Plugin Estimation of Smooth Optimal Transport Maps

Tudor Manole*, Sivaraman Balakrishnan*†, Jonathan Niles-Weed◇‡, Larry Wasserman*†

*Department of Statistics and Data Science, Carnegie Mellon University
†Machine Learning Department, Carnegie Mellon University
◇Center for Data Science, New York University
‡Courant Institute of Mathematical Sciences, New York University

tmanole@andrew.cmu.edu, jnw@cims.nyu.edu, {siva,larry}@stat.cmu.edu

May 17, 2022

arXiv:2107.12364v2 [math.ST] 16 May 2022

Abstract

We analyze a number of natural estimators for the optimal transport map between two distributions and show that they are minimax optimal. We adopt the plugin approach: our estimators are simply optimal couplings between measures derived from our observations, appropriately extended so that they define functions on $\mathbb{R}^d$. When the underlying map is assumed to be Lipschitz, we show that computing the optimal coupling between the empirical measures, and extending it using linear smoothers, already gives a minimax optimal estimator. When the underlying map enjoys higher regularity, we show that the optimal coupling between appropriate nonparametric density estimates yields faster rates. Our work also provides new bounds on the risk of corresponding plugin estimators for the quadratic Wasserstein distance, and we show how this problem relates to that of estimating optimal transport maps using stability arguments for smooth and strongly convex Brenier potentials. As an application of our results, we derive central limit theorems for plugin estimators of the squared Wasserstein distance, which are centered at their population counterpart when the underlying distributions have sufficiently smooth densities. In contrast to known central limit theorems for empirical estimators, this result easily lends itself to statistical inference for the quadratic Wasserstein distance.

1 Introduction

Optimal transport maps play a central role in the theory of optimal transport (Rachev and Rüschendorf, 1998; Villani, 2003; Santambrogio, 2015), and have received many recent methodological applications in statistics and machine learning (Kolouri et al., 2017; Panaretos and Zemel, 2019). Given two distributions P and Q with support contained in a set $\Omega \subseteq \mathbb{R}^d$, an optimal transport map T_0 from P to Q is any solution to the *Monge problem* (Monge, 1781),

$$\operatorname*{argmin}_{T \in \mathcal{T}(P,Q)} \int_\Omega \|x - T(x)\|^2 \, dP(x), \tag{1}$$

where $\mathcal{T}(P,Q)$ is the set of transport maps between P and Q, that is, the set of Borel-measurable functions $T : \Omega \to \Omega$ such that $T_\# P := P(T^{-1}(\cdot)) = Q$. Equivalently, we write $T_\# P = Q$ whenever $X \sim P$ implies $T(X) \sim Q$. As we shall see in Section 2, the Monge problem admits a solution T_0 as soon as P is absolutely continuous with respect to the Lebesgue measure.

A wide range of statistical applications involve transforming random variables to ensure they follow a desired distribution. Optimal transport maps form natural choices of such transformations when no other canonical choice is available. For instance, optimal transport maps form a useful tool for addressing label shift between train and test distributions in classification problems, and have more generally been applied to various domain adaptation and transfer learning problems (Courty et al., 2016; Redko et al., 2019; Rakotomamonjy et al., 2021; Zhu et al., 2021). A large body of recent work has also employed optimal transport maps for defining notions of multivariate ranks and quantiles (Chernozhukov et al., 2017; Hallin et al., 2021; Ghosal and Sen, 2022), and has applied them to a variety of nonparametric hypothesis testing problems (Shi et al., 2020; Deb and Sen, 2021; Deb et al., 2021a). We also note their recent uses in distributional regression (Ghodrati and Panaretos, 2021), generative modeling (Finlay et al., 2020; Onken et al., 2021), fairness in machine learning (Gordaliza et al., 2019; Black et al., 2020; de Lara et al., 2021), and in a wide range of statistical applications to the sciences (Read, 1999; Wang et al., 2011; Schiebinger et al., 2019; Komiske et al., 2020).

An important question arising in many of these applications is that of estimating the optimal transport map between unknown distributions, based on independent samples. The aim of this paper is to develop practical estimators of optimal transport maps achieving near-optimal risk. Specifically, given i.i.d. samples $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$, we derive estimators $\widehat{T}_{nm}$ which achieve the minimax rate of convergence[1], under the loss function

$$\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 = \int_\Omega \big\|\widehat{T}_{nm}(x) - T_0(x)\big\|^2 dP(x). \tag{2}$$

The theoretical study of such estimators was recently initiated by Hütter and Rigollet (2021), who proved that for any estimator $\widehat{T}_{nm}$ with $n = m$,

$$\sup_{P,Q} \mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|_{L^2(P)}^2 \gtrsim n^{-\frac{2\alpha}{2(\alpha-1)+d}} \vee \frac{1}{n}, \tag{3}$$

where the supremum is taken over all pairs of distributions (P, Q) admitting densities bounded away from zero over a compact set Ω, for which T_0 lies in an α-Hölder ball for some $\alpha \geq 1$, and satisfies a key curvature condition **A1**(λ) which we define below. The lower bound (3) is reminiscent of, but generally faster than, the classical $n^{-2\alpha/(2\alpha+d)}$ minimax rate of estimating an α-Hölder continuous nonparametric regression function (Tsybakov, 2008), and is shown by Hütter and Rigollet (2021) to be achievable up to a polylogarithmic factor. Nevertheless, their estimator is computationally intractable in general dimension, and their work leaves open the question of developing practical optimal transport map estimators which achieve comparable risk.

In this paper, we establish the minimax optimality of several natural and intuitive estimators of optimal transport maps, several of which have already been proposed in the statistical optimal transport literature, but have resisted sharp statistical analyses thus far. We focus on the following two classes of plugin estimators.

(i) **Empirical Estimators.** When no smoothness assumptions are placed on P and Q, it is natural to study the plugin estimator based on the empirical measures

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}, \quad \text{and} \quad Q_m = \frac{1}{m}\sum_{j=1}^m \delta_{Y_j}.$$

[1] Here and throughout, minimax rate-optimality is tacitly understood up to polylogarithmic factors.

In the special case $n = m$, there is an optimal transport map T_{nm} from P_n to Q_m, and more generally there is an optimal coupling of these measures. While the in-sample estimator T_{nm} is only defined over the support of P_n, we readily obtain estimators defined over the entire domain by casting the extension problem as one of nonparametric regression. We show how linear smoothers and least-squares estimators can be used to interpolate T_{nm}, leading to an estimator $\widehat{T}_{nm}$ defined over Ω. Under some conditions, $\widehat{T}_{nm}$ achieves the minimax rate for estimating Lipschitz optimal transport maps T_0.

(ii) **Smooth Estimators.** In order to obtain faster rates of convergence when P and Q admit smooth densities p and q, we next analyze the risk of the unique optimal transport map between kernel or wavelet density estimators of p and q. In contrast to our empirical optimal transport map estimators, we show that such smooth plugin estimators are able to take advantage of additional regularity of the densities p and q, and achieve minimax-optimal rates when these densities are Hölder smooth.

While our emphasis is on optimal transport maps, an equally important target of estimation is the optimal objective value in the Monge problem (1), which gives rise to the squared 2-Wasserstein distance $W_2^2(P, Q)$ (defined formally in Section 2). Our optimal transport map estimators naturally yield estimators for the Wasserstein distance, and we provide upper bounds on their risk, and derive limit laws, as a byproduct of our study.

Our Contributions. The primary contributions of this paper are summarized as follows.

(i) In Sections 3 and 4, we develop new stability bounds which relate the risk of plugin transport map estimators to the plugin density estimation risk, as measured in the Wasserstein distance. These stability bounds are quite general and enable the analysis of flexible, practical transport map estimators. The risk of density estimation under the Wasserstein distance has been extensively studied (Weed and Berthet, 2019; Divol, 2021), and our stability bounds enable us to leverage this past work. Additionally, our stability bounds enable the analysis of plugin estimators of the Wasserstein distance, once again relating the risk in this problem to the plugin density estimation risk.

(ii) We build on our stability bounds to analyze the risk of empirical, kernel-based and wavelet-based transport map estimators in both the one-sample setup (where the source distribution is known exactly, and the target distribution is sampled) and the two-sample setup (where both the source and target distributions are sampled). The rates we obtain are minimax optimal. For example, suppose that $\widehat{T}_n$ is the optimal transport map from P to $\widehat{Q}_n$, where $\widehat{Q}_n$ is a wavelet-estimator over the domain $[0, 1]^d$. Then, whenever P and Q admit $(\alpha - 1)$-Hölder densities and satisfy several additional conditions, we show that,

$$\mathbb{E}\big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \lesssim \begin{cases} n^{-\frac{2\alpha}{2(\alpha-1)+d}}, & d \geq 3 \\ (\log n)^2/n, & d = 2 \\ 1/n, & d = 1. \end{cases} \tag{4}$$

As we explain in Section 2, the Hölder smoothness of T_0 is typically expected to be of one degree greater than that of p and q, and thus our estimator achieves the minimax lower bound (3) when these densities are $(\alpha - 1)$-Hölder smooth, for any $\alpha > 1$[2]. In the two-sample setting, we develop analogous minimax-optimal analyses, for the empirical plugin estimator (Propositions 13–15) as well as for kernel-based and wavelet-based plugin estimators (Theorems 17–18) when P and Q admit Hölder-smooth densities. In the latter case, as we discuss further in the sequel, we avoid complications that arise in the optimal transport problem due to boundary effects by working over the d-dimensional flat torus.

(iii) In each of the above settings, we complement our results with upper bounds on the risk of plugin estimators of the Wasserstein distance. For instance, in the smooth setting discussed above, we show that,

$$\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \lesssim \left(\frac{1}{n}\right)^{\frac{2\alpha}{2(\alpha-1)+d}} \vee \frac{1}{\sqrt{n}}. \tag{5}$$

We also develop analogous results in the one and two-sample settings, for various empirical and smooth plugin estimators.

(iv) We build upon these estimation results to address *inference* for Wasserstein distances in the high-smoothness regime $2(\alpha + 1) > d$. We show in Section 5.1, under regularity conditions, that whenever $P \neq Q$, there exists $\sigma^2 > 0$ such that

$$\sqrt{n}\Big(W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\Big) \rightsquigarrow N(0, \sigma^2), \quad \text{as } n \to \infty. \tag{6}$$

We also develop analogous results in the two-sample setting. To the best of our knowledge, these are the first central limit theorems for a plugin estimator of the Wasserstein distance which is centered at its population counterpart, for absolutely continuous distributions P and Q in arbitrary dimension. We further show that the variance σ^2 of the limiting distribution can be estimated using our transport map estimators, leading to an asymptotic confidence interval for $W_2^2(P, Q)$.

(v) We also develop the semiparametric efficiency theory for the Wasserstein distance functional. In Section 5.2, we derive the efficient influence function of the Wasserstein distance, derive asymptotic local minimax lower bounds, and show that our plugin Wasserstein distance estimators are asymptotically efficient in the high-smoothness regime.

Related Work. The two recent works of Hütter and Rigollet (2021) and Gunsilius (2021) establish $L^2(P)$ convergence rates for transport map estimators. Gunsilius (2021) derives upper bounds on the risk of a plugin estimator for Brenier potentials, obtained via kernel density estimation of p and q. This analysis results in suboptimal convergence rates for the optimal transport map T_0 itself. We show in this work that such plugin estimators do in fact achieve the optimal convergence rate when the sampling domain is the d-dimensional torus.

Building upon a construction of del Barrio et al. (2020), a consistent estimator of T_0 was obtained by de Lara et al. (2021) under mild assumptions, by regularizing a piecewise constant approximation of the empirical optimal transport map T_n. We do not know if quantitative convergence rates can be obtained for their estimator under stronger assumptions. Beyond these works, a wide range of

[2]As discussed in Appendix E of Hütter and Rigollet (2021), the minimax lower bound (3) also holds under such smoothness conditions on the densities p and q, as opposed to smoothness conditions on T_0.

heuristic estimators have been proposed in the literature (Perrot et al., 2016; Nath and Jawanpuria, 2020; Makkuva et al., 2020), but their theoretical properties remain unknown to the best of our knowledge.

Rates of convergence for the problem of estimating Wasserstein distances have arguably received more attention than that of estimating optimal transport maps. Characterizing the convergence rate of the empirical measure under the Wasserstein distance is a classical problem (Dudley, 1969; Boissard and Le Gouic, 2014; Fournier and Guillin, 2015; Weed and Bach, 2019; Lei, 2020) which immediately leads to upper bounds on the convergence rate of the empirical plugin estimator of the Wasserstein distance. While such upper bounds are generally unimprovable (Liang, 2019; Niles-Weed and Rigollet, 2019), they have recently been sharpened by Chizat et al. (2020) and Manole and Niles-Weed (2021) when $P \neq Q$, and we employ these results to bound the convergence rates of our empirical optimal transport map estimators in Sections 3.2 and 4.2. Though the empirical plugin estimator of the Wasserstein distance is minimax optimal up to polylogarithmic factors under no assumptions on P and Q, it becomes suboptimal when P and Q are assumed to have smooth densities. Weed and Berthet (2019) derive the minimax rate of estimating smooth densities under the Wasserstein distance, and we build upon their results, together with those of Divol (2021), to characterize the risk of our density plugin estimators (cf. Sections 3.3, 4.3, and 4.4).

Central limit theorems for the empirical quadratic cost $W_2^2(P_n, Q_m)$ around its expectation have been derived by del Barrio and Loubes (2019) under mild conditions on the underlying distributions. As we discuss in Section 5.1, however, the centering sequence $\mathbb{E}W_2^2(P_n, Q_m)$ in these results cannot generally be replaced by its population counterpart $W_2^2(P, Q)$, which is a barrier to their use for statistical inference. Key exceptions are obtained when P and Q are one-dimensional (Munk and Czado, 1998; Freitag and Munk, 2005) or countable (Sommerfeld and Munk, 2018; Tameling et al., 2019), in which case non-degenerate limiting distributions for the process $W_2(P_n, Q_m) - W_2(P, Q)$ are known up to suitable scaling. In contrast, our work derives central limit theorems with desirable centering for any absolutely continuous distributions P and Q admitting sufficiently smooth densities.

Concurrent Work. During the final stages of preparation of the first version of our manuscript, we became aware of the recent independent work of Deb et al. (2021b), which bounds the risk of certain plugin optimal transport map estimators that are closely related to those in our work. In particular, assuming for simplicity that $n = m$, they show that an estimator derived from the empirical plugin optimal transport coupling achieves the $n^{-\left(\frac{1}{2} \wedge \frac{2}{d}\right)}$ convergence rate under the squared $L^2(P_n)$ loss up to polylogarithmic factors. Our work establishes an analogous result using a distinct proof, but further shows that empirical estimators achieve this rate in squared $L^2(P)$ norm, once suitably extended using nonparametric smoothers. We also sharpen this result to the rate $n^{-\left(1 \wedge \frac{2}{d}\right)}$ under additional conditions. Deb et al. (2021b) also analyze the convergence rate of plugin estimators based on wavelet and kernel density estimation. Their work shows that such estimators can achieve, for instance, the faster rate $n^{-\left(\frac{1}{2} \vee \frac{\alpha}{d+2(\alpha-1)}\right)}$, when the underlying densities lie in a $(\alpha - 1)$-Hölder ball for some $\alpha > 1$. While this upper bound illustrates an improvement over empirical estimators in the presence of smoothness, it scales at a quadratically slower rate than the minimax rate (3). In contrast, our work shows that wavelet density plugin estimators do in fact achieve the minimax rate $n^{-\left(1 \vee \frac{2\alpha}{d+2(\alpha-1)}\right)}$ (up to a polylogarithmic factor when $d = 2$). The current version of our manuscript extends this result to kernel density estimators, using a significantly different proof strategy than Deb et al. (2021b). Finally, we emphasize that our sharp analysis of estimators for the Wasserstein distance allows us to deduce that their bias is of lower order than their variance when $2(\alpha+1) > d$, which is a key component in our derivation of their limiting distribution. Indeed, our results in Section 5 on statistical inference for the 2-Wasserstein distance cannot be deduced from the work of Deb et al. (2021b).

Notation. For any $a, b \in \mathbb{R}$, we write $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$, $a_+ = a \vee 0$. If $a \geq 0$, $\lfloor a \rfloor$ and $\lceil a \rceil$ denote the respective floor and ceiling of a. The Euclidean norm on $\mathbb{R}^d$ is denoted $\|\cdot\|$, and the ℓ_p norm of a sequence $(a_n)_{n\geq 1} \subseteq \mathbb{R}$ is written $\|(a_n)_{n\geq 1}\|_{\ell_p} = (\sum_{n\geq 1} |a_n|^p)^{1/p}$ for all $1 \leq p \leq \infty$. Given a set Ω, which is either a closed subset of $\mathbb{R}^d$ or the d-dimensional flat torus $\Omega = \mathbb{T}^d := \mathbb{R}^d/\mathbb{Z}^d$, and given real numbers $\alpha > 0$, $s \in \mathbb{R} \setminus \{0\}$, $1 \leq p, q \leq \infty$, the Hölder spaces $\mathcal{C}^\alpha(\Omega)$, Besov spaces $\mathcal{B}^s_{p,q}(\Omega)$, homogeneous Sobolev spaces $\dot{H}^s(\Omega)$, inhomogeneous Sobolev spaces $H^s(\Omega)$, and their respective norms $\|\cdot\|_{\mathcal{C}^\alpha(\Omega)}$, $\|\cdot\|_{\mathcal{B}^s_{p,q}(\Omega)}$, $\|\cdot\|_{\dot{H}^s(\Omega)}$, $\|\cdot\|_{H^s(\Omega)}$, are defined in Appendix A. We drop the suffix Ω when the underlying space can be understood from context. We also define, for any $M, \gamma > 0$,

$$\mathcal{C}^\alpha(\Omega; M) := \left\{ f \in \mathcal{C}^\alpha(\Omega) : \|f\|_{\mathcal{C}^\alpha(\Omega)} \leq M \right\}, \tag{7}$$

$$\mathcal{C}^\alpha(\Omega; M, \gamma) := \left\{ f \in \mathcal{C}^\alpha(\Omega) : \|f\|_{\mathcal{C}^\alpha(\Omega)} \leq M, f \geq 1/\gamma \text{ over } \Omega \right\}. \tag{8}$$

Furthermore, $\mathcal{C}^\infty(\Omega)$ denotes the set of real-valued functions on Ω which are differentiable up to any order, and $\mathcal{C}^\infty_c(\Omega)$ denotes the set of compactly-supported functions in $\mathcal{C}^\infty(\Omega)$. Given a measure space $(\Omega, \mathcal{F}, \nu)$, $L^p(\nu)$ denotes the Lebesgue space of order $1 \leq p \leq \infty$, endowed with the norm $\|f\|_{L^p(\nu)} = (\int_\Omega |f(x)|^p d\nu(x))^{1/p}$, for any measurable function $f : \Omega \to \mathbb{R}$. We also write $L^p_0(\nu) = \{f \in L^p(\nu) : \int f d\nu = 0\}$. When ν is the Lebesgue measure $\mathcal{L}$ on $\Omega \subseteq \mathbb{R}^d$, we write $L^p(\Omega)$ (or $L^p_0(\Omega)$) instead of $L^p(\mathcal{L})$ (or $L^p_0(\mathcal{L})$). We adopt the same convention when $\Omega \subseteq \mathbb{T}^d$, in which case, by abuse of notation, $\mathcal{L}$ denotes the standard Haar measure over $\mathbb{T}^d$. We often write $\int f$ instead of $\int f d\mathcal{L}$. Given $T : \Omega \to \Omega$, we write by abuse of notation $\|T\|_{L^2(P)} = (\int \|T(x)\|^2 dP(x))^{1/2}$. For any set $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{R}$, we write $\|f\|_\infty = \sup_{x\in\mathcal{X}} |f(x)|$. The Fourier transform of a function $K \in L^1(\mathbb{R}^d)$ is denoted $\mathcal{F}[K](\xi) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i x^\top \xi} dx$ for all $\xi \in \mathbb{R}^d$. For any integer $B \geq 1$, the permutation group on $[B] = \{1, \dots, B\}$ is denoted S_B. The diameter of a set $\Omega \subseteq \mathbb{R}^d$ is denoted $\mathrm{diam}(\Omega) = \sup\{\|x - y\| : x, y \in \Omega\}$, and its interior and closure are respectively denoted Ω° and $\overline{\Omega}$. For all $x \in \mathbb{R}^d$ and $\epsilon > 0$, $B(x, \epsilon) = \{y \in \mathbb{R}^d : \|x - y\| \leq \epsilon\}$. Given sequences $(a_n)_{n=1}^\infty$ and $(b_n)_{n=1}^\infty$, we write $a_n \lesssim b_n$ if there exists $C > 0$ such that $a_n \leq C b_n$ for all $n \geq 1$, and we also write $a_n \asymp b_n$ if $b_n \lesssim a_n \lesssim b_n$. The constant C is always permitted to depend on the dimension d, the domain Ω, and additional problem parameters, whenever they are clear from context. We sometimes emphasize the latter case by using the symbols $\lesssim_{c_1,c_2,\dots}$ or $\asymp_{c_1,c_2,\dots}$, indicating that the suppressed constants depend on the problem parameters $c_1, c_2, \dots$.

2 Background on Optimal Transport

2.1 The Quadratic Optimal Transport Problem over $\mathbb{R}^d$

We provide a brief background on the optimal transport problem over $\mathbb{R}^d$ with respect to the squared Euclidean cost function, and direct the reader to Villani (2003); Santambrogio (2015) for further details. To simplify our exposition, we assume throughout the rest of the paper, except where otherwise specified, that all measures have support contained in a set $\Omega \subseteq \mathbb{R}^d$ satisfying the following condition.

(S1) Ω is a compact set such that $\Omega \subseteq [0,1]^d$.

Notice that once Ω is assumed compact, the final assumption in condition (S1) can always be guaranteed by rescaling. Let $\mathcal{P}(\Omega)$ denote the set of Borel probability measures with support contained in Ω, and $\mathcal{P}_{\mathrm{ac}}(\Omega)$ the subset of such measures which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. As we shall recall in Theorem 1 below, for any $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and $Q \in \mathcal{P}(\Omega)$ the Monge problem (1) admits a minimizer T_0, which is uniquely defined P-almost everywhere. The Monge problem may, however, be infeasible when the absolute continuity condition on P is removed. This observation motivated Kantorovich (1942, 1948) to develop the following convex relaxation of the Monge problem,

$$\operatorname*{argmin}_{\pi \in \Pi(P,Q)} \int_{\Omega} \|x-y\|^2 \, d\pi(x,y), \tag{9}$$

known as the Kantorovich problem, where $\Pi(P,Q)$ denotes the set of joint distributions on Ω^2 with marginal distributions P and Q, known as couplings of P and Q. That is,

$$\Pi(P,Q) = \{\pi \in \mathcal{P}(\Omega^2) : \pi(\cdot \times \Omega) = P, \pi(\Omega \times \cdot) = Q\}.$$

Notice that the Kantorovich problem is always feasible since $P \otimes Q \in \Pi(P,Q)$. It can be shown in our setting that a minimizer π in equation (9) always exists (Theorem 4.1, Villani (2008)), and is called an optimal coupling. In the special case where π is supported in the graph of a map $T_0 : \Omega \to \Omega$, it must be the case that $T_0 \in \mathcal{T}(P,Q)$ due to the marginal constraints in the definition of $\Pi(P,Q)$, and it must then follow that T_0 is precisely an optimal transport map from P to Q. As we shall elaborate below, this situation turns out to characterize all optimal couplings when $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and for such measures the Monge and Kantorovich problems yield equivalent solutions.

While an optimal coupling represents a transference plan for reconfiguring P into Q, the corresponding optimal value of the objective function (9) represents the optimal cost of such a reconfiguration, which provides an easily interpretable measure of divergence between P and Q. Specifically, it gives rise to the 2-Wasserstein distance,

$$W_2(P,Q) = \left(\inf_{\pi \in \Pi(P,Q)} \int \|x-y\|^2 \, d\pi(x,y) \right)^{\frac{1}{2}}. \tag{10}$$

The above problem is an (infinite-dimensional) convex program with linear constraints, and it admits a dual maximization problem, known as the Kantorovich dual problem, given by

$$W_2^2(P,Q) = \sup_{(\phi,\psi) \in \mathcal{K}} \int \phi dP + \int \psi dQ, \tag{11}$$

where $\mathcal{K}$ is the set of pairs $(\phi,\psi) \in L^1(\Omega) \times L^1(\Omega)$ such that $\phi(x) + \psi(y) \leq \|x-y\|^2$ for all $x, y \in \Omega$. In the present setting of the quadratic optimal transport problem over the compact set Ω, it can be shown that strong duality indeed holds in equation (11), and that the supremum is always achieved by some pair $(\phi_0, \psi_0) \in \mathcal{K}$. Any such pair of functions is called a pair of Kantorovich potentials. In this case, notice that (ϕ_0, ϕ_0^c), with $\phi_0^c(y) = \inf_{x \in \Omega} \left\{ \|x-y\|^2 - \phi_0(x) \right\}$, is itself a pair of Kantorovich potentials, since replacing ψ_0 by ϕ_0^c can only increase the objective value (11), while retaining the constraint $(\phi_0, \phi_0^c) \in \mathcal{K}$. If we define $\varphi_0 = \|\cdot\|^2 - 2\phi_0$, then $\phi_0^c = \|\cdot\|^2 - 2\varphi_0^*$, where for any $f : \Omega \to \mathbb{R}$,

$$f^*(y) = \sup_{x\in\Omega} \big\{ \langle x, y\rangle - f(x) \big\}, \quad y \in \Omega,$$

denotes the Legendre-Fenchel conjugate of f. Under this reparametrization, the Kantorovich dual problem is equivalent to the so-called semi-dual problem

$$\inf_{\varphi\in L^1(P)} \int \varphi dP + \int \varphi^* dQ, \tag{12}$$

in the sense that φ_0 is a solution to the semi-dual problem if and only if $(\|\cdot\|^2 - 2\varphi_0, \|\cdot\|^2 - 2\varphi_0^*)$ is a solution to the Kantorovich dual problem (11). The significance of the semi-dual problem is in part due to its connection to the Monge problem, as described by the following result due to Knott and Smith (1984); Brenier (1991).

Theorem 1 (Brenier's Theorem)**.** *Let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and $Q \in \mathcal{P}(\Omega)$.*

(i) There exists an optimal transport map T_0 between P and Q which takes the form $T_0 = \nabla\varphi_0$ for a convex function $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$ which solves the semi-dual problem (12). Furthermore, T_0 is uniquely determined P-almost everywhere.

(ii) If we further have $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, then $\nabla\varphi_0^*$ is the (Q-almost everywhere uniquely determined) gradient of a convex function such that $\nabla\varphi^*_{0\#}Q = P$, and solves the Monge problem for transporting Q onto P. Furthermore, for Lebesgue-almost every $x, y \in \Omega$

$$\nabla\varphi_0^* \circ \nabla\varphi_0(x) = x, \quad \nabla\varphi_0 \circ \nabla\varphi_0^*(y) = y.$$

Brenier's Theorem implies the aforementioned fact that a unique optimal transport map exists between any absolutely continuous distribution P and any distribution Q, where uniqueness is always understood in the Lebesgue-almost everywhere sense. It further characterizes this map as the gradient of an optimal semi-dual potential φ_0, which we refer to as a Brenier potential in the sequel.

The convexity of φ_0 already implies that it will be almost-everywhere twice differentiable. Further smoothness properties of Brenier potentials, and therefore of optimal transport maps, have been studied via the regularity theory of partial differential equations of the Monge-Ampère type, and we refer to De Philippis and Figalli (2014); Figalli (2017) for surveys. In short, denote by p, q the respective Lebesgue densities of $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume φ_0 is in fact everywhere twice continuously differentiable. Then, the constraint $\nabla\varphi_{0\#}P = Q$ implies by the change of variable formula that φ_0 solves the equation

$$\det\big(\nabla^2\varphi_0(x)\big) = \frac{p(x)}{q(\nabla\varphi_0(x))}, \quad x \in \Omega. \tag{13}$$

As a direct consequence of equation (13), notice that the Hessian $\nabla^2\varphi_0$ admits a uniformly bounded determinant whenever p and q are bounded, and bounded away from zero. This observation leads to the following simple result noted by Gigli (2011).

Lemma 2. *Assume $\varphi_0 \in \mathcal{C}^2(\Omega)$ and $\gamma^{-1} \le p, q \le \gamma$ for some $\gamma > 0$. Then, there exists a constant $\lambda > 0$, depending only on γ and $\|\varphi_0\|_{\mathcal{C}^2(\Omega)}$, such that φ_0 is λ-strongly convex.*

Lemma 2 shows that, whenever equation (13) has positive and bounded right-hand side, smooth Brenier potentials are also strongly convex. We shall require this property in Section 3.1 to derive stability bounds for the $L^2(P)$ loss. To further obtain sufficient conditions for the Hölder smoothness of φ_0, notice that the Monge-Ampère equation (13) suggests that φ_0 admits two degrees of smoothness more than the densities p and q. This intuition indeed turns out to hold true under suitable regularity conditions on Ω, as was established in a series of publications by Caffarelli (1991, 1992a,b, 1996). The following is a summary of these results, as stated by Villani (2008, Chapter 12).

Theorem 3 (Caffarelli's Regularity Theory)**.** *Assume Ω is convex and satisfies condition* ***(S1)****. Assume further that there exists $\gamma > 0$ such that $\gamma^{-1} \leq p, q \leq \gamma$ over Ω. Then, the Brenier potential φ_0 is unique up to an additive constant, and satisfies the following.*

(i) (Interior Regularity) Suppose there exists $\alpha > 1$, $\alpha \notin \mathbb{N}$, such that $p, q \in \mathcal{C}^{\alpha-1}(\Omega^\circ)$. Then $\varphi_0 \in \mathcal{C}^{\alpha+1}(\Omega^\circ)$. Moreover, for any open subdomain Ω' such that $\overline{\Omega'} \subseteq \Omega^\circ$, there exists a constant $C > 0$ depending on $\gamma, \alpha, \Omega, \Omega', \|\varphi_0\|_{L^\infty(\Omega)}, \|p\|_{\mathcal{C}^{\alpha-1}(\Omega^\circ)}, \|q\|_{\mathcal{C}^{\alpha-1}(\Omega^\circ)}$ such that

$$\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\Omega')} \leq C.$$

(ii) (Global Regularity) Assume Ω admits a $\mathcal{C}^2$ boundary and is uniformly convex. Assume further that there exists $\alpha > 1$, $\alpha \notin \mathbb{N}$, such that $p, q \in \mathcal{C}^{\alpha-1}(\Omega)$. Then, $\varphi_0 \in \mathcal{C}^{\alpha+1}(\Omega)$.

Theorem 3(ii) implies that, under suitable conditions, the optimal transport map T_0 inherits one degree of smoothness more than the densities p and q over Ω. Unlike the interior regularity result of Theorem 3(i), however, Theorem 3(ii) does not imply a uniform bound on $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\Omega)}$, and therefore does not preclude the possibility that the latter quantity diverges when p, q vary in a $\mathcal{C}^{\alpha-1}(\Omega)$ ball. Closely related global regularity results have also been established by Urbas (1997) under slightly stronger conditions, but we do not know if either of these results can be made uniform up to the boundary in an analogous way to the interior result of Theorem 3(i). Whenever global uniform regularity results are needed in our development, we sidestep this issue by working with the optimal transport problem over the torus, for which boundary considerations do not arise.

2.2 The Quadratic Optimal Transport Problem over the Flat Torus

Denote by $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$ the flat d-dimensional torus. Specifically, $\mathbb{T}^d$ is the set of equivalence classes $[x] = \{x + k : k \in \mathbb{Z}^d\}$, for all $x \in [0,1)^d$. Abusing notation, we typically write x instead of $[x]$. $\mathbb{T}^d$ is endowed with the standard metric

$$d_{\mathbb{T}^d}(x, y) = \min\{\|x - y + k\| : k \in \mathbb{Z}^d\}, \quad x, y \in \mathbb{T}^d.$$

We identify $\mathcal{P}(\mathbb{T}^d)$ with the set of Borel measures P on $\mathbb{R}^d$ such that $P([0,1)^d) = 1$ and which are $\mathbb{Z}^d$-periodic, in the sense that $P(B) = P(k + B)$ for all $k \in \mathbb{Z}^d$ and all Borel sets $B \subseteq \mathbb{R}^d$. Furthermore, $\mathcal{P}_{\text{ac}}(\mathbb{T}^d)$ denotes the subset of measures in $\mathcal{P}(\mathbb{T}^d)$ which are absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^d$. A function $f : \mathbb{T}^d \to \mathbb{R}$ is understood to be a function on $\mathbb{R}^d$ which is $\mathbb{Z}^d$-periodic, and we write $T : \mathbb{T}^d \to \mathbb{T}^d$ when T is a map from $\mathbb{R}^d$ to $\mathbb{R}^d$ such that $[T(x)] = [T(y)]$ whenever $[x] = [y]$.

The optimal transport problem over $\mathbb{T}^d$ with the quadratic cost $d^2_{\mathbb{T}^d}$ largely mirrors that of the squared Euclidean cost over $\mathbb{R}^d$. Define for all $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ the Monge problem

$$\operatorname*{argmin}_{T \in \mathcal{T}(P,Q)} \int_{\mathbb{T}^d} d^2_{\mathbb{T}^d}(x, T(x)) dP(x), \tag{14}$$

where the integral is understood as being taken over $[0,1)^d$. The Kantorovich problem and its dual give rise to the squared Wasserstein distance over $\mathcal{P}(\mathbb{T}^d)$,

$$W_2^2(P,Q) = \inf_{\pi \in \Pi(P,Q)} \int d^2_{\mathbb{T}^d}(x,y) d\pi(x,y) = \sup_{(\varphi,\psi) \in \mathcal{K}_T} \int \varphi dP + \int \psi dQ, \tag{15}$$

where $\mathcal{K}_T$ denotes the set of pairs of potentials $(\varphi, \psi) \in L^1(P) \times L^1(Q)$ satisfying the dual constraint $\varphi(x) + \psi(y) \le d^2_{\mathbb{T}^d}(x,y)$ for all $x, y \in \mathbb{T}^d$. We abuse notation by writing W_2 to denote both the 2-Wasserstein distance over $\mathbb{R}^d$ and $\mathbb{T}^d$. Whenever we speak of the optimal transport problem or Wasserstein distance between two measures $P, Q \in \mathcal{P}(\Omega)$, the underlying cost function is tacitly understood to be $\|\cdot\|^2$ when $\Omega \subseteq \mathbb{R}^d$, and $d^2_{\mathbb{T}^d}$ when $\Omega = \mathbb{T}^d$.

As in the Euclidean setting, the Kantorovich duality in the above display is equivalent to a semi-dual problem, whose solution characterizes the Monge problem. Indeed, the following result due to Cordero-Erausquin (1999) is an analogue of Brenier's Theorem, together with additional properties about the optimal transport problem over $\mathbb{T}^d$.

Proposition 4. *Let $P \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ and $Q \in \mathcal{P}(\mathbb{T}^d)$. Then, there exists a ($P$-almost everywhere uniquely determined) optimal transport map $T_0 = \nabla\varphi_0$ from P to Q which solves the Monge problem (14), where $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$ is a convex function satisfying the following properties.*

(i) $\|\cdot\|^2/2 - \varphi_0$ *is $\mathbb{Z}^d$-periodic.*

(ii) $T_0(x+k) = T_0(x) + k$ *for almost every $x \in \mathbb{R}^d$ and $k \in \mathbb{Z}^d$.*

(iii) For P-almost all $x \in \mathbb{R}^d$, $\|T_0(x) - x\| \le \mathrm{diam}(\mathbb{T}^d) = \sqrt{d}/2$ and $\|T_0(x) - x\| = d_{\mathbb{T}^d}(x, T_0(x))$.

Assume further that $Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$, and denote the respective densities of P, Q by p, q. Then,

(v) $\nabla\varphi_0^*$ *is the (Q-almost everywhere unique) optimal transport map from Q to P.*

(vi) $(\|\cdot\|^2 - 2\varphi_0, \|\cdot\|^2 - 2\varphi_0^*)$ *is a pair of optimal Kantorovich potentials in equation (15).*

(vii) If $\varphi_0 \in \mathcal{C}^2(\mathbb{R}^d)$, then it solves the Monge-Ampère equation

$$\det(\nabla^2\varphi_0(x)) q(\nabla\varphi_0(x)) = p(x), \quad x \in \mathbb{R}^d.$$

In particular, if $\gamma^{-1} \le p, q \le \gamma$ for some $\gamma > 0$, then φ_0 is λ-strongly convex, for some constant $\lambda > 0$ depending only on γ and $\|\varphi_0\|_{\mathcal{C}^2(\mathbb{R}^d)}$.

With Proposition 4 in place, regularity properties of Brenier potentials φ_0 may be deduced from smoothness conditions on p, q. The following result was stated by Cordero-Erausquin (1999) without explicit mention of the uniformity of the Hölder norms appearing therein, but can readily be made uniform using Caffarelli's interior regularity theory (Theorem 3(i); Figalli (2017), Chapter 4). We also note that this result was stated by Ambrosio et al. (2012) in the special case $d = 2$.

Theorem 5. *Let $P, Q \in \mathcal{P}(\mathbb{T}^d)$ be absolutely continuous with respect to the Lebesgue measure, with respective densities p, q satisfying $\gamma^{-1} \leq p, q \leq \gamma$ for some $\gamma > 0$. Assume further that $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d)$ for some $\alpha > 1$. Then, there exists a constant $C > 0$ depending only on $\alpha, \gamma, \|p\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)}$ and $\|q\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)}$ such that, $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{R}^d)} \leq C$.*

3 The One-Sample Problem

Throughout this section, we let $P \in \mathcal{P}_{\text{ac}}(\Omega)$ denote a known distribution, and $Q \in \mathcal{P}_{\text{ac}}(\mathcal{Y})$ denote an unknown distribution from which an i.i.d. sample $Y_1, \dots, Y_n \sim Q$ is observed. Let p and q denote their respective densities, and let $T_0 = \nabla\varphi_0$ denote the unique optimal transport map from P to Q, with respect to a convex Brenier potential φ_0. We also denote by $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ the Kantorovich potentials induced by φ_0. We assume condition **(S1)** holds throughout this section, and we may therefore assume without loss of generality that $-d \leq \phi_0 \leq 0$ and $0 \leq \psi_0 \leq d$ over Ω (Villani (2003), Remark 1.13).

Unlike the two-sample case which we discuss in Section 4, there exist canonical estimators of T_0 when the source distribution P is known. Indeed, since P is absolutely continuous, Brenier's Theorem implies that there exists a unique optimal transport map $\widehat{T}$ between P and any estimator $\widehat{Q}$ of Q, and we analyze two such examples below. We first take $\widehat{Q}$ to be the empirical measure of Q in Section 3.2, and show that the resulting estimator $\widehat{T}$ achieves the minimax risk of estimating Lipschitz optimal transport maps, under essentially no smoothness conditions on the underlying measures. In Section 3.3, we then take $\widehat{Q}$ to be a density estimator, leading to an estimator $\widehat{T}$ achieving faster rates of convergence when Q admits a smooth density. In both cases, our analysis will hinge upon known upper bounds on the risk of $\widehat{Q}$ under the Wasserstein distance, by invoking a key stability bound which we turn to first.

