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

Nabarun Deb
Columbia University

Promit Ghosal
MIT

Bodhisattva Sen
Columbia University

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 of Besov type or Sobolev type, wavelet based or kernel smoothed plug-in estimators respectively 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 the aforementioned 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, 48, 61, 112], dimension reduction [12, 66, 90], generative models [60, 81, 88, 110], 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 [97] 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 [29, 47, 107, 121], Bayesian statistics [46, 75, 80, 108], econometrics [15, 28, 45, 50, 54], nonparametric statistical inference [39–41, 113, 114]; also see [111, 128, 129] 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, 56, 91] and Definition 1.1 below). In practice, the main hurdle in constructing estimators for T_0 is that the explicit forms of the measures μ, ν are unknown; instead only random samples

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

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

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 [7, 30, 67, 93, 94, 102, 116].

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 $d \geq 4$ in the sense of (1.8). This rate happens to be minimax optimal under minimal smoothness assumptions (see [72, 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 wavelet or kernel based smoothing techniques 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 [51, 89], image analysis [18, 59], nonparametric testing [16, 106], generative modeling [10, 96], 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 [30, 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 [30].

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 [76, 77]) of the optimization problem in (1.1). The existence of a minimizer in (1.3) follows from [129, Theorem 4.1].

Proposition 1.1 (Brenier-McCann polar factorization theorem, see [91, 128])**.** *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:

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

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 of $X_1, \ldots, X_m$ 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 [104, 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 [26, 55].* ■

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 $\mathcal{P}_{\mathrm{ac}}(\mathbb{R}^d)$, 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 [82].

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 [72, 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 and Appendix A, we show that the curse of dimensionality in the above rates can be mitigated provided μ and ν admit (uniform) Sobolev smooth densities (see Section 2.2) or Besov smooth densities (see Appendix A). In Section 2.2, our plug-in estimator 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 Sobolev smoothness (see Theorem 2.5). 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 rates of convergence mentioned above also hold for (1.9) (see e.g., Proposition 2.6) which makes a strong case in favor of incorporating smoothness in the construction of plug-in estimators as was conjectured in [30]. In Appendix A, our plug-in estimator is obtained using natural wavelet based density estimators. The rate of convergence in (1.8) turns out to be $n^{-\frac{1+s}{d+2s}}$ where s denotes the degree of Besov smoothness (see Theorem A.1). Note that by choosing s large enough, the exponent in the rate can be made arbitrarily close to $1/2$, thereby reducing the curse of dimensionality.

3. In Section 2.3, we use a discretization technique from [131] to construct discrete approximations to the smoothed $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ from the previous paragraph that in turn yield computable plug-in estimators for T_0 (provided one can sample from $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$) that also achieve the same statistical guarantees as the smoothed plug-in estimator from Section 2.2 (see Theorem 2.7). However the number of atoms required in the discretizations and correspondingly the computational complexity increases with the degree of smoothness; this highlights a statistical and computational trade off.

4. We provide implications of our results in popular applications of OT such as estimating the barycenter of two multivariate probability distributions (see Theorem B.1 in Appendix B.1) and in nonparametric independence testing (see Theorem B.3 in Appendix B.2).

1.4 Related work

Many recent works have focused on obtaining consistent estimators of T_0 using the plug-in principle, see [26, 55] (in the semi-discrete problem) and [41, 68, 132] (in the discrete-discrete problem). In [55], 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 [9, Theorem 1.1] and [87, 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 [9, 87]. Perhaps the most closely related paper to ours would be [67]. In [67], 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 [67, Proposition 1] under the same smoothness assumptions (Sobolev type of order s, see Definition 2.4). In another line of work [72], 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, by themselves, are not transport maps between two probability measures, which makes them harder to interpret. In contrast, our focus is on obtaining rates of convergence for plug-in estimators, which are transport maps between natural aprroximations of μ and ν. Such plug-in type strategies are a lot more popular in computational OT [7, 30, 67, 93, 94, 102, 116].

In terms of obtaining rates of convergence for (1.9), some attempts include [109, 116] 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 [6, 42, 49, 100, 123, 131]. 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 [30] under no smoothness assumptions. Our rates coincide with the $n^{-2/d}$ rate from [30] 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.

