(d+2)/\alpha})$ or $O_p(r_d^{(m,m)} \times (\log(1+m))^{(d+2)/\alpha})$ respectively.*

A stronger result can be proved if both μ and ν are compactly supported.

Corollary 2.3. *Consider the setting from Theorem 2.2 and assume μ is compactly supported. Then,*

$$\mathbb{E}\left[\sup_{\gamma \in \widetilde{\Gamma}_{\min}} \int \|\widetilde{T}^{\gamma}_{m,n}(x) - T_0(x)\|^2 \, d\widetilde{\mu}_m(x)\right] \leq C r_d^{(m,n)},$$

for some constant $C > 0$, in both the discrete-discrete and semi-discrete settings from Theorem 2.2.

A brief description of the proof technique for Theorem 2.2 using Theorem 2.1 is provided in Remark B.3, and the actual proof is presented in Appendix A.1. As discussed in Section 1.3, the exponents obtained in Theorem 2.2 are minimax optimal under bare minimal smoothness assumptions (see [69, Theorem 6]). To the best of our knowledge, rates for the discrete-discrete case for $m \neq n$ and those for the semi-discrete case were not known previously in the literature. Our rates are also strictly better than those (for different estimators, based on space tessellations) obtained in [8, 82] and require less stringent assumptions than those in [31]. In the next section, we show how smoothness assumptions can be leveraged to mitigate the curse of dimensionality in Theorem 2.2.

2.2 Smooth plug-in estimator: mitigating the curse of dimensionality

For this section, assume that μ and ν admit Hölder smooth densities f_μ and f_ν (see Definition B.4). Given $\Omega \subseteq \mathbb{R}^d$ and $s > 0$, let $C^s(\Omega)$ denote the set of Hölder smooth functions on Ω of order s.

Assumption (A1) (Regularity of densities)**.** *Suppose that*

1. *f_μ and f_ν are supported on compact and convex subsets of $\mathbb{R}^d$, say $\mathcal{X}$ and $\mathcal{Y}$ respectively.*
2. *There exists $s \geq 0$ such that $f_\mu \in C^s(\mathcal{X}^\circ)$ and $f_\nu \in C^s(\mathcal{Y}^\circ)$, where $\mathcal{X}^\circ$ and $\mathcal{Y}^\circ$ are the interiors of $\mathcal{X}$ and $\mathcal{Y}$ respectively.*

We now define our estimators for μ and ν using the standard kernel density estimation technique (see [111, Section 1.2]). Set

$$\widehat{f}_\mu(x) := \frac{1}{nh_n^d} \sum_{i=1}^n K_d\left(\frac{X_i - x}{h_n}\right), \quad (2.5)$$

for some bandwidth parameter $h_n > 0$ and d-variate kernel $K_d(\cdot)$. We assume that $K_d(\cdot)$ is the d-fold product of a univariate kernels, i.e., there exists a kernel $K(\cdot)$ such that for $u = (u_1, \ldots, u_d) \in \mathbb{R}^d$, $K_d(u) = \prod_{i=1}^d K(u_i)$. We define $\widehat{f}_\nu(\cdot)$ similarly with the same univariate kernel and bandwidth.

Assumption (A2) (Regularity of the kernel)**.** *Assume that $K(\cdot)$ is a nonnegative, bounded, infinitely differentiable (in the interior of its support) kernel with all derivatives bounded and satisfies $\int uK(u)\,du = 0$, e.g., the popular Epanechnikov kernel given by $K(u) := (3/4)(1-u^2)\mathbb{1}(|u| \leq 1)$.*

Theorem 2.4. *Assume that the condition in Theorem 2.1 holds. Suppose that f_μ, f_ν are Lebesgue densities bounded above and below by positive constants on their supports, that also satisfy Assumption (A1). Let $K(\cdot)$ satisfy Assumption (A2). For the bandwidth parameter h_n, we will assume that $h_n \to 0$ and $n(\log n)^{-1}h_n^{d+2s} \to \infty$. Set $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ as the probability measures corresponding to the probability densities $\widehat{f}_\mu$ and $\widehat{f}_\nu$. Then the following holds:*

$$\sup_{\gamma \in \widetilde{\Gamma}_{\min}} \int \|\widetilde{T}^{\gamma}_{m,n}(x) - T_0(x)\|^2 \, d\widetilde{\mu}_m(x) = O_p(r_{d,s}^{(m,n)}),$$

$$\textit{where} \quad r_{d,s}^{(m,n)} := \begin{cases} m^{-1/2} + n^{-1/2} & \textit{for } d < 2(s+2), \\ m^{-1/2}\log(1+m) + n^{-1/2}\log(1+n) & \textit{for } d = 2(s+2), \\ m^{-\frac{s+2}{d}} + n^{-\frac{s+2}{d}} & \textit{for } d \geq 2(s+2). \end{cases}$$

The same bound also holds for $|W_2^2(\widetilde{\mu}_m, \widetilde{\nu}_n) - W_2^2(\mu, \nu)|$.