3.1 A General Stability Bound

The main technical result of this section will be stated under the following curvature condition.

A1(λ) The Brenier potential φ_0 is a closed convex function such that $\varphi_0 \in \mathcal{C}^2(\Omega)$ and $(1/\lambda)I_d \preceq \nabla^2\varphi_0(x) \preceq \lambda I_d$ for all $x \in \Omega$.

Condition **A1(λ)** implies in particular that T_0 is λ-Lipschitz over Ω. As noted in Lemma 2, whenever P and Q both admit densities satisfying $\gamma^{-1} \leq p, q \leq \gamma$ over Ω, for some $\gamma > 0$, the second inequality of **A1(λ)** is sufficient to imply the first, up to inflating λ by a factor depending on γ. Under this condition, we prove the following stability bounds in Appendix C.

Theorem 6. *Let $P, Q \in \mathcal{P}_{\text{ac}}(\Omega)$, and assume condition **A1(λ)** holds for some $\lambda > 0$. For any $\widehat{Q} \in \mathcal{P}(\Omega)$, let $\widehat{T} = \nabla\widehat{\varphi}$ be the unique optimal transport map from P to $\widehat{Q}$. Then,*

$$\frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 \leq W_2^2(P, \widehat{Q}) - W_2^2(P, Q) - \int \psi_0 d(\widehat{Q} - Q) \leq \lambda W_2^2(\widehat{Q}, Q).$$

We make several remarks regarding Theorem 6.

- Caffarelli's regularity theory (cf. Theorem 3) provides sufficient conditions on the smoothness of P, Q and $\partial\Omega$ for assumption **A1(λ)** to hold, albeit for a non-universal constant $\lambda > 0$. We note, however, that our assumption is considerably weaker. For instance, condition **A1(λ)** is satisfied whenever P and Q differ by a location transformation, irrespective of the regularity or positivity of their Lebesgue densities.

- We show in Section 5.2 that, under weaker assumptions than those of Theorem 6, the map $\psi_0 - \mathbb{E}_Q[\psi_0(Y)]$ is the efficient influence function of the functional $Q \in \mathcal{P}(\Omega) \mapsto W_2^2(P,Q)$ with respect to the tangent space $L_0^2(Q)$. It follows that the linear functional

$$L(\widehat{Q}) = \int \psi_0 d(\widehat{Q} - Q) \tag{16}$$

  is the first-order term in the von Mises expansion of $W_2^2(P,\widehat{Q})$ around $W_2^2(P,Q)$. The upper bound of Theorem 6 implies that the remainder of this expansion decays quadratically in the topology induced by W_2, a fact which we shall use to derive upper bounds and limit theorems for plugin estimators of the Wasserstein distance. This fact combined with the lower bound of Theorem 6 further implies the following remarkable equivalence,

$$\frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)} \leq W_2(\widehat{Q}, Q) \leq \|\widehat{T} - T_0\|_{L^2(P)}. \tag{17}$$

  Notice that the second inequality always holds due to the fact that $(\widehat{T}, T_0)_\# P$ is a coupling of $\widehat{Q}$ and Q. Equation (17) thus shows that the transport cost of this coupling is within a universal factor of being optimal, when the curvature condition **A1(λ)** is in force. We use this result to obtain upper bounds on the risk of one-sample plugin estimators $\widehat{T}$ by appealing to the corresponding risk of $\widehat{Q}$ under the Wasserstein distance. We note that weaker analogues of equation (17), in which the left-hand side admits an exponent greater than unity, have previously been derived by Mérigot et al. (2019); Delalande and Mérigot (2021). Those works adopted a weaker assumption than ours, however.

- Suppose that, in addition to the assumptions of Theorem 6, the measures Q and $\widehat{Q}$ are both absolutely continuous with respect to the Lebesgue measure, with respective densities q and $\widehat{q}$ which satisfy $\gamma^{-1} \leq q, \widehat{q} \leq \gamma$ over Ω, for some $\gamma > 0$. In this setting, it was shown by Peyre (2018) that the 2-Wasserstein distance is equivalent to the negative-order homogeneous Sobolev norm $\|\cdot\|_{\dot{H}^{-1}(\Omega)}$, in the sense that, under suitable conditions on Ω,

$$\gamma^{-1}\|\widehat{q} - q\|^2_{\dot{H}^{-1}(\Omega)} \lesssim W_2^2(\widehat{Q}, Q) \lesssim \gamma\|\widehat{q} - q\|^2_{\dot{H}^{-1}(\Omega)}. \tag{18}$$

  Theorem 6 and the above display then imply

$$\frac{1}{\lambda\gamma}\|\widehat{\varphi} - \varphi_0\|^2_{\dot{H}^1(\Omega)} \lesssim W_2^2(P,\widehat{Q}) - W_2^2(P,Q) - \int \psi_0 d(\widehat{Q} - Q) \lesssim \lambda\gamma\|\widehat{q} - q\|^2_{\dot{H}^{-1}(\Omega)}.$$

  It follows from the upper bound that $W_2^2(P,\cdot)$, when viewed as a functional of $\widehat{q}$, is Fréchet differentiable at q in the $\dot{H}^{-1}(\Omega)$ topology. It moreover implies that this functional is strongly convex and smooth with respect to the duality of the spaces $\dot{H}^{-1}(\Omega)$ and $\dot{H}^1(\Omega)$.

- Finally, one may also infer from Theorem 6 and the Kantorovich duality that,

$$\frac{1}{2\lambda}\|\nabla\widehat{\varphi} - \nabla\varphi_0\|^2_{L^2(P)} \leq \int (\varphi_0 - \widehat{\varphi})dP + \int (\varphi_0^* - \widehat{\varphi}^*)d\widehat{Q} \leq \frac{\lambda}{2}\|\nabla\widehat{\varphi} - \nabla\varphi_0\|^2_{L^2(P)}. \tag{19}$$

Equation (19) is a direct analogue of a stability bound proven by Hütter and Rigollet (2021, Proposition 10), who show that similar inequalities hold when the measure $\widehat{Q}$ appearing in the above display is replaced by Q. Their result assumes, however, that $\widehat{\varphi}$ itself satisfies condition **A1(λ)**. In contrast, we do not place any conditions on the estimator $\widehat{T}$ beyond it being the optimal transport map from P to $\widehat{Q}$. This will permit our study of transport map estimators which are potentially nonsmooth but easy to compute, as we show next.

3.2 Upper Bounds for the One-Sample Empirical Estimators

Recall that $Q_n = (1/n)\sum_{i=1}^n \delta_{Y_i}$ denotes the empirical measure. Since P is known and absolutely continuous, a natural estimator for T_0 is the optimal transport map T_n from P to Q_n, defined by

$$T_n = \operatorname*{argmin}_{T \in \mathcal{T}(P,Q_n)} \int \|x - T(x)\|^2 \, dP(x). \tag{20}$$

By Brenier's Theorem, the minimizer T_n in the above display exists and is uniquely determined P-almost everywhere. The optimization problem (20) is sometimes known as the semi-discrete optimal transport problem, for which efficient numerical solvers are well-studied (Mérigot, 2011; Levy and Schwindt, 2018).

In view of the stability bound in Theorem 6, the risk of T_n may be related to that of the empirical measure Q_n under the Wasserstein distance. For instance, from the work of Fournier and Guillin (2015) we obtain the following bound, under no assumptions beyond **(S1)**,

$$\mathbb{E}W_2^2(Q_n, Q) \lesssim \kappa_n := \begin{cases} n^{-1/2}, & d \le 3 \\ n^{-1/2}\log n, & d = 4 \\ n^{-2/d}, & d \ge 5. \end{cases} \tag{21}$$

The following bound on the risk of T_n is now an immediate consequence of Theorem 6, together with the fact that the functional L in equation (16) satisfies $\mathbb{E}[L(Q_n)] = 0$.

Corollary 7. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ and assume condition **A1(λ)** holds. Then,*

$$\mathbb{E}\big\|T_n - T_0\big\|_{L^2(P)}^2 \asymp_\lambda \mathbb{E}\big[W_2^2(P, Q_n) - W_2^2(P, Q)\big] \asymp_\lambda \mathbb{E}W_2^2(Q_n, Q) \lesssim \kappa_n.$$

When $d \ge 5$, Corollary 7 implies that the empirical estimator T_n achieves the minimax lower bound (3) for estimating Lipschitz transport maps T_0. On the other hand, when $1 \le d \le 4$, the rate κ_n does not improve beyond $n^{-1/2}$, unlike the minimax lower bound (3) of Hütter and Rigollet (2021), which scales as fast as $1/n$. This observation does not imply that the plugin estimator T_n is minimax suboptimal, since equation (3) holds under stronger assumptions than those of Corollary 7. In particular, it assumes that these distributions admit densities which are bounded away from zero, and thus have connected support. In contrast, Corollary 7 applies to measures P and Q with possibly disconnected support, for which our upper bound of κ_n cannot generally be improved up to a logarithmic factor—similar considerations are discussed for the convergence rate of the empirical measure by Bobkov and Ledoux (2019) when $d = 1$, and more generally by Weed and Berthet (2019).

Nevertheless, when we further assume that Q has a positive density, the result of Corollary 7 can be strengthened to match the minimax rate of Hütter and Rigollet (2021) even for $d \le 4$.

For instance, it was shown by Ledoux (2019) and references therein that, when Q is the uniform distribution on $[0,1]^d$, Q_n achieves the following faster rate,

$$\mathbb{E}W_2^2(Q_n, Q) \lesssim \begin{cases} n^{-1}, & d = 1 \\ n^{-1}\log n, & d = 2 \\ n^{-2/d}, & d \geq 3. \end{cases} \tag{22}$$

Such a result is also known to hold for any measure Q admitting positive density over a compact subset of the real line (Bobkov and Ledoux, 2019), or over the flat torus (Divol, 2021). Inspired by the latter result and by the work of Weed and Berthet (2019), we prove an analogous result for general measures supported on the unit hypercube, at the expense of an inflated polylogarithmic factor when $d = 2$.

Corollary 8. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$ and assume that condition **A1(λ)** holds. Assume further that $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$, for some $\gamma > 0$. Then,*

$$\mathbb{E}\big\|T_n - T_0\big\|_{L^2(P)}^2 \asymp \mathbb{E}\big[W_2^2(P, Q_n) - W_2^2(P, Q)\big] \asymp \mathbb{E}W_2^2(Q_n, Q) \lesssim \bar{\kappa}_n := \begin{cases} n^{-1}, & d = 1 \\ \frac{(\log n)^2}{n}, & d = 2 \\ n^{-2/d}, & d \geq 3. \end{cases}$$

Under the assumptions of Corollary 8, we deduce that the plugin estimator T_n is minimax optimal for all $d \geq 1$, up to a polylogarithmic factor when $d = 2$. The scale of this factor is further discussed following the statement of Theorem 18.

This result also provides a sharper bound on the bias of $W_2^2(P, Q_n)$ than could have been deduced from Chizat et al. (2020), who show that the risk of this estimator decays at the rate κ_n using distinct techniques. Indeed, Corollaries 7–8 can also be extended to recover the risk bounds of Chizat et al. (2020) under stronger conditions, though with an improved rate of convergence when P approaches Q in Wasserstein distance.

Corollary 9. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume condition **A1(λ)** holds. Then,*

$$\mathbb{E}\big|W_2^2(P, Q_n) - W_2^2(P, Q)\big| \lesssim_\lambda \mathbb{E}W_2^2(Q_n, Q) + n^{-\frac{1}{2}} \lesssim \kappa_n. \tag{23}$$

If we further assume that $\Omega = [0,1]^d$ and $\gamma^{-1} \leq q \leq \gamma$ over Ω for some $\gamma > 0$, then

$$\mathbb{E}\big|W_2^2(P, Q_n) - W_2^2(P, Q)\big| \lesssim_{\lambda,\gamma} \bar{\kappa}_n + W_2(P, Q) n^{-\frac{1}{2}}. \tag{24}$$

Equation (24) exhibits an upper bound on the risk of $W_2^2(P, Q_n)$ that interpolates between the fast rate $\bar{\kappa}_n$ when $W_2(P, Q) \lesssim n^{-1/2}$, and the rate κ_n of Chizat et al. (2020), which is minimax optimal when the distance between P and Q is unconstrained (Manole and Niles-Weed, 2021). We defer the proofs of Corollaries 8–9 to Appendix D.

3.3 Upper Bounds for One-Sample Wavelet Estimators

While the empirical estimator in the previous section achieves the minimax rate of estimating Lipschitz optimal transport maps, we do not generally expect it to achieve faster rates of convergence if T_0 is assumed to enjoy further regularity. We instead show that such improvements can be achieved when Q admits a smooth density q, and when the empirical measure Q_n is replaced by the distribution $\widehat{Q}_n$ of a density estimator $\widehat{q}_n$. Specifically, define

$$\widehat{T}_n = \operatorname*{argmin}_{T \in \mathcal{T}(P, \widehat{Q}_n)} \int \|x - T(x)\|^2 \, dP(x). \tag{25}$$

We focus on the case where $\widehat{q}_n$ is a wavelet density estimator, for which sharp risk estimates under the Wasserstein distance have been established by Weed and Berthet (2019). In order to appeal to their results, we assume that the sampling domain is the unit hypercube $\Omega = [0,1]^d$.

We briefly introduce notation from the theory of wavelets, and refer the reader to Appendix A for a detailed summary and references. To define a basis over the unit cube Ω, we focus on the boundary-corrected N-th Daubechies wavelet system, for an integer $N \geq 2$, as introduced by Cohen et al. (1993). In short, given an integer $j_0 \geq \log_2 N$, their construction leads to respective families of scaling and wavelet functions

$$\Phi^{\mathrm{bc}} = \{\zeta_{j_0 k}^{\mathrm{bc}} : 0 \leq k \leq 2^{j_0} - 1\}, \quad \Psi_j^{\mathrm{bc}} = \{\xi_{jk\ell}^{\mathrm{bc}} : 0 \leq k \leq 2^{j_0} - 1, \ell \in \{0,1\}^d \setminus \{0\}\}, \quad j \geq j_0,$$

such that $\Psi^{\mathrm{bc}} = \Phi^{\mathrm{bc}} \cup \bigcup_{j=j_0}^{\infty} \Psi_j^{\mathrm{bc}}$ forms an orthonormal basis of $L^2(\Omega)$, with the property that Φ^{bc} spans all polynomials of degree at most $N-1$ over Ω. Given a probability distribution $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admitting density $q \in L^2(\Omega)$, one then has

$$q = \sum_{\xi \in \Psi^{\mathrm{bc}}} \beta_\xi \xi = \sum_{\zeta \in \Phi^{\mathrm{bc}}} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} \beta_\xi \xi, \quad \text{where} \quad \beta_\xi = \int \xi dQ, \ \xi \in \Psi^{\mathrm{bc}},$$

where the series converges at least in $L^2(\Omega)$. The standard truncated wavelet estimator of q (Kerkyacharian and Picard, 1992) with a truncation level $J_n \geq j_0 > 0$ is then given by

$$\widetilde{q}_n^{(\mathrm{bc})} = \sum_{\xi \in \Psi^{\mathrm{bc}}} \widehat{\beta}_\xi \xi = \sum_{\zeta \in \Phi^{\mathrm{bc}}} \widehat{\beta}_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j^{\mathrm{bc}}} \widehat{\beta}_\xi \xi, \quad \text{where} \quad \widehat{\beta}_\xi = \int \xi dQ_n, \ \xi \in \Psi^{\mathrm{bc}}.$$

Notice that $\widetilde{q}_n^{(\mathrm{bc})}$ is permitted to take on negative values, in which case it does not define a probability density. We instead define the final estimator $\widehat{q}_n^{(\mathrm{bc})}$ in equation (25) by

$$\widehat{q}_n^{(\mathrm{bc})} = \frac{\widetilde{q}_n^{(\mathrm{bc})} I(\widetilde{q}_n^{(\mathrm{bc})} \geq 0)}{\int \widetilde{q}_n^{(\mathrm{bc})} I(\widetilde{q}_n^{(\mathrm{bc})} \geq 0)}, \quad \text{over } \Omega. \tag{26}$$

We drop all superscripts "bc" in the sequel whenever the choice of wavelet system is unambiguous. Weed and Berthet (2019) bounded the Wasserstein risk of a wavelet density estimator obtained from a distinct modification of $\widetilde{q}_n$. By appealing to L^∞ concentration inequalities for wavelet density estimators (Masry, 1997), we show in Appendix A.4.4 that their result carries over to the estimator $\widehat{q}_n$. Equipped with this result, we arrive at the following bound on the risk of the estimator $\widehat{T}_n \equiv \widehat{T}_n^{(\mathrm{bc})}$ defined in equation (25), and of the corresponding plugin estimator of the squared Wasserstein distance. Recall that the Hölder balls $\mathcal{C}^\alpha(\Omega; \cdot)$ and $\mathcal{C}^\alpha(\Omega; \cdot, \cdot)$ are defined in equations (7)–(8).

Theorem 10 (One-Sample Wavelet Estimators)**.** *Let $P, Q \in \mathcal{P}_{\rm ac}([0,1]^d)$ admit respective densities $p, q \in \mathcal{C}^{\alpha-1}([0,1]^d; M, \gamma)$, for some $\alpha > 1$ and $M, \gamma > 0$. Let $2^{J_n} \asymp n^{1/(d+2(\alpha-1))}$.*

(i) (Optimal Transport Maps) Assume φ_0 satisfies condition **A1(λ)** *for some $\lambda > 0$. Then, there exists a constant $C > 0$ depending on $M, \lambda, \gamma, \alpha$ such that the following hold,*

$$\mathbb{E}\big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \leq C R_{T,n}(\alpha), \quad \text{where } R_{T,n}(\alpha) := \begin{cases} 1/n, & d = 1 \\ (\log n)^2/n, & d = 2 \\ n^{-\frac{2\alpha}{2(\alpha-1)+d}}, & d \geq 3. \end{cases} \tag{27}$$

(ii) (Wasserstein Distances) Assume that for some $\lambda > 0$, $\varphi_0^* \in \mathcal{C}^{\alpha+1}([0,1]^d; \lambda)$. Then, there exists a constant $C > 0$ depending on $M, \lambda, \gamma, \alpha$ such that the following hold,

$$\big|\mathbb{E} W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \leq C R_{T,n}(\alpha),$$
$$\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big|^2 \leq \left[C R_{T,n}(\alpha) + \sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}} \right]^2.$$

Theorem 10 requires smoothness assumptions on both the density q and the potential φ_0^*; in particular, the assumption of Theorem 10(ii) requires both $q \in \mathcal{C}^{\alpha-1}(\Omega)$ and $\varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$. Caffarelli's regularity theory (Theorem 3) suggests that the former condition on q should be sufficient to imply the latter condition on φ_0^*, but such results cannot be invoked here due to the lack of smoothness of the boundary of the unit cube $[0,1]^d$. Even if the above analysis could be adapted to a domain Ω with smooth boundary, the lack of uniformity in Caffarelli's global regularity theory would prevent the bounds in Theorem 10 from holding uniformly in P and Q, in the absence of a smoothness condition on φ_0^*. We refer to Appendix E of Hütter and Rigollet (2021) for related discussions. In Proposition 43 of Appendix H, we will show that an analogue of Theorem 10 holds merely under smoothness conditions on p and q when Ω is the d-dimensional torus $\mathbb{T}^d$, which enjoys the global regularity result of Theorem 5. Here, instead impose smoothness conditions on both φ_0^* and q, in which case $\widehat{T}_n$ achieves the minimax rate (3) of estimating an α-Hölder optimal transport map.

Theorem 10(ii) also proves that the bias of $W_2^2(P, \widehat{Q}_n)$ achieves the same convergence rate, as does its risk when $d \geq 2(\alpha+1)$. In the high-smoothness regime $d < 2(\alpha+1)$, the risk of $W_2^2(P, \widehat{Q}_n)$, in squared loss, does not generally improve beyond the parametric rate $1/n$, except when $\mathrm{Var}_Q[\psi_0(Y)]$ vanishes. Using Lemma 34 in Appendix B, the latter quantity is bounded above by $W_2^2(P, Q)$ up to a constant, so Theorem 10(ii) also implies

$$\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \lesssim_{M,\gamma,\lambda,\alpha} R_{T,n}(\alpha) + \frac{W_2(P, Q)}{\sqrt{n}}. \tag{28}$$

We briefly highlight the main components of the proof of Theorem 10. Both assertions are proven by combining the stability results of Theorem 6 with the bound $\mathbb{E} W_2^2(\widehat{Q}_n, Q) \lesssim R_{T,n}(\alpha)$, which is stated formally in Lemma 30, and extends a result due to Weed and Berthet (2019). In particular, Theorem 10(i) follows immediately from the equivalence (17). Our proof of Theorem 10(ii) additionally requires us to analyze the evaluation $L(\widehat{Q}_n)$ of the linear functional L defined in equation (16), for which we prove the following.

Lemma 11. *Assume the same conditions as Theorem 10(ii). Then,*

$$\mathbb{E}[L(\widehat{Q}_n)] = O\left(2^{-2J_n\alpha}\right), \quad \mathrm{Var}\left[L(\widehat{Q}_n)\right] = \frac{1}{n}\mathrm{Var}_Q[\psi_0(Y)] + O\left(\frac{2^{-2J_n\alpha}}{n}\right),$$

where the implicit constants depend only on $M, \gamma, \lambda, \alpha$.

Lemma 11 shows that the bias of $L(\widehat{Q}_n)$ scales quadratically faster than the traditional bias of $\widehat{Q}_n$ in estimating an $(\alpha-1)$-Hölder density, which is known to be of order $2^{-J_n(\alpha-1)}$. We obtain the faster rate $2^{-2J_n\alpha}$ due to the assumed $(\alpha+1)$-Hölder smoothness of the potential φ_0^*. The proofs of Theorem 10 and Lemma 11 are deferred to Appendix E.

4 The Two-Sample Problem

In this section, we turn to analyzing two-sample estimators when both measures $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ are unknown. As in the one-sample case, we study two classes of plugin estimators. The first consists of estimators which interpolate the empirical in-sample optimal transport coupling using nonparametric smoothers. Such estimators will achieve the optimal rate of estimating T_0 when it is Lipschitz. The second class will consist of plugin estimators based on density estimates of P and Q, and will achieve faster rates of convergence when P and Q have smooth densities. As before, our proofs will rely on stability bounds for the two-sample problem, to which we turn our attention first.

4.1 Two-Sample Stability Bounds

The stability bounds of Theorem 6 admit the following one-sided extension when both measures P and Q are unknown.

Proposition 12. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume condition **A1(λ)** holds for some $\lambda > 0$. Then, for any measures $\widehat{P}, \widehat{Q} \in \mathcal{P}(\Omega)$,*

$$\begin{aligned} 0 \le W_2^2(\widehat{P}, \widehat{Q}) - W_2^2(P, Q) - \int \phi_0 d(\widehat{P} - P) - \int \psi_0 d(\widehat{Q} - Q) \\ \le \lambda \left[W_2(\widehat{P}, P) + W_2(\widehat{Q}, Q)\right]^2. \end{aligned} \tag{29}$$

The proof is deferred to Appendix F.1. Similarly to Theorem 6, this result shows that the remainder of a first-order expansion of $W_2^2(\widehat{P}, \widehat{Q})$ around $W_2^2(P, Q)$ decays quadratically in the W_2 topology. Unlike Theorem 6, however, we do not generally expect that the lower bound in Proposition 12 can be replaced by a squared distance between (P, Q) and $(\widehat{P}, \widehat{Q})$: for instance, the lower bound of zero is achieved in equation (29) when $\widehat{P} = \widehat{Q} \ne Q = P$, even though $\widehat{P}$ may be arbitrarily far from P in Wasserstein distance. This example shows more generally that the bivariate functional $W_2^2(\cdot, \cdot)$ is not strictly convex over $\mathcal{P}_{\mathrm{ac}}(\Omega) \times \mathcal{P}_{\mathrm{ac}}(\Omega)$, unlike the univariate functional $W_2^2(P, \cdot)$ for a fixed absolutely continuous measure P (cf. Theorem 6 and Proposition 7.19 of Santambrogio (2015)).

These observations do not preclude the possibility of replacing the lower bound in Proposition 12 by $\lambda^{-1}\|\widehat{T} - T_0\|^2_{L^2(P)}$, for $\widehat{T}$ the optimal transport map between $\widehat{P}$ and $\widehat{Q}$. We were not able to derive such a result under the stated assumptions, except when these estimators are taken to be empirical measures. We describe this special case next, and show how it may be used to derive estimators of Lipschitz optimal transport maps T_0.

4.2 Upper Bounds for Two-Sample Empirical Estimators

Let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ denote i.i.d. samples, and define the empirical measures $P_n = (1/n) \sum_{i=1}^n \delta_{X_i}$ and $Q_m = (1/m) \sum_{j=1}^m \delta_{Y_j}$. Though the Monge problem between P_n and Q_m can be infeasible when $n \neq m$, the Kantorovich problem is always feasible, and takes the following form

$$\widehat{\pi} \in \operatorname*{argmin}_{\pi \in \mathcal{Q}_{nm}} \sum_{i=1}^n \sum_{j=1}^m \pi_{ij} \left\| X_i - Y_j \right\|^2,$$

where $\mathcal{Q}_{nm}$ denotes the set of doubly stochastic matrices $\pi = (\pi_{ij} : 1 \leq i \leq n,\ 1 \leq j \leq m)$, satisfying $\pi_{ij} \geq 0$, $\sum_{i=1}^n \pi_{ij} = 1/m$ and $\sum_{j=1}^m \pi_{ij} = 1/n$. We shall formulate the main stability bound of this section in terms of the quantity

$$\Delta_{nm} = \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij} \left\| T_0(X_i) - Y_j \right\|^2.$$

Recall that (κ_n) and $(\bar{\kappa}_n)$ denote the sequences defined in equation (21) and Corollary 8 respectively. We obtain the following result, which we prove in Appendix F.3.

Proposition 13. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, and assume **A1($\lambda$)** holds for some $\lambda > 0$. Then,*

$$\mathbb{E}[\Delta_{nm}] \asymp_\lambda \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q) \Big] \lesssim \kappa_{n \wedge m}.$$

If, in addition, $\Omega = [0,1]^d$ and there exists $\gamma > 0$ such that $\gamma^{-1} \leq p, q \leq \gamma$ over Ω, then,

$$\mathbb{E}[\Delta_{nm}] \asymp_\lambda \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q) \Big] \lesssim_\gamma \bar{\kappa}_{n \wedge m}.$$

To gain intuition about Proposition 13, it is fruitful to consider the special case $n = m$. In this setting, there exists an optimal transport map T_n from P_n to Q_n, and we may take

$$\widehat{\pi}_{ij} = I(T_n(X_i) = Y_j)/n, \quad \text{for all } 1 \leq i, j \leq n.$$

We then have $\Delta_{nn} = \|T_n - T_0\|^2_{L^2(P_n)}$, and Proposition 13 implies

$$\mathbb{E} \left\| T_n - T_0 \right\|^2_{L^2(P_n)} \asymp \mathbb{E}\Big[W_2^2(P_n, Q_n) - W_2^2(P, Q) \Big]. \tag{30}$$

Equation (30) is a two-sample analogue of Corollary 7, and shows that the $L^2(P_n)$ risk of the in-sample transport map estimator is of same order as the bias of the two-sample empirical optimal transport cost. While the estimators T_n and $\widehat{\pi}$ are only defined over the support of P_n, we next show how they may be extended to the entire domain Ω. We begin with an estimator inspired by the classical method of nearest-neighbor nonparametric regression (Cover, 1968).

One-Nearest Neighbor Estimator. Define the Voronoi partition generated by $X_1, \dots, X_n$ as

$$V_j = \{ x \in \Omega : \|x - X_j\| \leq \|x - X_i\|,\ \forall i \neq j \}, \quad j = 1, \dots, n. \tag{31}$$

Then, we define the one-nearest neighbor estimator of T_0 by

$$\widehat{T}_{nm}^{\mathrm{1NN}}(x) = \sum_{i=1}^{n}\sum_{j=1}^{m}(n\widehat{\pi}_{ij})I(x \in V_i)Y_j, \quad x \in \Omega. \tag{32}$$

In order to state an upper bound on the convergence rate of $\widehat{T}_{nm}^{\mathrm{1NN}}$, we place the following mild condition on the support Ω. Recall that $\mathcal{L}$ denotes the Lebesgue measure on $\mathbb{R}^d$.

(S2) Ω is a standard set, in the sense that there exist $\epsilon_0, \delta_0 > 0$ such that for all $x \in \Omega$ and $\epsilon \in (0, \epsilon_0)$, we have $\mathcal{L}(B(x, \epsilon) \cap \Omega) \geq \delta_0 \mathcal{L}(B(x, \epsilon))$.

Condition (S2) arises frequently in the literature on statistical set estimation (Cuevas and Fraiman, 1997; Cuevas, 2009), and prevents Ω from admitting cusps. Under this condition, we arrive at the following upper bound, which we prove in Appendix G.1.

Proposition 14. *Let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit a density p such that $\gamma^{-1} \leq p \leq \gamma$ over Ω, for some $\gamma > 0$, and let $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$. Assume conditions **A1(λ)** and **(S1)–(S2)** hold. Then,*

$$\mathbb{E}\left\|\widehat{T}_{nm}^{\mathrm{1NN}} - T_0\right\|_{L^2(P)}^2 \lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)^2 \kappa_{n\wedge m}.$$

Furthermore, if $\Omega = [0,1]^d$ and we additionally assume that $\gamma^{-1} \leq q \leq \gamma$ over Ω, then

$$\mathbb{E}\left\|\widehat{T}_{nm}^{\mathrm{1NN}} - T_0\right\|_{L^2(P)}^2 \lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)^2 \overline{\kappa}_{n\wedge m}.$$

Proposition 14 proves that the one-nearest neighbor estimator achieves the minimax rate in equation (3), up to a polylogarithmic factor. This result is in stark contrast to standard risk bounds for K-nearest neighbor nonparametric regression, for which the number K of nearest neighbors is typically required to diverge in order to achieve the minimax estimation rate of a Lipschitz continuous regression function (Györfi et al., 2006). Though increasing K reduces the variance of such estimators, in our setting, Propositions 13–14 suggest that the variance of $\widehat{T}_{nm}^{\mathrm{1NN}}$ is already dominated by its large bias, stemming from that of the in-sample coupling $\widehat{\pi}$, thereby making it sufficient to use $K = 1$ to obtain a near-optimal rate. While the one-nearest neighbor estimator is simplest to analyze, it is natural to expect that any linear smoother with sufficiently small bandwidth may be used to smooth the in-sample coupling $\widehat{\pi}_{nm}$ and lead to a similar rate.

Convex Least Squares Estimator. Though nearly minimax optimal, the estimator $\widehat{T}_{nm}^{\mathrm{1NN}}$ is typically not the gradient of a convex function, and is therefore not an admissible optimal transport map in its own right. We next show how this property can be enforced using an estimator inspired by nonparametric least squares regression. Let $\mathcal{J}_\lambda$ denote the class of functions $\varphi : \Omega \to \mathbb{R}$ which are convex and have λ-Lipschitz gradients $\nabla\varphi$. Define the least squares estimator

$$\widehat{T}_{nm}^{\mathrm{LS}} = \nabla\widehat{\varphi}_{nm}^{\mathrm{LS}}, \quad \text{where } \widehat{\varphi}_{nm}^{\mathrm{LS}} \in \operatorname*{argmin}_{\varphi \in \mathcal{J}_\lambda} \sum_{i=1}^{n}\sum_{j=1}^{m} \widehat{\pi}_{ij} \left\|Y_j - \nabla\varphi(X_i)\right\|^2.$$

The computation of the above infinite-dimensional optimization problem can be reduced to that of solving a finite-dimensional quadratic program, by a direct extension of well-known solvers for shape-constrained nonparametric regression with Lipschitz and convex constraints (cf. Seijo and

Sen (2011), Mazumder et al. (2019), and references therein). We obtain the following upper bound by a simple extension of Proposition 14.

Proposition 15. *Proposition 14 continues to hold when* $\widehat{T}^{\rm 1NN}_{nm}$ *is replaced by* $\widehat{T}^{\rm LS}_{nm}$.

4.3 Upper Bounds for Two-Sample Wavelet Estimators

We next study two-sample estimators under stronger smoothness assumptions on P and Q. As discussed in Section 4.1, we do not know of a two-sample stability bound for the $L^2(P)$ loss which is analogous to Theorem 6, and only places regularity conditions on the population potential φ_0. Therefore, unlike Theorem 10, in which smoothness conditions on q and φ_0 were sufficient to obtain sharp upper bounds, in the two-sample case our analysis will also rely on the smoothness of *estimators* $\widehat{\varphi}_{nm}$ of the potential φ_0. In order to quantify their regularity, we shall require a uniform analogue of Caffarelli's global regularity theory (Theorem 3(ii)). Since we are unaware of such results for generic compact domains $\Omega \subseteq \mathbb{R}^d$, we instead assume throughout the remainder of this section that Ω is taken to be the d-dimensional torus $\mathbb{T}^d$, thus allowing us to appeal to Theorem 5. We note that such periodicity constraints are commonly imposed in nonparametric estimation problems to mitigate boundary issues (Efromovich, 1999; Krishnamurthy et al., 2014; Han et al., 2020). In many such cases, an alternative is to assume that the underlying probability measures place sufficiently small mass near the boundary. Such an assumption cannot be used in our context since, as before, we shall require all densities to be bounded away from zero throughout their support. Optimal estimation rates under Wasserstein distances differ dramatically in the absence of a density lower bound condition (Bobkov and Ledoux, 2019; Weed and Berthet, 2019), and we do not address this setting here.

Recall the background on the quadratic optimal transport problem over $\mathbb{T}^d$ in Section 2.2. Let $P, Q \in \mathcal{P}_{\rm ac}(\mathbb{T}^d)$ be absolutely continuous measures admitting respective $\mathbb{Z}^d$-periodic densities p and q. We now denote by T_0 the optimal transport map from P to Q, with respect to the cost $d^2_{\mathbb{T}^d}$. As outlined in Proposition 4, T_0 is the gradient of a convex potential $\varphi_0 : \mathbb{R}^d \to \mathbb{R}$, and is uniquely determined P-almost everywhere. We continue to denote by $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ a corresponding pair of Kantorovich potentials. Let $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ denote i.i.d. samples, which are independent of each other, and let $\widehat{P}_n, \widehat{Q}_m$ respectively denote the distributions induced by density estimators $\widehat{p}_n, \widehat{q}_m$ of p, q over $\mathbb{T}^d$, to be defined below. Our aim is to bound the risk of the estimator

$$\widehat{T}_{nm} = \nabla \widehat{\varphi}_{nm} = \operatorname*{argmin}_{T \in \mathcal{T}(\widehat{P}_n, \widehat{Q}_m)} \int d^2_{\mathbb{T}^d}(T(x), x) d\widehat{P}_n(x). \tag{33}$$

Note that $\widehat{P}_n$ and $\widehat{Q}_m$ are absolutely continuous, thus there indeed exists a unique solution to the above minimization problem, by Proposition 4. We continue to quantify the risk of $\widehat{T}_{nm}$ in terms of the $L^2(P)$ loss

$$\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P)} = \int_{\mathbb{T}^d} \big\|\widehat{T}_{nm}(x) - T_0(x)\big\|^2 dP(x).$$

Notice that the integrand on the right-hand side of the above display is $\mathbb{Z}^d$-periodic by Proposition 4(ii) and by the optimality of $\widehat{T}_{nm}$ and T_0, thus it indeed defines a map $\mathbb{T}^d \to \mathbb{R}$. As before, we shall also obtain upper bounds on the bias and risk of $W_2^2(\widehat{P}_n, \widehat{Q}_m)$ as a byproduct of our proofs. Indeed, our main results hinge upon the stability bounds derived in previous sections, which can easily be shown to hold in the present context.

Proposition 16. *Assume φ_0 satisfies condition **A1(λ)**, in the sense that φ_0 is a closed convex function over $\mathbb{R}^d$ satisfying $\lambda^{-1} I_d \preceq \nabla^2 \varphi_0(x) \leq \lambda I_d$ for all $x \in \mathbb{R}^d$. Then, Theorem 6 and Proposition 12 hold with $\Omega = \mathbb{T}^d$.*

As in the one-sample case, we focus on the situation where $\widehat{P}_n$ and $\widehat{Q}_m$ are the distributions of wavelet density estimators. Unlike the boundary-corrected wavelet system used in Section 3.3, it will be convenient to introduce a simpler basis which guarantees that the density estimators are periodic. Specifically, we describe in Appendix A.2.2 how the standard Daubechies wavelet system may be periodized to obtain a set of $\mathbb{Z}^d$-periodic functions

$$\Psi^{\mathrm{per}} = \{1\} \cup \bigcup_{j=0}^{\infty} \Psi_j^{\mathrm{per}}, \quad \text{where} \quad \Psi_j^{\mathrm{per}} = \left\{ \xi_{jk\ell}^{\mathrm{per}} : 0 \leq k \leq 2^{j-1}, \ell \in \{0,1\}^d \setminus \{0\} \right\},\ j \geq 0,$$

which forms an orthonormal basis of $L^2(\mathbb{T}^d)$ (Daubechies, 1992; Giné and Nickl, 2016). Whenever the densities p, q lie in $L^2(\mathbb{T}^d)$, they admit wavelet expansions of the form

$$p = 1 + \sum_{j=0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{per}}} \alpha_\xi \xi, \quad q = 1 + \sum_{j=0}^{\infty} \sum_{\xi \in \Psi_j^{\mathrm{per}}} \beta_\xi \xi,$$

where $\alpha_\xi = \int \xi dP$ and $\beta_\xi = \int \xi dQ$. We then define the wavelet density estimators

$$\widetilde{p}_n^{(\mathrm{per})} = 1 + \sum_{j=0}^{J_n} \sum_{\xi \in \Psi_j^{\mathrm{per}}} \widehat{\alpha}_\xi \xi, \quad \widetilde{q}_m^{(\mathrm{per})} = 1 + \sum_{j=0}^{J_m} \sum_{\xi \in \Psi_j^{\mathrm{per}}} \widehat{\beta}_\xi \xi,$$

where $\widehat{\alpha}_\xi = \int \xi dP_n$ and $\widehat{\beta}_\xi = \int \xi dQ_m$. By orthonormality of Ψ^{per}, it is straightforward to see that $\widetilde{p}_n^{(\mathrm{per})}, \widetilde{q}_m^{(\mathrm{per})}$ integrate to unity, but may nevertheless be negative. We therefore define the final density estimators by

$$\widehat{p}_n^{(\mathrm{per})} \propto \widetilde{p}_n^{(\mathrm{per})} I(\widetilde{p}_n^{(\mathrm{per})} \geq 0), \quad \widehat{q}_m^{(\mathrm{per})} \propto \widetilde{q}_m^{(\mathrm{per})} I(\widetilde{q}_m^{(\mathrm{per})} \geq 0), \tag{34}$$

where the proportionality constants are to be chosen such that $\widehat{p}_n^{(\mathrm{per})}$ and $\widehat{q}_m^{(\mathrm{per})}$ are probability densities, which respectively induce probability distributions $\widehat{P}_n^{(\mathrm{per})}, \widehat{Q}_m^{(\mathrm{per})} \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$. Once again, we drop all superscripts "per" whenever the choice of density estimator is unambiguous. As we state in Proposition 43, Appendix H, the one-sample results from Theorem 10 may readily be extended to the present periodic setting, by replacing the boundary-corrected wavelet estimator therein by the periodic wavelet estimator $\widehat{q}_m$ in the above display. Building upon this observation, we arrive at the following bound for the two-sample estimator $\widehat{T}_{nm} \equiv \widehat{T}_{nm}^{(\mathrm{per})}$ in equation (33), together with the associated plugin estimator of the squared Wasserstein distance. Recall the sequence $R_{T,n}(\alpha)$ defined in Theorem 10.