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,*

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

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

The proof of Theorem 2.1 (see Appendix C.1) starts along the same lines as the proof of the curvature estimate in [56, 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 (C.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 [92, Proposition 2] and [30, 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 [56, 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 [55, 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 [72, 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 [72] can obtain rates arbitrarily close to n^{-1}. This is a price we pay for analyzing natural and popular plug-in estimators as opposed to the (more intractable) wavelet based estimators in [72].*

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)*, $\Psi^*_{\widetilde{\mu}_m, \widetilde{\nu}_n}(\cdot)$, $\Psi^*_{\widetilde{\mu}_m, \nu^*_m}(\cdot) \in \mathcal{F}$. Further, by Caffarelli's regularity theory [20–22], 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 2.3 and 2.6) such that $\Psi^*_{\widetilde{\mu}_m, \widetilde{\nu}_n}(\cdot)$, $\Psi^*_{\widetilde{\mu}_m, \nu^*_m}(\cdot) \in \mathcal{F} \cap \mathcal{F}_s$. Thus, we can bound the first term on the right hand side of* (2.1) *as:*

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

The right hand side of (2.2) *can now be bounded using the corresponding Dudley's entropy integral bounds using empirical process techniques, see [126, 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:*

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

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

and

$$t_{d,\alpha} := \begin{cases} (4\alpha)^{-1}(4 + ((2\alpha + 2d\alpha - d) \vee 0)) & \text{for } d < 4, \\ (\alpha^{-1} \vee 7/2) - 1 & \text{for } d = 4, \\ 2(1 + d^{-1}) & \text{for } d > 4. \end{cases}$$

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 $O_p(r_d^{(n,n)} \times (\log(1+n))^{t_{d,\alpha}})$ or $O_p(r_d^{(m,m)} \times (\log(1+m))^{t_{d,\alpha}})$ 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 further that μ is compactly supported. Then, with $r_d^{(m,n)}$ defined as in* (2.4)*, we have:*

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

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 of Theorem 2.2 using Theorem 2.1 is provided in Remark 2.3 below, and the actual proof is presented in Appendix C.1.

Remark 2.3 (Proof technique)**.** *The proof of Theorem 2.2 proceeds using the strategy outlined in Remark 2.2. We first show that $\mathcal{F}_s$ (see Remark 2.2) can be chosen as a certain class of convex functions which are in $L^2(\nu)$. We then use Dudley's entropy integral type bounds which in turn requires the bracketing entropy [126, Page 270] of $\mathcal{F}_s$, recently proved in [83, Equation 26]. This strategy is slightly different from that used in the proof of [30, Theorem 2], where the authors assume that μ is compactly supported whereas we only assume the finiteness of $\mathbb{E}\exp(t\|X_1\|^{\alpha})$ for some $t > 0$, $\alpha > 0$. The compactness assumption on μ allows one to further restrict $\mathcal{F}_s$ to the class of Lipschitz functions. This additional restriction does not seem to be immediate without the compactness assumption.*

As discussed in Section 1.3, the exponents obtained in Theorem 2.2 are minimax optimal under bare minimal smoothness assumptions (see [72, 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 [9, 87] and require less stringent assumptions than those in [30]. 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 kernel based plug-in estimator: mitigating the curse of dimensionality

In this section, we focus on kernel based density estimators for the probability densities associated with μ and ν (see [57, 58, 99, 103, 115]). We will show, using Theorem 2.1, that the corresponding estimators of $T_0(\cdot)$ achieve (near) dimension-free rates under sufficient smoothness assumptions.

We first introduce the Sobolev class of functions which we will exploit in this subsection to construct estimators that achieve rates of convergence which mitigate the curse of dimensionality under sufficient smoothness.