As $\widetilde{\mu}_m, \widetilde{\nu}_n$ are smooth, $\widetilde{\Gamma}_{\min}$ is a singleton and the supremum in Theorem 2.4 can be dropped. A brief description of the proof technique for Theorem 2.4 is presented in Remark B.5 and the actual proof is given in Appendix A.1. Theorem 2.4 shows that, under enough smoothness, i.e., when $2(s+2) > d$, both the upper bounds for (1.8) and (1.9) are $O_p(n^{-1/2})$. This shows that, for large dimensions, provided μ and ν admit smooth enough densities, it is possible to construct plug-in estimators that mitigate the curse of dimensionality. Note that the same estimator was analyzed in [64, Proposition 1] when $m = n$. However, the rates obtained in Theorem 2.4 are *strictly better* than those in [64, Proposition 1]. For $m = n$, when $d < 2(s+2)$, [64] obtained a rate of $n^{-\frac{s+2}{2(s+2)+d}}$ which is *worse* than $n^{-1/2}$ obtained in Theorem 2.4. For the other regimes, [64] obtains rates (up to log factors) of $n^{-1/4}$ and $n^{-\frac{1}{(s+2)(d+2(s+2))}}$ which are both worse than the respective rates of $n^{-1/2}$ and $n^{-\frac{s+2}{d}}$ in Theorem 2.4. Another important implication of Theorem 2.4 is the bound obtained on $|W_2(\widetilde{\mu}_m, \widetilde{\nu}_n) - W_2(\mu, \nu)|$ when $\mu \neq \nu$. We first present the result and then describe the implication.

Proposition 2.5. *Consider the setting in Theorem 2.4. Then, provided $\mu \neq \nu$, the following holds:*

$$|W_2(\widetilde{\mu}_m, \widetilde{\nu}_n) - W_2(\mu, \nu)| = O_p(r_{d,s}^{(m,n)}).$$

Proposition 2.5 (see Appendix A.1 for a proof) shows an interesting distinction between the $\mu \neq \nu$ case and the $\mu = \nu$ case. For $\mu = \nu$, the best possible exponent is $n^{-\frac{1+s}{2s+d}}$ for $d \geq 3$ (see [117, Theorem 3] where the result was established under more general Besov smoothness assumptions). On the contrary, when $\mu \neq \nu$, Proposition 2.5 establishes a rate of $n^{-\frac{s+2}{d}}$ for the Wasserstein distance which is *strictly better* than the minimax achievable rate mentioned above when $\mu = \nu$.

2.3 Discretized plug-in estimator under smoothness assumptions

In Section 2.1, we discussed how smoothness can be incorporated into the plug-in procedure to get faster rates of convergence. Although such plug-in estimators are popular in the computational OT literature (see [6, 7, 26, 36]), $\widetilde{T}_{m,n}^{\gamma}$ as in Theorem 2.4 cannot be computed explicitly from data. This is in contrast to $\widetilde{T}_{m,n}^{\gamma}$ from Theorem 2.2 in the *discrete-discrete* case which is explicitly computable using a standard linear program, but achieves worse rates of convergence. This is not unexpected. Thanks to the *no free lunch* principle, better statistical accuracy is naturally accompanied by heavier computational challenges. Therefore, our goal here is to construct estimators, under smoothness assumptions as in Section 2.2, which are *explicitly* computable in polynomial time (with complexity increasing with smoothness) and also attain the rates of convergence as in Theorem 2.4.

Construction: Recall the kernel density estimators $\widehat{f}_\mu$ and $\widehat{f}_\nu$ (see (2.5)). Sample $M \geq 1$ random points from both $\widehat{f}_\mu$ and $\widehat{f}_\nu$. Let $\widehat{\mu}_{m,M}$ and $\widehat{\nu}_{n,M}$ denote the standard empirical measures on the M points sampled from $\widehat{f}_\mu$ and $\widehat{f}_\nu$ respectively. Finally construct $\widetilde{T}_{m,n} \equiv \widetilde{T}_{m,n}^{\gamma}$ as in Definition 1.2 with $\widetilde{\mu}_m = \widehat{\mu}_{m,M}$ and $\widetilde{\nu}_n = \widehat{\nu}_{n,M}$. It should be pointed out that a similar construction was also used in [117, Section 6]. Based on this construction, the main result of this section is as follows:

Theorem 2.6. *Consider the setting in Theorem 2.4 and the same construction of $\widetilde{T}_{m,n}^{\gamma}$ as above. For simplicity, let's also assume $m = n$. Accordingly set $M = n^{\frac{s+2}{2}}$. Then $\widetilde{\Gamma}_{\min}$ is a singleton and the following conclusion holds:*

$$\int \|\widetilde{T}_{m,n}(x) - T_0(x)\|^2 \, d\widetilde{\mu}_m(x) = O_p(r_{d,s}^{(n,n)}).$$

The same rates also hold for $|W_2^2(\widetilde{\mu}_m, \widetilde{\nu}_n) - W_2^2(\mu, \nu)|$.