Theorem 17 (Two-Sample Wavelet Estimators)**.** *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$ for some $\alpha > 1$ and $M, \gamma > 0$. Assume $2^{J_n} \asymp n^{\frac{1}{d+2(\alpha-1)}}$. Then, there exists a constant $C > 0$ depending only on M, γ, α such that the following statements hold.*

(i) (Optimal Transport Maps) We have,

$$\mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P)} \leq C R_{T,n\wedge m}(\alpha).$$

(ii) (Wasserstein Distances) When $\alpha \notin \mathbb{N}$, we have

$$\big|\mathbb{E}W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P,Q)\big| \leq C R_{T,n\wedge m}(\alpha),$$

$$\mathbb{E}\big|W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P,Q)\big|^2 \leq \left[C R_{T,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2.$$

The proof appears in Appendix H.1. Theorem 17 shows that the plugin estimators $\widehat{T}_{nm}$ and $W_2^2(\widehat{P}_n, \widehat{Q}_m)$ achieve the same convergence rates as in the one-sample setting. Unlike the latter results, we also note that Theorem 17 places no conditions on the regularity of T_0 or φ_0. Indeed, over $\mathbb{T}^d$, these can be inferred from the assumption $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$, due to Theorem 5. We exclude the case $\alpha \in \mathbb{N}$ from Theorem 17(ii) due in part to our use of this result. Nevertheless, even when $\alpha \in \mathbb{N}$, Theorem 17(ii) implies that

$$\big|\mathbb{E}W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P,Q)\big| \lesssim (n\wedge m)^{\epsilon} R_{T,n\wedge m}(\alpha),$$

for any $\epsilon > 0$, and similarly for the risk of $W_2^2(\widehat{P}_n, \widehat{Q}_m)$.

Similarly as in Section 3.3, we may deduce from Theorem 17(ii) and Lemma 34 that

$$\mathbb{E}\big|W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P,Q)\big| \lesssim_{M,\gamma,\alpha} R_{T,n\wedge m}(\alpha) + (n\wedge m)^{-1/2} W_2(P,Q).$$

Thus, in the high-smoothness regime $2(\alpha+1) > d$, the risk of $W_2^2(\widehat{P}_n, \widehat{Q}_m)$ decays at a rate which adapts to the magnitude of the Wasserstein distance between P and Q.

If one is willing to place assumptions on the regularity of the potentials φ_0 and φ_0^*, Theorem 17(ii) may be extended to the case where the sampling domain is taken to be the unit cube $[0,1]^d$. Such a result is stated in Proposition 44 of Appendix H, and is made possible by the fact that Proposition 12 does not require any regularity of the fitted potentials.

4.4 Upper Bounds for Two-Sample Kernel Estimators

Though our main focus has been on wavelet density estimators, the absence of a boundary on the sampling domain $\mathbb{T}^d$ also facilitates the analysis of kernel estimators, which we briefly discuss here.

Given $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$, we again denote by $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ two i.i.d. samples which are independent of each other. Given a kernel $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ and a bandwidth $h_n > 0$, write $K_{h_n} = h_n^{-d} K(\cdot/h_n)$, and define the kernel density estimators of p and q by

$$\widetilde{p}_n^{(\mathrm{ker})} = P_n \star K_{h_n} = \int_{\mathbb{R}^d} K_{h_n}(\cdot - z)dP_n(z), \quad \widetilde{q}_m^{(\mathrm{ker})} = Q_m \star K_{h_m} = \int_{\mathbb{R}^d} K_{h_m}(\cdot - z)dQ_m(z).$$

Recall that integration over $\mathbb{R}^d$ with respect to a measure in $\mathcal{P}(\mathbb{T}^d)$ is understood as integration with respect to this measure extended to $\mathbb{R}^d$ via translation by $\mathbb{Z}^d$-periodicity. The above estimators may take on negative values, thus we again define the final density estimators by

$$\widehat{p}_n^{(\mathrm{ker})} \propto \widetilde{p}_n^{(\mathrm{ker})} I(\widetilde{p}_n^{(\mathrm{ker})} \geq 0), \quad \widehat{q}_m^{(\mathrm{ker})} \propto \widetilde{q}_m^{(\mathrm{ker})} I(\widetilde{q}_m^{(\mathrm{ker})} \geq 0),$$

and their induced probability distributions by $\widehat{P}_n^{(\mathrm{ker})}$ and $\widehat{Q}_m^{(\mathrm{ker})}$. Furthermore, $\widehat{T}_{nm}^{(\mathrm{ker})}$ denotes the optimal transport map between these measures.

We shall require the following condition on the kernel K, for given real numbers $\zeta, \kappa > 0$.

K1(ζ, κ) $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$ is an even kernel, whose Fourier transform $\mathcal{F}[K]$ satisfies

$$\sup_{x \in \mathbb{R}^d \setminus \{0\}} |\mathcal{F}[K](x) - 1| \|x\|^{-\zeta} \leq \kappa. \tag{35}$$

A sufficient condition for equation (35) to hold is for $K \in L^\infty(\mathbb{R}^d)$ to be a kernel of order $\beta = \lceil \zeta - 1 \rceil$. Such a statement appears for instance in Tsybakov (2008) when $d = 1$, and can easily be generalized to $d > 1$. Multivariate kernels of order β which additionally lie in $\mathcal{C}_c^\infty(\mathbb{R}^d)$ can readily be defined; for example, one may start with a univariate even kernel $K_0 \in \mathcal{C}_c^\infty(\mathbb{R})$ of order β, constructed for instance using the procedure of Fan and Hu (1992), and then set $K(x) = \prod_{i=1}^d K_0(x_i)$ (Giné and Nickl, 2016).

Divol (2021) stated that their work may be used to show that $\widehat{P}_n^{(\mathrm{ker})}$ achieves a comparable rate of convergence as the boundary-corrected wavelet estimator $\widehat{P}_n^{(\mathrm{bc})}$, in Wasserstein distance. We provide a formal statement and proof of this fact in Lemma 46 of Appendix I, and use it to derive the following result.

Theorem 18 (Kernel Estimators)**.** *Assume the same conditions on P and Q as in the statement of Theorem 17. Assume further that K is a kernel satisfying condition* ***K1***$(2\alpha, \kappa)$ *for some $\kappa > 0$. Let $h_n \asymp n^{-1/(d+2(\alpha-1))}$. Then, there exists a constant $C > 0$ depending only on K, M, γ, α such that the following statements hold.*

(i) (Optimal Transport Maps) We have,

$$\mathbb{E}\big\|\widehat{T}_{nm}^{(\mathrm{ker})} - T_0\big\|_{L^2(P)}^2 \leq C R_{K, n\wedge m}(\alpha), \quad \text{where } R_{K,n}(\alpha) := \begin{cases} n^{-\frac{2\alpha}{2(\alpha-1)+d}}, & d \geq 3 \\ \log n / n, & d = 2 \\ 1/n, & d = 1. \end{cases}$$

(ii) (Wasserstein Distances) Assume further that $\alpha \notin \mathbb{N}$. Then,

$$\big|\mathbb{E} W_2^2(\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}) - W_2^2(P, Q)\big| \leq C R_{K, n\wedge m}(\alpha),$$

$$\mathbb{E}\big|W_2^2(\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}) - W_2^2(P, Q)\big|^2 \leq \left[C R_{K, n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2.$$

Theorem 18 shows that kernel plugin estimator of T_0 achieves the same minimax convergence rate as the wavelet estimators analyzed in Theorems 10 and 17, when $d \neq 2$. In contrast, when $d = 2$, Theorem 18 exhibits an improved convergence rate, scaling as $\log n / n$ instead of $(\log n)^2 / n$, which we now briefly discuss. This rate arises from our upper bound on $\mathbb{E} W_2^2(\widehat{P}_n^{(\mathrm{ker})}, P)$ in Lemma 46, which makes use of the inequality (18) comparing W_2 to a negative-order homogeneous Sobolev norm (Peyre, 2018). This last implies

$$W_2(\widehat{P}_n^{(\mathrm{ker})}, P) \lesssim \|\widehat{p}_n^{(\mathrm{ker})} - p\|_{\dot{H}^{-1}(\mathbb{T}^d)} \asymp \|\widehat{p}_n^{(\mathrm{ker})} - p\|_{\mathcal{B}^{-1}_{2,2}(\mathbb{T}^d)}. \tag{36}$$

In contrast, when $P \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$, our upper bounds for wavelet estimators (and implicitly for empirical estimators in Corollary 8) employed the following distinct relation, arising from the work of Weed and Berthet (2019),

$$W_2(\widehat{P}_n^{(\mathrm{bc})}, P) \lesssim \|\widehat{p}_n^{(\mathrm{bc})} - p\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)}, \tag{37}$$

and similarly for the estimator $\widehat{P}_n^{(\mathrm{per})}$. It can be seen that the $\mathcal{B}^{-1}_{2,2}$ norm is weaker than the $\mathcal{B}^{-1}_{2,1}$ norm. While either of these norms provide sufficiently tight upper bounds in equations (36) and (37) to obtain the minimax rate for density estimation in Wasserstein distance when $d \neq 2$, the former allows for a tighter logarithmic factor to be derived when $d = 2$. Inspired by the celebrated Ajtai–Komlós–Tusnády matching theorem (Ajtai et al., 1984; Talagrand, 1992), it is natural to conjecture that the rate $\log n/n$ in the definition of $R_{K,n}(\alpha)$ cannot be further improved when $d = 2$, for any of the conclusions of Theorem 18.

Theorem 18 is proved in Appendix I, where the main difficulty is to show that the evaluation $L(\widehat{Q}_m^{(\mathrm{ker})})$, of the linear functional L from equation (16), has bias decaying at the quadratic rate $h_m^{2\alpha}$. As for our analysis of wavelet estimators, this rate improves upon the naive upper bound $|\mathbb{E}L(\widehat{Q}_m^{(\mathrm{ker})})| \lesssim h_m^{\alpha-1}$, which could have been deduced from the traditional bias of kernel density estimators in estimating an $(\alpha-1)$-Hölder continuous density (Tsybakov, 2008). Similar considerations arise in the analysis of kernel-based estimators for other important functionals, such as the integral of a squared density (Giné and Nickl, 2008).

5 Efficient Statistical Inference for Wasserstein Distances

We now complement our results on estimation rates for Wasserstein distances by deriving limit laws, in Section 5.1, for the plugin estimators studied in Sections 3–4. We then derive lower bounds in Section 5.2, which show that these estimators are asymptotically efficient under suitable conditions.

5.1 Central Limit Theorems for Smooth Wasserstein Distances

Recall that we respectively denote by $P_n, \widehat{P}_n^{(\mathrm{bc})}, \widehat{P}_n^{(\mathrm{ker})}, \widehat{P}_n^{(\mathrm{per})}$ the empirical measure and the distributions induced by the boundary-corrected, periodic, and kernel density estimators of p (and similarly for q), as defined in Sections 3–4. Given a smoothness parameter $\alpha > 1$ to be specified, let their tuning parameters be chosen as $2^{J_n} \asymp h_n^{-1} \asymp n^{1/(d+2(\alpha-1))}$, and assume that the kernel K satisfies condition **K1**$(2\alpha, \kappa)$ for some $\kappa > 0$. Furthermore, in what follows, $X_1, \dots, X_n \sim P$ and $Y_1, \dots, Y_m \sim Q$ denote i.i.d. samples which are independent of each other, and we write

$$\sigma_\rho^2 = (1-\rho)\,\mathrm{Var}_P[\phi_0(X)] + \rho\,\mathrm{Var}_Q[\psi_0(Y)], \quad \text{for any } \rho \in [0,1],$$

where we recall that $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$, for any Brenier potential φ_0 in the optimal transport problem from P to Q. Our main result will be to derive one- and two-sample central limit theorems based on the various estimators $\widehat{P}_n$ and $\widehat{Q}_m$ under consideration, which take the following form:

$$\sqrt{n}\Big(W_2^2(\widehat{P}_n, Q) - W_2^2(P,Q)\Big) \rightsquigarrow N(0,\sigma_0^2), \quad \text{as } n \to \infty, \quad \text{and} \tag{38}$$

$$\sqrt{\frac{nm}{n+m}}\Big(W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P,Q)\Big) \rightsquigarrow N(0,\sigma_\rho^2), \quad \text{as } n,m \to \infty,\ \frac{n}{n+m} \to \rho, \tag{39}$$

for some $\rho \in [0,1]$. Our main result is the following.

Theorem 19 (Central Limit Theorems)**.** *Let Ω denote $\mathbb{T}^d$ or $[0,1]^d$. Assume that $P,Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit positive and bounded densities p,q over Ω. Then, the following assertions hold.*

(i) (Density Estimation over the Torus) Let $\Omega = \mathbb{T}^d$ and assume $p,q \in \mathcal{C}^{\alpha-1}(\Omega)$ for some $\alpha > 1$, $\alpha \notin \mathbb{N}$, satisfying $2(\alpha+1) > d$. Then, equations (38)–(39) hold when

$$(\widehat{P}_n, \widehat{Q}_m) = (\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}), \quad \text{or} \quad (\widehat{P}_n, \widehat{Q}_m) = (\widehat{P}_n^{(\mathrm{per})}, \widehat{Q}_m^{(\mathrm{per})}).$$

(ii) (Density Estimation over the Hypercube) Let $\Omega = [0,1]^d$, and assume $p,q \in \mathcal{C}^{\alpha-1}(\Omega)$ for some $\alpha > 1$ satisfying $2(\alpha+1) > d$. Assume additionally that $\varphi_0, \varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$. Then, equations (38)–(39) hold when

$$(\widehat{P}_n, \widehat{Q}_m) = (\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}).$$

(iii) (Empirical Measures) Let Ω be either $\mathbb{T}^d$ or $[0,1]^d$. Assume $d \le 3$, and that $\varphi_0, \varphi_0^* \in \mathcal{C}^2(\Omega)$. Then equations (38)–(39) hold when

$$(\widehat{P}_n, \widehat{Q}_m) = (P_n, Q_m).$$

To the best of our knowledge, Theorem 19 provides the first known central limit theorems for plugin estimators of the squared Wasserstein distance in arbitrary dimension $d \ge 1$ which are centered at their population counterpart $W_2^2(P,Q)$. We emphasize that the parametric scaling in the above result is made possible by the smoothness condition $2(\alpha+1) > d$. We do not generally expect that a central limit theorem for $W_2^2(\widehat{P}_n, Q)$ centered at $W_2^2(P,Q)$ can be obtained when $d > 2(\alpha+1)$, as we then expect the squared bias of this estimator to dominate its variance (cf. Theorems 10 and 17). In contrast, even in the absence of any smoothness conditions, del Barrio and Loubes (2019) derived limit laws of the form

$$\sqrt{n}\left(W_2^2(P_n, Q) - \mathbb{E}W_2^2(P_n,Q)\right) \rightsquigarrow N(0, \mathrm{Var}[\phi_0(X)]), \quad n \to \infty, \tag{40}$$

and two-sample analogues, for any $d \ge 1$. While such results are important and hold under milder regularity conditions than those of Theorem 19(v), their centering sequence is a barrier to their use for statistical inference for Wasserstein distances.

The proof of Theorem 19 is a consequence of the stability bounds in Theorem 6 and Proposition 12, which we use to show that $W_2^2(\widehat{P}_n, Q) - W_2^2(P,Q)$ asymptotically has same distribution as the linear functional $F(\widehat{P}_n) = \int \phi_0 d(\widehat{P}_n - P)$, under the stated smoothness conditions. We defer the proof to Appendix J. Though our arguments differ significantly from those used by del Barrio and Loubes (2019), this same linear functional also plays an important role in their work. Indeed, they prove that $n\,\mathrm{Var}[W_2^2(P_n,Q)-F(P_n)]=o(1)$ under mild conditions. In Appendix L, we provide an alternate proof of Theorem 19 which does not make use of our stability bounds, and which instead combines a generalization of the proof strategy of del Barrio and Loubes (2019), together with our convergence rates for optimal transport maps in Theorems 10, 17 and 18.

The variance σ_ρ^2 is positive if and only if ϕ_0 and ψ_0 are non-constant, which is equivalent to T_0 being different than the identity map. Thus, the distributional limits in Theorem 19 are non-degenerate whenever $P \neq Q$. When $P = Q$, it could already have been deduced from Lemma 30 that, for instance, the correct scaling for the process $W_2^2(\widehat{P}_n^{(\mathrm{bc})}, Q)$ is of larger order than $\sqrt{n}$, and we leave open the question of obtaining limit laws under this regime.

The variances appearing in Theorem 19 can be consistently estimated using estimators for the Kantorovich potentials ϕ_0 and ψ_0. Notice that the majority of our optimal transport map estimators in Sections 3–4 take the form $\widehat{T} = \nabla\widehat{\varphi}$ for a convex Brenier potential $\widehat{\varphi}$, in which case $(\|\cdot\|^2 - 2\widehat{\varphi}, \|\cdot\|^2 - 2\widehat{\varphi}^*)$ forms a natural estimator for the pair (ϕ_0,ψ_0). This estimator turns out to be consistent in most cases that we considered, and leads to the following result, which we only state for the setting of Theorem 19(i) in the interest of brevity.

Corollary 20 (Variance Estimation)**.** *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ be distributions with positive densities p, q over $\mathbb{T}^d$, and suppose $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d)$ for some $\alpha > 1$. Furthermore, let $\widehat{T}_{nm} = \nabla\widehat{\varphi}_{nm}$ denote the optimal transport map from $\widehat{P}_n^{(\mathrm{ker})}$ to $\widehat{Q}_m^{(\mathrm{ker})}$, and set $(\widehat{\phi}_{nm}, \widehat{\psi}_{nm}) = (\|\cdot\|^2 - 2\widehat{\varphi}_{nm}, \|\cdot\|^2 - 2\widehat{\varphi}_{nm}^*)$. Then, as $n, m \to \infty$,*

$$\widehat{\sigma}_{0,nm}^2 := \mathrm{Var}_{U\sim P_n}[\widehat{\phi}_{nm}(U)] \xrightarrow{p} \sigma_0^2, \quad \textit{and}, \quad \widehat{\sigma}_{1,nm}^2 := \mathrm{Var}_{V\sim Q_m}[\widehat{\psi}_{nm}(V)] \xrightarrow{p} \sigma_1^2.$$

Letting $\widehat{\sigma}_{nm}^2 = \frac{m\widehat{\sigma}_{0,nm}^2 + n\widehat{\sigma}_{1,nm}^2}{n+m}$, we deduce from Theorem 19(i) and Corollary 20 that

$$W_2^2(\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}) \pm \widehat{\sigma}_{nm} z_{\delta/2}\sqrt{\frac{n+m}{nm}}$$

is an asymptotic, two-sample $(1-\delta)$-confidence interval for $W_2^2(P,Q)$ under the conditions of Corollary 20, and under the additional condition $P \neq Q$. Here $z_{\delta/2}$ denotes the $\delta/2$ quantile of the standard Gaussian distribution, for any $\delta \in (0,1)$. To the best of our knowledge, this is the first practical confidence interval for the Wasserstein distance between absolutely continuous distributions in general dimension, albeit under the strong smoothness condition $2(\alpha+1) > d$. Similar confidence intervals can be obtained for the Wasserstein distance over the unit hypercube $[0,1]^d$, by replacing $\widehat{T}_{nm}$ with the estimator $\widehat{T}_{nm}^{\mathrm{LS}}$ defined in Section 4.2. In the sequel, we further derive efficiency lower bounds, showing that the asymptotic variances appearing in these results cannot be improved by any other regular estimator of $W_2^2(P,Q)$.

5.2 Efficiency Lower Bounds for Estimating the Wasserstein Distance

In discussing semiparametric efficiency theory, we follow the definitions and notation of van der Vaart (1998, 2002). We begin with a derivation of the efficient influence function of the functional

$$\Phi_Q : \mathcal{P}(\Omega) \to \mathbb{R}, \quad \Phi_Q(P) = W_2^2(P,Q),$$

where Ω is either $\mathbb{T}^d$ or a subset of $\mathbb{R}^d$, and $Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ is given. Santambrogio (2015, Proposition 7.17) has previously derived the first variation of this functional. The following result generalizes their argument, and states it in a language suitable for our development.

Lemma 21 (Efficient Influence Function)**.** *Let Ω be $\mathbb{T}^d$ or any connected and compact subset of $\mathbb{R}^d$, and let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$. Assume that the density of at least one of P and Q is positive over Ω. Let (ϕ_0, ψ_0) denote a pair of Kantorovich potentials in the optimal transport problem from P to Q, uniquely defined up to translation by a constant, and define the map*

$$\widetilde{\Phi}_{(P,Q)}(x) = \phi_0(x) - \int \phi_0 dP, \quad x \in \Omega.$$

Let $\dot{\mathcal{P}}_P \subseteq L_0^2(P)$ be any tangent set containing $\widetilde{\Phi}_{(P,Q)}$. Then, the functional Φ_Q is differentiable relative to $\dot{\mathcal{P}}_P$, with efficient influence function given by $\widetilde{\Phi}_{(P,Q)}$.

Lemma 21 is proved in Appendix K. The assumption that P or Q have support equal to Ω is only used to ensure that ϕ_0 is unique, up to translation by a constant (cf. Proposition 7.18 of Santambrogio (2015)). While this condition is not necessary (Staudt et al., 2022), we retain it for simplicity since we require it for our upper bounds.

By combining this result with the Convolution Theorem (van der Vaart (1998), Theorem 25.20), it immediately follows that any regular estimator sequence of $\Phi_Q(P)$ has asymptotic variance bounded below by $\mathrm{Var}_P[\phi_0(X)]/n$. The one-sample plugin estimators in Theorem 19 are thus optimal among regular estimators.

We next complement this result with an asymptotic minimax lower bound, which relaxes the assumption of regularity of such estimator sequences, at the expense of only comparing their worst-case risk. In this case, we also provide lower bounds for the two-sample functional

$$\Psi : \mathcal{P}(\Omega) \times \mathcal{P}(\Omega) \to \mathbb{R}, \quad \Psi(P, Q) = W_2^2(P, Q).$$

Using a construction of van der Vaart (1998), we fix two differentiable paths $(P_{t,h_1})_{t\geq 0}$ and $(Q_{t,h_2})_{t\geq 0}$, for any $(h_1, h_2) \in \mathbb{R}^2$, with respective score functions $h_1 \widetilde{\Phi}_{(P,Q)}$ and $h_2 \widetilde{\Psi}_{(P,Q)}$, where $\widetilde{\Psi}_{(P,Q)}(y) := \psi_0(y) - \int \psi_0 dQ$. These paths are defined in equations (104–105) of Appendix K, and we use them to obtain the following asymptotic minimax lower bound.

Theorem 22 (Asymptotic Minimax Lower Bound over $\mathbb{T}^d$)**.** *Given $M, \gamma > 0$ and $\alpha > 1$, let $P, Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$. Let (ϕ_0, ψ_0) denote a pair of Kantorovich potentials between P and Q, unique up to translation by a constant. Then, there exist $\overline{M}, \overline{\gamma}, \overline{u} > 0$, depending only on M, γ, α, such that P_{t,h_1} and Q_{t,h_2} admit densities in $\mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \overline{\gamma})$, for all $t > 0$ and $h_1, h_2 \in \mathbb{R}$ satisfying $t(|h_1| \vee |h_2|) \leq \overline{u}$. Furthermore,*

(i) (One Sample) For any estimator sequence $(U_n)_{n\geq 1}$, we have

$$\sup_{\substack{\mathcal{I} \subseteq \mathbb{R} \\ |\mathcal{I}| < \infty}} \liminf_{n\to\infty} \sup_{h \in \mathcal{I}} n\mathbb{E}_{n,h} \big| U_n - \Phi_Q(P_{n^{-1/2},h}) \big|^2 \geq \mathrm{Var}_P[\phi_0(X)].$$

where $\mathbb{E}_{n,h}$ denotes the expectation taken over the probability measure $P_{n^{-1/2},h}^{\otimes n}$.

(ii) (Two Sample) For any estimator sequence $(U_{nm})_{n,m\geq 1}$, we have

$$\sup_{\substack{\mathcal{I}\subseteq\mathbb{R}^2\\|\mathcal{I}|<\infty}} \liminf_{n,m\to\infty} \sup_{(h_1,h_2)\in\mathcal{I}} \frac{nm}{n+m}\mathbb{E}_{n,m,h_1,h_2}\left|U_{nm} - \Psi(P_{n^{-1/2},h_1}, Q_{m^{-1/2},h_2})\right|^2 \geq (1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho\operatorname{Var}_Q[\psi_0(Y)],$$

where the limit inferior is taken as $n/(n+m) \to \rho \in [0,1]$, and $\mathbb{E}_{n,m,h_1,h_2}$ denotes the expectation taken over the probability measure $P^{\otimes n}_{n^{-1/2},h_1} \otimes Q^{\otimes m}_{m^{-1/2},h_2}$.

The proof of Theorem 22 appears in Appendix K. For technical purposes, our statement assumes that P, Q admit densities lying in a strict subset $\mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$ of $\mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \overline{\gamma})$, the latter being the class in which our differentiable paths are shown to lie. With this caveat, our plugin estimators achieve the asymptotic minimax lower bounds of Theorem 22. For example, under the conditions of Theorems 18, when $2(\alpha+1) > d$ we deduce that

$$\sup_{\substack{\mathcal{I}\subseteq\mathbb{R}^2\\|\mathcal{I}|<\infty}} \liminf_{n,m\to\infty} \sup_{(h_1,h_2)\in\mathcal{I}} \frac{nm}{n+m}\mathbb{E}_{n,m,h_1,h_2}\left|\Psi(\widehat{P}_n^{(\mathrm{ker})}, \widehat{Q}_m^{(\mathrm{ker})}) - \Psi(P_{n^{-1/2},h_1}, Q_{m^{-1/2},h_2})\right|^2 = (1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho\operatorname{Var}_Q[\psi_0(Y)],$$

and a similar assertion holds for the periodic wavelet estimators $\widehat{P}_n^{(\mathrm{per})}, \widehat{Q}_m^{(\mathrm{per})}$.

It can be verified that Theorem 22 continues to hold with $\mathbb{T}^d$ replaced by $[0,1]^d$, under the additional condition that $\varphi_0, \varphi_0^* \in \mathcal{C}^{\alpha+1}([0,1]^d)$. We were unable, however, to derive differentiable paths $Q_{t,h} = (\nabla\varphi_{t,h})_{\#}P_{t,h}$ which simultaneously satisfy the Hölder continuity properties of Theorem 22 while also having Brenier potentials $\varphi_{t,h}, \varphi_{t,h}^*$ with uniformly bounded $\mathcal{C}^{\alpha+1}([0,1]^d)$ norm. We therefore leave open the question of whether our estimators over $[0,1]^d$ are asymptotically minimax optimal.

6 Discussion

We have shown that several families of plugin estimators for smooth optimal transport maps are minimax optimal. Our analysis hinged upon stability arguments which relate this problem to that of estimating the Wasserstein distance between two distributions, and, in turn, to that of estimating a distribution under the Wasserstein distance. The latter question is well-studied in the literature, and formed a key component in deriving convergence rates for the former two problems. As a byproduct of our stability results, we derived central limit theorems and efficiency lower bounds for estimating the Wasserstein distance between any two sufficiently smooth distributions. These results lead to the first practical confidence intervals for the Wasserstein distance in general dimension.

The estimators in this work are simple to compute and minimax optimal, but we make no claim that their computational efficiency is optimal. For example, our plugin estimators of the Wasserstein distance between $(\alpha-1)$-smooth densities can be approximated by sampling N observations from our density estimators, and computing the Wasserstein distance between the empirical measures formed by these observations, which can be done in polynomial time with respect to N (Peyré et al., 2019). In order for this approximation to achieve comparable risk to our theoretical estimators in the high-smoothness regime $\alpha \gtrsim d$, one must take $N \asymp n^{cd}$ for some $c \geq 1$. Our estimator thus requires computation time depending exponentially on d. Vacher et al. (2021) analyzed an alternative estimator with more favorable computational properties; though their estimator is not minimax optimal, it can be computed in polynomial time if $\alpha \gtrsim d$. It is an interesting open question to derive polynomial-time estimators in d which are also minimax optimal.

In our analysis of smooth two-sample optimal transport map estimators, we required the fitted optimal transport map to be twice Hölder-smooth, for which we appealed to Caffarelli's regularity theory. Since we do not know whether Caffarelli's boundary regularity estimates hold uniformly in the various problem parameters, we resorted to working over $\mathbb{T}^d$, where a uniform analogue of Caffarelli's theory is available (cf. Theorem 5). We leave open the question of bounding the risk of our estimators when the sampling domain is a subset of $\mathbb{R}^d$.

Finally, our work leaves open the question of estimating optimal transport maps when the ground cost function is not the squared Euclidean norm. While each of the plugin estimators in this paper can be naturally defined for generic cost functions, their theoretical analysis presents a breadth of challenges. For example, although the regularity theory of Caffarelli has been generalized to cover a large collection of cost functions (Ma et al., 2005), this collection does not include the costs $\|\cdot\|^p$ for $p \neq 2$ and $p > 1$, which are arguably most widely-used in statistical applications. For such costs, it remains unclear what regularity conditions are sensible to place on the population optimal transport map in order to obtain analogues of our risk bounds, and we hope to explore such questions in future work.

A Smoothness Classes and Wavelet Density Estimation

In this Appendix, we collect several definitions and properties of Hölder spaces, Besov spaces, and Sobolev Spaces, as well as properties of wavelet and kernel density estimators.

A.1 Hölder Spaces

Given a closed set $\Omega \subseteq \mathbb{R}^d$, let $\mathcal{C}_u(\Omega)$ denote the set of uniformly continuous real-valued functions on Ω. For any function $f : \Omega \to \mathbb{R}$ which is differentiable up to order $k \geq 1$ in the interior of Ω, and any multi-index $\gamma \in \mathbb{N}^d$, we write $|\gamma| = \sum_{i=1}^d \gamma_i$, and for all $|\gamma| \leq k$,

$$D^\gamma f = \frac{\partial^{|\gamma|} f}{\partial x_1^{\gamma_1} \dots \partial x_d^{\gamma_d}}.$$

Given $\alpha \geq 0$, the Hölder space $\mathcal{C}^\alpha(\Omega)$ is defined as the set of functions $f \in \mathcal{C}_u(\Omega)$ which are differentiable to order $\lfloor \alpha \rfloor$ in the interior of Ω, with derivatives extending continuously up to the boundary of Ω, and such that the Hölder norm

$$\|f\|_{\mathcal{C}^\alpha(\Omega)} = \sum_{j=0}^{\lfloor \alpha \rfloor} \sup_{|\gamma|=j} \|D^\gamma f\|_\infty + \sum_{|\gamma|=\lfloor \alpha \rfloor} \sup_{\substack{x,y \in \Omega^\circ \\ x \neq y}} \frac{|D^\gamma f(x) - D^\gamma f(y)|}{\|x-y\|^{\alpha - \lfloor \alpha \rfloor}}$$

is finite. Furthermore, for any $\alpha \geq 0$, $\mathcal{C}^\alpha(\mathbb{T}^d)$ (resp. $\mathcal{C}_u(\mathbb{T}^d)$) is defined as the set of $\mathbb{Z}^d$-periodic functions $f : \mathbb{R}^d \to \mathbb{R}$ such that $f \in \mathcal{C}^\alpha(\mathbb{R}^d)$ (resp. $f \in \mathcal{C}_u(\mathbb{R}^d)$).

Recall that $\mathcal{C}^\alpha(\Omega; \lambda)$ denotes the closed $\mathcal{C}^\alpha(\Omega)$ ball of radius $\lambda > 0$. We occasionally use the following simple observation.

Lemma 23. *Let Ω be $\mathbb{T}^d$ or a closed subset of $\mathbb{R}^d$. Then, for all $\lambda, \alpha > 0$, there exists a constant $C_{\lambda,\alpha} > 0$ such that*

$$\sup_{f,g \in \mathcal{C}^\alpha(\Omega;\lambda)} \|fg\|_{\mathcal{C}^\alpha(\Omega)} \le C_{\lambda,\alpha}.$$

Lemma 23 is stated for $\alpha < 1$ by Gilbarg and Trudinger (2001, equation (4.7)), and can easily be extended to all $\alpha \ge 1$ using the general Leibniz rule.

A.2 Wavelets and Besov Spaces

Recall that in Sections 3.3 and 4.3, we made use of the boundary-corrected wavelet system Ψ^{bc} over the unit cube $[0,1]^d$, and of the periodic wavelet system Ψ^{per} over the flat torus $\mathbb{T}^d$. In this section, we provide further descriptions and properties of these wavelet bases, before turning to definitions and characterizations of Besov spaces over $[0,1]^d$ and $\mathbb{T}^d$. For concreteness, we describe these constructions in terms of the compactly-supported N-th Daubechies scaling and wavelet functions $\zeta_0, \xi_0 \in \mathcal{C}^r(\mathbb{R}^d)$, where $r = 0.18(N-1)$ for an integer $N \ge 2$ (Daubechies (1988); Giné and Nickl (2016), Theorem 4.2.10). We also extend this definition to the case $N = 1$ by taking ζ_0, ξ_0 to be the (discontinuous) Haar functions (Giné and Nickl (2016), p. 298). Throughout the sequel and throughout the main manuscript, whenever we work with a Besov space $\mathcal{B}^s_{p,q}$ or a Hölder space $\mathcal{C}^s$ with $s > 0$, we tacitly assume that the parameter N is chosen such that the regularity r is strictly greater than the parameters $\lceil s \rceil$ or α, in which case it must at least hold that $N \ge 2$.

Our exposition closely follows that of Giné and Nickl (2016), and we also refer the reader to Cohen et al. (1993); Cohen (2003); Härdle et al. (2012) and references therein for further details.

A.2.1 Boundary-Corrected Wavelets on $[0,1]^d$

It is well-known that the N-th Daubechies wavelet system

$$\zeta_{0k} = \zeta_0(\cdot - k), \quad \xi_{0jk} = 2^{\frac{j}{2}} \xi_0(2^j(\cdot) - k), \quad j \ge 0, \ k \in \mathbb{Z},$$

forms a basis of $L^2(\mathbb{R})$, with the property that $\{\zeta_{0k} : k \in \mathbb{Z}\}$ spans all polynomials on $\mathbb{R}$ of degree at most $N-1$. While this family may easily be periodized to obtain a basis for $L^2([0,1])$, as in the following subsection, doing so may not accurately reflect the regularity of functions in $L^2([0,1])$ via the decay of their wavelet coefficients, near the boundaries of the interval. This consideration motivated Meyer (1991) and Cohen et al. (1993) to introduce the so-called boundary-corrected wavelet system on $[0,1]$, which preserves the standard Daubechies scaling functions lying sufficiently far from the boundaries of the interval, and adds edge scaling functions such that their union continues to span all polynomials up to degree $N-1$ on $[0,1]$. In short, given a fixed integer $j_0 \ge \log_2 N$, the construction of Cohen et al. (1993) leads to smooth scaling and wavelet edge basis functions

$$\begin{aligned} &\zeta^{\text{left}}_{0j_0k}, \xi^{\text{left}}_{0j_0k} && \text{supported in } [0, (2N-1)/2^{j_0}], \\ &\zeta^{\text{right}}_{0j_0k}, \xi^{\text{right}}_{0j_0k} && \text{supported in } [1-(2N-1)/2^{j_0}, 1], \quad k = 0, \dots, N-1. \end{aligned}$$

In this case, if one defines,

$$\zeta^a_{0jk} = 2^{\frac{j-j_0}{2}} \zeta^a_{0j_0k}\left(2^{j-j_0}(\cdot)\right), \quad \xi^a_{0jk} = 2^{\frac{j-j_0}{2}} \xi^a_{0j_0k}\left(2^{j-j_0}(\cdot)\right), \quad \text{for all } j \ge j_0, \ a \in \{\text{left}, \text{right}\},$$

then the family

$$\begin{aligned}\Phi_0^{\mathrm{bc}}&=\{\zeta_{0j_0k}^{\mathrm{bc}}:0\le k\le 2^{j_0}-1\}=\left\{\zeta_{0j_0k}^{\mathrm{left}},\zeta_{0j_0k}^{\mathrm{right}},\zeta_{0m}:0\le k\le N-1, N\le m\le 2^{j_0}-N-1\right\},\\ \Psi_0^{\mathrm{bc}}&=\{\xi_{0jk}^{\mathrm{bc}}:0\le k\le 2^j-1, j\ge j_0\}\\ &=\left\{\xi_{0jk}^{\mathrm{left}},\xi_{0jk}^{\mathrm{right}},\xi_{0jm}:0\le k\le N-1, N\le m\le 2^{j_0}-N-1, j\ge j_0\right\},\end{aligned}$$

form an orthonormal basis of $L^2([0,1])$, with the property that Φ^{bc} spans all polynomials on $[0,1]$ of degree at most $N-1$. We then define a tensor product wavelet basis of $L^2([0,1]^d)$ by setting for all $j\ge j_0$ and all $\ell=(\ell_1,\dots,\ell_d)\in\{0,1\}^d\setminus\{0\}$,

$$\zeta_{j_0k}^{\mathrm{bc}}(x)=\prod_{i=1}^d\zeta_{0j_0k_i}^{\mathrm{bc}}(x_i),\quad\text{and}\quad \xi_{jk\ell}^{\mathrm{bc}}(x)=\prod_{i:\ell_i=0}\zeta_{0jk_i}^{\mathrm{bc}}(x_i)\prod_{i:\ell_i=1}\xi_{0jk_i}^{\mathrm{bc}}(x_i),\quad x\in[0,1]^d,$$

where in the definition of $\zeta_{j_0k}^{\mathrm{bc}}$, the index $k=(k_1,\dots,k_d)$ ranges over $\mathcal{K}(j_0):=\{1,\dots,2^{j_0}-1\}^d$, while in the definition of $\xi_{jk\ell}^{\mathrm{bc}}$, k ranges over $\mathcal{K}(j)$. In this case, the wavelet system

$$\Psi^{\mathrm{bc}}=\Phi^{\mathrm{bc}}\cup\bigcup_{j=j_0}^\infty\Psi_j^{\mathrm{bc}},\quad \Phi^{\mathrm{bc}}=\{\zeta_{j_0k}^{\mathrm{bc}}:k\in\mathcal{K}(j_0)\},\quad \Psi_j^{\mathrm{bc}}=\{\xi_{jk\ell}^{\mathrm{bc}}:k\in\mathcal{K}(j)\},\quad j\ge j_0,$$

announced in Section 3.3 forms a basis of $L^2([0,1]^d)$. We sometimes make use of the abbreviation $\Psi_{j_0-1}=\Phi$.