Definition 2.4 (Uniform Sobolev class of functions)**.** *Let $\Omega \subseteq \mathbb{R}^d$ and $f(\cdot)$ be uniformly continuous on Ω and admits uniformly continuous derivatives up to order s on Ω for some $s \in \mathbb{N}$. For any $\mathfrak{m} := (m_1, \ldots, m_d) \in \mathbb{N}^d$, let*

$$\partial^{\mathfrak{m}} f := \frac{\partial}{\partial_{x_1}^{m_1}} \cdots \frac{\partial}{\partial_{x_d}^{m_d}} f, \quad |\mathfrak{m}| := \sum_{i=1}^{d} m_i.$$

For any $k \leq s$, we further define,

$$\|f\|_{C^k(\Omega)} := \sum_{|\mathfrak{m}| \leq k} \|\partial^{\mathfrak{m}} f\|_{L^\infty(\Omega)}.$$

The space $C^s(\Omega)$ is defined as the set of functions $f(\cdot)$ for which $\|f\|_{C^k(\Omega)} < \infty$ for all $k \leq s$.

For this subsection, assume that μ and ν admit Sobolev smooth densities $f_\mu(\cdot)$ and f_ν in the uniform norm (see Definition 2.4 above). Given $\Omega \subseteq \mathbb{R}^d$ and $s \in \mathbb{N}$, let $C^s(\Omega)$ denote the set of Sobolev smooth functions on Ω of order s.

Assumption (A1) (Regularity of the 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 $s, M > 0$ such that $f_\mu(\cdot) \in C^s(\mathcal{X}; M)$ and $f_\nu(\cdot) \in C^s(\mathcal{Y}; M)$ where $C^s(\mathcal{X}; M)$ is the space of real valued functions supported on $\mathcal{X}$ such that for all $f(\cdot) \in C^s(\mathcal{X}; M)$, we have $M^{-1} \leq f(x) \leq M$ for all $x \in \mathcal{X}$ and $\|f\|_{C^s(\mathcal{X})} \leq M$. Here $\|\cdot\|_{C^s(\mathcal{X})}$ is the standard uniform Sobolev norm as defined in Definition 2.4. The space $C^s(\mathcal{Y}; M)$ is defined analogously.*

We now define our estimators for $f_\mu(\cdot)$ and $f_\nu(\cdot)$ using the standard kernel density estimation technique (see [125, Section 1.2]). Set

$$\widehat{f}_\mu(x) := \frac{1}{m h_m^d} \sum_{i=1}^{m} K_d\left(\frac{X_i - x}{h_m}\right), \tag{2.5}$$

for some bandwidth parameter $h_m > 0$ and d-variate kernel $K_d(\cdot)$. We assume that $K_d(\cdot)$ is the d-fold product of 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 symmetric, bounded, $s+1$ times differentiable kernel on $\mathbb{R}^d$ with all $s+1$ derivatives bounded and integrable. Further, suppose that $K(\cdot)$ is of order $2s+2$, i.e.,*

$$\int u^j K(u)\, du = \mathbb{1}(j = 0), \qquad \text{for } j = \{0, 1, 2, \ldots, 2s+1\}, \qquad \text{and} \int |u|^{2s+2} |K(u)|\, du < \infty.$$

The above assumptions on $K(\cdot)$ are standard for estimating smooth densities and their derivatives of different orders in the kernel density estimation literature; see e.g. [4, 57, 58, 69, 125]. There are several natural ways to construct kernels satisfying Assumption (A2), see [125, Section 1.2.2]; an example is also provided in Example 2.4 below.

Example 2.4 (Example of a kernel satisfying Assumption (A2))**.** *Let $\psi_m(\cdot)$ be the m-th Hermite polynomial on $\mathbb{R}$ (see [84]). Then the kernel function defined as*

$$K(u) := \sum_{m=0}^{2s+2} \psi_m(0) \psi_m(u) \exp(-u^2/2)$$

satisfies Assumption (A2). ■

It is evident from Assumption (A2) that $K(\cdot)$ may take some negative values, in which case, $\widehat{f}_\mu(\cdot)$ (respectively $\widehat{f}_\nu(\cdot)$) may not be a probability density. Consequently the barycentric projection (see Definition 1.2) between $\widehat{f}_\mu(\cdot)$ and $\widehat{f}_\nu(\cdot)$ is not well-defined. We get around this by resorting

to the same (approximate) projection technique we used in (A.2) for the wavelet based estimators. In this case however, instead of (approximately) projecting using an appropriate Besov norm as in (A.2), we use a integral probability metric (see Definition 2.5; also see [98, 105, 117] for examples, computational procedures and applications of such metrics). The corresponding measure is defined below:

Definition 2.5 (Integral probability metric)**.** *Given a class $\mathcal{F}$ of bounded functions on $\mathbb{R}^d$ and two probability densities $g_1(\cdot)$ and $g_2(\cdot)$ on $\mathbb{R}^d$, the integral probability metric/distance between $g_1(\cdot)$ and $g_2(\cdot)$ with respect to $\mathcal{F}$ is defined as*

$$d_{\mathrm{IP}}(g_1, g_2; \mathcal{F}) := \sup_{\psi(\cdot)\in\mathcal{F}} \left| \int \psi(x)(g_1(x) - g_2(x))\, dx \right|.$$

Sufficient conditions on $\mathcal{F}$ for $d_{\mathrm{IP}}(\cdot, \cdot; \mathcal{F})$ to be a metric on the space of probability measures (not on the space of probability densities as they can be altered on set of Lebesgue measure 0 without altering the underlying probability measures) on $\mathbb{R}^d$ have been discussed in [98]. Observe that the measure $d_{\mathrm{IP}}(g_1, g_2; \mathcal{F})$ is well defined even when $g_1(\cdot)$ and $g_2(\cdot)$ are not probability densities.

In Theorem 2.5 below, we use $\mathcal{F} = C^{s+2}(\mathcal{X}, M')$. Note that any function in $C^{s+2}(\mathcal{X}, M')$ can be extended to a function in $C^{s+2}(\mathbb{R}^d; M')$ (see [72, Theorem 23] and [124, Theorem 1.105]). The fact that this choice of $\mathcal{F}$ results in a metric follows from the argument in [98, Page 8].

We are now in a position to describe the projection estimators for $f_\mu(\cdot)$ and $f_\nu(\cdot)$, and the rates achieved by the corresponding plug-in estimator.

Theorem 2.5. *Assume that $T_0(\cdot)$ is L-Lipschitz and f_μ, f_ν are Lebesgue densities satisfying Assumption (A1). Also suppose that $K(\cdot)$ satisfies Assumption (A2). Define $h_m := m^{-\frac{1}{d+2s}} \log m$, $h_n := n^{-\frac{1}{d+2s}} \log n$ and $T := \int |K_d(u)|\, du + 1$. Fix any $M' > 0$. Consider any probability density $\widetilde{f}_\mu^{M'}(\cdot) \in C^s(\mathcal{X}; TM)$ (where M is defined as in Assumption (A1)) which satisfies*

$$d_{\mathrm{IP}}\left(\widetilde{f}_\mu^{M'}, \widehat{f}_\mu; C^{s+2}(\mathcal{X}; M')\right) \le \inf_{\substack{f(\cdot)\in C^s(\mathcal{X};TM)\\ f\ge 0,\ \int f = 1}} d_{\mathrm{IP}}\left(\widehat{f}_\mu, f; C^{s+2}(\mathcal{X}; M')\right) + r_{d,s}^{(m,n)} \tag{2.6}$$

where $r_{d,s}^{(m,n)}$ is defined as in (2.7) and $d_{\mathrm{IP}}(\cdot, \cdot; C^{s+2}(\mathcal{X}; M'))$ is the integral probability metric defined in Definition 2.5. We define $\widetilde{f}_\nu^{M'}(\cdot)$ analogously as in (2.6) with $\mathcal{X}$, $\widehat{f}_\mu(\cdot)$ replaced by $\mathcal{Y}$, $\widehat{f}_\nu(\cdot)$. Then the following conclusions hold.