The proof of Theorem 2.6 is given in Appendix A.1. An explicit computation of $\widetilde{T}_{m,n}$ as described above requires $O(M^3) = O(n^{\frac{3(s+2)}{2}})$ steps using the *Hungarian algorithm*, see [70]. This highlights the statistical versus computational tradeoff, i.e., in order to mitigate the curse of dimensionality in convergence rates by exploiting smoothness, the computational complexity gets progressively worse by polynomial factors in n. It should be mentioned that faster algorithms approximating $\widetilde{T}_{m,n}$ can be found in [1, 36, 52] to name a few. Due to space constraints, we avoid a detailed discussion on this.

In the above construction, sampling from the smoothed kernel densities $\widehat{f}_\mu$ and $\widehat{f}_\nu$ is crucial. If we would simply draw M bootstrap samples from the empirical distributions $\widehat{\mu}_m$ and $\widehat{\nu}_n$, the rates of convergence wouldn't improve from those observed in Theorem 2.2 no matter how large M is chosen.

3 Applications

In this section, we will apply our results to two popular problems, namely — estimating the Wasserstein barycenter between two probability distributions (see [2, 13, 25, 37]) in Section 3.1, and obtaining detection thresholds in some recent optimal transport based independence testing procedures (see [40, 41, 54, 102, 103]) in Section 3.2.

3.1 Wasserstein barycenter estimation

Let $\mu, \nu \in \mathcal{P}_2(\mathbb{R}^d)$. The Wasserstein barycenter between μ and ν is then given by:

$$\rho_0 := \min_{\rho \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)} \left(\frac{1}{2} W_2^2(\mu, \rho) + \frac{1}{2} W_2^2(\rho, \nu) \right). \tag{3.1}$$

In fact, by Proposition 1.1, there exists an optimal transport map T_0 from μ to ν and by [2, 13, 16], an alternative characterization of ρ_0 is as follows:

$$\rho_0 = \left(\frac{1}{2}\mathrm{Id} + \frac{1}{2} T_0 \right) \#\mu, \qquad \text{where} \qquad \mathrm{Id}(x) = x. \tag{3.2}$$

Estimating ρ_0 as in (3.1) has attracted significant attention over the past few years in economics [24, 29], Bayesian learning [106, 107], dynamic formulations [28, 33], algorithmic fairness [32, 59], etc. The most natural strategy employed in estimating ρ_0 is to use the empirical plug-in estimator, i.e., replacing μ, ν in (3.1) with $\widehat{\mu}_m, \widehat{\nu}_n$. This strategy has been used, approximated and analyzed extensively in e.g., [16, 25, 37, 80]. Based on (3.2), the natural plug-in estimator of ρ_0 would be:

$$\widehat{\rho}_0^\gamma = \left(\frac{1}{2}\mathrm{Id} + \frac{1}{2} \widetilde{T}_{m,n}^\gamma \right) \#\widetilde{\mu}_m \tag{3.3}$$

where $\widetilde{T}_{m,n}^\gamma$ is the plug-in estimator of T_0 obtained by solving (1.7), with μ and ν replaced by $\widetilde{\mu}_m$ and $\widehat{\nu}_n$ respectively and $\gamma \in \widetilde{\Gamma}_{\min}$. While the consistency of $\widehat{\rho}_0^\gamma$ has been analyzed for $m = n$ in [80] and rates have been obtained for $d = 1$ in [12], the more general question of obtaining rates of convergence for $\widehat{\rho}_0^\gamma$ for general dimensions $d \geq 1$ is yet unanswered. We address this question in the following result (see Appendix A.2 for a proof).

Theorem 3.1. *Suppose that the same assumptions from Theorem 2.2 hold. Then, with $\widehat{\rho}_0^\gamma$, as defined in* (3.3) *and $r_d^{(m,n)}$ defined via* (2.4)*, the following holds:*

$$\sup_{\gamma \in \widetilde{\Gamma}_{\min}} W_2^2(\widehat{\rho}_0^\gamma, \rho_0) = O_p\left(r_d^{(m,n)} \times (\log(1 + \max\{m, n\}))^{(d+2)/\alpha} \right).$$

3.2 Nonparametric independence testing: Optimal transport based Hilbert-Schmidt independence criterion

Let $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{i.i.d.}{\sim} \pi$, a probability measure on $\mathbb{R}^{d_1+d_2}$, with marginals $\mu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^{d_1})$ and $\nu \in \mathcal{P}_{\mathrm{ac}}(\mathbb{R}^{d_2})$. Our problem of interest is the following hypothesis testing problem, given as:

$$\mathrm{H}_0 : \pi = \mu \otimes \nu \qquad \text{versus} \qquad \mathrm{H}_1 : \pi \neq \mu \otimes \nu. \tag{3.4}$$

This is the classical nonparametric independence testing problem which has received a lot of attention in the statistics and machine learning literature (see [10, 60, 67, 109], and [43, 71] for a review).

In keeping with the overall theme of this paper, our focus here will be on a large class of OT based independence testing procedures, introduced first in [41] followed by recent developments in [40, 102, 103]. These tests bear resemblance to the Hilbert-Schmidt independence criterion (HSIC); see [60–62] and have attractive properties such as distribution-freeness (see Proposition 3.2), consistency without moment assumptions and robustness against heavy-tailed distributions and against contamination [41, 102]. Below, we describe this class of tests, see (3.5) and (3.7). Our main theoretical contribution of this section will be to provide detection thresholds of these OT based tests.