A.2.2 Periodic Wavelets on $\mathbb{T}^d$

When working over $\mathbb{T}^d$, a simpler construction may be used due to the periodicity of the functions involved. Denote the periodization on $\mathbb{T}$ of dilations of the maps ζ_0,ξ_0 by

$$\zeta_0^{\mathrm{per}}=\sum_{k\in\mathbb{Z}}\zeta_0(\cdot-k)=1,\quad \xi_{0j}^{\mathrm{per}}=\sum_{k\in\mathbb{Z}}2^{j/2}\xi_0(2^j(\cdot-k)),\quad j\ge 0.$$

In this case, the collection

$$\Psi_0^{\mathrm{per}}=\left\{1,\xi_{0jk}^{\mathrm{per}}=\xi_{0j}^{\mathrm{per}}(\cdot-2^{-j}k):0\le k\le 2^j-1, j\ge 0\right\}$$

forms an orthonormal basis of $L^2(\mathbb{T})$, which may again be extended to $L^2(\mathbb{T}^d)$ using tensor product wavelets. Specifically, if $\xi_{jk\ell}^{\mathrm{per}}=\prod_{i=1}^d(\xi_{jk}^{\mathrm{per}})^{\ell_i}$ for all $\ell=(\ell_1,\dots,\ell_d)\in\{0,1\}^d\setminus\{0\}$, then

$$\Psi^{\mathrm{per}}=\{1\}\cup\bigcup_{j=0}^\infty\Psi_j^{\mathrm{per}},\quad\text{with}\quad \Psi_j^{\mathrm{per}}=\left\{\xi_{jk\ell}:k\in\mathcal{K}(j),\ell\in\{0,1\}^d\setminus\{0\}\right\},\ j\ge 0,$$

forms an orthonormal basis of $L^2(\mathbb{T}^d)$ (Daubechies (1992), Section 9.3; Giné and Nickl (2016), Section 4.3).

A.2.3 Properties of Boundary-Corrected and Periodic Wavelet Systems

In both of the preceding constructions, one obtains a family Φ of scaling functions and a sequence of families $(\Psi_j)_{j\geq j_0}$ of wavelet functions, such that

$$\Phi = \Psi_{j_0-1} = \{\zeta_k : k \in \mathcal{K}(j_0)\} = \begin{cases} \Phi^{\mathrm{bc}}, & \Psi = \Psi^{\mathrm{bc}} \\ \{1\}, & \Psi = \Psi^{\mathrm{per}}, \end{cases}$$

$$\Psi_j = \{\xi_{jk\ell} : k \in \mathcal{K}(j), \ell \in \{0,1\}^d \setminus \{0\}\} = \begin{cases} \Psi_j^{\mathrm{bc}}, & \Psi = \Psi^{\mathrm{bc}} \\ \Psi_j^{\mathrm{per}}, & \Psi = \Psi^{\mathrm{per}}, \end{cases} \qquad j \geq j_0,$$

$$j_0 = \begin{cases} \lceil \log_2 N \rceil, & \Psi = \Psi^{\mathrm{bc}} \\ 0, & \Psi = \Psi^{\mathrm{per}}, \end{cases}$$

$$\mathcal{K}(j) = \{0, \dots, 2^j - 1\}^d, \quad j \geq j_0.$$

In both cases, the wavelet system is defined over a domain Ω, which is to be understood as either $[0,1]^d$ in the boundary-corrected case, or as $\mathbb{T}^d$ (which itself may be identified with $(0,1]^d$) in the periodic case. In either of these settings, the wavelet system

$$\Psi = \Phi \cup \bigcup_{j=j_0}^{\infty} \Psi_j \tag{41}$$

forms a basis of $L^2(\Omega)$. The following simple result collects several properties and definitions which are common to both of the above bases.

Lemma 24. *Let $N \geq 1$. There exist constants $C_1, C_2 \geq 1$ depending only on N, d and on the choice of basis $\Psi \in \{\Psi^{\mathrm{bc}}, \Psi^{\mathrm{per}}\}$ such that the following properties hold.*

(i) $|\Phi| \leq C_1$, $|\Psi_j| \leq C_2 2^{dj}$ *for all* $j \geq j_0$.

(ii) *For all* $j \geq j_0$ *and* $\xi \in \Psi_j$, *there exists a rectangle* $I_\xi \subseteq \Omega$ *such that* $\mathrm{diam}(I_\xi) \leq C_1 2^{-j}$, $\mathrm{supp}(\xi_j) \subseteq I_\xi$, *and* $\left\| \sum_{\xi \in \Psi_j} I(\cdot \in I_\xi) \right\|_{L^\infty} \leq C_2$.

(iii) $\xi \in \mathcal{C}^r(\Omega)$ *for all* $\xi \in \Psi$.

(iv) *Polynomials of degree at most* $N-1$ *over* Ω *lie in* $\mathrm{Span}(\Phi)$.

(v) *If* $N \geq 2$, *we have,*

$$\sup_{0\leq|\gamma|\leq\lfloor r\rfloor} \sup_{\zeta\in\Phi} \|D^\gamma \zeta\|_{L^\infty} \leq C_1, \qquad \sup_{0\leq|\gamma|\leq\lfloor r\rfloor} \sup_{j\geq j_0} \sup_{\xi\in\Psi_j} 2^{-j\left(\frac{d}{2}+|\gamma|\right)} \|D^\gamma \xi\|_{L^\infty} \leq C_2.$$

A.2.4 Besov Spaces

We next define the Besov spaces $\mathcal{B}^s_{p,q}(\Omega)$, for $s > 0$, $p, q \geq 1$. Once again, Ω is understood to be one of $[0,1]^d$ or $\mathbb{T}^d$, and Ψ is understood to be the corresponding wavelet basis as in equation (41).

Let $f \in L^p(\Omega)$ admit the wavelet expansion

$$f = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \beta_\xi \xi, \quad \text{over } \Omega,$$

with convergence in $L^p(\Omega)$, where $\beta_\xi = \int \xi f$ for all $\xi \in \Psi$. Then, the Besov norm of f may be defined by

$$\|f\|_{\mathcal{B}^s_{p,q}(\Omega)} = \|(\beta_\zeta)_{\zeta \in \Phi}\|_{\ell_p} + \left\| \left(2^{j(s+\frac{d}{2}-\frac{d}{p})} \left\|(\beta_\xi)_{\xi \in \Psi_j}\right\|_{\ell_p} \right)_{j \geq j_0} \right\|_{\ell_q}, \tag{42}$$

and we define

$$\mathcal{B}^s_{p,q}(\Omega) = \begin{cases} \left\{ f \in L^p(\Omega) : \|f\|_{\mathcal{B}^s_{p,q}(\Omega)} < \infty \right\}, & 1 \leq p < \infty \\ \left\{ f \in \mathcal{C}_u(\Omega) : \|f\|_{\mathcal{B}^s_{p,q}(\Omega)} < \infty \right\}, & p = \infty. \end{cases}$$

We extend the above definition to $s < 0$ by the duality $\mathcal{B}^s_{p',q'}(\Omega) = \left(\mathcal{B}^{-s}_{p,q}(\Omega)\right)^*$, where $\frac{1}{p'} + \frac{1}{p} = \frac{1}{q'} + \frac{1}{q} = 1$. It can be shown that the resulting norm on the space $\mathcal{B}^s_{p',q'}(\Omega)$ is equivalent to the sequence norm $\|\cdot\|_{\mathcal{B}^s_{p',q'}(\Omega)}$ in equation (42) (cf. Cohen (2003), Theroem 3.8.1), thus we extend its definition to $s < 0$.

We shall often make use of Besov spaces in order to characterize Hölder continuous functions in terms of the decay of their wavelet coefficients, via the following classical result.

Lemma 25. *For all $0 < s < r$, and $d \geq 1$, we have*

$$\mathcal{C}^s([0,1]^d) \subseteq \mathcal{B}^s_{\infty,\infty}([0,1]^d), \quad \mathcal{C}^s(\mathbb{T}^d) \subseteq \mathcal{B}^s_{\infty,\infty}(\mathbb{T}^d), \tag{43}$$

and there exist $C_1, C_2 > 0$ such that

$$\|\cdot\|_{\mathcal{B}^s_{\infty,\infty}([0,1]^d)} \leq C_1 \|\cdot\|_{\mathcal{C}^s([0,1]^d)}, \quad \|\cdot\|_{\mathcal{B}^s_{\infty,\infty}(\mathbb{T}^d)} \leq C_2 \|\cdot\|_{\mathcal{C}^s(\mathbb{T}^d)}.$$

If $s \notin \mathbb{N}$, then equation (43) holds with equalities, and with equivalent norms.

An analogue of Lemma 25 is well-known to hold for the Daubechies wavelet system over $\mathbb{R}^d$, in which case it can readily be proven using an equivalent characterization of Besov spaces in terms of moduli of smoothness (Giné and Nickl (2016), Section 4.3.1). Such characterizations are also available for the periodized and boundary-corrected wavelet systems (Giné and Nickl (2016), Theorem 4.3.26 and discussions in Sections 4.3.5–4.3.6), and at least in the periodized case can be shown to lead to Lemma 25 (Giné and Nickl (2016), equation (4.167)). For the boundary-corrected case, Lemma 25 is known to hold in the special case $d = 1$ (Cohen et al. (1993), Theorem 4; Giné and Nickl (2016), equation (4.152)), but we do not know of a reference stating this precise result when $d > 1$, in part due to the potential ambiguity of defining the Hölder space $\mathcal{C}^s([0,1]^d)$ over the closed set $[0,1]^d$. We thus provide a self-contained proof of Lemma 25 in the boundary-corrected case for completeness, using standard arguments.

Proof of Lemma 25 (Boundary-Corrected Case). Let $\Omega = [0,1]^d$. Suppose first that $f \in \mathcal{B}^s_{\infty,\infty}(\Omega)$ for some $s \notin \mathbb{N}$, with wavelet expansion

$$f = \sum_{\zeta \in \Phi^{\text{bc}}} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j^{\text{bc}}} \beta_\xi \xi.$$

We wish to show that $\|f\|_{\mathcal{C}^s(\Omega)} \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. By Lemma 24, $\xi \in \mathcal{C}^r(\Omega)$ for all $\xi \in \Psi^{\text{bc}}$, where recall that $s < r$, thus we may define the map

$$f_\gamma = \sum_{\zeta\in\Phi^{\text{bc}}} \beta_\zeta D^\gamma \zeta + \sum_{j=j_0}^\infty \sum_{\xi\in\Psi_j^{\text{bc}}} \beta_\xi D^\gamma \xi, \quad \text{for all } 0 \le |\gamma| \le \lfloor s \rfloor.$$

Notice that $\|D^\gamma\zeta\|_{L^\infty} \lesssim 1$ for all $\zeta \in \Phi^{\text{bc}}$, and for all $j \ge j_0, k \in \mathcal{K}(j)$, $\ell \in \{0,1\}^d \setminus \{0\}$,

$$D^\gamma \xi^{\text{bc}}_{jk\ell} = 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} D^\gamma \xi^{\text{bc}}_{j_0k\ell}(2^{j-j_0}(\cdot))$$

Then, it follows from Lemma 24 that for all $x \in \Omega^\circ$,

$$\begin{aligned}
|f_\gamma(x)| &\le \sum_{\zeta\in\Phi^{\text{bc}}} |\beta_\zeta D^\gamma \zeta(x)| + \sum_{j=j_0}^\infty \sum_{\xi\in\Psi_j^{\text{bc}}} |\beta_\xi D^\gamma \xi(x)| \\
&\lesssim \left\|(\beta_\zeta)_{\zeta\in\Phi^{\text{bc}}}\right\|_{\ell_\infty} |\Phi^{\text{bc}}| + \sum_{j=j_0}^\infty \|(\beta_\xi)_{\xi\in\Psi_j^{\text{bc}}}\|_{\ell_\infty} 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} \sum_{\xi\in\Psi_j^{\text{bc}}} I(|\xi(x)|>0) \\
&\lesssim \left\|(\beta_\zeta)_{\zeta\in\Phi^{\text{bc}}}\right\|_{\ell_\infty} + \sum_{j=j_0}^\infty 2^{j\left(\frac{d}{2}+|\gamma|\right)} \|(\beta_\xi)_{\xi\in\Psi_j^{\text{bc}}}\|_{\ell_\infty} \\
&\lesssim \left\|(\beta_\zeta)_{\zeta\in\Phi^{\text{bc}}}\right\|_{\ell_\infty} + \left\|\left(2^{j\left(\frac{d}{2}+s\right)} \|(\beta_\xi)_{\xi\in\Psi_j^{\text{bc}}}\|_{\ell_\infty}\right)_{j\ge j_0}\right\|_{\ell_\infty} \sum_{j=j_0}^\infty 2^{\frac{(|\gamma|-s)j}{2}} \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}, \qquad (44)
\end{aligned}$$

where on the final line, we used the fact that s is not an integer, thus $|\gamma| < s$. An analogous calculation reveals that the series defining f_γ converges uniformly for any $0 \le |\gamma| \le \lfloor s \rfloor$, thus it must follow that f is differentiable up to order $\lfloor s \rfloor$ with derivatives given by $D^\gamma f = f_\gamma$, which by equation (44) must satisfy $|D^\gamma f(x)| \le C \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$ for all $x \in \Omega^\circ$, for a constant $C > 0$ depending only on d and r We next prove that $D^\gamma f$ is uniformly $(s - \lfloor s \rfloor)$-Hölder continuous over Ω°, for all $|\gamma| = \lfloor s \rfloor$. For all $x, y \in \Omega^\circ$, we have,

$$|D^\gamma f(x) - D^\gamma f(y)| \le \sum_{\zeta\in\Phi^{\text{bc}}} |\beta_\zeta||D^\gamma\zeta(x) - D^\gamma\zeta(y)| + \sum_{j=j_0}^\infty \sum_{\xi\in\Psi_j^{\text{bc}}} |\beta_\xi||D^\gamma\xi(x) - D^\gamma\xi(y)|.$$

Since $\zeta \in \mathcal{C}^r(\Omega)$ for all $\zeta \in \Phi^{\text{bc}}$,

$$\sum_{\zeta\in\Phi^{\text{bc}}} |\beta_\zeta||D^\gamma\zeta(x) - D^\gamma\zeta(y)| \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} |\Phi^{\text{bc}}| \|x-y\| \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} \|x-y\|.$$

Furthermore, using the definition of the boundary-corrected wavelet basis and its locality property in Lemma 24(ii), we have

$$\sum_{j=j_0}^\infty \sum_{\xi\in\Psi_j^{\text{bc}}} |\beta_\xi||D^\gamma\xi(x) - D^\gamma\xi(y)|$$

$$
\begin{aligned}
&= \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1}\sum_{l\in\{0,1\}^d\setminus\{0\}} |\beta_{\xi_{jk\ell}}| 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} |D^\gamma \xi_{j_0 k\ell}(2^{j-j_0}(x)) - D^\gamma \xi_{j_0 k\ell}(2^{j-j_0}(y))| \\
&\lesssim \sum_{j=j_0}^{\infty} \|(\beta_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} \left(\|2^{j-j_0}x - 2^{j-j_0}y\| \wedge 1\right) \sum_{\xi\in\Psi_j^{\mathrm{bc}}} I(|\xi(x)| \vee |\xi(y)| > 0) \\
&\lesssim \sum_{j=j_0}^{\infty} \|(\beta_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\|_{\ell_\infty} 2^{(j-j_0)\left(\frac{d}{2}+|\gamma|\right)} \left(2^{j-j_0}\|x-y\| \wedge 1\right) \\
&\lesssim \sum_{j=0}^{\infty} \|(\beta_\xi)_{\xi\in\Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} 2^{j\left(\frac{d}{2}+|\gamma|\right)} \left(2^{j}\|x-y\| \wedge 1\right) \\
&\lesssim \sum_{j=0}^{J(x,y)} \|(\beta_\xi)_{\xi\in\Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} 2^{j\left(\frac{d}{2}+|\gamma|+1\right)} \|x-y\| + \sum_{j=J(x,y)}^{\infty} \|(\beta_\xi)_{\xi\in\Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} 2^{j\left(\frac{d}{2}+|\gamma|\right)},
\end{aligned}
$$

where $J(x,y)$ is the smallest integer $j \geq 0$ such that $2^j|x-y| \geq 1$; in particular,

$$
2^{-J(x,y)} \leq \|x-y\| \leq 2^{-J(x,y)+1} \tag{45}
$$

Now, since $2^{j(\frac{d}{2}+s)}\|(\beta_\xi)_{\xi\in\Psi_{j+j_0}^{\mathrm{bc}}}\|_{\ell_\infty} \leq \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} < \infty$, and since $|\gamma| < s \notin \mathbb{N}$, we obtain

$$
\begin{aligned}
\|f\|^{-1}_{\mathcal{B}^s_{\infty,\infty}(\Omega)} |D^\gamma f(x) - D^\gamma f(y)| &\lesssim \|x-y\| \sum_{j=0}^{J(x,y)} 2^{j(|\gamma|-s+1)} + \sum_{j=J(x,y)}^{\infty} 2^{j(|\gamma|-s)} \\
&\lesssim \|x-y\| 2^{J(x,y)(|\gamma|-s+1)} + 2^{J(x,y)(|\gamma|-s)} \lesssim \|x-y\|^{s-|\gamma|},
\end{aligned}
$$

where the final inequality is due to equation (45). It readily follows that $\|f\|_{\mathcal{C}^s(\Omega)} \lesssim \|f\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. Furthermore, since $D^\gamma f$ is uniformly Hölder continuous over $(0,1)^d$, it is in particular uniformly continuous and hence extends to a continuous function over $[0,1]^d$, thus $f \in \mathcal{C}^s([0,1]^d)$. We next show that $\mathcal{C}^s([0,1]^d) \subseteq \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ for all $s > 0$, with the requisite Hölder norms. Assume $\|f\|_{\mathcal{C}^s(\Omega)} < \infty$, and let $\beta_\xi = \int f\xi$ for all $\xi \in \Psi^{\mathrm{bc}}$. By definition of the Besov norm, it will suffice to prove that

$$
\left\|(\beta_\zeta)_{\zeta\in\Phi^{\mathrm{bc}}}\right\| \lesssim \|f\|_{\mathcal{C}^s([0,1]^d)}, \quad \left\|(\beta_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\right\| \lesssim \|f\|_{\mathcal{C}^s([0,1]^d)} 2^{-j\left(\frac{d}{2}+s\right)}, \quad j \geq j_0.
$$

The first bound is immediate, since f is bounded above by $\|f\|_{\mathcal{C}^s([0,1]^d)}$ over $[0,1]^d$. To prove the second bound, let $x_0 \in (0,1)^d$, and let $\underline{s}$ denote the largest integer strictly less than s. By a Taylor expansion to order $\underline{s}$, there exists $c_s > 0$ such that

$$
\left| f(x) - \sum_{0\leq|\gamma|\leq\underline{s}} D^\gamma f(x_0)(x-x_0)^\gamma \right| \leq c_s \|f\|_{\mathcal{C}^s(\Omega)} \|x-x_0\|^s, \quad x \in \Omega, \tag{46}
$$

where $(x-x_0)^\gamma = \prod_{i=1}^d (x_i - x_{0i})^{\gamma_i}$. In particular, for any given $\xi \in \Psi_j^{\mathrm{bc}}$, $j \geq j_0$, choose $x_0 \in I_\xi \cap (0,1)^d$, where $\mathrm{diam}(I_\xi) \lesssim 2^{-j}$ and I_ξ is a set containing the support of ξ, as defined in

Lemma 24(ii). We then have,

$$\left|\int \xi f\right| \lesssim \left|\int \xi(x) \sum_{0\leq|\gamma|\leq \underline{s}} D^\gamma f(x_0)(x-x_0)^\gamma dx\right| \\ + \|f\|_{\mathcal{C}^s(\Omega)} \int |\xi(x)| \, \|x-x_0\|^s \, dx = \|f\|_{\mathcal{C}^s(\Omega)} \int |\xi(x)| \, \|x-x_0\|^s \, dx,$$

where the final equality uses the fact that polynomials of degree at most $\lfloor r \rfloor$ lie in $\mathrm{Span}(\Phi^{\mathrm{bc}})$ by Lemma 24(iv), and are therefore orthogonal to ξ. We thus have,

$$\begin{aligned} |\beta_\xi| &\lesssim \|f\|_{\mathcal{C}^s(\Omega)} \int_\Omega |\xi(x)| \, \|x-x_0\|^s \, dx \\ &= \|f\|_{\mathcal{C}^s(\Omega)} \int_{I_\xi} |\xi(x)| \, \|x-x_0\|^s \, dx \\ &\lesssim \|f\|_{\mathcal{C}^s(\Omega)} 2^{dj/2} \operatorname{diam}(I_\xi)^s \mathcal{L}(I_\xi) \lesssim \|f\|_{\mathcal{C}^s(\Omega)} 2^{-j\left(\frac{d}{2}+s\right)}. \end{aligned}$$

The claim readily follows. □

A.3 Sobolev Spaces

For our analysis of the kernel plugin estimators appearing in Section 4.3, we briefly recall a Fourier analytic description of the Sobolev spaces $H^s(\mathbb{T}^d) = \mathcal{B}^s_{2,2}(\mathbb{T}^d)$ over the torus, and refer the reader to Roe (1999); Grafakos (2009); Bahouri et al. (2011) for further details. Given a function $\phi \in L^2(\mathbb{T}^d)$, denote its sequence of Fourier coefficients by

$$\mathcal{F}[\phi](\xi) = \int \phi(x) \exp(-2\pi i x^\top \xi) dx, \quad \xi \in \mathbb{Z}^d.$$

If instead $\phi \in L^1(\mathbb{R}^d)$, we continue to denote by $\mathcal{F}[\phi]$ the Fourier transform of ϕ, now defined for all $\xi \in \mathbb{R}^d$. The inhomogeneous Sobolev norm of order $s \in \mathbb{R}$ is defined by

$$\|\phi\|_{H^s(\mathbb{T}^d)} = \left\|\langle\cdot\rangle^s \mathcal{F}[\phi](\cdot)\right\|_{\ell^2(\mathbb{Z}^d)},$$

where $\langle\xi\rangle = (1+\|\xi\|^2)^{1/2}$, and the inhomogeneous Sobolev space $H^s(\mathbb{T}^d)$ is then defined as the completion of $\mathcal{C}^\infty(\mathbb{T}^d)$ in the above norm. In the special case where $s \in \mathbb{N}$, one may equivalently write

$$H^s(\Omega) \equiv W^{s,2}(\mathbb{T}^d) = \left\{ f \in L^2(\mathbb{T}^d) : D^\gamma f \in L^2(\mathbb{T}^d), 0 \leq |\gamma| \leq s \right\},$$

where differentiation is understood in the distributional sense, and the norm $\|\cdot\|_{H^s(\mathbb{T}^d)}$ is then equivalent to the norm

$$\|\phi\|_{W^{s,2}(\mathbb{T}^d)} = \sum_{0\leq|\gamma|\leq s} \|D^\gamma \phi\|_{L^2(\mathbb{T}^d)}.$$

We also denote the homogeneous Sobolev seminorm of a map $\phi \in L^2(\mathbb{T}^d)$ by

$$\|\phi\|_{\dot{H}^s(\mathbb{T}^d)} = \left\|\|\cdot\|^s \mathcal{F}[\phi](\cdot)\right\|_{\ell^2(\mathbb{Z}^d)}.$$

for any $s \in \mathbb{R}$, with the convention $0/0 = 0$. $\|\cdot\|_{\dot{H}^s(\mathbb{T}^d)}$ is in fact a norm on $L_0^2(\mathbb{T}^d)$, and we define the homogeneous Sobolev space $\dot{H}^s(\mathbb{T}^d)$ as the completion of $L_0^2(\mathbb{T}^d) \cap \mathcal{C}^\infty(\mathbb{T}^d)$ under this norm. As before, one may equivalently write for $s \in \mathbb{N}$,

$$\dot{H}^s(\mathbb{T}^d) = \left\{ f \in L_0^2(\mathbb{T}^d) : D^\gamma f \in L^2(\mathbb{T}^d), |\gamma| = s \right\}.$$

The following result summarizes some elementary identities (cf. Theorem 1.122 of Triebel (2006) and Section 4.3.6 of Giné and Nickl (2016)).

Lemma 26. *Let $s > 0$. Then, there exists a constant $C > 0$ depending only on d and s such that $\|\cdot\|_{H^s(\mathbb{T}^d)} \leq C\|\cdot\|_{\mathcal{C}^s(\mathbb{T}^d)}$, and hence $\mathcal{C}^s(\mathbb{T}^d) \subseteq H^s(\mathbb{T}^d)$. Also, for any $s \in \mathbb{R}$,*

$$H^s(\mathbb{T}^d) = \mathcal{B}^s_{2,2}(\mathbb{T}^d),$$

with equivalent norms.

A.4 Wavelet Density Estimation

We next state several properties of wavelet density estimators over $\Omega \in \{\mathbb{T}^d, [0,1]^d\}$, with the corresponding basis $\Psi \in \{\Psi^{\mathrm{per}}, \Psi^{\mathrm{bc}}\}$ as in Section A.2.3. Let $q \in L^2(\Omega)$ denote a probability density with corresponding probability distribution Q, and with corresponding wavelet expansion

$$q = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \beta_\xi \xi.$$

Given an i.i.d. sample $Y_1, \dots, Y_n \sim Q$ with corresponding empirical measure $Q_n = (1/n)\sum_{i=1}^n \delta_{Y_i}$, define the unnormalized and normalized wavelet density estimators of the density q of Q,

$$\widetilde{q}_n = \sum_{\zeta \in \Phi} \widehat{\beta}_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \widehat{\beta}_\xi \xi, \qquad \widehat{q}_n = \frac{\widetilde{q}_n I(\widetilde{q}_n \geq 0)}{\int_{\widetilde{q}_n \geq 0} \widetilde{q}_n d\mathcal{L}}, \tag{47}$$

where $J_n \geq j_0$ is a deterministic threshold, and $\widehat{\beta}_\xi = \int \xi dQ_n$ for all $\xi \in \Psi_j$, $j_0 \leq j \leq J_n$. The following simple result guarantees that $\widetilde{q}_n$ integrates to unity since q is a probability density.

Lemma 27. *We have $\int_\Omega \widetilde{q}_n = 1$. In particular, it follows that $\sum_{\zeta \in \Phi} \widehat{\beta}_\zeta \int_\Omega \zeta = 1$.*

The proof of Lemma 27 appears in Appendix A.4.1. In the special case of the periodic wavelet system, for which Φ^{per} consists only of the constant function 1, Lemma 27 implies that the corresponding estimated coefficient satisfies $\widehat{\beta}_1 = 1$ deterministically, thus the definition of $\widetilde{q}_n$ in equation (47) coincides with that given in Section 4.3.

With this result in place, we turn to L^∞ concentration results for $\widetilde{q}_n$, as well as for Besov norms of $\widetilde{q}_n$, which we frequently use throughout our proofs. In what follows, write

$$q_{J_n}(y) = \mathbb{E}[\widetilde{q}_{J_n}(y)] = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \beta_\xi \xi, \quad y \in \Omega.$$

Lemma 28. *Let $N \geq 2$ and $q \in \mathcal{B}^s_{\infty,\infty}(\Omega)$ for some $s > 0$. Then, there exist constants $v, b > 0$ depending only on the choice of wavelet system, such that for any $J_n \geq j_0$, and all $u > 0$,*

$$\mathbb{P}\left(\sup_{\zeta\in\Phi} |\widehat{\beta}_\zeta - \beta_\zeta| \geq u\right) \lesssim \exp\left\{-\frac{nu^2}{b}\right\}, \tag{48}$$

$$\mathbb{P}\left(\sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \geq u\right) \lesssim 2^{\frac{dj}{2}} \exp\left\{-\frac{nu^2}{v + 2^{jd/2}bu}\right\}, \quad j_0 \leq j \leq J_n. \tag{49}$$

Furthermore, if $2^{J_n} = c_0 n^{\frac{1}{d+2s}}$ for some $c_0 > 0$, then there exists a constant $C > 0$, depending on c_0 and on the choice of wavelet system Ψ, such that the following assertions hold.

(i) For all $0 < u \leq 1$,

$$\mathbb{P}\Big(\|\widetilde{q}_n\|_{\mathcal{B}^{s/2}_{\infty,\infty}(\Omega)} \geq u + \|q\|_{\mathcal{B}^{s/2}_{\infty,\infty}(\Omega)}\Big) \leq CJ_n 2^{dJ_n} \exp\big(-u^2 2^{sJ_n}/C\big).$$

(ii) For all $2^{-J_n} \leq u \leq 1$,

$$\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \geq u\right) \leq CJ_n 2^{J_n d(d+3)} \exp\big(-nu^2 2^{-dJ_n}/C\big).$$

Lemma 28(ii) is implicit in the proofs of almost sure L^∞ bounds for wavelet estimators by Masry (1997) and Guo and Kou (2019), as well as Giné and Nickl (2009) when $d = 1$. While these results are based on wavelet estimators over $\mathbb{R}^d$, they can readily be adapted to the wavelet systems considered here, as consequences of inequalities (48)–(49). For completeness, we provide a proof of Lemma 28(ii), along with the remaining assertions of Lemma 28, in Appendix A.4.2.

Using Lemmas 27 and 28(ii), the following result is now straightforward.

Lemma 29. *Let $N \geq 2$. Assume there exist $\gamma, s > 0$ such that $q \geq 1/\gamma$ over Ω, and such that $q \in \mathcal{B}^s_{\infty,\infty}(\Omega)$. Then, there exists $c_1 > 0$ depending on $\gamma, \|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$ such that with probability at least $1 - c_1/n^2$, $\widetilde{q}_n$ is a valid probability density over Ω, and hence $\widehat{q}_n = \widetilde{q}_n$. If we instead have $N = 1$, then under no conditions on q it holds that $\widehat{q}_n = \widetilde{q}_n$ almost surely.*

Having now established that $\widetilde{q}_n$ is a valid density with high probability, we may speak of its convergence in Wasserstein distance. Weed and Berthet (2019) previously derived upper bounds on the risk, in Wasserstein distance over $[0,1]^d$, of a projection of $\widetilde{q}_n$ onto the set of probability densities. Using Lemma 29, we are able to extend their result to the estimator $\widehat{Q}_n$, i.e. the distribution function of the density $\widehat{q}_n$ defined in equation (47). We also state this result for a general exponent of the 2-Wasserstein distance.

Lemma 30. *Let $\Psi = \Psi^{\mathrm{bc}}$ with $N \geq 2$. Assume that $q \in \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ for some $s > 0$. Assume further that $q \geq 1/\gamma$ over $[0,1]^d$ for some $\gamma > 0$. Let $2^{J_n} \asymp n^{1/(d+2s)}$. Then, for any $\rho \geq 0$, there exists a constant $C > 0$ depending on M, γ, ρ, s such that*

$$\mathbb{E}W_2^\rho(\widehat{Q}_n, Q) \leq C \begin{cases} n^{-\frac{\rho(s+1)}{2s+d}}, & d \geq 3 \\ (\log n/\sqrt{n})^\rho, & d = 2 \\ 1/n^{\rho/2}, & d = 1. \end{cases} \tag{50}$$

Furthermore, when $N = 1$, equation (50) continues to hold with $s = 0$ for any density satisfying $\gamma^{-1} \le q \le \gamma$ over $[0,1]^d$, for some $\gamma > 0$.

The proof appears in Appendix A.4.4.

A.4.1 Proof of Lemma 27

Recall that Span(Φ) contains all polynomials of degree at most $N-1$ over Ω, by Lemma 24(iv). In particular, it contains the constant function 1, thus if $\beta'_\zeta = \int_\Omega \zeta$, we obtain $1 = \sum_{\zeta\in\Phi} \beta'_\zeta \zeta$. It then follows by orthonormality of Ψ that

$$\int_\Omega \widetilde{q}_n = \int_\Omega \left(\sum_{\zeta\in\Phi} \beta'_\zeta \zeta\right)\left(\sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \xi\right) = \sum_{\zeta\in\Phi} \beta'_\zeta \widehat{\beta}_\zeta = \int \left(\sum_{\zeta\in\Phi} \beta'_\zeta \zeta\right) dQ_n = 1.$$

This proves the claim. □

A.4.2 Proof of Lemma 28

Throughout the proof, $b, v, c > 0$ denote constants depending only on c_0 and the choice of wavelet system, whose value may change from line to line. To prove inequality (48), recall first from Lemma 24(v) that

$$\sup_{\zeta\in\Phi} \|\zeta\|_{L^\infty(\Omega)} \le b, \qquad \sup_{j\ge j_0} 2^{-jd/2} \sup_{\xi\in\Psi_j} \|\xi\|_{L^\infty(\Omega)} \le b. \tag{51}$$

By Hoeffding's inequality, equation (51) implies that for all $u > 0$,

$$\mathbb{P}\left(\sup_{\zeta\in\Phi} |\widehat{\beta}_\zeta - \beta_\zeta| \ge u\right) \le \sum_{\zeta\in\Phi} \mathbb{P}\left(\left|\int \zeta d(Q_n - Q)\right| \ge u\right) \lesssim \exp\left\{-\frac{nu^2}{b^2}\right\}, \tag{52}$$

where we have used the fact that $|\Phi| \lesssim 1$ by Lemma 24(i). To prove equation (49), notice that for all $\xi \in \Psi_j$ and $j \ge j_0$, given $Y \sim Q$,

$$\mathrm{Var}[\xi(Y)] \le \int \xi^2(y) q(y) dy \le \|q\|_{L^\infty(\Omega)} \int \xi^2(y) dy = \|q\|_{L^\infty(\Omega)} \le v,$$

where we used the fact that $q \in \mathcal{B}^s_{\infty,\infty}(\Omega) \subseteq L^\infty(\Omega)$. Therefore, by Bernstein's inequality, we have for all $u > 0$ and $j_0 \le j \le J_n$,

$$\mathbb{P}\left(\sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \ge u\right) \le \sum_{\xi\in\Psi_j} \mathbb{P}\Big(|\widehat{\beta}_\xi - \beta_\xi| \ge u\Big) \lesssim 2^{dj} \exp\left\{-\frac{nu^2}{v + 2^{jd/2} bu}\right\}. \tag{53}$$

Here, the last inequality uses the fact that $|\Psi_j| \lesssim 2^{dj}$ by Lemma 24(i) for all $j \ge j_0$.

To prove part (i) from here, let $0 < u \le 1$. A union bound combined with the above display leads to

$$\mathbb{P}\left(\sup_{j_0\le j\le J_n} \sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \ge u\right) \lesssim J_n 2^{dJ_n} \exp\left\{-\frac{nu^2}{v + 2^{J_n d/2} bu}\right\}, \tag{54}$$

whence, since $2^{J_n} \asymp n^{\frac{1}{d+2s}}$,

$$\mathbb{P}\left(2^{\frac{J_n(s+d)}{2}} \sup_{j_0 \le j \le J_n} \left\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi \in \Psi_j}\right\|_{\ell_\infty} \ge u\right) \tag{55}$$
$$\lesssim J_n 2^{dJ_n} \exp\left\{-\frac{nu^2 2^{-J_n(s+d)}}{v + b2^{\frac{dJ_n}{2}} 2^{-\frac{J_n s}{2} - \frac{dJ_n}{2}} u}\right\} \le J_n 2^{dJ_n} \exp\left\{-cu^2 2^{J_n s}\right\}.$$

Combining this fact with equation (52), we have

$$\begin{aligned}
&\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{\mathcal{B}^{s/2}_{\infty,\infty}} \ge u\right) \\
&\le \mathbb{P}\left(\left\|(\widehat{\beta}_\zeta - \beta_\zeta)_{\zeta \in \Phi}\right\|_{\ell_\infty} \ge u/2\right) + \mathbb{P}\left(2^{\frac{J_n(d+s)}{2}} \sup_{j_0 \le j \le J_n} \left\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi \in \Psi_j}\right\|_{\ell_\infty} \ge u/2\right) \\
&\le C J_n 2^{dJ_n} \exp\{-u^2 2^{J_n s}/C\},
\end{aligned}$$

for a large enough constant $C > 0$. Thus, we have

$$\|\widetilde{q}_n\|_{\mathcal{B}^{s/2}_{\infty,\infty}} \le \|\widetilde{q}_n - q_{J_n}\|_{\mathcal{B}^{s/2}_{\infty,\infty}} + \|q_{J_n}\|_{\mathcal{B}^{s/2}_{\infty,\infty}} \le u + \|q\|_{\mathcal{B}^{s/2}_{\infty,\infty}}$$

with probability at least $1 - CJ_n 2^{dJ_n} \exp\{-u^2 2^{J_n s}/C\}$. Part (i) thus follows. To prove part (ii), let $\delta_n \le 2^{J_n(d+2)}/(4C_0)$, for a constant $C_0 > 0$ to be specified below. Notice that for all $x, y \in \Omega$,

$$\begin{aligned}
|\widetilde{q}_n(x) - \widetilde{q}_n(y)| &\le \left|\sum_{\zeta \in \Phi} \widehat{\beta}_\zeta(\zeta(x) - \zeta(y))\right| + \left|\sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \widehat{\beta}_\xi(\xi(x) - \xi(y))\right| \\
&\lesssim \sum_{\zeta \in \Phi} |\widehat{\beta}_\zeta| \|x - y\| + \sum_{j=j_0}^{J_n} 2^{j(\frac{d}{2}+1)} |\widehat{\beta}_\xi| \|x - y\| \sum_{\xi \in \Psi_j} I(\xi(x) \wedge \xi(y) > 0) \\
&\lesssim \|x - y\| + \sum_{j=j_0}^{J_n} 2^{j(\frac{d}{2}+1)} \|\xi\|_{L^\infty(\Omega)} \|x - y\| \\
&\lesssim \sum_{j=j_0}^{J_n} 2^{j(d+1)} \|x - y\| \lesssim 2^{J_n(d+1)} \|x - y\|,
\end{aligned}$$

where we have again used the properties appearing in Lemma 24. Upon repeating an analogous calculation, we deduce that both $\widetilde{q}_n$ and q_{J_n} are $C_0 2^{J_n(d+1)}$-Lipschitz.