1. *There exists $M' > 0$ (depending on M) such that, if, $\widetilde{\mu}_m$ and $\widetilde{\nu}_n$ are the probability measures corresponding to the probability densities $\widetilde{f}_\mu^{M'}(\cdot)$ and $\widetilde{f}_\nu^{M'}(\cdot)$, then the following holds for some constant $C > 0$:*

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

$$\text{where} \quad r_{d,s}^{(m,n)} := \begin{cases} m^{-1/2} + n^{-1/2} & \text{for } d < 2(s+2), \\ m^{-1/2}\left(\log(1+m)\right)^d + n^{-1/2}\left(\log(1+n)\right)^d & \text{for } d = 2(s+2), \\ m^{-\frac{s+2}{d}} + n^{-\frac{s+2}{d}} & \text{for } d \ge 2(s+2). \end{cases} \tag{2.7}$$

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

2. *$\widehat{f}_\mu(\cdot)$ satisfies*

$$\lim_{n\to\infty} \max\left\{\mathbb{P}\left(\|\widehat{f}_\mu\|_{C^s(\widetilde{\mathcal{X}})} \ge TM\right), \mathbb{P}\left(\sup_{x\in\widetilde{\mathcal{X}}} |\widehat{f}_\mu(x) - f_\mu(x)| \ge \varepsilon\right)\right\} = 0 \tag{2.8}$$

   for any $\varepsilon > 0$, where $\widetilde{\mathcal{X}}$ is any compact subset of $\mathcal{X}^o$. The same conclusion holds for $\widehat{f}_\nu(\cdot)$ with $\mathcal{X}$ replaced by $\mathcal{Y}$.

In Theorem 2.5, $\widetilde{f}_{\mu}^{M'}(\cdot)$ can be viewed as an approximate minimizer of $d_{\mathrm{IP}}(\widehat{f}_{\mu}, \cdot; C^{s+2}(\mathcal{X}, M'))$ over an appropriate class of Sobolev smooth probability densities. This is carried out because $\widehat{f}_{\mu}(\cdot)$ by itself may not be a probability density.

Further note that $\widetilde{\mu}_m, \widetilde{\nu}_n$ as specified in Theorem 2.5 are both smooth, and consequently $\widetilde{\Gamma}_{\min}$ is a singleton and the supremum in Theorem 2.5 can be dropped. A brief description of the proof technique for Theorem 2.5 is presented in Remark 2.6 below and the actual proof is given in Appendix C.1.

Remark 2.6 (Proof technique)**.** *The proof of Theorem 2.5 proceeds along the same lines as Remark 2.3. We first show that $\mathcal{F}_s$ (see Remark 2.2) can be chosen as a certain subset of $C^{s+2}(\mathcal{Y}^{\circ})$. We then use Dudley's entropy integral type bounds which in turn requires the bracketing entropy [126, Page 270] of the class of compactly supported Sobolev smooth functions which can be found in [127, Corollary 2.7.2].*

We now explain the implications of both the parts of Theorem 2.5 in the following two remarks.

Remark 2.7 (Mitigating the curse of dimensionality)**.** *Theorem 2.5 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 a similar estimator was analyzed in [67, Proposition 1] when $m = n$. However, the rates obtained in Theorem 2.5 are* strictly better *than those in [67, Proposition 1]. For $m = n$, when $d < 2(s+2)$, [67] obtained a rate of $n^{-\frac{s+2}{2(s+2)+d}}$ which is* worse *than $n^{-1/2}$ obtained in Theorem 2.5. For the other regimes, [67] 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 2.5. In fact, the rates obtained in Theorem A.1 are also strictly better than the rates obtained in [67, Proposition 1] (for every fixed d) described above, but they are strictly worse than the rates obtained in Theorem 2.5. When the degree of smoothness s is large, both Theorems A.1 and 2.5 lead to rates of (approximately) $n^{-1/2}$. However when s is small, the rate in Theorem 2.5 is much faster than that in Theorem A.1, e.g., if s is close to 0, the rate in Theorem A.1 is approximately $n^{-\frac{1}{d}}$ whereas that in Theorem 2.5 is the faster rate of $n^{-\frac{2}{d}}$. It must be noted however that the smoothness assumptions are different in Theorems A.1 and 2.5.*