Construction: Suppose υ_1, υ_2 be two compactly supported probability distributions on $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ respectively (e.g., $\upsilon_1 \equiv \text{Unif}[0,1]^{d_1}$, $\upsilon_2 \equiv \text{Unif}[0,1]^{d_2}$). Let $U_1, \ldots, U_n \overset{i.i.d.}{\sim} \upsilon_1$, $V_1, \ldots, V_n \overset{i.i.d.}{\sim} \upsilon_2$, $\widehat{u}_n := n^{-1}\sum_{i=1}^n \delta_{U_i}$ and $\widehat{v}_n := n^{-1}\sum_{j=1}^n \delta_{V_j}$. Recall the definitions of $\widehat{\mu}_n$ (with $m = n$) and $\widehat{\nu}_n$ from (1.2). Let $\widehat{T}_{1,n}$ ($\widehat{T}_{2,n}$) be obtained by solving (1.7), with μ and ν replaced by $\widehat{\mu}_n$ and $\widehat{u}_n$ ($\widehat{\nu}_n$ and $\widehat{v}_n$) respectively. Consider two nonnegative definite, continuous, characteristic kernels (see [51, 105] for definitions) $K_1(\cdot,\cdot)$ and $K_2(\cdot,\cdot)$ on $(\text{supp}(\upsilon_1))^2$ and $(\text{supp}(\upsilon_2))^2$. Set $\widehat{x}_{ij} := K_1(\widehat{T}_{1,n}(X_i), \widehat{T}_{1,n}(X_j))$ and $\widehat{y}_{ij} := K_2(\widehat{T}_{2,n}(Y_i), \widehat{T}_{2,n}(Y_j))$. Our test statistic is as follows:

$$\widehat{\text{rHSIC}} := n^{-2}\sum_{i,j}\widehat{x}_{ij}\widehat{y}_{ij} + n^{-4}\sum_{i,j,r,s}\widehat{x}_{ij}\widehat{y}_{rs} - 2n^{-3}\sum_{i,j,r}\widehat{x}_{ij}\widehat{y}_{ir}. \tag{3.5}$$

Proposition 3.2 (See [41, 103])**.** *1.* ***(Distribution-freeness)*** *When X_1 and Y_1 are independent, the distribution of $n \times \widehat{\text{rHSIC}}$ is universal, i.e., it does not depend on μ and ν for every fixed n.*

2. ***(Consistency against fixed alternatives)*** *Let $c_{n,\alpha}$ be the upper $(1-\alpha)$-th quantile from the universal distribution in part 1 above. Then $\widehat{\text{rHSIC}} \xrightarrow{P} \text{rHSIC}(\pi|\mu\otimes\nu)$ where*

$$\begin{aligned}\text{rHSIC}(\pi|\mu\otimes\nu) &:= \mathbb{E}[K_1(T_1(X_1),T_1(X_2))K_2(T_2(Y_1),T_2(Y_2))] + \mathbb{E}[K_1(T_1(X_1),T_1(X_2))] \\ &\times \mathbb{E}[K_2(T_2(Y_1),T_2(Y_2))] - 2\mathbb{E}[K_1(T_1(X_1),T_1(X_2))K_2(T_2(Y_1),T_2(Y_3))],\end{aligned} \tag{3.6}$$

where $T_1(\cdot)$ (respectively $T_2(\cdot)$) is the optimal transport map from μ (ν) to υ_1 (υ_2); see Definition 1.1. Further $\text{rHSIC}(\pi|\mu\otimes\nu) = 0$ if and only if $\pi = \mu\otimes\nu$. Define the following test function:

$$\phi_{n,\alpha} := \mathbb{1}(n \times \widehat{\text{rHSIC}} \geq c_{n,\alpha}). \tag{3.7}$$

Then $\mathbb{E}[\phi_{n,\alpha}] \to 1$ as $n \to \infty$ under H_1, i.e., when $\pi \neq \mu\otimes\nu$.

Proposition 3.2 shows that the test based on $\widehat{\text{rHSIC}}$ (see (3.5)), i.e., $\phi_{n,\alpha}$ (see (3.7)), can be carried out without resorting to the permutation principle as is necessary for the usual HSIC based test (see [61]). Further, when the sampling distribution is fixed, Proposition 3.2 shows that $\widehat{\text{rHSIC}}$ consistently estimates $\text{rHSIC}(\pi|\mu\otimes\nu)$, a quantity which equals 0 if and only if $\pi = \mu\otimes\nu$ (this yields the consistency of $\phi_{n,\alpha}$) against fixed alternatives.