Let $K_n = O(1/\delta_n^d) = O(2^{-d(d+2)})$ denote the δ_n-covering number of the unit cube $[0,1]^d$ with respect to the Euclidean norm, and let $\{x_{0k} : 1 \le k \le K_n\}$ be a corresponding δ_n-cover. Letting $I_k = \{x \in [0,1]^d : \|x - x_{0k}\| \le \delta_n\}$, we have (for both $\Omega \in \{[0,1]^d, \mathbb{T}^d\}$),

$$\begin{aligned}
\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} &\le \max_{1 \le k \le K_n} \sup_{x \in I_j} |\widetilde{q}_n(x) - q_{J_n}(x)| \\
&\le \max_{1 \le k \le K_n} \sup_{x \in I_j} |\widetilde{q}_n(x) - \widetilde{q}_n(x_{0k})|
\end{aligned}$$

$$
\begin{aligned}
&\quad + \max_{1\le k\le K_n} \sup_{x\in I_j} |q_{J_n}(x_{0k}) - q_{J_n}(x)| + \max_{1\le k\le K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})| \\
&\le 2C_0 2^{J_n(d+1)}\delta_n + \max_{1\le k\le K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})| \\
&\le 2^{-J_n}/2 + \max_{1\le k\le K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})|.
\end{aligned}
$$

Thus, for any $2^{-J_n} \le u \le 1$, using Lemma 24 and the bounds (52)–(54), we have

$$
\begin{aligned}
&\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \ge u\right) \\
&\le \mathbb{P}\left(\max_{1\le k\le K_n} |\widetilde{q}_n(x_{0k}) - q_{J_n}(x_{0k})| \ge u/2\right) \\
&\le \sum_{k=1}^{K_n} \mathbb{P}\left(\left|\sum_{\zeta\in\Phi}(\widehat{\beta}_\zeta - \beta_\zeta)\zeta(x_{0k}) + \sum_{j=j_0}^{J_n}\sum_{\xi\in\Psi_j}(\widehat{\beta}_\xi - \beta_\xi)\xi(x_{0k})\right| \ge u/2\right) \\
&\le \sum_{k=1}^{K_n} \mathbb{P}\left(\left|\sum_{\zeta\in\Phi}(\widehat{\beta}_\zeta - \beta_\zeta)\zeta(x_{0k})\right| \ge u/4\right) + \sum_{k=1}^{K_n} \mathbb{P}\left(\left|\sum_{j=j_0}^{J_n}\sum_{\xi\in\Psi_j}(\widehat{\beta}_\xi - \beta_\xi)\xi(x_{0k})\right| \ge u/4\right) \\
&\le K_n \mathbb{P}\left(\sup_{\zeta\in\Phi} |\widehat{\beta}_\zeta - \beta_\zeta| \ge cu\right) + K_n \mathbb{P}\left(J_n 2^{\frac{dJ_n}{2}} \sup_{j_0\le j\le J_n}\sup_{\xi\in\Psi_j} |\widehat{\beta}_\xi - \beta_\xi| \ge cu\right) \\
&\lesssim K_n \exp(-nc^2u^2/b^2) + J_n K_n 2^{dJ_n} \exp\left(-nc^2u^2 2^{-dJ_n}/(J_n^2 v + cbJ_n u)\right).
\end{aligned}
$$

It follows that, for a sufficiently large constant $C > 0$,

$$
\mathbb{P}\left(\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \ge u\right) \le C J_n 2^{J_n d(d+3)} \exp\left(-nu^2 2^{-dJ_n}/(J_n C)\right),
$$

for all $2^{-J_n} < u \le 1$. The claim readily follows. □

A.4.3 Proof of Lemma 29

The claim for $N = 1$ follows by definition of the Haar system, thus we assume $s > 0$ and $N \ge 2$. Recall that $\widetilde{q}_n$ integrates to unity by Lemma 27, thus it suffices to show that $\widetilde{q}_n \ge 0$ with high probability. Apply Lemma 28 to deduce that

$$
\|\widetilde{q}_n - q_{J_n}\|_{L^\infty(\Omega)} \ge \gamma^{-1}/4,
$$

except on an event with probability at most c_1/n^2, for some $c_1 > 0$ depending on γ^{-1} and $\|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. Furthermore, using Lemma 24, the bias of $\widetilde{q}_n$ satisfies

$$
\begin{aligned}
\|q_{J_n} - q\|_{L^\infty(\Omega)} &= \sum_{j\ge J_n+1} 2^{\frac{dj}{2}} \|(\beta_\xi)_{\xi\in\Psi_j}\|_{\ell_\infty} \\
&\le \|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} \sum_{j\ge J_n+1} 2^{\frac{dj}{2} - j(\frac{d}{2}+s)} \lesssim \|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} 2^{-J_n s} \le \gamma^{-1}/4,
\end{aligned}
$$

for all n larger than a universal constant depending only on $\|q\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)}$. Therefore, after possibly increasing $c_1 > 0$, we have with probability at least $1 - c_1/n^2$ that for all $n \geq 1$,

$$\|\widetilde{q}_n - q\|_{L^\infty(\Omega)} \leq \gamma^{-1}/2.$$

Since $q \geq \gamma^{-1}$, we deduce that $\widetilde{q}_n \geq \gamma^{-1}/2 \geq 0$, over the same high probability event. □

A.4.4 Proof of Lemma 30

By Jensen's inequality, it suffices to assume that $\rho \geq 1$. It is straightforward to verify from Lemma 24 that the wavelet system Ψ^{bc} satisfies Assumptions E.1–E.6 of Weed and Berthet (2019), except Assumption E.2 in the special case $N = 1$. We also have $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$. These conditions are sufficient to invoke their Theorem 4 for any $N \geq 1$, leading to

$$W_2(\widehat{Q}_n, Q) \lesssim_\gamma \|\widehat{q}_n - q\|_{\mathcal{B}^{-1}_{2,1}([0,1]^d)}\,.$$

Furthermore, it follows from Lemma 29 that the event $A_n = \{\widehat{q}_n = \widetilde{q}_n\}$ satisfies $\mathbb{P}(A_n^{\mathsf{c}}) \lesssim n^{-2}$. Let $q_{J_n} = \mathbb{E}[\widetilde{q}_n]$, so that

$$\mathbb{E}W_2^\rho(\widehat{Q}_n, Q) = \mathbb{E}\Big[W_2^\rho(\widehat{Q}_n, Q) I_{A_n}\Big] + \mathbb{E}\Big[W_2^\rho(\widehat{Q}_n, Q) I_{A_n^{\mathsf{c}}}\Big] \lesssim \mathbb{E}\|\widetilde{q}_n - q\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} + n^{-2}.$$

Now, we make use of the following result which can be deduced from the proof of Theorem 1 and Proposition 4 of Weed and Berthet (2019).

Lemma 31 (Weed and Berthet (2019))**.** *Let q be a density satisfying $\gamma^{-1} \leq q \leq \gamma$ over $[0,1]^d$. Assume further that $q \in \mathcal{B}^s_{\infty,\infty}([0,1]^d)$ for some $s \geq 0$. Then,*

$$\|q_{J_n} - q\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \lesssim 2^{-\rho J_n (s+1)},$$
$$\mathbb{E}\big\|(\widehat{\beta}_\zeta - \beta_\zeta)_{\zeta \in \Phi^{\mathrm{bc}}}\big\|^\rho_{\ell_2} \lesssim 1/n^{\rho/2}, \quad \mathbb{E}\big\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\big\|^\rho_{\ell_2} \lesssim 2^{dj}/n^{\rho/2}, \quad j \geq j_0.$$

Let $\rho' \geq 1$ satisfy $\frac{1}{\rho} + \frac{1}{\rho'} = 1$. Lemma 31 implies,

$$\begin{aligned}
&\mathbb{E}\|\widetilde{q}_n - q\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \\
&\lesssim \mathbb{E}\|\widetilde{q}_n - q_{J_n}\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} + \|q_{J_n} - q\|^\rho_{\mathcal{B}^{-1}_{2,1}([0,1]^d)} \\
&\lesssim \mathbb{E}\big\|(\widehat{\beta}_\zeta - \beta_\zeta)_{\zeta \in \Phi^{\mathrm{bc}}}\big\|^\rho_{\ell_2} + \mathbb{E}\left(\sum_{j=j_0}^{J_n} 2^{-j}\big\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\big\|_{\ell_2}\right)^\rho + 2^{-\rho J_n(s+1)} \\
&\lesssim n^{-\frac{\rho}{2}} + \left(\sum_{j=j_0}^{J_n} 2^{\rho(\eta-1)j}\mathbb{E}\big\|(\widehat{\beta}_\xi - \beta_\xi)_{\xi \in \Psi_j^{\mathrm{bc}}}\big\|^\rho_{\ell_2}\right)\left(\sum_{j=j_0}^{J_n} 2^{-\rho'\eta j}\right)^{\frac{\rho}{\rho'}} + 2^{-\rho J_n(s+1)} \\
&\lesssim n^{-\frac{\rho}{2}} + n^{-\frac{\rho}{2}}\left(\sum_{j=j_0}^{J_n} 2^{\rho(\eta+\frac{d}{2}-1)j}\right)\left(\sum_{j=j_0}^{J_n} 2^{-\rho'\eta j}\right)^{\frac{\rho}{\rho'}} + 2^{-\rho J_n(s+1)},
\end{aligned}$$

for any $\eta \in \mathbb{R}$. Now, when $d \geq 3$, choose $1 - \frac{d}{2} < \eta < 0$. In this case, the above display is of order

$$n^{-\frac{\rho}{2}} 2^{[\rho(\eta + \frac{d}{2} - 1) - \rho\eta] J_n} + 2^{-\rho J_n (s+1)} = 2^{\rho(\frac{d}{2} - 1) J_n} + 2^{-\rho J_n (s+1)} \lesssim n^{-\frac{\rho(s+1)}{2s+d}},$$

which proves the claim for $d \geq 3$. When $d \leq 2$, choose $\eta = 0$. Then, the penultimate display is dominated by its second term, which is of order $n^{-\rho/2}$ when $d = 1$ and of order $(\log n/\sqrt{n})^\rho$ when $d = 2$. The claim follows. $\square$

A.5 Kernel Density Estimation

We close this appendix with several properties of kernel density estimators. We adopt the same notation as in Section 4.3. Specifically, $K : \mathbb{R}^d \to \mathbb{R}$ denotes an even kernel, we write $K_{h_n} = h_n^{-d} K(\cdot/h_n)$ for some bandwidth $h_n > 0$, and we consider the kernel density estimator

$$\widetilde{q}_n = Q_n \star K_{h_n} = \int_{\mathbb{R}^d} K_{h_n}(\cdot - z) dQ_n(z),$$

where $Q_n \in \mathcal{P}(\mathbb{T}^d)$ denotes the empirical measure based on an i.i.d. sample $Y_1, \dots, Y_n \sim Q \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$. In the above display, recall that integration over $\mathbb{R}^d$ with respect to the measure Q_n is understood as integration with respect to its extension to $\mathbb{R}^d$ by $\mathbb{Z}^d$-periodicity, namely the measure

$$\frac{1}{n} \sum_{k \in \mathbb{Z}^d} \sum_{i=1}^n \delta_{Y_i + k}.$$

Equivalently, we may write

$$\widetilde{q}_n = \int_{\mathbb{T}^d} K_{h_n}^{(\mathrm{per})}(\cdot - z) dQ_n(z), \quad \text{where } K_{h_n}^{(\mathrm{per})} = \sum_{k \in \mathbb{Z}^d} K_{h_n}(\cdot + k).$$

With the same conventions, we define

$$q_{h_n}(y) = \mathbb{E}[\widetilde{q}_n(y)] = Q \star K_{h_n}(y), \quad y \in \mathbb{T}^d.$$

We begin by proving an $L^\infty(\mathbb{T}^d)$ concentration inequality for the estimator $\widetilde{q}_n$ about its mean. Though such concentration inequalities have previously been established by Giné and Guillou (2002) under very general conditions on K, the following simple result will suffice for our purposes.

Lemma 32. *Assume $q \leq \gamma$ over $\mathbb{T}^d$ for some $\gamma > 0$, and that $K \in \mathcal{C}^1(\mathbb{R}^d)$. Then, there exists a constant $C > 0$ depending only on $\gamma, \|K\|_{\mathcal{C}^1(\mathbb{R}^d)}$ such that for all $h_n \leq u \leq 1$,*

$$\mathbb{P}\Big(\|\widetilde{q}_n - q_{h_n}\|_{L^\infty(\mathbb{T}^d)} \geq u \Big) \leq C h_n^{-d(d+2)} \exp\big(-n u^2 h_n^d / C \big).$$

The proof appears in Appendix A.5.1. When the true density q is Hölder continuous with any positive exponent, and bounded below by a positive constant, it is easy to infer from this result that $\widetilde{q}_n$ defines a valid density except on an event with exponentially small probability. We shall additionally require the following result, which ensures that the fitted density $\widetilde{q}_n$ enjoys a nonzero amount of Hölder regularity.

Lemma 33. *Assume $nh_n^d \asymp n^a$ for some $a \in (0,1)$. Assume further that $q \in \mathcal{C}^s(\mathbb{T}^d)$ for some $s > 0$, and that $K \in \mathcal{C}^1(\mathbb{R}^d) \cap L^1(\mathbb{R}^d)$. Then, there exist constants $C, c_1 > 0, \beta \in (0, s \wedge 1)$ depending only on $\|q\|_{\mathcal{C}^s(\mathbb{T}^d)}, \|K\|_{\mathcal{C}^1(\mathbb{R}^d)}, \|K\|_{L^1(\mathbb{R}^d)}, a, s, d$ such that for all $n \geq 1$, with probability at least $1 - c_1/n^2$,*

$$\|\widetilde{q}_n\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \leq C.$$

The proof appears in Appendix A.5.2.

A.5.1 Proof of Lemma 32

Let $\delta_n \leq h_n^{d+2}/(4\|K\|_{\mathcal{C}^1(\mathbb{R}^d)})$. By a direct calculation, it can be seen that

$$\|\widetilde{q}_n\|_{\mathcal{C}^1(\mathbb{T}^d)} \vee \|q_{h_n}\|_{\mathcal{C}^1(\mathbb{T}^d)} \leq \|K\|_{\mathcal{C}^1(\mathbb{R}^d)} h_n^{-(d+1)}.$$

Let $J_n = O(1/\delta_n^d) = O(h_n^{-d(d+2)})$ denote the δ_n-covering number of the unit cube $[0,1]^d$ with respect to the Euclidean norm, and let $\{x_{0j} : 1 \leq j \leq J_n\}$ be a corresponding δ_n-cover. Letting $I_j = \{x \in [0,1]^d : \|x - x_{0j}\| \leq \delta_n\}$, we have,

$$\begin{aligned}
\|\widetilde{q}_n - q_{h_n}\|_{L^\infty(\mathbb{T}^d)} &\leq \max_{1\leq j\leq J_n} \sup_{x\in I_j} |\widetilde{q}_n(x) - q_{h_n}(x)| \\
&\leq \max_{1\leq j\leq J_n} \sup_{x\in I_j} |\widetilde{q}_n(x) - \widetilde{q}_n(x_{0j})| \\
&\quad + \max_{1\leq j\leq J_n} \sup_{x\in I_j} |q_{h_n}(x_{0j}) - q_{h_n}(x)| + \max_{1\leq j\leq J_n} |\widetilde{q}_n(x_{0j}) - q_{h_n}(x_{0j})| \\
&\leq 2\|K\|_{\mathcal{C}^1(\mathbb{R}^d)} h_n^{-(d+1)}\delta_n + \max_{1\leq j\leq J_n} |\widetilde{q}_n(x_{0j}) - q_{h_n}(x_{0j})| \\
&\leq h_n/2 + \max_{1\leq j\leq J_n} |\widetilde{q}_n(x_{0j}) - q_{h_n}(x_{0j})|.
\end{aligned}$$

Thus, for any $h_n \leq u \leq 1$,

$$\begin{aligned}
&\mathbb{P}\left(\|\widetilde{q}_n - q_{h_n}\|_{L^\infty(\mathbb{T}^d)} \geq u\right) \\
&\leq \mathbb{P}\left(\max_{1\leq j\leq J_n} |\widetilde{q}_n(x_{0j}) - q_{h_n}(x_{0j})| \geq u/2\right) \\
&\leq \sum_{j=1}^{J_n} \mathbb{P}\left(|\widetilde{q}_n(x_{0j}) - q_{h_n}(x_{0j})| \geq u/2\right) \\
&\leq \sum_{j=1}^{J_n} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n \left[K_{h_n}^{(\text{per})}(\|x_{0j} - X_i\|) - \mathbb{E}\left\{K_{h_n}^{(\text{per})}(\|x_{0j} - X_i\|)\right\}\right]\right| \geq u/2\right) \\
&\leq 2J_n \exp\left(-\frac{nu^2}{8\left(\gamma\|K^{(\text{per})}\|_{L^\infty(\mathbb{T}^d)}h_n^{-d} + uh_n^{-d}\|K^{(\text{per})}\|_{L^\infty(\mathbb{R}^d)}/3\right)}\right),
\end{aligned}$$

where we invoked Bernstein's inequality by noting that

$$\left\|K_{h_n}^{(\text{per})}(\|x - \cdot\|)\right\|_{L^\infty(\mathbb{T}^d)} \leq h_n^{-d}\|K^{(\text{per})}\|_{L^\infty(\mathbb{T}^d)},$$

$$\text{and,}\quad \mathrm{Var}\left[K_{h_n}^{(\mathrm{per})}(\|x - X_i\|)\right] \leq h_n^{-2d} \int \left[K^{(\mathrm{per})}\left(\frac{\|x-y\|}{h_n}\right)\right]^2 q(y)dy$$
$$\leq \gamma \|K^{(\mathrm{per})}\|_{L^\infty(\mathbb{T}^d)} h_n^{-d} \int K_{h_n}^{(\mathrm{per})}(\|x-y\|)\,dy$$
$$= \gamma \|K^{(\mathrm{per})}\|_{L^\infty(\mathbb{T}^d)} h_n^{-d}.$$

It follows that, for a sufficiently large constant $C > 0$ depending on γ and $\|K\|_{\mathcal{C}^1(\mathbb{R}^d)}$, we have

$$\mathbb{P}\left(\|\widetilde{q}_n - q_{h_n}\|_{L^\infty(\mathbb{T}^d)} \geq u\right) \leq C h_n^{-d(d+2)} \exp\left(-nu^2 h_n^d / C\right).$$

The claim readily follows. □

A.5.2 Proof of Lemma 33

By Lemma 32, there is a constant $c_1 > 0$ and an event A_n satisfying $\mathbb{P}(A_n) \geq 1 - 1/n^2$ such that

$$\|\widetilde{q}_n - q_{h_n}\|_{L^\infty(\mathbb{T}^d)} \leq \gamma_n = c_1 \sqrt{\frac{\log n}{n h_n^d}}.$$

All subsequent statements are made over the event A_n. Now, given $\beta \in (0, s \wedge 1)$ to be specified below, and $x, y \in \mathbb{T}^d$, we have

$$\begin{aligned}
|\widetilde{q}_n(x) - \widetilde{q}_n(y)| &\leq 2\|\widetilde{q}_n - q_{h_n}\|_{L^\infty(\mathbb{T}^d)} + |q_{h_n}(x) - q_{h_n}(y)| \\
&\leq 2\gamma_n + \int_{\mathbb{R}^d} \left|K(z)\left[q(x - h_n z) - q(y - h_n z)\right]\right| dz \\
&\leq 2\gamma_n + \|q\|_{\mathcal{C}^\beta(\mathbb{T}^d)} \|K\|_{L^1(\mathbb{R}^d)} \|x-y\|^\beta \\
&\leq C_1(\gamma_n + \|x-y\|^\beta),
\end{aligned}$$

for a large enough constant $C_1 > 0$. If $\|x-y\|^\beta \geq \gamma_n$, then $\widetilde{q}_n$ already satisfies the condition of β-Hölder continuity, thus it suffices to assume $\|x-y\|^\beta < \gamma_n$. Recall that

$$\|\widetilde{q}_n\|_{\mathcal{C}^1(\mathbb{T}^d)} \leq \|K\|_{\mathcal{C}^1(\mathbb{R}^d)} h_n^{-(d+1)}.$$

We deduce that for all x, y such that $\|x-y\|^\beta < \gamma_n$,

$$|\widetilde{q}_n(x) - \widetilde{q}_n(y)| \lesssim \frac{\|x-y\|}{h_n^{d+1}} \leq \frac{\gamma_n^{\frac{1}{\beta}-1}}{h_n^{d+1}} \|x-y\|^\beta \lesssim \frac{(n^a/\log n)^{\frac{1}{2}\left(1-\frac{1}{\beta}\right)}}{n^{(a-1)(d+1)}} \|x-y\|^\beta \lesssim \|x-y\|^\beta,$$

for any small enough choice of β. The claim follows. □

B On the Variance of Kantorovich Potentials

We state a straightforward technical result which will be used throughout our proofs.

Lemma 34. *Let Ω be equal to $[0,1]^d$ or $\mathbb{T}^d$. Given $P, Q \in \mathcal{P}_{\mathrm{ac}}(\Omega)$, let (ϕ_0, ψ_0) be a pair of Kantorovich potentials in the optimal transport transport problem from P to Q. Assume further that the density q of Q satisfies $\gamma^{-1} \le q \le \gamma$ over Ω, for some $\gamma > 0$. Define $\overline{\psi}_0 = \psi_0 - \int_\Omega \psi_0$. Then, there exists a constant $C > 0$ depending only on d such that*

$$\|\overline{\psi}_0\|_{L^2(Q)} \le C\gamma W_2(P, Q).$$

In particular,

$$\mathrm{Var}_Q[\psi_0(Y)] \le (C\gamma)^2 W_2^2(P, Q).$$

The proof will follow from Poincaré inequalities over $[0,1]^d$ and $\mathbb{T}^d$, which we recall here as they will be needed again in the sequel. The following is a special case of the Poincaré inequality for convex domains (see for instance Leoni (2017), Theorem 12.30).

Lemma 35. *Let $0 < a < b < \infty$ and $\Omega = [a, b]^d$. Then, there exists a constant $C > 0$ depending only on d such that for all $f \in H^1(\Omega)$ satisfying $\int_\Omega f = 0$,*

$$\|f\|_{L^2(\Omega)} \le C(b-a) \|\nabla f\|_{L^2(\Omega)}.$$

We also state the following classical periodic Poincaré inequality (see for instance Steinerberger (2016) for a simple proof).

Lemma 36. *Let $f \in H^1(\mathbb{T}^d)$ satisfy $\int_{\mathbb{T}^d} f = 0$. Then, $\|f\|_{L^2(\mathbb{T}^d)} \le \|\nabla f\|_{L^2(\mathbb{T}^d)}$.*

Proof of Lemma 34. Since $\psi_0 \in H^1(\Omega)$ by definition, we may apply the Poincaré inequality over Ω (namely, Lemma 35 when $\Omega = [0,1]^d$, or Lemma 36 when $\Omega = \mathbb{T}^d$). This fact, together with the assumption $\gamma^{-1} \le q \le \gamma$, implies

$$\|\overline{\psi}_0\|_{L^2(Q)}^2 \le \gamma \|\overline{\psi}_0\|_{L^2([0,1]^d)}^2 \le C^2 \gamma \|\nabla \psi_0\|_{L^2([0,1]^d)}^2 \le (C\gamma)^2 \|\nabla \psi_0\|_{L^2(Q)}^2 = (C\gamma)^2 W_2^2(P, Q),$$

which then also implies

$$\mathrm{Var}_Q[\psi_0(Y)] = \mathrm{Var}_Q[\overline{\psi}_0(Y)] \le \|\overline{\psi}_0\|_{L^2(Q)}^2 \le (C\gamma)^2 W_2^2(P, Q),$$

as claimed. □

C Proofs of Stability Bounds

C.1 Proof of Theorem 6

Recall that φ_0 denotes a Brenier potential from P to Q, while $\phi_0 = \|\cdot\|^2 - 2\varphi_0$ and $\psi_0 = \|\cdot\|^2 - 2\varphi_0^*$ denote the corresponding Kantorovich potentials. Since we have assumed that both P and Q are absolutely continuous distributions, Brenier's Theorem implies that $S_0 = \nabla \varphi_0^*$ is the optimal transport map from Q to P. Since φ_0 is closed, the assumption

$$\frac{1}{\lambda} I_d \preceq \nabla^2 \varphi_0 \preceq \lambda I_d,$$

from condition **A1(λ)** also implies (Hiriart-Urruty and Lemaréchal (2004), Theorem 4.2.2),

$$\frac{1}{\lambda} I_d \preceq \nabla^2 \varphi_0^* \preceq \lambda I_d.$$

Combining this bound with a second-order Taylor expansion of φ_0^* leads to the following inequalities

$$\frac{1}{2\lambda} \|x-y\|^2 \le \varphi_0^*(y) - \varphi_0^*(x) - \big\langle S_0(x), y-x \big\rangle \le \frac{\lambda}{2} \|x-y\|^2, \quad x, y \in \Omega. \tag{56}$$

With these facts in place, we turn to proving the theorem, namely that

$$\frac{1}{\lambda} \|\widehat{T} - T_0\|_{L^2(P)}^2 \le W_2^2(P, \widehat{Q}) - W_2^2(P, Q) - \int \psi_0 d(\widehat{Q} - Q) \le \lambda W_2^2(\widehat{Q}, Q). \tag{57}$$

We begin with the first inequality. Since $\widehat{T}$ is the optimal transport map from P to $\widehat{Q}$, we have,

$$\begin{aligned}
W_2^2(P, \widehat{Q}) &= \int \|\widehat{T}(x) - x\|^2 dP(x) \\
&= \int \|T_0(x) - x\|^2 dP(x) \\
&+ \int 2\big\langle T_0(x) - x, \widehat{T}(x) - T_0(x) \big\rangle dP(x) + \int \|\widehat{T}(x) - T_0(x)\|^2 dP(x) \\
&= W_2^2(P, Q) + \int 2\big\langle T_0(x) - x, \widehat{T}(x) - T_0(x) \big\rangle dP(x) + \|\widehat{T} - T_0\|_{L^2(P)}^2.
\end{aligned}$$

To bound the cross term, notice that equation (56) implies

$$\begin{aligned}
2 \int &\big\langle T_0(x) - x, \widehat{T}(x) - T_0(x) \big\rangle dP(x) \\
&= 2 \int \big\langle T_0(x) - S_0(T_0(x)), \widehat{T}(x) - T_0(x) \big\rangle dP(x) \\
&\ge 2 \int \Big[\big\langle T_0(x), \widehat{T}(x) - T_0(x) \big\rangle \\
&\qquad\quad + \varphi_0^*(T_0(x)) - \varphi_0^*(\widehat{T}(x)) + \frac{1}{2\lambda} \|\widehat{T}(x) - T_0(x)\|^2 \Big] dP(x) \\
&= \int \Big[\|\widehat{T}(x)\|^2 - \|T_0(x)\|^2 - \|\widehat{T}(x) - T_0(x)\|^2 \\
&\qquad\quad + 2\varphi_0^*(T_0(x)) - 2\varphi_0^*(\widehat{T}(x)) + \frac{1}{\lambda} \|\widehat{T}(x) - T_0(x)\|^2 \Big] dP(x) \\
&= \left(\frac{1}{\lambda} - 1 \right) \|\widehat{T} - T_0\|_{L^2(P)}^2 + \int \psi_0 d(\widehat{Q} - Q).
\end{aligned}$$

We deduce

$$W_2^2(P, \widehat{Q}) \ge W_2^2(P, Q) + \frac{1}{\lambda} \|\widehat{T} - T_0\|_{L^2(P)}^2 + \int \psi_0 d(\widehat{Q} - Q),$$

To prove the second inequality in equation (57), let $\widehat{\pi}$ denote an optimal coupling between Q and $\widehat{Q}$. Then, the measure $\widehat{\pi}_{S_0} = (S_0, Id)_{\#}\widehat{\pi}$ is a (possibly suboptimal) coupling between P and $\widehat{Q}$, thus

$$W_2^2(P, \widehat{Q}) \leq \int \|x - z\|^2 \, d\widehat{\pi}_{S_0}(x, z) = \int \|S_0(y) - z\|^2 \, d\widehat{\pi}(y, z). \tag{58}$$

The claim is now a consequence of the following technical Lemma, which will be used again in the sequel.

Lemma 37. *We have,*

$$\int \|S_0(y) - z\|^2 \, d\widehat{\pi}(y, z) \leq W_2^2(P, Q) + \int \psi_0 d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q).$$

C.2 Proof of Lemma 37

We have,

$$\begin{aligned}
&\int \|S_0(y) - z\|^2 \, d\widehat{\pi}(y, z) \\
&\quad = \int \|S_0(y) - y\|^2 \, dQ(y) + \int \|y - z\|^2 \, d\widehat{\pi}(y, z) + 2 \int \langle S_0(y) - y, y - z \rangle d\widehat{\pi}(y, z) \\
&\quad = W_2^2(P, Q) + W_2^2(\widehat{Q}, Q) + 2 \int \langle S_0(y) - y, y - z \rangle d\widehat{\pi}(y, z).
\end{aligned}$$

Now, notice that by (56),

$$\begin{aligned}
2 \int \langle S_0(y), y - z \rangle d\widehat{\pi}(y, z) &\leq 2 \int \left[\varphi_0^*(y) - \varphi_0^*(z) + \frac{\lambda}{2} \|y - z\|^2 \right] d\widehat{\pi}(y, z) \\
&= 2 \int \varphi_0^* d(Q - \widehat{Q}) + \lambda W_2^2(\widehat{Q}, Q),
\end{aligned}$$

and,

$$\begin{aligned}
&2 \int \langle -y, y - z \rangle d\widehat{\pi}(y, z) \\
&\quad = \int \left[\|z\|^2 - \|z - y\|^2 - \|y\|^2 \right] d\widehat{\pi}(y, z) = \int \|\cdot\|^2 \, d(\widehat{Q} - Q) - W_2^2(\widehat{Q}, Q).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&W_2^2(P, \widehat{Q}) - W_2^2(P, Q) \\
&\quad \leq \int \left(\|\cdot\|^2 - 2\varphi_0^* \right) d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q) = \int \psi_0 d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q),
\end{aligned}$$

and the claim follows. □

D Proofs of Upper Bounds for One-Sample Empirical Estimators

In this Appendix, we prove Corollaries 8 and 9.

D.1 Proof of Corollary 8

We shall make use of the notation introduced in Section 3.3 and Appendix A.2.4, regarding wavelet density estimation over $[0,1]^d$. In particular, let $\Psi = \Psi^{\text{bc}}$ with $N = 1$, so that Ψ is the Haar wavelet basis on $[0,1]^d$.

Lemma 38. *Let $J \geq 1$ be an integer. For any $\mu \in \mathcal{P}([0,1]^d)$, let $\mu_J \in \mathcal{P}_{\text{ac}}([0,1]^d)$ denote the measure admitting density*

$$q_J = 1 + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \xi \int \xi d\mu,$$

with respect to the Lebesgue measure on $[0,1]^d$. Then, $W_2(\mu, \mu_J) \leq \sqrt{d} 2^{-J}$.

Proof. The Lemma is a straightforward consequence of dyadic partitioning arguments which have previously been used by Boissard and Le Gouic (2014); Fournier and Guillin (2015); Weed and Bach (2019); Lei (2020). In particular, for all $j \geq 0$, let $\mathcal{Q}_j$ denote the natural partition (up to intersections on Lebesgue null sets) of $[0,1]^d$ into 2^{dj} cubes of length 2^{-j}. Then, Proposition 1 of Weed and Bach (2019) implies

$$W_2^2(\mu, \mu_J) \leq d \left[2^{-2J} + \sum_{j=1}^{J} 2^{-2(j-1)} \sum_{S \in \mathcal{Q}_j} |\mu(S) - \mu_J(S)| \right].$$

To prove the claim, it thus suffices to show that $\mu(S) = \mu_J(S)$ for all $S \in \mathcal{Q}_j$ and $j = 1, \dots, J$.

Let $j \geq 0$, $S \in \mathcal{Q}_j$, and recall that I_S is the indicator function of S. Denote its expansion in the Haar basis by

$$I_S = \gamma_\zeta + \sum_{\ell=0}^{\infty} \sum_{\xi \in \Psi_\ell} \gamma_\xi \xi, \quad \text{where } \gamma_\psi = \int I_S \psi, \psi \in \Psi.$$

Notice that for any $\ell \geq j$ and $\xi \in \Psi_\ell$, we have

$$\text{supp}(\xi) \subseteq I_S, \quad \text{or} \quad \text{supp}(\xi) \cap I_S = \emptyset.$$

Furthermore, since $\zeta = I_{[0,1]^d}$, and the Haar basis is orthonormal, we must have $\int_{[0,1]^d} \xi = 0$ for any $\xi \in \Psi_j$, $j \geq 0$. It must follow that

$$\gamma_\xi = \int I_S \xi = 0, \quad \text{for all } \xi \in \Psi_\ell, \ell \geq j,$$

that is, $I_S \in \text{Span}\left(\Phi \cup \bigcup_{\ell=0}^{j-1} \Psi_\ell\right)$. We therefore have,

$$\mu_J(S) = \int I_S(y) q_J(y) dy$$

$$
\begin{aligned}
&= \mathcal{L}(S) + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \left(\int \xi d\mu \right) \left(\int I_S(y)\xi(y)dy \right) \\
&= \mathcal{L}(S) + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \left(\int \xi d\mu \right) \gamma_\xi \\
&= \int \left(\mathcal{L}(S) + \sum_{j=0}^{J} \sum_{\xi \in \Psi_j} \xi \gamma_\xi \right) d\mu = \int I_S d\mu = \mu(S).
\end{aligned}
$$

The claim follows. □

To prove the claim from here, let $2^{J_n} \asymp n^{1/d}$, and let $\widehat{Q}_n$ be the distribution with density

$$
\widehat{q}_n(y) = 1 + \sum_{j=0}^{J_n} \sum_{\xi \in \Psi_j} \left(\int \xi dQ_n \right) \xi, \quad y \in [0,1]^d.
$$

Apply Lemma 38 to the measure $\mu = Q_n$ to obtain

$$
W_2^2(Q_n, Q) \lesssim W_2^2(Q_n, \widehat{Q}_n) + W_2^2(\widehat{Q}_n, Q) \lesssim 2^{-2J_n} + W_2^2(\widehat{Q}_n, Q) \lesssim n^{-2/d} + W_2^2(\widehat{Q}_n, Q).
$$

Furthermore, recall that $\gamma^{-1} \le q \le \gamma$, thus we may apply Lemma 30 to deduce

$$
\mathbb{E} W_2^2(\widehat{Q}_n, Q) \lesssim \begin{cases} n^{-2/d}, & d \ge 3 \\ (\log n)^2/n, & d = 2 \\ 1/n, & d = 1. \end{cases}
$$

The claim follows. □

D.2 Proof of Corollary 9

By Theorem 6,

$$
\mathbb{E}\big| W_2^2(P, Q_n) - W_2^2(P, Q) \big| \le \mathbb{E} W_2^2(Q_n, Q) + \mathbb{E} \left| \int \psi_0 d(Q_n - Q) \right|.
$$

By Jensen's inequality, the final term satisfies

$$
\mathbb{E} \left| \int \psi_0 d(Q_n - Q) \right| \le n^{-\frac{1}{2}} \sqrt{\mathrm{Var}_Q[\psi_0(Y)]}.
$$

Since ψ_0 is uniformly bounded by a constant depending only on d, the right-hand side of the above display is of order $n^{-1/2}$. Furthermore, by equation (21), we have $\mathbb{E} W_2^2(Q_n, Q) \lesssim \kappa_n$, thus the first part of the claim follows.

Under the assumptions of the second part of the claim, we may instead use Corollary 8 to obtain the stronger bound $\mathbb{E} W_2^2(Q_n, Q) \lesssim \overline{\kappa}_n$, as well as Lemma 34 to derive $\mathrm{Var}_Q[\psi_0(Y)] \lesssim W_2^2(P, Q)$. The claim then follows. □

E Proofs of Upper Bounds for One-Sample Wavelet Estimators

E.1 Proof of Theorem 10

Under the assumptions of part (i), we may apply Theorem 6 and Lemma 30 to obtain,

$$\mathbb{E}\big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \lesssim_\lambda \mathbb{E} W_2^2(\widehat{Q}_n, Q) \lesssim_{M,\gamma,\alpha} R_{T,n}(\alpha),$$

which immediately leads to the first claim. To prove the second claim, recall that we have assumed $\alpha > 1$, whence the assumption on φ_0 implies in particular that $\|\varphi_0\|_{\mathcal{C}^2(\Omega)} \leq \lambda$. Since the densities p, q are bounded from below by γ^{-1} over $[0,1]^d$, and also bounded from above due to their Hölder continuity and the compactness of $[0,1]^d$, it follows by Lemma 2 that φ_0 satisfies condition **A1(λ)**, after possibly modifying the value of λ in terms of γ. We may therefore invoke Theorem 6 to obtain,

$$L(\widehat{Q}_n) \leq W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q) \leq \lambda W_2^2(\widehat{Q}_n, Q) + L(\widehat{Q}_n).$$

Let $C > 0$ be a constant depending only on $M, \lambda, \gamma, \alpha$, whose value may change from line to line. By Lemma 30, we have

$$\mathbb{E} W_2^2(\widehat{Q}_n, Q) \leq C R_{T,n}(\alpha), \quad \text{and} \quad \mathbb{E} W_2^4(\widehat{Q}_n, Q) \leq C R_{T,n}^2(\alpha).$$

Furthermore, by Lemma 11, we have

$$\begin{aligned}
\big|\mathbb{E} L(\widehat{Q}_n)\big| &\leq C R_{T,n}(\alpha) \\
\mathrm{Var}\left[L(\widehat{Q}_n)\right] &\leq \frac{1}{n}\left(\mathrm{Var}_Q[\psi_0(Y)] + 2^{-2J_n(\alpha+1)}\right) \leq \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} + C R_{T,n}^2(\alpha) \\
\mathbb{E}\big|L(\widehat{Q}_n)\big|^2 = \big|\mathbb{E} L(\widehat{Q}_n)\big|^2 + \mathrm{Var}\left[L(\widehat{Q}_n)\right] &\leq \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} + C R_{T,n}^2(\alpha).
\end{aligned}$$

Combining the preceding three displays, we deduce that

$$\big|\mathbb{E} W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big| \leq \lambda \mathbb{E} W_2^2(\widehat{Q}_n, Q) + \big|\mathbb{E} L(\widehat{Q}_n)\big| \leq C R_{T,n}(\alpha), \tag{59}$$

and,

$$\begin{aligned}
&\mathbb{E}\big|W_2^2(P, \widehat{Q}_n) - W_2^2(P, Q)\big|^2 \\
&\quad\leq \mathbb{E}\left[\left(\lambda W_2^2(\widehat{Q}_n, Q) + \big|L(\widehat{Q}_n)\big|\right)^2\right] \\
&\quad\leq \lambda^2 \mathbb{E} W_2^4(\widehat{Q}_n, Q) + 2\lambda \mathbb{E}\left[W_2^2(\widehat{Q}_n, Q)\big|L(\widehat{Q}_n)\big|\right] + \mathbb{E}\big|L(\widehat{Q}_n)\big|^2 \\
&\quad\leq \lambda^2 \mathbb{E} W_2^4(\widehat{Q}_n, Q) + 2\lambda\sqrt{\left(\mathbb{E} W_2^4(\widehat{Q}_n, Q)\right)\mathbb{E}\big|L(\widehat{Q}_n)\big|^2} + \mathbb{E}\big|L(\widehat{Q}_n)\big|^2 \\
&\quad\leq C\lambda^2 R_{T,n}^2(\alpha) + 2\lambda\sqrt{C R_{T,n}^2(\alpha)\left(C R_{T,n}^2(\alpha) + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}\right)} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n} \\
&\quad\leq C^2 R_{T,n}^2(\alpha) + 2C R_{T,n}(\alpha)\sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}
\end{aligned}$$

$$\leq \left(CR_{T,n}(\alpha) + \sqrt{\frac{\mathrm{Var}_Q[\psi_0(Y)]}{n}} \right)^2.$$

The claim follows $\square$

It thus remains to prove Lemma 11.