Remark 2.8 (Computational aspects of Theorem 2.5)**.** *In Theorem 2.5, we have shown that the plug-in estimator for $T_0(\cdot)$ using $\widetilde{f}_{\mu}^{M'}(\cdot)$ and $\widetilde{f}_{\nu}^{M'}(\cdot)$ achieve rates that mitigate the curse of dimensionality under sufficient smoothness. However, as is evident, $\widetilde{f}_{\mu}^{M'}(\cdot)$ is hard to compute whereas $\widehat{f}_{\mu}(\cdot)$ is computable easily in linear time. Note that if $\widehat{f}_{\mu}(\cdot)$ itself were a probability density in $C^s(\mathcal{X}; TM)$, then we would have $\widehat{f}_{\mu} = \widetilde{f}_{\mu}^{M'}$. While Theorem 2.5 does not establish that, it does come close in part 2, from which we can easily derive the following:*

$$\lim_{n\to\infty} \mathbb{P}(\widehat{f}_{\mu}(\cdot) \notin C^s(\widetilde{\mathcal{X}}; TM)) = 0.$$

The above shows that $\widehat{f}_{\mu}(\cdot)$ is indeed bounded below by $(TM)^{-1}$ on $\widetilde{\mathcal{X}}$ (any compact subset of the interior of $\mathcal{X}$), and additionally belongs to $C^s(\widetilde{\mathcal{X}}; TM)$ with probability converging to 1. This leads us to conjecture that the natural density version of $\widehat{f}_{\mu}(\cdot)$, i.e.,

$$\frac{\max\{\widehat{f}_{\mu}(\cdot), 0\}}{\int \max\{\widehat{f}_{\mu}(x), 0\}\, dx}$$

should serve as a good proxy for $\widetilde{f}_{\mu}^{M'}(\cdot)$ and lead to rates of convergence that mitigate the curse of dimensionality. From a computational perspective, the density specified above is easy to simulate from using an accept-reject algorithm without computing the integral in the denominator (see [101, Algorithm 4.3]). However, our current proof technique does not provide rates of convergence for the above density estimator based on $\widehat{f}_{\mu}(\cdot)$.

Another important implication of Theorem 2.5 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 2.6. *Consider the setting in Theorem 2.5. 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 2.6 (see Appendix C.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 [131, Theorem 3] where the result was established under more general Besov smoothness assumptions). On the contrary, when $\mu \neq \nu$, Proposition 2.6 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 $\mu = \nu$. This observation complements [30, Corollary 1] where the authors make a similar observation for the special case of $s = 0$.

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. Such plug-in estimators are popular in the computational OT literature (see [7, 8, 25, 36]). However, even after $\widetilde{f}_\mu(\cdot)$, $\widetilde{f}_\nu(\cdot)$ are calculated, $\widetilde{T}^\gamma_{m,n}$ as in Theorem 2.5 cannot be computed explicitly from data if $\widetilde{f}_\mu(\cdot)$ and $\widetilde{f}_\nu(\cdot)$ are continuous densities. This is in contrast to $\widetilde{T}^\gamma_{m,n}$ 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 computable in polynomial time (with complexity increasing with smoothness) provided $\widetilde{f}_\mu(\cdot)$ and $\widetilde{f}_\nu(\cdot)$ can be sampled from, and also attain rates that mitigate the curse of dimensionality.

Construction: We will illustrate the discretized estimator using the kernel based estimator from Section 2.2. Similar results also hold for the wavelet based estimator from Appendix A. Recall the kernel density estimators $\widetilde{f}_\mu(\cdot)$ and $\widetilde{f}_\nu(\cdot)$ (see (2.6)). Sample $M \geq 1$ random points from both $\widetilde{f}_\mu(\cdot)$ and $\widetilde{f}_\nu(\cdot)$. Let $\widehat{\mu}_{m,M}$ and $\widehat{\nu}_{n,M}$ denote the standard empirical measures on the M points sampled from $\widetilde{f}_\mu(\cdot)$ and $\widetilde{f}_\nu(\cdot)$ respectively. Finally construct $\widetilde{T}_{m,n} \equiv \widetilde{T}^\gamma_{m,n}$ 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 [131, Section 6] for estimating probability densities under the Wasserstein loss. Based on this construction, the main result of this section is as follows:

Theorem 2.7. *Consider the setting in Theorem 2.5 and the same construction of $\widetilde{T}^\gamma_{m,n}$ 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 consequently the following conclusion holds for some constant $C > 0$:*

$$\mathbb{E}\left[\int \|\widetilde{T}_{m,n}(x) - T_0(x)\|^2 \, d\widetilde{\mu}_m(x)\right] \leq C r_{d,s}^{(n,n)}.$$