While consistency against fixed alternatives is an attractive feature of $\phi_{n,\alpha}$, a more intricate question of statistical interest is to understand the local power of $\phi_{n,\alpha}$ under "changing sequence of alternatives converging to the null" as $n \to \infty$. To study the local power of $\phi_{n,\alpha}$, we need to consider a triangular array setting, where the data distribution changes with n, i.e., $(X_1,Y_1),\ldots,(X_n,Y_n) \overset{i.i.d.}{\sim} \pi^{(n)}$, a probability measure on $\mathbb{R}^{d_1+d_2}$, with marginals $\mu^{(n)} \in \mathcal{P}_{\text{ac}}(\mathbb{R}^{d_1})$ and $\nu^{(n)} \in \mathcal{P}_{\text{ac}}(\mathbb{R}^{d_2})$. As $\text{rHSIC}(\cdot|\cdot)$ characterizes independence, a mathematical formulation of "alternatives converging to null" would be to say $\text{rHSIC}(\pi^{(n)}|\mu^{(n)}\otimes\nu^{(n)}) \to 0$ as $n \to \infty$. Similar questions have attracted a lot of attention in modern statistics, featuring measures (other than $\text{rHSIC}(\cdot|\cdot)$) which characterize independence, see e.g., [4, 10, 75, 81]. In the following result (see Appendix A.2 for a proof), we show that if $\text{rHSIC}(\pi^{(n)}|\mu^{(n)}\otimes\nu^{(n)}) \to 0$ slowly enough with n, then $\phi_{n,\alpha}$ yields a consistent sequence of tests for problem (3.4).

Theorem 3.3. *Consider problem* (3.4) *with $\pi^{(n)}$, $\mu^{(n)}$ and $\nu^{(n)}$ (changing with n). Suppose that $T_{1,n}(\cdot)$ and $T_{2,n}(\cdot)$ are both L-Lipschitz (L is free of n). Also assume that $K_1(\cdot)$, $K_2(\cdot)$ are Lipschitz, $\mu^{(n)}$, $\nu^{(n)}$ are supported on fixed compact sets (the supports don't change with n). Then,*

$$\mathbb{E}[\phi_{n,\alpha}] \to 1 \qquad \textit{if} \qquad (r_d^{(n,n)})^{-1/2} \times \text{rHSIC}(\pi^{(n)}|\mu^{(n)}\otimes\nu^{(n)}) \to \infty.$$

References

[1] Pankaj K Agarwal and R Sharathkumar. Approximation algorithms for bipartite matching with metric and geometric costs. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 555–564, 2014.

[2] Martial Agueh and Guillaume Carlier. Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, 43(2):904–924, 2011.

[3] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows in metric spaces and in the space of probability measures*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008.

[4] Arnab Auddy, Nabarun Deb, and Sagnik Nandy. Exact detection thresholds for chatterjee's correlation. *arXiv preprint arXiv:2104.15140*, 2021.

[5] Franck Barthe and Charles Bordenave. Combinatorial optimization over two random point sets. In *Séminaire de Probabilités XLV*, volume 2078 of *Lecture Notes in Math.*, pages 483–535. Springer, Cham, 2013.

[6] Jean-David Benamou and Yann Brenier. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

[7] Jean-David Benamou, Brittany D Froese, and Adam M Oberman. Numerical solution of the optimal transportation problem using the monge–ampère equation. *Journal of Computational Physics*, 260:107–126, 2014.

[8] Robert J Berman. Convergence rates for discretized monge–ampère equations and quantitative stability of optimal transport. *Foundations of Computational Mathematics*, pages 1–42, 2020.

[9] Espen Bernton, Pierre E Jacob, Mathieu Gerber, and Christian P Robert. Inference in generative models using the Wasserstein distance. *arXiv preprint arXiv:1701.05146*, 1(8):9, 2017.

[10] T. B. Berrett and R. J. Samworth. Nonparametric independence testing via mutual information. *Biometrika*, 106(3):547–566, 2019.

[11] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo López, et al. Geodesic PCA in the Wasserstein space by convex PCA. In *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, volume 53, pages 1–26. Institut Henri Poincaré, 2017.

[12] Jérémie Bigot, Raúl Gouet, Thierry Klein, Alfredo Lopez, et al. Upper and lower risk bounds for estimating the wasserstein barycenter of random measures on the real line. *Electronic journal of statistics*, 12(2):2253–2289, 2018.

[13] Jérémie Bigot and Thierry Klein. Characterization of barycenters in the wasserstein space by averaging optimal transport maps. *ESAIM: Probability and Statistics*, 22:35–57, 2018.

[14] Adrien Blanchet and Guillaume Carlier. Optimal transport and cournot-nash equilibria. *Mathematics of Operations Research*, 41(1):125–145, 2016.

[15] Melf Boeckel, Vladimir Spokoiny, and Alexandra Suvorikova. Multivariate brenier cumulative distribution functions and their application to non-parametric testing. *arXiv preprint arXiv:1809.04090*, 2018.

[16] Emmanuel Boissard, Thibaut Le Gouic, Jean-Michel Loubes, et al. Distribution's template estimate with wasserstein metrics. *Bernoulli*, 21(2):740–759, 2015.

[17] 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.

[18] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities*. Oxford University Press, Oxford, 2013. A nonasymptotic theory of independence, With a foreword by Michel Ledoux.

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

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

[21] Luis A. Caffarelli. Boundary regularity of maps with convex potentials. *Comm. Pure Appl. Math.*, 45(9):1141–1151, 1992.