E.2 Proof of Lemma 11

In order to bound the bias of $\int \psi_0 \widehat{q}_n$, recall from Lemma 29 that the event $A_n = \{\widehat{q}_n = \widetilde{q}_n\}$ satisfies $\mathbb{P}(A_n^{\mathsf{c}}) \lesssim n^{-2}$. Since ψ_0 is bounded by a constant depending only on d, we have,

$$\mathbb{E}\left| \int \psi_0(\widehat{q}_n - \widetilde{q}_n) \right| \leq \mathbb{E}\left(\left| \int \psi_0(\widehat{q}_n - \widetilde{q}_n) \right| I_{A_n^{\mathsf{c}}} \right) \lesssim \mathbb{P}(A_n^{\mathsf{c}}) \lesssim 1/n^2.$$

We deduce that

$$\left| \mathbb{E}\big[L(\widehat{Q}_n)\big] \right| \lesssim \left| \int \psi_0(\widetilde{q}_n - q) \right| + \frac{1}{n^2},$$

thus we are left with bounding the bias of $\int \psi_0 \widetilde{q}_n$. Recall that $\widehat{\beta}_\xi$ is an unbiased estimator of β_ξ for all $\xi \in \Psi$, so that

$$q_{J_n} := \mathbb{E}[\widetilde{q}_n] = \sum_{\zeta \in \Phi} \beta_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi \in \Psi_j} \beta_\xi \xi.$$

Write the expansion of ψ_0 in the basis Ψ as

$$\psi_0 = \sum_{\zeta \in \Phi} \gamma_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \gamma_\xi \xi, \quad \text{where } \gamma_\xi = \int \psi_0 \xi \text{ for all } \xi \in \Psi,$$

where the series converges uniformly due to the Hölder regularity of ψ_0, so that,

$$\int \psi_0(q - q_{J_n}) = \int \left(\sum_{\zeta \in \Phi} \gamma_\zeta \zeta + \sum_{j=j_0}^{\infty} \sum_{\xi \in \Psi_j} \gamma_\xi \xi \right) \left(\sum_{j=J_n+1}^{\infty} \sum_{\xi \in \Psi_j} \beta_\xi \xi \right) = \sum_{j=J_n+1}^{\infty} \sum_{\xi \in \Psi_j} \gamma_\xi \beta_\xi,$$

by orthonormality of the basis Ψ. By Lemma 24(i) in Appendix A.2, we have $|\Psi_j| \lesssim 2^{dj}$, therefore

$$\left| \int \psi_0(q - q_{J_n}) \right| \leq \sum_{j=J_n+1}^{\infty} \sum_{\xi \in \Psi_j} |\gamma_\xi \beta_\xi| \lesssim \sum_{j=J_n+1}^{\infty} 2^{dj} \|(\gamma_\xi)_{\xi \in \Psi_j}\|_\infty \|(\beta_\xi)_{\xi \in \Psi_j}\|_\infty. \tag{60}$$

On the other hand, we have $\|\cdot\|_{\mathcal{B}^s_{\infty,\infty}(\Omega)} \lesssim \|\cdot\|_{\mathcal{C}^s(\Omega)}$ for all $s > 0$ by Lemma 25. Therefore, by assumption on q and φ_0^*, we obtain

$$\begin{aligned} \|(\beta_\xi)_{\xi \in \Psi_j}\|_{\ell_\infty} &\leq \|q\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}(\Omega)} 2^{-j[(\alpha-1)+\frac{d}{2}]} \lesssim 2^{-j[(\alpha-1)+\frac{d}{2}]}, \\ \|(\gamma_\xi)_{\xi \in \Psi_j}\|_{\ell_\infty} &\leq \|\psi_0\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}(\Omega)} 2^{-j[(\alpha+1)+\frac{d}{2}]} \lesssim 2^{-j[(\alpha+1)+\frac{d}{2}]}, \end{aligned} \tag{61}$$

for all $j \geq j_0$. Combine equations (60)–(61) to deduce

$$\left|\mathbb{E}L(\widehat{Q}_n)\right| \lesssim \sum_{j=J_n+1}^{\infty} 2^{dj} 2^{-j[(\alpha+1)+\frac{d}{2}]} 2^{-j[(\alpha-1)+\frac{d}{2}]} \lesssim \sum_{j=J_n+1}^{\infty} 2^{-2j\alpha} \lesssim 2^{-2J_n\alpha} \asymp n^{-\frac{2\alpha}{2(\alpha-1)+d}}.$$

We next bound the variance $\mathrm{Var}_Q[L(\widehat{Q}_n)]$. Denote by

$$\psi_{J_n} = \sum_{\zeta\in\Phi} \gamma_\zeta \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \xi\gamma_\xi$$

the projection of ψ_0 onto $\mathrm{Span}\left(\Phi \cup \bigcup_{j=j_0}^{J_n} \Psi_j\right)$. By again applying Lemma 29, it is a straightforward observation that

$$\left|\mathrm{Var}\left[\int \psi_0 \widehat{q}_n\right] - \mathrm{Var}\left[\int \psi_0 \widetilde{q}_n\right]\right| \lesssim n^{-2},$$

thus it suffices to show that $\mathrm{Var}\left[\int \psi_0 \widetilde{q}_n\right] = \mathrm{Var}_Q[\psi_0(Y)]/n + O(2^{-2J_n\alpha}/n)$. Notice that

$$\begin{aligned}
\int \psi_0 \widetilde{q}_n &= \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \int \psi_0\zeta + \sum_{j=j_0-1}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \int \psi_0 \xi \\
&= \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta \gamma_\zeta + \sum_{j=j_0-1}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi \gamma_\xi = \frac{1}{n}\sum_{i=1}^{n}\left[\sum_{\zeta\in\Phi} \zeta(Y_i)\gamma_\zeta + \sum_{j=j_0-1}^{J_n} \sum_{\xi\in\Psi_j} \xi(Y_i)\gamma_\xi\right] = \int \psi_{J_n} dQ_n,
\end{aligned} \tag{62}$$

whence,

$$\begin{aligned}
\mathrm{Var}\left[\int \psi_0 \widetilde{q}_n\right] &= \frac{1}{n}\mathrm{Var}_Q[\psi_{J_n}(Y)] \\
&= \frac{1}{n}\mathrm{Var}_Q[\psi_0(Y)] + \frac{1}{n}(\mathrm{Var}_Q[\psi_{J_n}(Y)] - \mathrm{Var}_Q[\psi_0(Y)]).
\end{aligned}$$

It thus remains to bound the final term. Notice that

$$\begin{aligned}
&\left|\mathrm{Var}_Q[\psi_{J_n}(Y)] - \mathrm{Var}_Q[\psi_0(Y)]\right| \\
&\qquad \leq \left|\mathbb{E}_Q[\psi_{J_n}^2(Y) - \psi_0^2(Y)]\right| + \left|\mathbb{E}_Q[\psi_{J_n}(Y) - \psi_0(Y)]\right| = (I) + (II).
\end{aligned}$$

We begin by bounding (I). Letting $g_n = (\psi_{J_n} + \psi_0)q$, we have,

$$(I) = \left|\int (\psi_{J_n} - \psi_0)(\psi_{J_n} + \psi_0)q\right| = \left|\int (\psi_{J_n} - \psi_0) g_n\right|.$$

It is clear that $\|\psi_{J_n}\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}([0,1]^d)} \leq \|\psi_0\|_{\mathcal{B}^{\alpha+1}_{\infty,\infty}([0,1]^d)} \lesssim \lambda$, thus for any fixed $\epsilon > 0$ sufficiently small, the map $\psi_{J_n} + \psi_0$ lies in $\mathcal{C}^{\alpha+1-\epsilon}([0,1]^d)$ with uniformly bounded norm, by Lemma 25. Note that one may take $\epsilon = 0$ if α is not an integer. On the other hand, we also have $\|q\|_{\mathcal{C}^{\alpha-1}([0,1]^d)} \leq M$.

Deduce that $\sup_{n\geq 1} \|g_n\|_{\mathcal{C}^{\alpha-1}([0,1]^d)} \lesssim 1$, by Lemma 23. By following the same argument as in the first part of this proof, and using again the fact that $\|\psi_0\|_{\mathcal{C}^{\alpha+1}([0,1]^d)} \lesssim \lambda$, we may deduce that

$$(I) \leq \sum_{j=J_n+1}^{\infty} 2^{dj} 2^{-j[(\alpha+1)+\frac{d}{2}]} 2^{-j[(\alpha-1)+\frac{d}{2}]} \lesssim 2^{-2J_n\alpha}.$$

Likewise, we have

$$(II) = \left| \int (\psi_{J_n} - \psi_0) q \right| \lesssim 2^{-2J_n\alpha},$$

and the claim follows from here. □

F Proofs of Two-Sample Stability Bounds

F.1 Proof of Proposition 12

Due to the absolute continuity of P and Q, the optimal transport map from Q to P is given by $S_0 = \nabla\varphi_0^*$. Furthermore, by absolute continuity of P, there exists an optimal transport map $\widehat{\sigma}$ from P to $\widehat{P}$. We clearly have,

$$(\widehat{\sigma} \circ S_0)_{\#} Q = \widehat{P}.$$

Also let $\widehat{\pi} \in \Pi(Q, \widehat{Q})$ be the optimal coupling between Q and $\widehat{Q}$, so that

$$(\widehat{\sigma} \circ S_0, Id)_{\#}\widehat{\pi} \in \Pi(\widehat{P}, \widehat{Q}).$$

We deduce,

$$\begin{aligned} W_2^2(\widehat{P}, \widehat{Q}) &\leq \int \|\widehat{\sigma} \circ S_0(y) - z\|^2 \, d\widehat{\pi}(y,z) \\ &= \int \|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 + \|S_0(y) - z\|^2 \, d\widehat{\pi}(y,z) \\ &+ 2 \int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y) - z \rangle d\widehat{\pi}(y,z). \end{aligned} \tag{63}$$

Notice that

$$\int \|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 \, d\widehat{\pi}(y,z) = \int \|\widehat{\sigma}(x) - x\|^2 dP(x) = W_2^2(\widehat{P}, P). \tag{64}$$

Furthermore, we have

$$\int \|S_0(y) - z\|^2 d\widehat{\pi}(y,z) \leq W_2^2(P,Q) + \int \psi_0 d(\widehat{Q} - Q) + \lambda W_2^2(\widehat{Q}, Q), \tag{65}$$

by Lemma 37. Additionally, the cross term in equation (63) is bounded as follows.

Lemma 39. *We have,*

$$2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y) - z\rangle d\widehat{\pi}(y,z)$$
$$\leq \int \phi_0 d(\widehat{P} - P) + 2W_2(\widehat{P}, P)W_2(\widehat{Q}, Q) + (\lambda - 1)W_2^2(\widehat{P}, P).$$

We prove Lemma 39 in Appendix F.2 below. By equations (63–65) and Lemma 39, we obtain

$$\begin{aligned}
W_2^2(\widehat{P}, \widehat{Q}) &\leq W_2^2(P, Q) + \lambda W_2^2(\widehat{P}, P) + \lambda W_2^2(\widehat{Q}, Q) \\
&\quad + \int \psi_0 d(\widehat{Q} - Q) + \int \phi_0 d(\widehat{P} - P) + 2W_2(\widehat{P}, P)W_2(\widehat{Q}, Q) \\
&\leq W_2^2(P, Q) + \lambda \left[W_2(\widehat{P}, P) + W_2(\widehat{Q}, Q)\right]^2 + \int \psi_0 d(\widehat{Q} - Q) + \int \phi_0 d(\widehat{P} - P).
\end{aligned}$$

This proves the upper bound of the claim. To prove the lower bound, notice that, by the Kantorovich duality,

$$\begin{aligned}
W_2^2(\widehat{P}, \widehat{Q}) &\geq \int \phi_0 d\widehat{P} + \int \psi_0 d\widehat{Q} \\
&= \int \phi_0 dP + \int \psi_0 dQ + \int \phi_0 d(\widehat{P} - P) + \int \psi_0 d(\widehat{Q} - Q) \\
&= W_2^2(P, Q) + \int \phi_0 d(\widehat{P} - P) + \int \psi_0 d(\widehat{Q} - Q).
\end{aligned}$$

The claim follows. □

F.2 Proof of Lemma 39

Write

$$2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y) - z\rangle d\widehat{\pi}(y,z) = (I) + (II) + (III), \tag{66}$$

where

$$\begin{aligned}
(I) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), y - z\rangle d\widehat{\pi}(y,z) \\
(II) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), -y\rangle d\widehat{\pi}(y,z) \\
(III) &= 2\int \langle \widehat{\sigma} \circ S_0(y) - S_0(y), S_0(y)\rangle d\widehat{\pi}(y,z).
\end{aligned}$$

Regarding (I), the Cauchy-Schwarz inequality implies

$$\begin{aligned}
(I) &\leq 2\left(\int \|\widehat{\sigma} \circ S_0(y) - S_0(y)\|^2 \, d\widehat{\pi}(y,z)\right)^{\frac{1}{2}} \left(\int \|y - z\|^2 \, d\widehat{\pi}(y,z)\right)^{\frac{1}{2}} \\
&= 2\left(\int \|\widehat{\sigma}(x) - x\|^2 \, dP(x)\right)^{\frac{1}{2}} \left(\int \|y - z\|^2 \, d\widehat{\pi}(y,z)\right)^{\frac{1}{2}}
\end{aligned}$$

$$= 2W_2(\widehat{P}, P)W_2(\widehat{Q}, Q). \tag{67}$$

Regarding term (II), recall that φ_0 satisfies assumption **A1(λ)**, thus we have

$$\frac{1}{2\lambda}\|x-y\|^2 \le \varphi_0(y) - \varphi_0(x) - \langle T_0(x), y-x\rangle \le \frac{\lambda}{2}\|x-y\|^2, \quad x, y \in \Omega,$$

We deduce that,

$$\begin{aligned}
(II) &= 2\int \langle \widehat{\sigma}\circ S_0(y) - S_0(y), -T_0 \circ S_0(y)\rangle d\widehat{\pi}(y,z) \\
&\le 2\int \left[\varphi_0(S_0(y)) - \varphi_0(\widehat{\sigma}\circ S_0(y)) + \frac{\lambda}{2}\|S_0(y) - \widehat{\sigma}\circ S_0(y)\|^2\right] d\widehat{\pi}(y,z) \\
&= \int 2\varphi_0 d(P - \widehat{P}) + \lambda W_2^2(\widehat{P}, P). \qquad (68)
\end{aligned}$$

Finally, term (III) satisfies

$$\begin{aligned}
(III) &= \int \left[\|\widehat{\sigma}\circ S_0(y)\|^2 - \|\widehat{\sigma}\circ S_0(y) - S_0(y)\|^2 - \|S_0(y)\|^2\right] d\widehat{\pi}(y,z) \\
&= \int \|\cdot\|^2 d(\widehat{P} - P) - W_2^2(\widehat{P}, P). \qquad (69)
\end{aligned}$$

Combine equations (67)–(69) with equation (66) to deduce the claim. □

F.3 Proof of Proposition 13

Once again, denote by $S_0 = \nabla\varphi_0^*$ the optimal transport map from Q to P. Recall from the proof of Theorem 6 (equation (56)) that, due to assumption **A1(λ)**,

$$\frac{1}{2\lambda}\|x-y\|^2 \le \varphi_0^*(y) - \varphi_0^*(x) - \langle S_0(x), y-x\rangle \le \frac{\lambda}{2}\|x-y\|^2,$$

for all $x, y \in \Omega$. Now, we have,

$$\begin{aligned}
W_2^2(P_n, Q_m) &= \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\|X_i - Y_j\|^2 \\
&= \sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\Big[\|T_0(X_i) - X_i\|^2 \\
&\qquad + 2\langle T_0(X_i) - X_i, Y_j - T_0(X_i)\rangle + \|Y_j - T_0(X_i)\|^2\Big].
\end{aligned}$$

Notice that

$$\mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^m \widehat{\pi}_{ij}\|T_0(X_i) - X_i\|^2\right] = \mathbb{E}\left[\sum_{i=1}^n \left(\sum_{j=1}^m \widehat{\pi}_{ij}\right)\|T_0(X_i) - X_i\|^2\right]$$

$$= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|T_0(X_i) - X_i\|^2\right] = W_2^2(P,Q),$$

where we have used the marginal constraint on the coupling $\widehat{\pi}$ in the first equality of the above display. Recalling that $\Delta_{nm} = \sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{\pi}_{ij}\|X_i - Y_j\|^2$, thus we obtain,

$$\begin{aligned}
&\mathbb{E}\Big[W_2^2(P_n,Q_m) - W_2^2(P,Q)\Big] \\
&\qquad = \mathbb{E}[\Delta_{nm}] + 2\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{\pi}_{ij}\langle T_0(X_i) - X_i, Y_j - T_0(X_i)\rangle\right] \\
&\qquad = \mathbb{E}[\Delta_{nm}] + 2\mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{\pi}_{ij}\langle T_0(X_i) - S_0(T_0(X_i)), Y_j - T_0(X_i)\rangle\right].
\end{aligned}$$

Now,

$$2\langle -S_0(T_0(X_i)), Y_j - T_0(X_i)\rangle \geq 2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) + \frac{1}{\lambda}\|T_0(X_i) - Y_j\|^2, \qquad (70)$$

whence, we obtain,

$$\begin{aligned}
\mathbb{E}\Big[W_2^2(P_n,Q_m) - W_2^2(P,Q)\Big] \geq \mathbb{E}[\Delta_{nm}] + \mathbb{E}\Bigg[\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{\pi}_{ij}\Big(&2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) \\
&+ \frac{1}{\lambda}\|T_0(X_i) - Y_j\|^2 + 2\langle T_0(X_i), Y_j - T_0(X_i)\rangle\Big)\Bigg]
\end{aligned}$$

Now, notice that

$$2\langle T_0(X_i), Y_j - T_0(X_i)\rangle = -\|T_0(X_i) - Y_j\|^2 + \|Y_j\|^2 - \|T_0(X_i)\|^2.$$

Thus, continuing from before, we have

$$\begin{aligned}
&\mathbb{E}\Big[W_2^2(P_n,Q_m) - W_2^2(P,Q)\Big] \\
&\qquad \geq \frac{1}{\lambda}\mathbb{E}[\Delta_{nm}] + \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\widehat{\pi}_{ij}\Big(2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) + \|Y_j\|^2 - \|T_0(X_i)\|^2\Big)\right] \\
&\qquad = \frac{1}{\lambda}\mathbb{E}[\Delta_{nm}] + \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}\Big(\|Y_j\|^2 - 2\varphi_0^*(Y_j)\Big)\right] \\
&\qquad\qquad - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\Big(\|T_0(X_i)\|^2 - 2\varphi_0^*(T_0(X_i))\Big)\right] = \frac{1}{\lambda}\mathbb{E}[\Delta_{nm}].
\end{aligned}$$

This proves one of the inequalities of the claim. To obtain the other, return to equation (70) and notice that one also has

$$2\langle -S_0(T_0(X_i)), Y_j - T_0(X_i)\rangle \le 2\varphi_0^*(T_0(X_i)) - 2\varphi_0^*(Y_j) + \lambda \|T_0(X_i) - Y_j\|^2 .$$

The proof then proceeds analogously. This proves that

$$\mathbb{E}[\Delta_{nm}] \asymp_\lambda \mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q)\Big].$$

To conclude, apply Proposition 12 to deduce

$$\begin{aligned}
&\mathbb{E}\Big[W_2^2(P_n, Q_m) - W_2^2(P, Q)\Big] \\
&\qquad \le \mathbb{E}\int \phi_0 d(P_n - P) + \mathbb{E}\int \psi_0 d(Q_m - Q) + 2\lambda\Big[\mathbb{E}W_2^2(P_n, P) + \mathbb{E}W_2^2(Q_m, Q)\Big] \\
&\qquad = 2\lambda\Big[\mathbb{E}W_2^2(P_n, P) + \mathbb{E}W_2^2(Q_m, Q)\Big].
\end{aligned}$$

The above display is of the order $\kappa_{n\wedge m}$ due to equation (21). When we additionally assume that $\Omega = [0,1]^d$ and $\gamma^{-1} \le p, q \le \gamma$, we may instead bound it from above by $\bar{\kappa}_{n\wedge m}$, due to Corollary 8. The claim follows. □

F.4 Proof of Proposition 16

The claim follows along the same lines as the proofs of Theorem 6 and Proposition 12, thus we only provide a brief proof of the analogue of Theorem 6 over the torus. It will suffice to prove

$$\frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 \le W_2^2(P, \widehat{Q}) - W_2^2(P, Q) - \int \psi_0 d(\widehat{Q} - Q) \le \lambda W_2^2(\widehat{Q}, Q). \tag{71}$$

Recall that $\widehat{T}$ is the optimal transport map from P to $\widehat{Q}$. By Proposition 4(iii), we therefore have

$$d_{\mathbb{T}^d}(\widehat{T}(x), x) = \|\widehat{T}(x) - x\|, \quad d_{\mathbb{T}^d}(T_0(x), x) = \|T_0(x) - x\|, \quad x \in \mathbb{T}^d.$$

It follows that

$$W_2^2(P, \widehat{Q}) - W_2^2(P, Q) = \int \|\widehat{T}(x) - x\|^2 dP(x) - \int \|T_0(x) - x\|^2 dP(x).$$

From here, it follows identically as in the proof of Theorem 6 that

$$W_2^2(P, \widehat{Q}) - W_2^2(P, Q) \ge \frac{1}{\lambda}\|\widehat{T} - T_0\|_{L^2(P)}^2 + \int \psi_0 d(\widehat{Q} - Q).$$

To prove the second inequality in equation (71), let $\widehat{\pi}$ denote an optimal coupling between Q and $\widehat{Q}$ with respect to the cost $d_{\mathbb{T}^d}^2$. Notice similarly as before that Proposition 4(iii) implies

$$W_2^2(P, Q) = \int \|S_0(y) - y\|^2 dQ(y), \quad W_2^2(Q, \widehat{Q}) = \int \|y - z\|^2 d\widehat{\pi}(y, z),$$

thus, since $(S_0, Id)_{\#}\widehat{\pi} \in \Pi(P, \widehat{Q})$, and using the fact that $d_{\mathbb{T}^d} \leq \|\cdot\|$, we have

$$
\begin{aligned}
&W_2^2(P, \widehat{Q}) \\
&\quad \leq \int d^2_{\mathbb{T}^d}(S_0(y), z) d\widehat{\pi}(y,z) \\
&\quad \leq \int \|S_0(y) - z\|^2 d\widehat{\pi}(y,z) \\
&\quad = \int \|S_0(y) - y\|^2 \, dQ(y) + \int \|y - z\|^2 \, d\widehat{\pi}(y,z) + 2\int \big\langle S_0(y) - y, y - z \big\rangle d\widehat{\pi}(y,z) \\
&\quad = W_2^2(P, Q) + W_2^2(\widehat{Q}, Q) + 2\int \big\langle S_0(y) - y, y - z \big\rangle d\widehat{\pi}(y,z).
\end{aligned}
$$

By the same argument as in Theorem 6, the cross term is bounded above by $(\lambda - 1)W_2^2(\widehat{Q}, Q) + \int \psi_0 d(\widehat{Q} - Q)$, thus the claim follows. □

G Proofs of Upper Bounds for Two-Sample Empirical Estimators

In this Appendix, we prove Propositions 14 and 15. We begin with the following result.

Lemma 40. *Let Ω satisfy conditions (S1)–(S2). Let $P \in \mathcal{P}_{\mathrm{ac}}(\Omega)$ admit a density p such that $\gamma^{-1} \leq p \leq \gamma$ for some $\gamma > 0$. Let $V_1, \dots, V_n$ denote the Voronoi partition in equation (31), based on an i.i.d. sample $X_1, \dots, X_n \sim P$. Then, there exist constants $C_1, C_2 > 0$ depending only on $d, \gamma, \epsilon_0, \delta_0$ such that,*

(i) For all $\delta \in (0,1)$, we have,

$$
\mathbb{P}\left(\max_{1 \leq i \leq n} P(V_i) \geq \frac{C_1}{n}\Big[d \log n + \log(1/\delta) \Big] \right) \leq \delta.
$$

(ii) We have,

$$
\mathbb{E}\left[\max_{1 \leq i \leq n} \mathrm{diam}(V_i)^2 \right] \leq C_2 \left(\frac{\log n}{n} \right)^{\frac{2}{d}}.
$$

Proof of Lemma 40. We shall make use of the relative Vapnik-Chervonenkis inequality (Vapnik, 2013; Bousquet et al., 2003), in the following form stated by Chaudhuri and Dasgupta (2010).

Lemma 41. *Let $\mathcal{B}$ denote the set of balls in $\mathbb{R}^d$. Then, there exists a universal constant $C > 0$ such that for every $\delta \in (0,1)$, we have with probability at least $1 - \delta$ that for all $B \in \mathcal{B}$,*

$$
P(B) \geq \frac{C}{n}\left[d \log n + \log\left(\frac{1}{\delta}\right) \right] \implies P_n(B) > 0.
$$

We now turn to the proof. Recall that Ω is a standard set by condition (S2), and recall the constants $\epsilon_0, \delta_0 > 0$ therein. Fix $1 \leq i \leq n$. Pick $x_i \in V_i$ for all $1 \leq i \leq n$, and let $\rho_i = (\epsilon_0/2d)\|x_i - X_i\|$. Since $\mathrm{diam}(\Omega) \leq \sqrt{d}$ by condition (S1), we have $\rho_i \leq \epsilon_0$. We also have

$\rho_i < \|x_i - X_i\|$, thus the balls $B(x_i, \rho_i)$ of radius ρ_i centered at x_i contain no sample points. Therefore, by Lemma 41, we have that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\max_{1\le i\le n} P(B(x_i,\rho_i)) \le \frac{C}{n}\left[d\log n + \log\left(\frac{1}{\delta}\right)\right]. \tag{72}$$

Now, since $\gamma^{-1} \le p \le \gamma$, the assumption of standardness on Ω leads to the bound

$$P(B(x_i,\rho_i)) \ge \gamma^{-1}\mathcal{L}(B(x_i,\rho_i)\cap\Omega) \ge \delta_0\gamma^{-1}\mathcal{L}(B(x_i,\rho_i)) \asymp \rho_i^d,$$

thus equation (72) reduces to

$$\max_{1\le i\le n} \rho_i^d \le \frac{C}{n}\left[d\log n + \log\left(\frac{1}{\delta}\right)\right].$$

By definition of ρ_i, we deduce that for some constant $C_1 > 0$ not depending on δ, we have with probability at least $1-\delta$,

$$\max_{1\le i\le n} \mathrm{diam}(V_i) \le C_1\left[\frac{d\log n + \log(1/\delta)}{n}\right]^{\frac{1}{d}}.$$

To prove claim (i), notice that since the density of P is bounded from above, we also have with probability at least $1-\delta$,

$$\max_{1\le i\le n} P(V_i) \le \gamma \max_{1\le i\le n}\mathcal{L}(V_i) \lesssim \max_{1\le i\le n}\mathrm{diam}(V_i)^d \lesssim \frac{1}{n}\Big[d\log n + \log(1/\delta)\Big].$$

To prove claim (ii), let $t_n = 2C_1^d((d+2)\log n/n)^{2/d}$. Set $\delta = n^d\exp\left(-\frac{u^d n}{C_1^d}\right)$ for any $u>0$ to obtain

$$\begin{aligned}
\mathbb{E}\left[\max_{1\le i\le n}\mathrm{diam}(V_i)^2\right] &= \int_0^\infty \mathbb{P}\left(\max_{1\le i\le n}\mathrm{diam}(V_i)^2 \ge u\right)du \\
&\le t_n + n^d\int_{t_n}^\infty \exp\left(-\frac{u^{\frac{d}{2}}n}{C_1^d}\right)du \\
&= t_n + \frac{4n^d}{d}\int_{t_n^{d/4}}^\infty \exp\left(-\frac{v^2 n}{C_1^d}\right)v^{\frac{4}{d}-1}dv \\
&\lesssim t_n + n^d\int_{t_n^{d/4}}^\infty \exp\left(-\frac{v^2 n}{2C_1^d}\right)dv \\
&\lesssim t_n + \frac{n^d}{\sqrt{n}}\exp\left(-\frac{t_n^{d/2}n}{2C_1^d}\right) \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}}.
\end{aligned}$$

The claim follows. ☐

G.1 Proof of Proposition 14

Abbreviate $\widehat{T}_{nm}^{\mathrm{1NN}}$ by $\widehat{T}_{nm}$. We have,

$$\begin{aligned}\left\|\widehat{T}_{nm} - T_0\right\|_{L^2(P)}^2 &= \sum_{i=1}^n \int_{V_i} \left\|\widehat{T}_{nm}(x) - T_0(X_i) + T_0(X_i) - T_0(x)\right\|^2 dP(x) \\ &\lesssim \sum_{i=1}^n \int_{V_i} \left[\|\widehat{T}_{nm}(x) - T_0(X_i)\|^2 + \|T_0(X_i) - T_0(x)\|^2\right] dP(x).\end{aligned}$$

To bound the first term, notice that,

$$\begin{aligned}\sum_{i=1}^n \int_{V_i} \|\widehat{T}_{nm}(x) - T_0(X_i)\|^2 dP(x) &= \sum_{i=1}^n \int_{V_i} \left\| \sum_{j=1}^m (n\widehat{\pi}_{ij}) Y_j - T_0(X_i)\right\|^2 dP(x) \\ &= \sum_{i=1}^n P(V_i) \left\| \sum_{j=1}^m (n\widehat{\pi}_{ij}) Y_j - T_0(X_i)\right\|^2 \\ &\leq \sum_{i=1}^n P(V_i) \sum_{j=1}^m (n\widehat{\pi}_{ij}) \|Y_j - T_0(X_i)\|^2,\end{aligned}$$

by convexity of $\|\cdot\|^2$. Therefore, setting $M_n = \max_{1\leq i\leq n} P(V_i)$, we obtain

$$\left\|\widehat{T}_{nm} - T_0\right\|_{L^2(P)}^2 \leq n\Delta_{nm}\left(\max_{1\leq i\leq n} P(V_i)\right) + \sum_{i=1}^n \int_{V_i} \|T_0(X_i) - T_0(x)\|^2 dP(x).$$

Since T_0 is λ-Lipschitz by condition **A1(λ)**, the claim is now a consequence of the following simple Lemma, which we isolate for future reference.

Lemma 42. *Under the conditions of Proposition 14, we have for any λ-Lipschitz map $F : \Omega \to \Omega$,*

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|F(X_i) - F(x)\|^2 dP(x)\right] &\lesssim_{\lambda,\gamma} (\log n/n)^{2/d}, \\ \mathbb{E}\left[n\Delta_{nm}\left(\max_{1\leq i\leq n} P(V_i)\right)\right] &\lesssim_{\lambda,\gamma,\epsilon_0,\delta_0} (\log n)\kappa_{n\wedge m}.\end{aligned}$$

G.1.1 Proof of Lemma 42

The first quantity is easily bounded as follows,

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|F(X_i) - F(x)\|^2 dP(x)\right] &\leq \lambda^2 \mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|X_i - x\|^2 dP(x)\right] \\ &\leq \lambda^2 \mathbb{E}\left[\sum_{i=1}^n P(V_i)\,\mathrm{diam}(V_i)^2\right]\end{aligned}$$

$$\leq \lambda^2 \mathbb{E}\left[\max_{1\leq i\leq n} \mathrm{diam}(V_i)^2\right] \lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}},$$

where the final inequality is due to Lemma 40(ii). To bound the second quantity, let $M_n = \max_{1\leq i\leq n} P(V_i)$. By Lemma 40(i) with $\delta = 1/n^2$, there is a large enough constant $c > 0$ such that if $m_n = c\log n/n$, then $\mathbb{P}(M_n \geq m_n) \leq 1/n^2$. We have,

$$\mathbb{E}\left[nM_n\Delta_{nm}\right] = \mathbb{E}\left[nM_n I(M_n \geq m_n)\Delta_{nm}\right] + \mathbb{E}\left[nM_n I(M_n < m_n)\Delta_{nm}\right].$$

Notice that Δ_{nm} is bounded above by $\mathrm{diam}(\Omega)^2$, and $0 \leq M_n \leq 1$, thus, by Proposition 13,

$$\mathbb{E}\left[nM_n\Delta_{nm}\right] \lesssim n\mathbb{P}(M_n \geq m_n) + m_n n\mathbb{E}\left[\Delta_{nm}\right] \lesssim \frac{1}{n} + (\log n)\mathbb{E}\left[\Delta_{nm}\right] \lesssim (\log n)\kappa_{n\wedge m}.$$

This proves the claim. □

G.2 Proof of Proposition 15

Abbreviate $\widehat{T}^{\mathrm{LS}}_{nm}$ by $\widehat{T}_{nm}$. Notice first that we have

$$\begin{aligned}
\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P_n)} &= \frac{1}{n}\sum_{i=1}^n \big\|\widehat{T}_{nm}(X_i) - T_0(X_i)\big\|^2 \\
&= \sum_{i=1}^n\sum_{j=1}^m \widehat{\pi}_{ij}\big\|\widehat{T}_{nm}(X_i) - T_0(X_i)\big\|^2 \\
&\lesssim \sum_{i=1}^n\sum_{j=1}^m \widehat{\pi}_{ij}\big\|\widehat{T}_{nm}(X_i) - Y_j\big\|^2 + \sum_{i=1}^n\sum_{j=1}^m \widehat{\pi}_{ij}\big\|Y_j - T_0(X_i)\big\|^2 \\
&\leq 2\sum_{i=1}^n\sum_{j=1}^m \widehat{\pi}_{ij}\big\|Y_j - T_0(X_i)\big\|^2 = 2\Delta_{nm}, \qquad (73)
\end{aligned}$$

where the final inequality follows by definition of $\widehat{T}_{nm}$, since $\varphi_0 \in \mathcal{J}_\lambda$ under assumption **A1(λ)**. Therefore,

$$\begin{aligned}
\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P)} = \sum_{i=1}^n \int_{V_i} \big\|\widehat{T}_{nm} - T_0\big\|^2 dP \lesssim \sum_{i=1}^n \int_{V_i} \Big[&\|\widehat{T}_{nm}(x) - \widehat{T}_{nm}(X_i)\|^2 \\
&+ \|\widehat{T}_{nm}(X_i) - T_0(X_i)\|^2 + \|T_0(X_i) - T_0(x)\|^2\Big] dP(x).
\end{aligned}$$

By definition of $\mathcal{J}_\lambda$ and by assumption **A1(λ)**, $\widehat{T}_{nm}$ and T_0 are both λ-Lipschitz, thus by Lemma 42,

$$\begin{aligned}
\mathbb{E}\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P)} &\lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}} + \mathbb{E}\left[\sum_{i=1}^n \int_{V_i} \|\widehat{T}_{nm}(X_i) - T_0(X_i)\|^2 dP(x)\right] \\
&\leq \left(\frac{\log n}{n}\right)^{\frac{2}{d}} + \mathbb{E}\left[n\left(\max_{1\leq i\leq n} P(V_i)\right)\big\|\widehat{T}_{nm} - T_0\big\|^2_{L^2(P_n)}\right]
\end{aligned}$$

$$\lesssim \left(\frac{\log n}{n}\right)^{\frac{2}{d}} + \mathbb{E}\left[n\left(\max_{1\leq i\leq n} P(V_i)\right)\Delta_{nm}\right],$$

where we used equation (73). Lemma 42 may now be applied to bound the right-hand term in the above display, leading to the claim. □

H Proofs of Upper Bounds for Two-Sample Wavelet Estimators

The aim of this section is to prove Theorem 17. We first note that the one-sample results from Section 3.3 may readily be extended to the optimal transport problem over $\mathbb{T}^d$.

Proposition 43. *Assume $P, Q \in \mathcal{P}_{\text{ac}}(\mathbb{T}^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$ for some $\alpha > 1$, $\alpha \notin \mathbb{N}$, and $M, \gamma > 0$. Let $\widehat{q}_m = \widehat{q}_m^{(\text{per})}$ be the periodic wavelet estimator defined in equation (34), and let $\widehat{Q}_m$ be the induced probability distribution. Let*

$$\overline{T}_m = \operatorname*{argmin}_{T\in\mathcal{T}(P,\widehat{Q}_m)} \int d^2_{\mathbb{T}^d}(x, T(x))dP(x).$$

Furthermore, let $2^{J_m} \asymp m^{\frac{1}{2(\alpha-1)+d}}$. Then, there exists a constant $C > 0$ depending only on M, γ, α such that the following statements hold.

(i) We have $\mathbb{E}W_2^2(\widehat{Q}_m, Q) \leq CR_{T,m}(\alpha)$ and $\mathbb{E}W_2^4(\widehat{Q}_m, Q) \leq CR^2_{T,m}(\alpha)$.

(ii) We have,

$$\begin{aligned}
\left|\mathbb{E}\int \psi_0 d(\widehat{Q}_m - Q)\right| &\leq C2^{-2J_m\alpha} \\
\left|\text{Var}\left[\int \psi_0 d(\widehat{Q}_m - Q)\right] - \frac{\text{Var}_Q[\psi_0(Y)]}{m}\right| &\leq \frac{C2^{-2J_m\alpha}}{m}.
\end{aligned}$$

(iii) We have,

$$\begin{aligned}
\mathbb{E}\|\overline{T}_m - T_0\|^2_{L^2(P)} &\leq CR_{T,m}(\alpha), \\
\left|\mathbb{E}W_2^2(P, \widehat{Q}_m) - W_2^2(P, Q)\right| &\leq CR_{T,m}(\alpha), \\
\mathbb{E}\left|W_2^2(P, \widehat{Q}_m) - W_2^2(P, Q)\right|^2 &\leq \left[CR^2_{T,m}(\alpha) + \sqrt{\frac{\text{Var}_Q[\psi_0(Y)]}{m}}\right]^2.
\end{aligned}$$

Notice that the only properties of the boundary-correct wavelet basis used in the proofs of Lemma 30 and Theorem 10 are those contained in Lemmas 24 and Lemma 28 of Appendix A.2, which are also stated to hold for the periodic wavelet basis. The proof of Proposition 43 is therefore a direct extension of these results. Notice that, unlike Theorem 10, we no longer require any conditions on the smoothnes of φ_0 itself, due to the torus regularity result of Theorem 5. Indeed, under the assumptions of Proposition 43, the latter implies that there exists a constant $C' > 0$ depending only on α, γ, M such that $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq C'$, assuming $\alpha \notin \mathbb{N}$.