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

The proof of Theorem 2.7 is given in Appendix C.1. Once the empirical measures $\widehat{\mu}_{m,M}$ and $\widehat{\nu}_{n,M}$ have been obtained, 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 [73]. This highlights the statistical versus computational trade off, 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 (approximate) algorithms faster than the Hungarian algorithm stated above, can be found in [1, 36, 53] 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 $\widetilde{f}_\mu(\cdot)$ and $\widetilde{f}_\nu(\cdot)$ 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.

Acknowledgments and Disclosure of Funding

We would like to thank the reviewers for their constructive suggestions that greatly helped improve the quality of the paper. The third author is supported by NSF Grant DMS-2015376.

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] Ery Arias-Castro, David Mason, and Bruno Pelletier. On the estimation of the gradient lines of a density and the consistency of the mean-shift algorithm. *The Journal of Machine Learning Research*, 17(1):1487–1514, 2016.

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

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

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

[12] 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.

[13] 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.

[14] 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.

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

[16] 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.

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

[18] 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.

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

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

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

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

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

[24] 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.

[25] 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.

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

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] Albert Cohen. *Numerical analysis of wavelet methods*, volume 32 of *Studies in Mathematics and its Applications*. North-Holland Publishing Co., Amsterdam, 2003.

[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] David L. Donoho, Iain M. Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

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

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

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

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

[48] 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.

[49] 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.

[50] 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.

[51] 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.

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

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

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

[55] 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.

[56] 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.

[57] Evarist Giné and Richard Nickl. Uniform central limit theorems for kernel density estimators. *Probab. Theory Related Fields*, 141(3-4):333–387, 2008.

[58] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, 2016.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

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

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

[68] 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.

[69] Bruce E. Hansen. Uniform convergence rates for kernel estimation with dependent data. *Econometric Theory*, 24(3):726–748, 2008.

[70] Wolfgang Härdle, Gerard Kerkyacharian, Dominique Picard, and Alexander Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

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

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

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

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

[75] 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.

[76] 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.

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

[78] G. Kerkyacharian and D. Picard. Density estimation in Besov spaces. *Statist. Probab. Lett.*, 13(1):15–24, 1992.

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

[80] 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.

[81] 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.

[82] 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.

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

[84] Pierre Simon Laplace. *Théorie analytique des probabilités*. Courcier, 1820.

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

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

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

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

[89] 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.

[90] 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.

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

[92] 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.

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

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

[95] Yves Meyer. *Ondelettes et opérateurs. I.* Actualités Mathématiques. [Current Mathematical Topics]. Hermann, Paris, 1990. Ondelettes. [Wavelets].

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

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

[98] Alfred Müller. Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29(2):429–443, 1997.

[99] È. A. Nadaraja. On non-parametric estimates of density functions and regression. *Teor. Verojatnost. i Primenen.*, 10:199–203, 1965.

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

[101] Nadia Oudjane and Christian Musso. L2-density estimation with negative kernels. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 34–39. IEEE, 2005.

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

[103] Emanuel Parzen. On estimation of a probability density function and mode. *Ann. Math. Statist.*, 33:1065–1076, 1962.

[104] 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.

[105] Svetlozar T Rachev. The monge–kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985.

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

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

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

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

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

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

[112] 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.

[113] 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.

[114] 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.

[115] Bernard W. Silverman. Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Statist.*, 6(1):177–184, 1978.

[116] 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.

[117] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electron. J. Stat.*, 6:1550–1599, 2012.

[118] 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.

[119] 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.

[120] 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.

[121] 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.

[122] 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.

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

[124] Hans Triebel. *Theory of function spaces. III*, volume 100 of *Monographs in Mathematics*. Birkhäuser Verlag, Basel, 2006.

[125] 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.

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

[127] 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.

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

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

[130] Gilbert G. Walter. Approximation of the delta function by wavelets. *J. Approx. Theory*, 71(3):329–343, 1992.

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

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

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 B.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 [72].

(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]