[22] Luis A. Caffarelli. The regularity of mappings with a convex potential. *J. Amer. Math. Soc.*, 5(1):99–104, 1992.

[23] Luis A. Caffarelli. Boundary regularity of maps with convex potentials. II. *Ann. of Math. (2)*, 144(3):453–496, 1996.

[24] Guillaume Carlier and Ivar Ekeland. Matching for teams. *Economic theory*, 42(2):397–418, 2010.

[25] Guillaume Carlier, Adam Oberman, and Edouard Oudet. Numerical methods for matching for teams and wasserstein barycenters. *ESAIM: Mathematical Modelling and Numerical Analysis*, 49(6):1621–1642, 2015.

[26] Rick Chartrand, Brendt Wohlberg, Kevin Vixie, and Erik Bollt. A gradient descent solution to the monge-kantorovich problem. *Applied Mathematical Sciences*, 3(22):1071–1080, 2009.

[27] Victor Chernozhukov, Alfred Galichon, Marc Hallin, and Marc Henry. Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.*, 45(1):223–256, 2017.

[28] Sinho Chewi, Tyler Maunu, Philippe Rigollet, and Austin J Stromme. Gradient descent algorithms for bures-wasserstein barycenters. In *Conference on Learning Theory*, pages 1276–1304. PMLR, 2020.

[29] Pierre-André Chiappori, Robert J McCann, and Lars P Nesheim. Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness. *Economic Theory*, 42(2):317–354, 2010.

[30] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. Scaling algorithms for unbalanced optimal transport problems. *Mathematics of Computation*, 87(314):2563–2609, 2018.

[31] Lenaic Chizat, Pierre Roussillon, Flavien Léger, François-Xavier Vialard, and Gabriel Peyré. Faster Wasserstein distance estimation with the sinkhorn divergence. *Advances in Neural Information Processing Systems*, 33, 2020.

[32] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. In *NeurIPS 2019 - 33th Annual Conference on Neural Information Processing Systems*, Vancouver, Canada, December 2019.

[33] Sebastian Claici, Edward Chien, and Justin Solomon. Stochastic wasserstein barycenters. In *International Conference on Machine Learning*, pages 999–1008. PMLR, 2018.

[34] Nicolas Courty, Rémi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Joint distribution optimal transportation for domain adaptation. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 3733–3742. Curran Associates Inc., 2017.

[35] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2016.

[36] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

[37] Marco Cuturi and Arnaud Doucet. Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR, 2014.

[38] Bharath Bhushan Damodaran, Benjamin Kellenberger, Rémi Flamary, Devis Tuia, and Nicolas Courty. Deepjdot: Deep joint distribution optimal transport for unsupervised domain adaptation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 447–463, 2018.

[39] Nabarun Deb, Bhaswar B Bhattacharya, and Bodhisattva Sen. Efficiency lower bounds for distribution-free hotelling-type two-sample tests based on optimal transport. *arXiv preprint arXiv:2104.01986*, 2021.

[40] Nabarun Deb, Promit Ghosal, and Bodhisattva Sen. Measuring association on topological spaces using kernels and geometric graphs. *arXiv preprint arXiv:2010.01768*, 2020.

[41] Nabarun Deb and Bodhisattva Sen. Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*, (just-accepted):1–45, 2021.

[42] Steffen Dereich, Michael Scheutzow, and Reik Schottstedt. Constructive quantization: approximation by empirical measures. *Ann. Inst. Henri Poincaré Probab. Stat.*, 49(4):1183–1203, 2013.

[43] Dominique Drouet Mari and Samuel Kotz. *Correlation and dependence*. Imperial College Press, London; distributed by World Scientific Publishing Co., Inc., River Edge, NJ, 2001.

[44] Ivar Ekeland, Alfred Galichon, and Marc Henry. Optimal transportation and the falsifiability of incompletely specified economic models. *Economic Theory*, 42(2):355–374, 2010.

[45] Tarek A El Moselhy and Youssef M Marzouk. Bayesian inference with optimal maps. *Journal of Computational Physics*, 231(23):7815–7850, 2012.

[46] Sira Ferradans, Nicolas Papadakis, Gabriel Peyré, and Jean-François Aujol. Regularized discrete optimal transport. *SIAM Journal on Imaging Sciences*, 7(3):1853–1882, 2014.

[47] Aden Forrow, Jan-Christian Hütter, Mor Nitzan, Philippe Rigollet, Geoffrey Schiebinger, and Jonathan Weed. Statistical optimal transport via factored couplings. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 2454–2465. PMLR, 2019.

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

[49] Terry L Friesz and J Enrique Fernandez. A model of optimal transport maintenance with demand responsiveness. *Transportation Research Part B: Methodological*, 13(4):317–339, 1979.

[50] Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. Learning with a Wasserstein loss. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2053–2061, Cambridge, MA, USA, 2015. MIT Press.

[51] Kenji Fukumizu, Bharath K Sriperumbudur, Arthur Gretton, and Bernhard Schölkopf. Characteristic kernels on groups and semigroups. In *NIPS*, pages 473–480, 2008.