H.1 Proof of Theorem 17

Throughout the proof, we use the abbreviations

$$F(\widehat{P}_n) = \int \phi_0 d(\widehat{P}_n - P), \quad L(\widehat{Q}_m) = \int \psi_0 d(\widehat{Q}_m - Q).$$

We begin by proving part (ii). Under the assumptions of this case, Theorem 5 implies that $\|\varphi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq M_0$ for a universal constant $M_0 > 0$ depending only on α, γ and M. In particular, it also follows from Proposition 4(vii) that φ_0 is strongly convex, and thus satisfies condition **A1(λ)** for some $\lambda > 0$ depending only on M_0 and γ. We may therefore invoke the two-sample stability bound over $\mathbb{T}^d$ in Proposition 13 (arising from Proposition 12) to deduce

$$\begin{aligned} F(\widehat{P}_n) + L(\widehat{Q}_m) &\leq W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P, Q) \\ &\leq F(\widehat{P}_n) + L(\widehat{Q}_m) + 2\lambda \left[W_2^2(\widehat{P}_n, P) + W_2^2(\widehat{Q}_m, Q) \right]. \end{aligned}$$

From Proposition 43(ii), it can be deduced that

$$\left|\mathbb{E}F(\widehat{P}_n)\right| \vee \left|\mathbb{E}L(\widehat{Q}_m)\right| \lesssim R_{T,n\wedge m}(\alpha) \tag{74}$$

$$\operatorname{Var}\left[F(\widehat{P}_n)\right] \leq \frac{\operatorname{Var}_P[\phi_0(X)]}{n} + CR_{T,n}^2(\alpha) \tag{75}$$

$$\operatorname{Var}\left[L(\widehat{Q}_m)\right] \leq \frac{\operatorname{Var}_Q[\psi_0(Y)]}{m} + CR_{T,m}^2(\alpha), \tag{76}$$

for a constant $C > 0$ depending only on M, γ, α, whose value we allow to change from line to line in the remainder of the proof. Thus, recalling Proposition 43(i),

$$\begin{aligned} &\left|\mathbb{E}W_2^2(\widehat{P}_n,\widehat{Q}_m) - W_2^2(P, Q)\right| \\ &\quad\lesssim \left|\mathbb{E}F(\widehat{P}_n)\right| + \left|\mathbb{E}L(\widehat{Q}_m)\right| + \mathbb{E}W_2^2(\widehat{P}_n, P) + \mathbb{E}W_2^2(\widehat{Q}_m, Q) \lesssim R_{T,n\wedge m}(\alpha). \end{aligned}$$

Furthermore,

$$\begin{aligned} &\mathbb{E}\left|W_2^2(\widehat{P}_n, \widehat{Q}_m) - W_2^2(P, Q)\right|^2 \\ &\quad\leq \mathbb{E}\left[\left(\left|F(\widehat{P}_n)\right| + \left|L(\widehat{Q}_m)\right| + 2\lambda\big(W_2^2(\widehat{P}_n, P) + W_2^2(\widehat{Q}_m, Q)\big)\right)^2\right] =: (I) + (II) + (III), \end{aligned}$$

where

$$\begin{aligned} (I) &= \mathbb{E}\left[\left(\left|F(\widehat{P}_n)\right| + \left|L(\widehat{Q}_m)\right|\right)^2\right] \\ (II) &= 4\lambda^2\mathbb{E}\left[\left(W_2^2(\widehat{P}_n, P) + W_2^2(\widehat{Q}_m, Q)\right)^2\right] \\ (III) &= 4\lambda\mathbb{E}\left[\left(W_2^2(\widehat{P}_n, P) + W_2^2(\widehat{Q}_m, Q)\right)\left(\left|F(\widehat{P}_n)\right| + \left|L(\widehat{Q}_m)\right|\right)\right]. \end{aligned}$$

Regarding term (I), recall that we have assumed that X_i is independent of Y_j for all $i, j = 1, \dots, n$. Therefore, using equations (74–76),

$$\begin{aligned}
(I) &= \mathbb{E}\big[F^2(\widehat{P}_n)\big] + \mathbb{E}\big[L^2(\widehat{Q}_m)\big] + 2\mathbb{E}\big|F(\widehat{P}_n)L(\widehat{Q}_m)\big| \\
&= \mathbb{E}\big[F^2(\widehat{P}_n)\big] + \mathbb{E}\big[L^2(\widehat{Q}_m)\big] + 2\mathbb{E}\big|F(\widehat{P}_n)\big|\mathbb{E}\big|L(\widehat{Q}_m)\big| \\
&= \mathrm{Var}\left[F(\widehat{P}_n)\right] + \mathrm{Var}\left[L(\widehat{Q}_m)\right] + \left|\mathbb{E}F(\widehat{P}_n)\right|^2 + \left|\mathbb{E}L(\widehat{Q}_m)\right|^2 + 2\mathbb{E}\big|F(\widehat{P}_n)\big|\mathbb{E}\big|L(\widehat{Q}_m)\big| \\
&\leq \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m} + CR^2_{T,n\wedge m}(\alpha).
\end{aligned}$$

Furthermore, by Proposition 43(i) we have

$$(II) \leq 8\lambda^2 \left(\mathbb{E}W_2^4(\widehat{P}_n, P) + \mathbb{E}W_2^4(\widehat{Q}_m, Q)\right) \leq CR^2_{T,n\wedge m}(\alpha),$$

and, using the Cauchy-Schwarz inequality and equations (74–76), we obtain

$$\begin{aligned}
(III) &\leq C\sqrt{\left(\mathbb{E}W_2^4(\widehat{P}_n, P) + \mathbb{E}W_2^4(\widehat{Q}_m, Q)\right)\left(\mathbb{E}\big|F(\widehat{P}_n)\big|^2 + \mathbb{E}\big|L(\widehat{Q}_m)\big|^2\right)} \\
&\leq C\sqrt{R^2_{T,n\wedge m}(\alpha)\left(CR^2_{T,n\wedge m}(\alpha) + \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}\right)} \\
&\leq CR^2_{T,n\wedge m}(\alpha) + CR_{T,n\wedge m}(\alpha)\sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}.
\end{aligned}$$

Deduce that

$$(I) + (II) + (III) \leq \left(CR_{T,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right)^2.$$

Claim (ii) follows from here.

To prove part (i), we shall make use of the one-sample optimal transport problem from P to $\widehat{Q}_m$. Denote by $\bar{\varphi}_m$ an optimal Brenier potential for this problem, so that $\overline{T}_m = \nabla\bar{\varphi}_m$ is the optimal transport map pushing P forward onto $\widehat{Q}_m$, with respect to the cost function $d^2_{\mathbb{T}^d}$. Furthermore, denote by

$$\bar{\phi}_m = \|\cdot\|^2 - 2\bar{\varphi}_m, \quad \bar{\psi}_m = \|\cdot\|^2 - 2\bar{\varphi}^*_m,$$

a corresponding pair of optimal Kantorovich potentials. We proceed with three steps.

Step 1: Regularity of the Fitted Potentials. Recall that $\alpha > 1$, and fix $\epsilon \in (0, 1 \wedge \frac{\alpha-1}{2})$. By Lemma 28(i) and Lemma 29, under our choice of threshold J_n, notice that the event

$$\begin{aligned}
E_{nm} = \{\widetilde{p}_n = \widehat{p}_n\} \cap \{\widetilde{q}_m = \widehat{q}_m\} \\
\cap \left\{\|\widetilde{p}_n\|_{\mathcal{B}^\epsilon_{\infty,\infty}(\mathbb{T}^d)} \leq 2\|p\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}(\mathbb{T}^d)}\right\} \cap \left\{\|\widetilde{q}_m\|_{\mathcal{B}^\epsilon_{\infty,\infty}(\mathbb{T}^d)} \leq 2\|q\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}(\mathbb{T}^d)}\right\}
\end{aligned}$$

satisfies $\mathbb{P}(E^{\mathsf{c}}_{nm}) \lesssim (n \wedge m)^{-2}$. Note that $\epsilon \notin \mathbb{N}$, thus by Lemma 25, we have on the event E_{nm},

$$\|\widehat{q}_m\|_{\mathcal{C}^\epsilon(\mathbb{T}^d)} \lesssim \|\widehat{q}_m\|_{\mathcal{B}^\epsilon_{\infty,\infty}(\mathbb{T}^d)} \lesssim \|q\|_{\mathcal{B}^{\alpha-1}_{\infty,\infty}(\mathbb{T}^d)} \lesssim \|q\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)} \leq M,$$

and similarly for $\widehat{p}_n$. Thus, there exists $M_1 > 0$ depending only on M, γ such that

$$\|\widehat{p}_n\|_{\mathcal{C}^\epsilon(\mathbb{T}^d)}, \|\widehat{q}_m\|_{\mathcal{C}^\epsilon(\mathbb{T}^d)} \leq M_1, \quad \text{on } E_{nm}.$$

Furthermore, by Lemma 28(ii), there exists $\gamma_0 > 0$ depending only on γ such that

$$\gamma_0^{-1} \leq \widehat{p}_n, \widehat{q}_m \leq \gamma_0, \quad \text{on } E_{nm}.$$

Under the preceding two displays, together with the smoothness assumptions on the population densities p, q themselves, we may apply the regularity Theorem 5 to deduce that there exists a constant $M_2 > 0$ depending only on M_0, M_1, γ such that

$$\|\varphi_0\|_{\mathcal{C}^{2+\epsilon}(\mathbb{R}^d)} \vee \|\widehat{\varphi}_{nm}\|_{\mathcal{C}^{2+\epsilon}(\mathbb{R}^d)} \vee \|\overline{\varphi}_m\|_{\mathcal{C}^{2+\epsilon}(\mathbb{R}^d)} \leq M_2, \tag{77}$$

on E_{nm}. Apply Proposition 4(vii) to deduce that $\widehat{\varphi}_{nm}$ and $\overline{\varphi}_m$ satisfy the curvature condition **A1(λ)** almost surely, up to modifying the value of $\lambda > 0$ in terms of M_2 and γ, namely:

$$\lambda^{-1} I_d \preceq \nabla^2\varphi_0(x), \nabla^2\overline{\varphi}_m(x), \nabla^2\widehat{\varphi}_{nm}(x) \preceq \lambda I_d, \quad \text{for all } x \in \mathbb{R}^d;\ n, m \geq 1, \tag{78}$$

on the event E_{nm}.

Step 2: Reduction to Optimal Transport Problems with Same Source Distribution. In order to appeal to the one-sample stability bounds of Theorem 6, write

$$\|\widehat{T}_{nm} - T_0\|^2_{L^2(P)} \lesssim \|\widehat{T}_{nm} - \overline{T}_m\|^2_{L^2(P)} + \|\overline{T}_m - T_0\|^2_{L^2(P)}. \tag{79}$$

The first term in the above display compares transport maps which are optimal for distinct source distributions. We therefore proceed with the following reduction, over the event E_{nm}:

$$\begin{aligned}
\|\widehat{T}_{nm} - \overline{T}_m\|^2_{L^2(P)} &= \int_{\mathbb{T}^d} \|\widehat{T}_{nm}(x) - \overline{T}_m(x)\|^2 dP(x) \\
&= \int_{\mathbb{T}^d} \|\widehat{T}_{nm}(\overline{T}_m^{-1}(y)) - y\|^2 d\widehat{Q}_m(y) \\
&= \int_{\mathbb{T}^d} \|\widehat{T}_{nm}(\overline{T}_m^{-1}(y)) - \widehat{T}_{nm}(\widehat{T}_{nm}^{-1}(y))\|^2 d\widehat{Q}_m(y),
\end{aligned} \tag{80}$$

where the second line follows from the fact that $(\overline{T}_m)_\# P = \widehat{Q}_m$, and the third follows from invertibility of $\widehat{T}_{nm}$, which is ensured by the strong convexity of $\widehat{\varphi}_{nm}$ in equation (78). This same equation implies that, over the event E_{nm}, $\widehat{T}_{nm} = \nabla\widehat{\varphi}_{nm}$ is Lipschitz with a universal constant. It follows that

$$\|\widehat{T}_{nm} - \overline{T}_m\|^2_{L^2(P)} \lesssim \int_{\mathbb{T}^d} \|\widehat{T}_{nm}^{-1}(y) - \overline{T}_m^{-1}(y)\|^2 d\widehat{Q}_m(y) = \|\widehat{T}_{nm}^{-1} - \overline{T}_m^{-1}\|^2_{L^2(\widehat{Q}_m)}. \tag{81}$$

Step 3: Stability Bounds. Due to the inequalities (78), the stability bounds of Proposition 16 (arising from Theorem 6) imply

$$\|\overline{T}_m^{-1} - \widehat{T}_{nm}^{-1}\|^2_{L^2(\widehat{Q}_m)} \leq \lambda^2 W_2^2(\widehat{P}_n, P), \quad \|\overline{T}_m - T_0\|^2_{L^2(P)} \leq \lambda^2 W_2^2(\widehat{Q}_m, Q). \tag{82}$$

Thus, combined with equations (79) and (81), we have on the event E_{nm},

$$\|\widehat{T}_{nm} - T_0\|_{L^2(P)}^2 \lesssim W_2^2(\widehat{P}_n, P) + W_2^2(\widehat{Q}_m, Q).$$

We deduce,

$$\begin{aligned}
\mathbb{E}\|\widehat{T}_{nm} - T_0\|_{L^2(P)}^2 &= \mathbb{E}\left[\|\widehat{T}_{nm} - T_0\|_{L^2(P)}^2 I_{E_{nm}}\right] + \mathbb{E}\left[\|\widehat{T}_{nm} - T_0\|_{L^2(P)}^2 I_{E_{nm}^{\mathsf{c}}}\right] \\
&\lesssim \mathbb{E}\left[W_2^2(\widehat{P}_n, P) I_{E_{nm}}\right] + \mathbb{E}\left[W_2^2(\widehat{Q}_m, Q) I_{E_{nm}}\right] + \mathbb{P}(E_{nm}^{\mathsf{c}}) \\
&\leq \mathbb{E}\left[W_2^2(\widehat{P}_n, P)\right] + \mathbb{E}\left[W_2^2(\widehat{Q}_m, Q)\right] + \mathbb{P}(E_{nm}^{\mathsf{c}}) \\
&\lesssim R_{T,n}(\alpha) + R_{T,m}(\alpha) + (n \wedge m)^{-2} \lesssim R_{T,n\wedge m}(\alpha),
\end{aligned}$$

where we made use of Proposition 43(i) on the final line. The claim follows. □

We now state an extension of Theorem 17(ii) to the unit hypercube $[0,1]^d$. Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$, and denote by $\widehat{P}_n^{(\mathrm{bc})}$ and $\widehat{Q}_m^{(\mathrm{bc})}$ the boundary corrected wavelet estimators defined in Section 3.3.

Proposition 44. *Let $P, Q \in \mathcal{P}_{\mathrm{ac}}([0,1]^d)$ admit densities $p, q \in \mathcal{C}^{\alpha-1}([0,1]^d; M, \gamma)$ for some $\alpha > 1$ and $M, \gamma > 0$. Assume further that for some $\lambda > 0$,*

$$\varphi_0, \varphi_0^* \in \mathcal{C}^{\alpha+1}([0,1]^d; \lambda). \tag{83}$$

Let $2^{J_n} \asymp n^{\frac{1}{d+2(\alpha-1)}}$. Then, there exists a constant $C > 0$ depending only on $M, \lambda, \gamma, \alpha$ such that,

$$\begin{aligned}
\left|\mathbb{E}W_2^2(\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}) - W_2^2(P,Q)\right| &\leq C R_{T,n\wedge m}(\alpha), \\
\mathbb{E}\left|W_2^2(\widehat{P}_n^{(\mathrm{bc})}, \widehat{Q}_m^{(\mathrm{bc})}) - W_2^2(P,Q)\right|^2 &\leq \left[C R_{T,n\wedge m}(\alpha) + \sqrt{\frac{\mathrm{Var}_P[\phi_0(X)]}{n} + \frac{\mathrm{Var}_Q[\psi_0(Y)]}{m}}\right]^2.
\end{aligned}$$

The proof follows along similar lines as that of Theorem 17(ii), and is therefore omitted.

Condition (83) places a smoothness assumption on φ_0^* in addition to φ_0. If our analysis could be carried out over a domain $\Omega \subseteq \mathbb{R}^d$ with smooth boundary, then, under appropriate boundary conditions on the potentials and under the assumptions made on p, q, the Evans-Krylov Theorem for uniformly elliptic fully nonlinear partial differential equations could be applied to the Monge-Ampère equation to obtain that $\varphi_0^* \in \mathcal{C}^{\alpha+1}(\Omega)$ as soon as $\varphi_0 \in \mathcal{C}^2(\Omega)$, with uniform Hölder norms (see, for instance, the discussion in Appendix A.5.5 of Figalli (2017)). We do not know whether analogues of such results can be applied over the hypercube $[0,1]^d$, thus we have placed assumptions both on φ_0 and its convex conjugate.

I Proofs of Upper Bounds for Kernel Estimators

The goal of this Appendix is to prove Theorem 18. For ease of notation, we omit the superscript "ker" in all kernel-based estimators, and write

$$p_{h_n}(x) = \mathbb{E}[\widetilde{p}_n(x)] = (p \star K_{h_n})(x), \quad q_{h_m}(y) = \mathbb{E}[\widetilde{q}_m(y)] = (q \star K_{h_m})(y), \quad x, y \in \mathbb{T}^d.$$

We begin with the following technical result.

Lemma 45. *Let $t, s > 0$, and assume $p \in H^s(\mathbb{T}^d)$. Assume further that the kernel K satisfies condition **K1**$(s+t,\kappa)$ for some $\kappa > 0$. Then, for any $h_n > 0$,*

$$\|p_{h_n} - p\|_{\dot{H}^{-t}(\mathbb{T}^d)} \leq \kappa \|p\|_{H^s(\mathbb{T}^d)} h_n^{s+t}$$

Proof of Lemma 45. By definition of the $\dot{H}^{-t}(\mathbb{T}^d)$ norm,

$$\|p_{h_n} - p\|_{\dot{H}^{-t}} = \big\| \|\cdot\|^{-t} \mathcal{F}[p_{h_n} - p](\cdot) \big\|_{\ell^2(\mathbb{Z}^d)}.$$

Furthermore, using standard properties of the Fourier transform,

$$\mathcal{F}[p_{h_n} - p] = \mathcal{F}[p \star K_{h_n} - p] = (\mathcal{F}[K](h_n \cdot) - 1)\mathcal{F}[p].$$

Thus, using assumption **K1**$(s+t,\kappa)$, we have

$$\begin{aligned}
\|p_{h_n} - p\|^2_{\dot{H}^{-t}(\mathbb{T}^d)} &= \sum_{\xi \in \mathbb{Z}^d} \frac{1}{\|\xi\|^{2t}} \big|\mathcal{F}[K](h_n \xi) - 1\big|^2 \mathcal{F}[p]^2(\xi) \\
&\leq \kappa^2 \sum_{\xi \in \mathbb{Z}^d} \frac{\|h_n \xi\|^{2(s+t)}}{\|\xi\|^{2t}} \mathcal{F}[p]^2(\xi) \\
&= \kappa^2 h_n^{2(s+t)} \sum_{\xi \in \mathbb{Z}^d} \|\xi\|^{2s} \mathcal{F}[p]^2(\xi) \leq \kappa^2 h_n^{2(s+t)} \|p\|^2_{H^s(\mathbb{T}^d)},
\end{aligned}$$

as claimed. □

While Lemma 45 will be needed in the proof of Theorem 18 below, we begin by showing how it may also be used to derive a rate of convergence of $\widehat{Q}_n$ under the Wasserstein distance. The following result was anticipated by Divol (2021), who derived a Fourier-analytic proof of the convergence rate of the empirical measure under the Wasserstein distance on $\mathbb{T}^d$. Our proof follows along similar lines, and is simplified by the fact that we work only with the Wasserstein distance of second order, but is complicated by the fact that we require general exponents $\rho \geq 1$.

Lemma 46. *Let $s > 0$. Assume $P \in \mathcal{P}_{\mathrm{ac}}(\mathbb{T}^d)$ admits a density p such that*

$$\|p\|_{H^s(\mathbb{T}^d)} \leq R < \infty, \qquad 0 < \gamma^{-1} \leq p \leq \gamma < \infty.$$

*Assume further that the kernel K satisfies condition **K1**$(s+1,\kappa)$ for some $\kappa > 0$. Set $h_n \asymp n^{\frac{1}{2s+d}}$. Then, for any $\rho \geq 0$,*

$$\mathbb{E}W_2^\rho(\widehat{P}_n, P) \lesssim_{R,\rho,\gamma,s} \begin{cases} n^{-\frac{\rho(s+1)}{2s+d}}, & d \geq 3 \\ (\log n/n)^{\rho/2}, & d = 2 \\ (1/n)^{\rho/2}, & d = 1. \end{cases}$$

Proof. By Jensen's inequality, it suffices to prove the claim for $\rho \geq 2$. It is a direct consequence of Lemma 32 and the assumption $\gamma^{-1} \leq p \leq \gamma$ that the event $A_n = \{\widehat{p}_n = \widetilde{p}_n\}$ satisfies $\mathbb{P}(A_n) \lesssim 1/n^2$. Furthermore, recall from equation (18), arising from the work of Peyre (2018), that

$$W_2(\widehat{P}_n, P) \lesssim \|\widehat{p}_n - p\|_{\dot{H}^{-1}}.$$

We therefore have,

$$\begin{aligned}
\mathbb{E}W_2^\rho(\widehat{P}_n, P) &= \mathbb{E}\Big[W_2^\rho(\widehat{P}_n, P)I_{A_n}\Big] + \mathbb{E}\Big[W_2^\rho(\widehat{P}_n, P)I_{A_n^\complement}\Big] \\
&\lesssim \mathbb{E}\Big[\|\widehat{p}_n - p\|_{\dot{H}^{-1}}^\rho I_{A_n}\Big] + 1/n^2 \\
&= \mathbb{E}\Big[\|\widetilde{p}_n - p\|_{\dot{H}^{-1}}^\rho I_{A_n}\Big] + 1/n^2 \\
&\lesssim \|p_{h_n} - p\|_{\dot{H}^{-1}}^\rho + \mathbb{E}\|\widetilde{p}_n - p_{h_n}\|_{\dot{H}^{-1}}^\rho + 1/n^2 \\
&\lesssim h_n^{\rho(s+1)} + \mathbb{E}\|\widetilde{p}_n - p_{h_n}\|_{\dot{H}^{-1}}^\rho + 1/n^2, \qquad (84)
\end{aligned}$$

where we used Lemma 45 on the final line, together with the assumption **K1**$(s+1,\kappa)$. To bound the variance term, write $\mathbb{E}\|\widetilde{p}_n - p_{h_n}\|_{\dot{H}^{-1}}^\rho \lesssim S_{n,1} + S_{n,2}$, where

$$S_{n,1} := \mathbb{E}\left[\left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\|\le 1} \|\xi\|^{-2}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^2\right)^{\frac{\rho}{2}}\right],$$

$$S_{n,2} := \mathbb{E}\left[\left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\|> 1} \|\xi\|^{-2}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^2\right)^{\frac{\rho}{2}}\right].$$

We begin by bounding $S_{n,1}$. Recall that

$$\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi) = \mathcal{F}[K](h_n\xi)\frac{1}{n}\sum_{j=1}^n \left(e^{-2\pi i\langle X_j,\xi\rangle} - \mathcal{F}[p](\xi)\right), \quad \xi\in\mathbb{Z}^d,$$

where $i^2=-1$. In fact, since $\widetilde{p}_n$ and p_{h_n} integrate to the same constant, we have $\mathcal{F}[\widetilde{p}_n - p_{h_n}](0) = 0$. Furthermore, let $\rho'\in\mathbb{R}$ satisfy $\frac{1}{\rho}+\frac{1}{\rho'}=\frac{1}{2}$. Then, for any $\eta\in\mathbb{R}$, we have by Hölder's inequality,

$$\begin{aligned}
S_{n,1} &= \mathbb{E}\left[\left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\|\le 1} \|\xi\|^{-2\eta}\|\xi\|^{2(\eta-1)}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^2\right)^{\frac{\rho}{2}}\right] \\
&\le \left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{-\rho'\eta}\right)^{\frac{\rho}{\rho'}} \mathbb{E}\left[\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{\rho(\eta-1)}\left|\mathcal{F}[\widetilde{p}_n - p_{h_n}](\xi)\right|^\rho\right] \\
&= \left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{-\rho'\eta}\right)^{\frac{\rho}{\rho'}} \sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\le 1}} \|\xi\|^{\rho(\eta-1)}\left|\mathcal{F}[K](h_n\xi)\right|^\rho \mathbb{E}\left|\frac{1}{n}\sum_{j=1}^n Z_j(\xi)\right|^\rho,
\end{aligned}$$

where $Z_j(\xi) = e^{-2\pi i\langle X_j,\xi\rangle} - \mathcal{F}[p](\xi)$, for all $j=1,\dots,n$ and $\xi\in\mathbb{Z}^d$. Since $\rho\ge 2$, it can be deduced from Rosenthal's inequalities (Rosenthal, 1970, 1972) that,

$$\mathbb{E}\left|\frac{1}{n}\sum_{j=1}^{n} Z_j(\xi)\right|^{\rho} \lesssim n^{-\frac{\rho}{2}} \left(\mathbb{E}|Z_1(\xi)|^2\right)^{\rho} + n^{1-\rho}\mathbb{E}|Z_1(\xi)|^{\rho}.$$

Notice that $|Z_1(\xi)| \leq 2$ for any $\xi \in \mathbb{Z}^d$, and $\rho/2 \leq \rho - 1$, thus we deduce from the previous two displays that,

$$S_{n,1} \lesssim n^{-\frac{\rho}{2}} \left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\leq 1}} \|\xi\|^{-\rho'\eta} \right)^{\frac{\rho}{\rho'}} \sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\leq 1}} \|\xi\|^{\rho(\eta-1)} \left|\mathcal{F}[K](h_n\xi)\right|^{\rho} \tag{85}$$

$$\lesssim n^{-\frac{\rho}{2}} \left(\sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\leq 1}} \|\xi\|^{-\rho'\eta} \right)^{\frac{\rho}{\rho'}} \sum_{\substack{\xi\in\mathbb{Z}^d,\xi\neq 0\\ \|h_n\xi\|\leq 1}} \|\xi\|^{\rho(\eta-1)}, \tag{86}$$

where the final inequality follows from the fact that the Fourier transform of K is bounded over the unit ball, since $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$. When $d \geq 3$, due to the condition $\frac{1}{\rho} + \frac{1}{\rho'} = \frac{1}{2}$, we may choose η satisfying

$$d\left(\frac{1}{d} - \frac{1}{\rho}\right) < \eta < \frac{d}{\rho'}. \tag{87}$$

In particular, we then have $-d < \rho(\eta - 1)$ and $-d < -\rho'\eta$, so that

$$S_{n,1} \lesssim n^{-\frac{\rho}{2}} \left(h_n^{\rho'\eta-d}\right)^{\frac{\rho}{\rho'}} h_n^{-\rho(\eta-1)-d} = n^{-\frac{\rho}{2}} h_n^{\rho-d(\frac{\rho}{\rho'}+1)} = n^{-\frac{\rho}{2}} h_n^{\rho(1-\frac{d}{2})}.$$

If $d = 2$, we choose η such that the strict inequalities in equation (87) both hold with equality. In this case, we have $\rho'\eta = d$ and $\rho(\eta - 1) = -d$, thus

$$S_{n,1} \lesssim n^{-\frac{\rho}{2}} \left(\sum_{\xi\in\mathbb{Z}^d, \|h_n\xi\|\leq 1} \|\xi\|^{-d} \right)^{\frac{\rho}{\rho'}+1} \lesssim n^{-\frac{\rho}{2}} \log(h_n^{-1})^{\frac{\rho}{\rho'}+1} = \left(\log(h_n^{-1})/n\right)^{\frac{\rho}{2}}.$$

Finally, if $d = 1$, choose η such that

$$1 - \frac{1}{\rho} > \eta > \frac{1}{\rho'}. \tag{88}$$

In this case, both sequences in equation (86) are summable, and we obtain $S_{n,1} \lesssim n^{-\rho/2}$. In summary, we deduce

$$S_{n,1} \lesssim \beta_n := n^{-\frac{\rho}{2}} \begin{cases} h_n^{\rho(1-\frac{d}{2})}, & d \geq 3 \\ \big(\log(h_n^{-1})\big)^{\rho/2}, & d = 2 \\ 1, & d = 1. \end{cases} \tag{89}$$

We next bound $S_{n,2}$. Let $\eta < d/\rho'$. Apply a similar reduction as in equation (85), to obtain

$$S_{n,2} \lesssim n^{-\frac{\rho}{2}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{-\rho'\eta} \right)^{\frac{\rho}{\rho'}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{\rho(\eta-1)} \big|\mathcal{F}[K](h_n\xi)\big|^{\rho} \right).$$

Since $K \in \mathcal{C}_c^\infty(\mathbb{R}^d)$, notice that K and $\mathcal{F}[K]$ are Schwartz functions. In particular, $\mathcal{F}[K](\xi) \lesssim \|\xi\|^{-\ell}$ for any $\ell > 0$. Choose $\ell > 0$ such that $\rho(\eta - 1 - \ell) < -d$. We then have,

$$\begin{aligned} S_{n,2} &\lesssim n^{-\frac{\rho}{2}} \left(\sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{-\rho'\eta} \right)^{\frac{\rho}{\rho'}} \left(h_n^{-\rho\ell} \sum_{\xi \in \mathbb{Z}^d, \|h_n \xi\| > 1} \|\xi\|^{\rho(\eta-1-\ell)} \right) \\ &\lesssim n^{-\frac{\rho}{2}} \left(h_n^{\rho'\eta - d} \right)^{\frac{\rho}{\rho'}} h_n^{-\rho\ell} h_n^{-\rho(\eta-1-\ell)-d} \lesssim n^{-\frac{\rho}{2}} h_n^{\rho(1-\frac{d}{2})} \lesssim \beta_n. \end{aligned}$$

Combine this bound with those of equations (84) and (89)

$$\mathbb{E} W_2^\rho(\widehat{P}_n, P) \lesssim h_n^{\rho(s+1)} + \beta_n + 1/n^2 \lesssim \begin{cases} n^{-\frac{\rho(s+1)}{2s+d}}, & d \geq 3 \\ (\log n/n)^{\rho/2}, & d = 2 \\ (1/n)^{\rho/2}, & d = 1. \end{cases}$$

The claim follows. □

We are now in a position to prove Theorem 18.

I.1 Proof of Theorem 18

In view of Lemmas 32, 33, 45, and 46, the proof of the claim is analogous to that of Theorem 17, thus we only provide brief justifications.

Regarding part (i), apply Lemmas 25 and 32–33 to deduce that there exists $\epsilon \in (0, 1 \wedge \frac{\alpha-1}{2})$ and an event of probability at least $1 - 1/n^2$ over which $\widehat{p}_n, \widehat{q}_m$ coincide with $\widetilde{p}_n, \widetilde{q}_m$ respectively, and are of class $\mathcal{C}^\epsilon(\mathbb{T}^d)$, with Hölder norm uniformly bounded in n. By Theorem 5, it follows that, over this same high-probability event, any mean-zero Brenier potential in the optimal transport problem from P to $\widehat{Q}_m$, or from $\widehat{P}_n$ to $\widehat{Q}_m$, is of class $\mathcal{C}^{2+\epsilon}(\mathbb{T}^d)$, again with a uniformly bounded Hölder norm. Arguing as in Step 1 of the proof of Theorem 17(i), we deduce that these potentials achieve the conclusion of equation (78) therein. The same argument as in Steps 2–3 of that proof, coupled with Lemma 46 stating the convergence rate of the kernel density estimator in Wasserstein distance, can then be used to deduce that the optimal transport map $\widehat{T}_{nm}$ from $\widehat{P}_n$ to $\widehat{Q}_m$ satisfies

$$\mathbb{E}\big\|\widehat{T}_{nm} - \widehat{T}_0\big\|_{L^2(P)}^2 \lesssim \mathbb{E} W_2^2(\widehat{P}_n, P) + \mathbb{E} W_2^2(\widehat{Q}_m, Q) + \frac{1}{(n \wedge m)^2} \lesssim R_{K, n \wedge m}(\alpha).$$

In applying Lemma 46, we note that our stated assumption **K1**($2\alpha, \kappa$) implies **K1**($\alpha + 1, \kappa'$) for a constant $\kappa' > 0$ depending only on α and κ. This proves part (i). To prove part (ii), we use the following observation.

Lemma 47. *Under the assumptions of Theorem 18, we have*

$$\mathbb{E}\left[\int \phi_0(\widehat{p}_n - p)\right] = O(h_n^{2\alpha}), \quad \mathrm{Var}\left[\int \phi_0(\widehat{p}_n - p)\right] = \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + O\left(\frac{h_n^{2\alpha}}{n}\right), \tag{90}$$

where the implicit constants depend only on M, γ, α.

Using Lemmas 46–47, the same argument as in the proof of Theorem 17(ii) leads to the claim of part (ii). □

It thus remains to prove Lemma 47.

I.2 Proof of Lemma 47

Using Lemma 32, and arguing similarly as in the proof of Lemma 11, it will suffice to prove that

$$\int \phi_0(p - p_{h_n}) = O(h_n^{2\alpha}), \quad \mathrm{Var}\left[\int \phi_0(\widetilde{p}_n - p_{h_n})\right] = \frac{\mathrm{Var}_P[\phi_0(X)]}{n} + O\left(\frac{h_n^{2\alpha}}{n}\right). \tag{91}$$

Under the condition $\alpha \notin \mathbb{N}$, $\alpha > 1$, we deduce from Theorem 5 that there exists $\lambda > 0$ depending only on M, γ, α such that

$$\phi_0, \psi_0 \in \mathcal{C}^{\alpha+1}(\mathbb{T}^d; \lambda). \tag{92}$$

Now, by Parseval's Theorem,

$$\begin{aligned}
\left|\int_{\mathbb{T}^d} \phi_0(p - p_{h_n})\right| &= \left|\sum_{\xi \in \mathbb{Z}^d} \mathcal{F}[\phi_0](\xi)\mathcal{F}[p - p_{h_n}](\xi)\right| \\
&\le \big\|\|\cdot\|^{\alpha+1}\mathcal{F}[\phi_0](\cdot)\big\|_{\ell^2(\mathbb{Z}^d)}\big\|\|\cdot\|^{-(\alpha+1)}\mathcal{F}[p - p_{h_n}](\cdot)\big\|_{\ell^2(\mathbb{Z}^d)} \\
&= \|\phi_0\|_{\dot{H}^{\alpha+1}(\mathbb{T}^d)}\|p - p_{h_n}\|_{\dot{H}^{-(\alpha+1)}(\mathbb{T}^d)} \lesssim \|p - p_{h_n}\|_{\dot{H}^{-(\alpha+1)}(\mathbb{T}^d)},
\end{aligned}$$

where we used equation (92) and the fact that $\|\phi_0\|_{\dot{H}^{\alpha+1}(\mathbb{T}^d)} \le \|\phi_0\|_{H^{\alpha+1}(\mathbb{T}^d)} \lesssim \|\phi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)}$ (cf. Lemma 26). Apply Lemma 45, under the assumption **K1**($2\alpha, \kappa$), to deduce

$$\left|\int_{\mathbb{T}^d} \phi_0 d(p - p_{h_n})\right| \lesssim h_n^{2\alpha} \lesssim R_{K,n\wedge m}(\alpha).$$

To prove the variance bound, notice that

$$\mathrm{Var}\left[\int \phi_0(\widetilde{p}_n - p_{h_n})\right] = \mathrm{Var}\left[\int (\phi_0 \star K_{h_n}) d(P_n - P)\right] = \frac{1}{n}\mathrm{Var}_P[\phi_{h_n}(X)],$$

where $\phi_{h_n} = \phi_0 \star K_{h_n}$. Thus, reasoning as in the proof of Lemma 11, we have

$$\left|\mathrm{Var}\left[\int \phi_0(\widetilde{p}_n - p_{h_n})\right] - \frac{1}{n}\mathrm{Var}_P[\phi_0(X)]\right|$$

$$\leq \frac{1}{n}\Big| \mathrm{Var}_P[\phi_{h_n}(X)] - \mathrm{Var}_P[\phi_0(X)]\Big|$$
$$\leq \frac{1}{n}\Big|\mathbb{E}[\phi_{h_n}^2(X) - \phi_0^2(X)]\Big| + \frac{1}{n}\Big|\mathbb{E}[\phi_{h_n}(X) - \phi_0(X)]\Big| = \frac{1}{n}\big[(I) + (II)\big].$$

We shall again bound term (I), and a similar proof can be used for term (II). Notice that

$$(I) = \left|\int (\phi_{h_n} - \phi_0)(\phi_{h_n} + \phi_0)p\right| \leq \|\phi_{h_n} - \phi_0\|_{\dot{H}^{-(\alpha-1)}(\mathbb{T}^d)} \|(\phi_{h_n} + \phi_0)p\|_{\dot{H}^{\alpha-1}(\mathbb{T}^d)}.$$

It is a straightforward observation that $\|\phi_{h_n}\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq \|\phi_0\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)}$ for all $n \geq 1$, thus the function $(\phi_{h_n} + \phi_0)p$ has uniformly bounded $\mathcal{C}^{\alpha-1}(\mathbb{T}^d)$ norm, by Lemma 23. Since $\phi_0 \in \mathcal{C}^{\alpha+1}(\mathbb{T}^d; \lambda)$, we deduce that

$$(I) \lesssim \|\phi_{h_n} - \phi_0\|_{\dot{H}^{-(\alpha-1)}(\mathbb{T}^d)} \lesssim h_n^{2\alpha},$$

by Lemma 45. The claim follows from here. □

J Proofs of Central Limit Theorems

The aim of this appendix is to prove Theorem 19 and Corollary 20. We begin by deriving limit laws for the functional $\int \phi_0(\widehat{p}_n - p)$. Here, once again, $\widehat{p}_n$ is one of the estimators $\widehat{p}_n^{(\mathrm{bc})}, \widehat{p}_n^{(\mathrm{per})}, \widehat{p}_n^{(\mathrm{ker})}$, which respectively arise from the classical boundary-corrected, periodic, and kernel density estimators $\widetilde{p}_n^{(\mathrm{bc})}, \widetilde{p}_n^{(\mathrm{per})}, \widetilde{p}_n^{(\mathrm{ker})}$. We also write

$$p_{J_n}^{(\mathrm{bc})} = \mathbb{E}[\widetilde{p}_n^{(\mathrm{bc})}], \quad p_{J_n}^{(\mathrm{per})} = \mathbb{E}[\widetilde{p}_n^{(\mathrm{per})}], \quad p_{h_n}^{(\mathrm{ker})} = \mathbb{E}[\widetilde{p}_n^{(\mathrm{ker})}].$$

Lemma 48. *Let $\epsilon, s > 0$, and let $h_n^{-1} \asymp 2^{J_n} \uparrow \infty$.*

(i) (Unit Hypercube) Let $p \in \mathcal{C}^\epsilon([0,1]^d)$ be positive over $[0,1]^d$. Assume that $\phi_0 \in \mathcal{C}^s([0,1]^d)$ satisfies $\mathrm{Var}_P[\phi_0(X)] > 0$. Then, as $n \to \infty$,

$$\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) \rightsquigarrow N(0, \mathrm{Var}_P[\phi_0(X)]).$$

(ii) (Flat Torus) Let $p \in \mathcal{C}^\epsilon(\mathbb{T}^d)$ be positive over $\mathbb{T}^d$. Assume that $\phi_0 \in \mathcal{C}^s(\mathbb{T}^d)$ satisfies $\mathrm{Var}_P[\phi_0(X)] > 0$. Then, as $n \to \infty$,

$$\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{per})} - p_{J_n}^{(\mathrm{per})}) \rightsquigarrow N(0, \mathrm{Var}_P[\phi_0(X)]),$$
$$\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{ker})} - p_{h_n}^{(\mathrm{ker})}) \rightsquigarrow N(0, \mathrm{Var}_P[\phi_0(X)]).$$