[52] Harold N Gabow and Robert E Tarjan. Faster scaling algorithms for network problems. *SIAM Journal on Computing*, 18(5):1013–1036, 1989.

[53] Alfred Galichon. *Optimal transport methods in economics*. Princeton University Press, 2016.

[54] Promit Ghosal and Bodhisattva Sen. Multivariate ranks and quantiles using optimal transportation and applications to goodness-of-fit testing. *arXiv preprint arXiv:1905.05340*, 2019.

[55] Nicola Gigli. On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proc. Edinb. Math. Soc. (2)*, 54(2):401–409, 2011.

[56] 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.

[57] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Advances in Neural Information Processing Systems 27*, pages 2672–2680, 2014.

[58] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *2011 International Conference on Computer Vision*, pages 999–1006. IEEE, 2011.

[59] Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pages 2357–2365. PMLR, 2019.

[60] A. Gretton, R. Herbrich, A. Smola, O. Bousquet, and B. Schölkopf. Kernel methods for measuring independence. *J. Mach. Learn. Res.*, 6:2075–2129 (electronic), 2005.

[61] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[62] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, Alexander J Smola, et al. A kernel statistical test of independence. In *Nips*, volume 20, pages 585–592. Citeseer, 2007.

[63] Florian Gunsilius and Susanne M Schennach. Independent nonlinear component analysis. Technical report, cemmap working paper, 2019.

[64] Florian F Gunsilius. On the convergence rate of potentials of brenier maps. *Econometric Theory*, pages 1–37, 2021.

[65] Adityanand Guntuboyina and Bodhisattva Sen. L1 covering numbers for uniformly bounded convex functions. In *Conference on Learning Theory*, pages 12–1. JMLR Workshop and Conference Proceedings, 2012.

[66] Marc Hallin, Eustasio Del Barrio, Juan Cuesta-Albertos, and Carlos Matrán. Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49(2):1139–1165, 2021.

[67] Ruth Heller, Yair Heller, and Malka Gorfine. A consistent multivariate test of association based on ranks of distances. *Biometrika*, 100(2):503–510, 2013.

[68] Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms. II*, volume 306 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993. Advanced theory and bundle methods.

[69] Jan-Christian Hütter and Philippe Rigollet. Minimax estimation of smooth optimal transport maps. *The Annals of Statistics*, 49(2):1166–1194, 2021.

[70] Roy Jonker and Anton Volgenant. A shortest augmenting path algorithm for dense and sparse linear assignment problems. *Computing*, 38(4):325–340, 1987.

[71] Julie Josse and Susan Holmes. Measuring multivariate association and beyond. *Stat. Surv.*, 10:132–167, 2016.

[72] Kirthevasan Kandasamy, Willie Neiswanger, Jeff Schneider, Barnabás Póczos, and Eric P. Xing. Neural architecture search with bayesian optimisation and optimal transport. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2020–2029, Red Hook, NY, USA, 2018. Curran Associates Inc.

[73] L. V. Kantorovich. On a problem of Monge. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 312(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 11):15–16, 2004.

[74] L. Kantorovitch. On the translocation of masses. *C. R. (Doklady) Acad. Sci. URSS (N.S.)*, 37:199–201, 1942.

[75] Ilmun Kim, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimality of permutation tests. *arXiv preprint arXiv:2003.13208*, 2020.

[76] Sanggyun Kim, Rui Ma, Diego Mesa, and Todd P Coleman. Efficient bayesian inference methods via convex optimization and optimal transport. In *2013 IEEE International Symposium on Information Theory*, pages 2259–2263. IEEE, 2013.

[77] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[78] J. Komlós, P. Major, and G. Tusnády. An approximation of partial sums of independent RV's, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 34(1):33–58, 1976.

[79] Gil Kur, Fuchang Gao, Adityanand Guntuboyina, and Bodhisattva Sen. Convex regression in multidimensions: Suboptimality of least squares estimators. *arXiv preprint arXiv:2006.02044*, 2020.

[80] Thibaut Le Gouic and Jean-Michel Loubes. Existence and consistency of wasserstein barycenters. *Probability Theory and Related Fields*, 168(3):901–917, 2017.

[81] Tong Li and Ming Yuan. On the optimality of gaussian kernel based nonparametric tests against smooth alternatives. *arXiv preprint arXiv:1909.03302*, 2019.

[82] Wenbo Li and Ricardo H Nochetto. Quantitative stability and error estimates for optimal transport plans. *IMA Journal of Numerical Analysis*, 2020.

[83] Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *International Conference on Machine Learning*, pages 1718–1727. PMLR, 2015.

[84] Giulia Luise, Alessandro Rudi, Massimiliano Pontil, and Carlo Ciliberto. Differential properties of sinkhorn approximation for learning with Wasserstein distance. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 5864–5874, Red Hook, NY, USA, 2018. Curran Associates Inc.

[85] Valentina Masarotto, Victor M Panaretos, and Yoav Zemel. Procrustes metrics on covariance operators and optimal transportation of gaussian processes. *Sankhyā A*, 81(1):172–213, 2019.

[86] Robert J. McCann. Existence and uniqueness of monotone measure-preserving maps. *Duke Math. J.*, 80(2):309–323, 1995.

[87] Gonzalo Mena and Jonathan Niles-Weed. Statistical bounds for entropic optimal transport: Sample complexity and the central limit theorem. *Advances in Neural Information Processing Systems*, 32, 2019.

[88] Quentin Mérigot. A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592. Wiley Online Library, 2011.

[89] Quentin Merigot and Boris Thibert. Optimal transport: discretization and algorithms. *Handbook of Numerical Analysis 22 – Geometric PDES (arXiv preprint arXiv:2003.00855)*, 2020.

[90] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. In *Proceedings of the International Conference in Learning Representations*, 2017.

[91] Gaspard Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, pages 666–704, 1781.

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

[93] Nicolas Papadakis, Gabriel Peyré, and Edouard Oudet. Optimal transport with proximal splitting. *SIAM Journal on Imaging Sciences*, 7(1):212–238, 2014.

[94] Gabriel Peyré, Marco Cuturi, et al. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

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

[96] Sebastian Reich. A dynamical systems framework for intermittent data assimilation. *BIT Numerical Mathematics*, 51(1):235–249, 2011.

[97] Sebastian Reich. A nonparametric ensemble transform method for bayesian inference. *SIAM Journal on Scientific Computing*, 35(4):A2013–A2024, 2013.

[98] Thomas Rippl, Axel Munk, and Anja Sturm. Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.*, 151:90–109, 2016.

[99] Tim Salimans, Han Zhang, Alec Radford, and Dimitris Metaxas. Improving GANs using optimal transport. *arXiv preprint arXiv:1803.05573*, 2018.

[100] Filippo Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, 55(58-63):94, 2015.

[101] Vivien Seguy, Bharath Bhushan Damodaran, Rémi Flamary, Nicolas Courty, Antoine Rolet, and Mathieu Blondel. Large-scale optimal transport and mapping estimation. In *Proceedings of the International Conference in Learning Representations*, 2018.

[102] Hongjian Shi, Mathias Drton, and Fang Han. Distribution-free consistent independence tests via center-outward ranks and signs. *Journal of the American Statistical Association*, pages 1–16, 2020.

[103] Hongjian Shi, Marc Hallin, Mathias Drton, and Fang Han. Rate-optimality of consistent distribution-free tests of independence based on center-outward ranks and signs. *arXiv preprint arXiv:2007.02186*, 2020.

[104] Max Sommerfeld and Axel Munk. Inference for empirical Wasserstein distances on finite spaces. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(1):219–238, 2018.

[105] Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R. G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561, 2010.

[106] Sanvesh Srivastava, Volkan Cevher, Quoc Dinh, and David Dunson. Wasp: Scalable bayes via barycenters of subset posteriors. In *Artificial Intelligence and Statistics*, pages 912–920. PMLR, 2015.

[107] Sanvesh Srivastava, Cheng Li, and David B Dunson. Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, 19(1):312–346, 2018.

[108] Zhengyu Su, Yalin Wang, Rui Shi, Wei Zeng, Jian Sun, Feng Luo, and Xianfeng Gu. Optimal mass transport for shape matching and comparison. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2246–2259, 2015.

[109] G. J. Székely, M. L. Rizzo, and N. K. Bakirov. Measuring and testing dependence by correlation of distances. *Ann. Statist.*, 35(6):2769–2794, 2007.

[110] M. Talagrand. The transportation cost from the uniform measure to the empirical measure in dimension ≥ 3. *Ann. Probab.*, 22(2):919–959, 1994.

[111] Alexandre B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.

[112] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[113] Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. With applications to statistics.

[114] Cédric Villani. *Topics in optimal transportation*, volume 58 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2003.

[115] Cédric Villani. *Optimal transport*, volume 338 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 2009. Old and new.

[116] Martin J. Wainwright. *High-dimensional statistics*, volume 48 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2019. A non-asymptotic viewpoint.

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

[118] Yoav Zemel and Victor M. Panaretos. Fréchet means and Procrustes analysis in Wasserstein space. *Bernoulli*, 25(2):932–976, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] In fact, in the Contributions Section (see Section 1.3), we have referenced the exact Theorems/Propositions where the claims made in the abstract and the introduction have been substantiated.
   (b) Did you describe the limitations of your work? [Yes] See Remark 2.1 where we compare our main result Theorem 3.1 with existing results in the literature. There, we clearly discuss the advantages and disadvantages of our result compared to, e.g., a similar result in [69].
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is exclusively of a theoretical nature and does not have any negative societal impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] All our results are accompanied with the full set of assumptions required for them to hold.
   (b) Did you include complete proofs of all theoretical results? [Yes] The proofs of all our results can be found in the supplementary file.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A] This work is exclusively of a theoretical nature. We have not run any experiments for this paper.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...

(a) If your work uses existing assets, did you cite the creators? [N/A] This work is exclusively of a theoretical nature. We have not used any existing or curating assets in this paper.

(b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] This work is exclusively of a theoretical nature. We have not used any crowdsourcing and neither have we conducted research with human subjects.

(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]