J.1 Proof of Lemma 48

The proof is standard, thus we only prove claim (i). The remaining claims can be proven similarly. For simplicity, we write $\Psi_{j_0-1}^{\mathrm{bc}} = \Phi^{\mathrm{bc}}$ throughout the proof. Reasoning as in the proof of Lemma 11, and in particular using Lemma 29, it holds that

$$\begin{aligned}\sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) &= \sqrt{n}\int \phi_0(\widetilde{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) + \sqrt{n}\int f(\widehat{p}_n^{(\mathrm{bc})} - \widetilde{p}_n^{(\mathrm{bc})}) \\ &= \sqrt{n}\int \phi_0(\widetilde{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})}) + o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n (Z_{n,i} - \mathbb{E}[Z_{n,i}]),\end{aligned}$$

where we write

$$Z_{n,i} = \sum_{j=j_0-1}^{J_n} \sum_{\xi\in\Psi_j^{\mathrm{bc}}} \xi(X_i)\gamma_\xi,$$

and where $\gamma_\xi = \int \phi_0\xi$ for all $\xi \in \Psi^{\mathrm{bc}}$. By Lyapunov's central limit theorem (Billingsley (1968), Theorem 7.3), it holds that

$$\frac{1}{\sqrt{\sum_{i=1}^n \mathrm{Var}[Z_{n,i}]}}\sum_{i=1}^n (Z_{n,i} - \mathbb{E}[Z_{n,i}]) \rightsquigarrow N(0,1), \tag{93}$$

provided that for some $p > 2$,

$$\frac{\sum_{i=1}^n \mathbb{E}\left[|Z_{n,i} - \mathbb{E}Z_{n,i}|^p\right]}{\left(\sum_{i=1}^n \mathrm{Var}[Z_{n,i}]\right)^{p/2}} \to 0. \tag{94}$$

Now, using Lemma 24, it holds that

$$\begin{aligned}\sup_{n\geq 1}\sup_{1\leq i\leq n}|Z_{n,i}| &\leq \sup_{n\geq 1}\left\|\sum_{j=j_0-1}^{J_n}\sum_{\xi\in\Psi_j^{\mathrm{bc}}}\xi\gamma_\xi\right\|_{L^\infty([0,1]^d)} \\ &\leq \sum_{j=J_n+1}^\infty \|(\gamma_\xi)_{\xi\in\Psi_j^{\mathrm{bc}}}\|_{\ell_\infty}\left(\sup_{\xi\in\Psi_j^{\mathrm{bc}}}\|\xi\|_{L^\infty([0,1]^d)}\right)\left\|\sum_{\xi\in\Psi_j^{\mathrm{bc}}} I(|\xi|>0)\right\|_{L^\infty([0,1]^d)} \\ &\lesssim \sum_{j=J_n+1}^\infty 2^{-j(\frac{d}{2}+s)}2^{\frac{dj}{2}} \lesssim 2^{-J_n s} = o(1).\end{aligned} \tag{95}$$

On the other hand, under the stated conditions, it follows from Lemma 11 that

$$\sum_{i=1}^n \mathrm{Var}[Z_{n,i}] = n(\mathrm{Var}_P[\phi_0(X)] + o(1)). \tag{96}$$

Since $\mathrm{Var}_P[\phi_0(X)] > 0$, the denominator in equation (94) is of the order $n^{p/2}$, while the numerator is of order n by equation (95). It follows that Lyapunov's condition (94) holds for all $p > 2$. The claim thus follows from equations (93) and (96). □

J.2 Proof of Theorem 19

Assume first that $\sigma_0, \sigma_1 > 0$. We begin with part (i). Apply the stability bound of Theorem 6 to obtain,

$$0 \le W_2^2(\widehat{P}_n^{(\mathrm{per})}, Q) - W_2^2(P, Q) - \int \phi_0 d(\widehat{P}_n^{(\mathrm{per})} - P) \le W_2^2(\widehat{P}_n^{(\mathrm{per})}, P).$$

Using the convergence rate of $\widehat{P}_n^{(\mathrm{per})}$ under W_2^2 in Proposition 43(i), and Proposition 43(ii) regarding the bias of $\int \phi_0 d\widehat{P}_n^{(\mathrm{per})}$, we obtain

$$W_2^2(\widehat{P}_n^{(\mathrm{per})}, Q) - W_2^2(P, Q) = \int \phi_0(\widehat{p}_n^{(\mathrm{per})} - p_{J_n}^{(\mathrm{per})}) + O_p\left(n^{-\frac{2\alpha}{2(\alpha-1)+d}} \vee \frac{(\log n)^2}{n}\right).$$

Using the assumption $2(\alpha+1) > d$, deduce that

$$\sqrt{n}\left(W_2^2(\widehat{P}_n^{(\mathrm{per})}, Q) - W_2^2(P, Q)\right) = \sqrt{n}\int \phi_0(\widehat{p}_n^{(\mathrm{per})} - p_{J_n}^{(\mathrm{per})}) + o_p(1).$$

Apply Lemma 48 to deduce that

$$\sqrt{n}\left(W_2^2(\widehat{P}_n^{(\mathrm{per})}, Q) - W_2^2(P, Q)\right) \rightsquigarrow N(0, \sigma_0^2), \quad \text{as } n \to \infty.$$

By the same reasoning, but now using the two-sample stability bound of Proposition 12, we also have

$$\begin{aligned}&\sqrt{\frac{nm}{n+m}}\left(W_2^2(\widehat{P}_n^{(\mathrm{per})}, Q_m^{(\mathrm{per})}) - W_2^2(P, Q)\right)\\ &\qquad = \sqrt{(1-\rho)n}\int \phi_0(\widehat{p}_n^{(\mathrm{per})} - p_{J_n}^{(\mathrm{per})}) + \sqrt{\rho m}\int \psi_0(\widehat{q}_m^{(\mathrm{per})} - q_{J_m}^{(\mathrm{per})}) + o_p(1),\end{aligned}$$

as $n, m \to \infty$ such that $n/(n+m) \to \rho \in [0,1]$. By Lemma 48 and the independence of $X_1, \dots, X_n, Y_1, \dots, Y_m$, we deduce that

$$\sqrt{\frac{nm}{n+m}}\left(W_2^2(\widehat{P}_n^{(\mathrm{per})}, Q_m^{(\mathrm{per})}) - W_2^2(P, Q)\right) \rightsquigarrow N(0, \sigma_\rho^2),$$

as $n, m \to \infty$ such that $n/(n+m) \to \rho$. This proves claim (i) for the periodic wavelet estimators. Claim (i) for kernel estimators (resp. claim (ii) for boundary-corrected wavelet estimators) follows analogously by now using Lemma 47 (resp. Lemma 11) to bound the bias of $\int \phi_0 d\widehat{P}_n^{(\mathrm{ker})}$ (resp. $\int \phi_0 d\widehat{P}_n^{(\mathrm{bc})}$), Lemma 46 (resp. Lemma 30) to bound the convergence rate of $\widehat{P}_n^{(\mathrm{ker})}$ (resp. $\widehat{P}_n^{(\mathrm{bc})}$) in Wasserstein distance, and Lemma 48 to obtain the limiting distribution of $\int \phi_0(\widehat{p}_n^{(\mathrm{ker})} - p_{h_n}^{(\mathrm{ker})})$ (resp. $\int \phi_0(\widehat{p}_n^{(\mathrm{bc})} - p_{J_n}^{(\mathrm{bc})})$).

Finally, to prove part (v), apply Corollary 8 and the result of Divol (2021) to deduce that, for $\Omega \in \{[0,1]^d, \mathbb{T}^d\}$, since the densities p and q are bounded and bounded away from zero, we have

$$\sqrt{n}W_2^2(P_n, P) = o_p(1), \quad \sqrt{m}W_2^2(Q_m, Q) = o_p(1),$$

as $n, m \to \infty$, whenever $d \le 3$. Therefore, using Theorem 6, Proposition 12, and Proposition 16, we obtain

$$\sqrt{n}\Big(W_2^2(P_n,Q) - W_2^2(P,Q)\Big) = \sqrt{n}\int \phi_0 d(P_n - P) + o_p(1),$$

$$\sqrt{\frac{nm}{n+m}}\Big(W_2^2(P_n,Q_m) - W_2^2(P,Q)\Big) = \sqrt{(1-\rho)n}\int \phi_0 d(P_n - P) + \sqrt{\rho m}\int \psi_0 d(Q_m - Q) + o_p(1).$$

Claim (v) then follows by the classical central limit theorem. It thus remains to consider the situation where $\sigma_1 = 0$ or $\sigma_0 = 0$. Notice that the Kantorovich potentials ϕ_0 and ψ_0 are almost everywhere constant if and only if $P = Q$. As a result, the statements "$\sigma_0 = 0$", "$\sigma_1 = 0$", and "$P = Q$" are equivalent, thus it remains to prove the claim when $P = Q$. In this case, it suffices to show that $\sqrt{n}W_2^2(\widehat{P}_n, P) = o_p(1)$ and $\sqrt{\frac{nm}{n+m}}W_2^2(\widehat{P}_n, \widehat{Q}_m) = o_p(1)$ for the various estimators $\widehat{P}_n$ and $\widehat{Q}_m$ under consideration. But these assertions are a direct consequence of the aforementioned convergence rates of these estimators in Wasserstein distance, under the assumptions of each of parts (i)–(iii). The claim thus follows. □

J.3 Proof of Corollary 20

Throughout the proof, we write

$$\widehat{\phi}_{nm}^o = \widehat{\phi}_{nm} - \int_{\mathbb{T}^d} (\widehat{\phi}_{nm} - \phi_0), \quad \widehat{\psi}_{nm}^o = \widehat{\psi}_{nm} - \int_{\mathbb{T}^d} (\widehat{\psi}_{nm} - \psi_0).$$

Recall from Proposition 4 that the Kantorovich potentials $\phi_0, \widehat{\phi}_{nm}, \psi_0, \widehat{\psi}_{nm}$ are $\mathbb{Z}^d$-periodic, and may be taken to be uniformly bounded over $\mathbb{R}^d$ due to the compactness of $\mathbb{T}^d$. The same is then also true of $\widehat{\phi}_{nm}^o, \widehat{\psi}_{nm}^o$, and we obtain,

$$\begin{aligned}
|\widehat{\sigma}_{0,nm}^2 - \sigma_0^2| &= \Big|\mathrm{Var}_{P_n}[\widehat{\phi}_{nm}(U)] - \mathrm{Var}_P[\phi_0(X)]\Big| \\
&= \Big|\mathrm{Var}_{P_n}[\widehat{\phi}_{nm}^o(U)] - \mathrm{Var}_P[\phi_0(X)]\Big| \\
&\lesssim \left|\int_{\mathbb{T}^d} (\widehat{\phi}_{nm}^o)^2 dP_n - \int_{\mathbb{T}^d} \phi_0^2 dP\right| + \left|\int_{\mathbb{T}^d} \widehat{\phi}_{nm}^o dP_n - \int_{\mathbb{T}^d} \phi_0 dP\right| \\
&\lesssim \left|\int_{\mathbb{T}^d} (\widehat{\phi}_{nm}^o)^2 d(P_n - P)\right| + \left|\int_{\mathbb{T}^d} \widehat{\phi}_{nm}^o d(P_n - P)\right| + \big\|\widehat{\phi}_{nm}^o - \phi_0\big\|_{L^2(P)} \\
&= \left|\int_{[0,1]^d} (\widehat{\phi}_{nm}^o)^2 d(P_n - P)\right| + \left|\int_{[0,1]^d} \widehat{\phi}_{nm}^o d(P_n - P)\right| + \big\|\widehat{\phi}_{nm}^o - \phi_0\big\|_{L^2(P)}.
\end{aligned} \tag{97}$$

Since $\widehat{\phi}_{nm}^o$ is convex up to translation by a quadratic function, and uniformly bounded, it must be Lipschitz with respect to $\|\cdot\|$ over the compact set $[0,1]^d$, with a uniform constant depending only on the diameter $\sqrt{d}$ of this set (Hiriart-Urruty and Lemaréchal (2004), Lemma 3.1.1, p. 102). Thus, $(\widehat{\phi}_{nm}^o)^2$ is also Lipschitz over $[0,1]^d$ with uniform constant. The set of Lipschitz functions with a uniformly bounded Lipschitz constant, over any given compact domain, forms a Glivenko-Cantelli class (van der Vaart and Wellner (1996), Theorem 2.7.1), thus the first two terms on the right-hand side of the above display vanish in probability.

We next bound the third term. Since p is positive over $\mathbb{T}^d$ and lies in $\mathcal{C}^{\alpha-1}(\mathbb{T}^d)$, implying that it is continuous, there must exist a real number $\gamma > 0$ (depending on p) such that $2/\gamma \leq p \leq \gamma/2$ over $\mathbb{T}^d$. Thus,

$$\|\widehat{\phi}^o_{nm} - \phi_0\|^2_{L^2(P)} \leq \frac{\gamma}{2}\|\widehat{\phi}^o_{nm} - \phi_0\|^2_{L^2(\mathbb{T}^d)}.$$

Now, notice that $\int_{\mathbb{T}^d} \widehat{\phi}^o_{nm} = \int_{\mathbb{T}^d} \phi_0$ for all $n, m \geq 1$. As a result, we may apply the Poincaré inequality over $\mathbb{T}^d$ (Lemma 36) to deduce that

$$\|\widehat{\phi}^o_{nm} - \phi_0\|^2_{L^2(\mathbb{T}^d)} \leq \|\nabla(\widehat{\phi}^o_{nm} - \phi_0)\|^2_{L^2(\mathbb{T}^d)} = \|\widehat{T}_{nm} - T_0\|^2_{L^2(\mathbb{T}^d)} \lesssim \|\widehat{T}_{nm} - T_0\|^2_{L^2(P)}. \tag{98}$$

By Theorem 18(i), $\mathbb{E}\|\widehat{T}_{nm} - T_0\|^2_{L^2(P)} \to 0$. Combine this fact with equations (97–98) to deduce that $|\widehat{\sigma}^2_{0,nm} - \sigma^2_0| = o_p(1)$ as $n, m \to \infty$, as claimed.

To prove an analogous result for $\widehat{\sigma}^2_{1,nm}$, apply the same reduction as in equation (97) to obtain that, as $n, m \to \infty$,

$$|\widehat{\sigma}^2_{1,m} - \sigma^2_1| \lesssim \|\widehat{\psi}_{nm} - \psi_0\|^2_{L^2(Q)} + o_p(1). \tag{99}$$

By the same reasoning as before, up to modifying the value of $\gamma > 0$, it holds that $2/\gamma \leq q \leq \gamma/2$ over $\mathbb{T}^d$. Therefore,

$$\|\widehat{\psi}_{nm} - \psi_0\|^2_{L^2(Q)} \lesssim \|\widehat{\psi}_{nm} - \psi_0\|^2_{L^2(\mathbb{T}^d)} \leq \|\widehat{T}^{-1}_{nm} - T^{-1}_0\|^2_{L^2(\mathbb{T}^d)},$$

where the final inequality again follows from the Poincaré inequality (Lemma 36), since $\int_{\mathbb{T}^d} \widehat{\psi}_{nm} = \int_{\mathbb{T}^d} \psi_0$. It can further be deduced from Lemma 32 that $\mathbb{P}(\gamma^{-1} \leq \widehat{p}_n, \widehat{q}_m \leq \gamma \text{ over } \mathbb{T}^d) \to 1$ as $n, m \to \infty$, thus the above display also leads to

$$\|\widehat{\psi}_{nm} - \psi_0\|^2_{L^2(Q)} \lesssim \|\widehat{T}^{-1}_{nm} - T^{-1}_0\|^2_{L^2(\widehat{Q}_m)},$$

with probability tending to one. Now, since $2/\gamma \leq p, q \leq \gamma/2$ and $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d)$, we deduce from Theorem 5 that $\varphi_0 \in \mathcal{C}^{\alpha+1}(\mathbb{T}^d)$. In particular, T^{-1}_0 is Lipschitz, thus continuing from the above display, we have over the event $\gamma^{-1} \leq \widehat{p}_n, \widehat{q}_m \leq \gamma$, which has probability tending to one, that,

$$\begin{aligned}
\|\widehat{\psi}_{nm} - \psi_0\|^2_{L^2(Q)} &\lesssim \|Id - T^{-1}_0 \circ \widehat{T}_{nm}\|^2_{L^2(\widehat{P}_n)} \\
&= \|T^{-1}_0 \circ T_0 - T^{-1}_0 \circ \widehat{T}_{nm}\|^2_{L^2(\widehat{P}_n)} \\
&\lesssim \|T_0 - \widehat{T}_{nm}\|^2_{L^2(\widehat{P}_n)} \lesssim \|T_0 - \widehat{T}_{nm}\|^2_{L^2(P)} = o_p(1),
\end{aligned}$$

where the final order assessment again follows from Theorem 18(i). The claim follows. □

K Proofs of Efficiency Lower Bounds

Throughout this appendix, given $Q \in \mathcal{P}_{\text{ac}}(\Omega)$, we abbreviate the functional Φ_Q by Φ, and the influence functions $\widetilde{\Phi}_{(P,Q)}$ and $\widetilde{\Psi}_{(P,Q)}$ by $\widetilde{\Phi}$ and $\widetilde{\Psi}$, respectively.

We begin by defining the differentiable paths $(P_{t,h_1})_{t\geq 0}$ and $(Q_{t,h_2})_{t\geq 0}$, for all $(h_1, h_2) \in \mathbb{R}^2$, as announced in Section 5.2. We follow a construction from Example 1.12 of van der Vaart (2002).

Recall $P, Q \in \mathcal{P}_{\rm ac}(\Omega)$ admit respective densities p, q. Let $\zeta \in \mathcal{C}^\infty(\mathbb{R}) \cap \mathcal{C}^2(\mathbb{R})$ be a bounded non-negative map, which is bounded away from zero over $\mathbb{R}$ by a positive constant, and which satisfies $\zeta(0) = \zeta'(0) = \zeta''(0) = 1$. For any functions $f \in L_0^2(P)$ and $g \in L_0^2(Q)$, define $P_t^f, Q_t^g \in \mathcal{P}_{\rm ac}(\Omega)$ to be the distributions with densities

$$p_t^f(x) \propto \zeta(tf(x))p(x), \quad q_t^g(y) \propto \zeta(tg(y))q(y), \tag{100}$$

for all $x, y \in \Omega$ and $t \geq 0$. Since ζ is bounded away from zero over $\mathbb{R}$, notice that the implicit normalizing constants in the above display are bounded from above by a constant which does not depend on t, f, g, p, q. We now turn to the proofs of Lemma 21 and Theorem 22.

K.1 Proof of Lemma 21

To prove the claim, we shall make use of the following stability result for Kantorovich potentials over compact metric spaces, due to Santambrogio (2015, Theorem 1.52), which we only state in the generality required for our proof.

Lemma 49 (Santambrogio (2015))**.** *Let Ω be equal to $\mathbb{T}^d$, or to a compact and connected subset of $\mathbb{R}^d$. Let $P, Q \in \mathcal{P}_{\rm ac}(\Omega)$, and assume that at least one of P and Q has support equal to Ω. Let $(P_k)_{k\geq 1}, (Q_k)_{k\geq 1} \subseteq \mathcal{P}(\Omega)$ be sequences which respectively converge to P, Q weakly. Let (ϕ_k, ψ_k) denote a pair of Kantorovich potentials in the optimal transport problem from P_k to Q_k, for all $k \geq 1$. Then, up to taking subsequences, it holds that $\phi_k \to f_0$ and $\psi_k \to g_0$ as $k \to \infty$, the convergence being uniform over Ω, for some pair of Kantorovich potentials (f_0, g_0) in the optimal transport problem from P to Q, which is uniquely defined up to translation by constants.*

Returning to the proof, let $f \in \dot{\mathcal{P}}_P$ be an arbitrary score function, and abbreviate the differentiable path $P_t := P_t^f$, and its density $p_t := p_t^f$, for all $t \geq 0$. Here, we use the definition in equation (100). Let (ϕ_t, ψ_t) denote a pair of Kantorovich potentials in the optimal transport problem from P_t to Q, which we may and do choose to be uniformly bounded by $\mathrm{diam}(\Omega)$, and hence uniformly bounded in t. By the Kantorovich duality, one has

$$\begin{aligned}
\Phi(P_t) - \Phi(P) &= \sup_{(\phi,\psi)\in\mathcal{K}} \left[\int \phi dP_t + \int \psi dQ\right] - \int \phi_0 dP - \int \psi_0 dQ \\
&\geq \int \phi_0 dP_t + \int \psi_0 dQ - \int \phi_0 dP - \int \psi_0 dQ = \int \phi_0 d(P_t - P), \qquad (101)\\
\Phi(P_t) - \Phi(P) &= \int \phi_t dP_t + \int \psi_t dQ - \sup_{(\phi,\psi)\in\mathcal{K}} \left[\int \phi dP - \int \psi dQ\right] \leq \int \phi_t d(P_t - P).
\end{aligned}$$

By construction, the map $t \in [0, \infty) \mapsto p_t(x)$ is differentiable for every $x \in \Omega$, and letting $\Delta_t(x) = (p_t(x) - p(x))/t$, we have

$$\lim_{t\to 0} \Delta_t(x) = \frac{\partial}{\partial t} p_t(x)\bigg|_{t=0} = f(x)p(x). \tag{102}$$

Now, notice that for all $t \geq 0$,

$$|\Delta_t(x)| \lesssim \left|\frac{\zeta(tf(x)) - 1}{t}\right| p(x) = \left|\frac{\zeta(tf(x)) - \zeta(0)}{t}\right| p(x) \leq \|\zeta\|_{\mathcal{C}^1(\mathbb{R}^d)} f(x)p(x).$$

Since $f \in L_0^2(P) \subseteq L_0^1(P)$, we deduce that $\Delta_t(x)$ is dominated by an integrable function, uniformly in t. Since ϕ_0 is uniformly bounded, we also deduce that the map $|\phi_0||\Delta_t - fp|$ is dominated by an integrable function. We then have, by equation (102) and the Dominated Convergence Theorem,

$$\begin{aligned}
\liminf_{t\to 0} \frac{\Phi(P_t) - \Phi(P)}{t} &\geq \liminf_{t\to 0} \int_\Omega \phi_0 \Delta_t d\mathcal{L} \\
&= \int_\Omega \phi_0 f dP + \liminf_{t\to 0} \int_\Omega \phi_0 \left[\Delta_t - fp\right] d\mathcal{L} \\
&\geq \int_\Omega \phi_0 f dP - \limsup_{t\to 0} \int_\Omega |\phi_0| \left|\Delta_t - fp\right| d\mathcal{L} \\
&\geq \int_\Omega \phi_0 f dP - \int_\Omega |\phi_0| \limsup_{t\to 0} \left|\Delta_t - fp\right| d\mathcal{L} = \int \phi_0 f dP, \qquad (103)
\end{aligned}$$

and similarly,

$$\begin{aligned}
\limsup_{t\to 0} \frac{\Phi(P_t) - \Phi(P)}{t} &\leq \limsup_{t\to 0} \int \phi_t \Delta_t d\mathcal{L} \\
&\leq \limsup_{t\to 0} \int \phi_t f dP + \left(\sup_{t\geq 0} \|\phi_t\|_{L^\infty(\Omega)}\right) \limsup_{t\to 0} \int \left|\Delta_t - fp\right| d\mathcal{L} \\
&= \limsup_{t\to 0} \int \phi_t f dP.
\end{aligned}$$

Let $t_k \downarrow 0$ be a sequence achieving the limit superior, in the sense that $\lim_{k\to\infty} \int \phi_{t_k} f dP = \limsup_{t\to 0} \int \phi_t dP$. Up to taking a subsequence of (t_k), Lemma 49 implies that ϕ_{t_k} converges uniformly to a Kantorovich potential f_0 from P to Q, which is unique up to translation by a constant, and which therefore takes the form $f_0 = \phi_0 + a$ for some $a \in \mathbb{R}$. The limit superior clearly continues to be achieved along this subsequence, thus we replace it by (ϕ_{t_k}) without loss of generality. We thus have

$$\limsup_{t\to 0} \int \phi_t f dP = \lim_{k\to\infty} \int \phi_{t_k} f dP = \int \left(\lim_{k\to\infty} \phi_{t_k}\right) f dP = \int (\phi_0 + a) f dP = \int \phi_0 f dP,$$

where the interchange of limit and integration holds again by the Dominated Convergence Theorem, since ϕ_t are uniformly bounded, and $f \in L_0^2(P)$. Combine this fact with equation (103) to deduce that

$$\lim_{t\to 0} \frac{\Phi(P_t) - \Phi(P)}{t} = \int \phi_0 f dP.$$

It follows that $\widetilde{\Phi} = \phi_0 - \int \phi_0 dP$ is an influence function of Φ with respect to $\dot{\mathcal{P}}_P$. Since we assumed that $\widetilde{\Phi} \in \dot{\mathcal{P}}_P$, it must in fact be the case that $\widetilde{\Phi}$ is the unique efficient influence function of Φ with respect to $\dot{\mathcal{P}}_P$ (van der Vaart, 2002), and the claim follows. □

K.2 Proof of Theorem 22

We shall use the following abbrevations of the differentiable paths defined in equation (100). For any $h \in \mathbb{R}$ and $t \geq 0$, if $f = h\widetilde{\Phi}$ and $g = h\widetilde{\Psi}$, we write

$$P_{t,h} := P_t^f, \quad p_{t,h}(x) := p_t^f(x) = c_h(t)\zeta(th\widetilde{\Phi}(x))p(x), \tag{104}$$

$$Q_{t,h} := Q_t^g, \quad q_{t,h}(y) := q_t^g(y) = k_h(t)\zeta(th\widetilde{\Psi}(y))q(y), \tag{105}$$

for all $x, y \in \mathbb{T}^d$, where the normalizing constants are explicitly denoted

$$c_h(t) = \left(\int_{\mathbb{T}^d} \zeta(th\widetilde{\Phi}(x))dP(x)\right)^{-1}, \quad k_h(t) = \left(\int_{\mathbb{T}^d} \zeta(th\widetilde{\Psi}(y))dQ(y)\right)^{-1}.$$

In this case, the collections $\{(P_{t,h})_{t\geq 0} : h \in \mathbb{R}\}$ and $\{(Q_{t,h})_{t\geq 0} : h \in \mathbb{R}\}$ respectively have score functions given by the tangent spaces

$$\dot{\mathcal{P}}_P = \{h\widetilde{\Phi} : h \in \mathbb{R}\}, \quad \dot{\mathcal{P}}_Q = \{h\widetilde{\Psi} : h \in \mathbb{R}\}.$$

We begin by showing that there exist $\overline{M}, \bar{\gamma} > 0$ such that $p_{t,h} \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \bar{\gamma})$ uniformly in t, h. An identical argument may then be used to show that $q_{t,h} \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; \overline{M}, \bar{\gamma})$ for all appropriate t, h. Our proof then proceeds by proving parts (i) and (ii).

Since $p \geq \gamma^{-1}$, and since ζ is bounded from below by positive constants, it is clear that there must exist $\bar{\gamma} > 0$ depending on M and ζ such that

$$\bar{\gamma}^{-1} \leq p_{t,h} \quad \text{over } \mathbb{T}^d, \text{ for all } t \geq 0, h \in \mathbb{R}. \tag{106}$$

We next prove the uniform Hölder continuity of $p_{t,h}$. We begin by studying the Hölder continuity of the map $\zeta(th\widetilde{\Phi}(\cdot))$. Since $p, q \in \mathcal{C}^{\alpha-1}(\mathbb{T}^d; M, \gamma)$, and since we assumed $\alpha \notin \mathbb{N}$, we have by Theorem 5 that, for some constant $\lambda > 0$ depending only on M, γ, α,

$$\|\widetilde{\Phi}\|_{\mathcal{C}^{\alpha+1}(\mathbb{T}^d)} \leq \lambda. \tag{107}$$

Furthermore, recall that $\zeta \in \mathcal{C}^\infty(\mathbb{R})$. Thus, by the multivariate Faà di Bruno formula (see, for instance, Encinas and Masque (2003)), it holds that for all multi-indices $1 \leq |\beta| \leq \lfloor \alpha + 1 \rfloor$,

$$D^\beta \zeta(th\widetilde{\Phi}(\cdot)) = \beta! \sum_{\ell=0}^{|\beta|} (th)^\ell \zeta^{(\ell)}(th\widetilde{\Phi}(\cdot)) \sum_{(e_\tau)_\tau} \prod_\tau \frac{1}{e_\tau!}\left(\frac{1}{\tau!}D^\tau \widetilde{\Phi}(\cdot)\right)^{e_\tau},$$

where the second summation is taken over all tuples $(e_\tau)_{1\leq|\tau|\leq|\beta|} \subseteq \mathbb{N}$ such that $\sum_\tau e_\tau = \ell$, and the product is taken over all $1 \leq |\tau| \leq \ell$ such that $\sum_\tau e_\tau \tau = \beta$. Furthermore, $\beta! = \beta_1! \dots \beta_d!$. Since $\zeta \in \mathcal{C}^\infty(\mathbb{R})$, its derivatives of all orders less than $\alpha + 1$ are uniformly bounded over any fixed compact set. Since $\widetilde{\Phi}$ is bounded, we deduce for any $\bar{u} > 0$,

$$\sup_{0\leq\ell\leq\lfloor\alpha+1\rfloor} \sup_{\substack{t\geq 0, h\in\mathbb{R} \\ t|h|\leq\bar{u}}} \|(th)^\ell \zeta^{(\ell)}(th\widetilde{\Phi}(\cdot))\|_\infty \lesssim_{\bar{u},\alpha} 1.$$

Furthermore, we have $\|D^\tau\widetilde{\Phi}\|_\infty \leq \lambda$ for any $0 \leq |\tau| \leq \lfloor \alpha+1 \rfloor$. This fact together with the preceding two displays implies

$$\sup_{0\le|\beta|\le\lfloor\alpha+1\rfloor}\sup_{\substack{t\ge 0,h\in\mathbb{R}\\ t|h|\le\bar{u}}}\|D^\beta\zeta(th\widetilde{\Phi}(\cdot))\|_\infty \lesssim_{\lambda,\bar{u},\alpha} 1. \tag{108}$$

Now, recall that $p_{t,h}(\cdot)=c_h(t)\zeta(th\widetilde{\Phi}(\cdot))p(\cdot)$, and that $c_h(t)$ is uniformly bounded in h and t because ζ is bounded away from zero by a positive constant. Thus, using the above display, the fact that $p\in\mathcal{C}^{\alpha-1}(\mathbb{T}^d;M)$, and Lemma 23, we deduce there exists a constant $\overline{M}>0$, depending only on $M,\bar{u},\alpha$ and the choice of ζ, such that

$$\sup_{\substack{t\ge 0,h\in\mathbb{R}\\ t|h|\le\bar{u}}}\|p_{t,h}\|_{\mathcal{C}^{\alpha-1}(\mathbb{T}^d)}\le\overline{M}.$$

Combine this fact with equation (106) to deduce that

$$p_{t,h}\in\mathcal{C}^{\alpha-1}(\mathbb{T}^d;\overline{M},\bar{\gamma}),\quad \text{for all } t\ge 0, h\in\mathbb{R}, t|h|\le\bar{u}.$$

We now prove part (i). Since $\widetilde{\Phi}\in\dot{\mathcal{P}}_P$, it follows from Lemma 21 that $\widetilde{\Phi}$ is the efficient influence function of Φ relative to $\dot{\mathcal{P}}_P$. Since $\dot{\mathcal{P}}_P$ is a vector space, it follows from Theorem 25.21 of van der Vaart (1998) that for any estimator sequence U_n,

$$\sup_{\substack{\mathcal{I}\subseteq\mathbb{R}\\|\mathcal{I}|<\infty}}\liminf_{n\to\infty}\sup_{h\in\mathcal{I}} n\mathbb{E}_{n,h}\big|U_n-\Phi_Q(P_{n^{-1/2},h})\big|^2\ge \mathrm{Var}_P[\phi_0(X)],$$

where the infimum is over all estimator sequences.

We next prove part (ii). Inspired by the proof of Theorem 11 of Berrett and Samworth (2019), our goal will be to invoke a more general version of Theorem 25.21 of van der Vaart (1998), given in Theorem 3.11.5 of van der Vaart and Wellner (1996), whose statement we briefly summarize here. Let H be a Hilbert space with inner product $\langle\cdot,\cdot\rangle_H$, and norm $\|\cdot\|_H$. Let $(\mathcal{X}_n,\mathcal{A}_n,\mu_{n,h}:h\in H)$ be a sequence of asymptotically normal experiments (as defined in Section 3.11 of van der Vaart and Wellner (1996)). A parameter sequence $(\kappa_n(h):h\in H)\subseteq\mathbb{R}$ is said to be regular if there exists a nonnegative sequence (r_n) such that

$$r_n\big(\kappa_n(h)-\kappa_n(0)\big)\to\dot{\kappa}(h),\quad h\in H,$$

for a continuous linear map $\dot{\kappa}:H\to\mathbb{R}$. Denote by $\dot{\kappa}^*:\mathbb{R}\to H$ the adjoint of $\dot{\kappa}$, namely the map satisfying $\langle\dot{\kappa}^*(b^*),h\rangle_H=b^*\dot{\kappa}(h)$ for all $h\in H$.

Lemma 50 (van der Vaart and Wellner (1996), Theorem 3.11.5)**.** *Let the sequence of experiments $(\mathcal{X}_n,\mathcal{A}_n,\mu_{n,h}:h\in H)$ be asymptotically normal, and let the parameter sequence $(\kappa_n(h):h\in H)$ be regular. Suppose there exists a Gaussian random variable G such that for all $b^*\in\mathbb{R}$, $b^*G\sim N(0,\|\dot{\kappa}^*(b^*)\|_H^2)$. Then, for any estimator sequence $(U_n)_{n\ge1}$,*

$$\sup_{\substack{\mathcal{I}\subseteq H\\|\mathcal{I}|<\infty}}\liminf_{n\to\infty}\sup_{h\in\mathcal{I}} r_n^2\mathbb{E}_{\mu_{n,h}}(U_n-\kappa_n(h))^2\ge\mathrm{Var}[G].$$

Returning to the proof, define the Hilbert space $H = \mathbb{R}^2$ with inner product

$$\langle (h_1, h_2), (h_1', h_2') \rangle_H = h_1 h_1' \operatorname{Var}_P[\phi_0(X)] + h_2 h_2' \operatorname{Var}_Q[\psi_0(Y)], \quad (h_1, h_2), (h_1', h_2') \in H,$$

and the sequence of experiments

$$\mu_{n,h} = P_{n^{-1/2},h_1}^{\otimes n} \otimes Q_{m^{-1/2},h_2}^{\otimes m}, \quad h = (h_1, h_2) \in \mathbb{R}^2,$$

endowed with the standard Borel σ-algebra. Here, m is viewed as a function of n which satisfies $n/(n+m) \to \rho \in [0,1]$ as $n \to \infty$. The following result can be deduced from Section 7.5 of Berrett and Samworth (2019) with minor modifications, using our assumptions placed on ζ.

Lemma 51 (Berrett and Samworth (2019))**.** *The sequence of experiments* $(\mu_{n,h} : h \in H)$ *is asymptotically normal.*

For all $h = (h_1, h_2) \in H$, let $\kappa_n(h) = \Psi(P_{n^{-1/2},h_1}, Q_{m^{-1/2},h_2})$, where again m is treated as a function of n. Notice that $\kappa_n(0) = \Psi(P, Q)$ for any $n \geq 1$. By following the same argument as in the proof of Lemma 21, using the Kantorovich duality and the stability result for Kantorovich potentials in Lemma 49, one has

$$\kappa_n(h) - \kappa_n(0) = \int \phi_0 d(P_{n^{-1/2},h_1} - P) + \int \psi_0 d(Q_{m^{-1/2},h_2} - Q) + o(1).$$

Now, since $\zeta(0) = \zeta'(0) = 1$ and $\int \widetilde{\Phi} dP = 0$, we have for all $t \geq 0$,

$$\left| \frac{1}{c_{h_1}(t)} - 1 \right| = \left| \int \left[\zeta(th_1 \widetilde{\Phi}(x)) - 1 - th_1 \widetilde{\Phi}(x) \right] dP(x) \right| \lesssim \|\zeta\|_{\mathcal{C}^2(\mathbb{R})} t^2 h_1^2 \|\widetilde{\Phi}\|_{L^2(P)}.$$

Recall that p and ζ are bounded, and that $c_{n^{-1/2}}(h_1)$ is uniformly bounded in h_1 and n, thus for all $x \in \mathbb{T}^d$,

$$\begin{aligned} p_{n^{-1/2},h_1}(x) - p(x) &= p(x) \left[c_{h_1}(n^{-1/2}) \zeta\big(h_1 n^{-1/2} \widetilde{\Phi}(x)\big) - 1 \right] \\ &= p(x) \left[\zeta\big(h_1 n^{-1/2} \widetilde{\Phi}(x)\big) - 1 \right] + O\left(\frac{\|p\|_\infty \|\zeta\|_\infty h_1^2}{n} \right) \\ &= p(x) h_1 n^{-1/2} \widetilde{\Phi}(x) + O\left(h_1^2/n\right), \end{aligned}$$

where we again used the fact that $\zeta(0) = \zeta'(0) = 1$. Similarly, for all $y \in \mathbb{T}^d$,

$$q_{m^{-1/2},h_2}(y) - q(y) = q(y) h_2 m^{-1/2} \widetilde{\Psi}(y) + O\left(h_2^2/m\right),$$

implying that,

$$\begin{aligned} \kappa_n(h) - \kappa_n(0) &= h_1 n^{-1/2} \int \phi_0 \widetilde{\Phi} dP + h_2 m^{-1/2} \int \psi_0 \widetilde{\Psi} dQ + O(h_1^2/n + h_2^2/m) \\ &= h_1 n^{-1/2} \operatorname{Var}_P[\phi_0(X)] + h_2 m^{-1/2} \operatorname{Var}_Q[\psi_0(Y)] + O(h_1^2/n + h_2^2/m). \end{aligned}$$

We deduce that

$$\sqrt{\frac{nm}{m+m}}(\kappa_n(h) - \kappa_n(0)) \longrightarrow \dot{\kappa}(h) := \langle(\sqrt{1-\rho}, \sqrt{\rho}), (h_1, h_2)\rangle_H,$$

as $n, m \to \infty$ such that $n/(n+m) \to \rho$. It follows that the sequence of parameters $(\kappa_n(h) : h \in H)$ is regular. Furthermore, the adjoint of $\dot{\kappa}$ is easily seen to be $\dot{\kappa}^*(b^*) = b^*(\sqrt{1-\rho}, \sqrt{\rho})$, for all $b^* \in \mathbb{R}$, and one has

$$\|\dot{\kappa}^*(b^*)\|_H^2 = b^*\Big((1-\rho)\operatorname{Var}_P[\phi_0(X)] + \rho \operatorname{Var}_Q[\psi_0(Y)]\Big).$$

The claim now follows from Lemma 50. □

L Alternate Proof of Central Limit Theorems

In this Section, we provide an alternate proof of Theorem 19 which does not rely on our stability bounds in Theorem 6 and Proposition 12. We instead follow the strategy developed by del Barrio and Loubes (2019) for obtaining limit laws of the distinct process $\sqrt{n}(W_2^2(P_n, Q) - W_2^2(P, Q))$. For the sake of brevity, we only prove the one-sample case of Theorem 19(ii), and the remaining assertions of Theorem 19 can be handled similarly. Throughout this section, we abbreviate $\Psi = \Psi^{\mathrm{bc}}$ and $\widehat{P}_n = \widehat{P}_n^{(\mathrm{bc})}$.

We shall make use of the classical Efron-Stein inequality (see for instance Boucheron et al. (2013), Theorem 3.1) for bounding the variance of functions of independent random variables, stated as follows.

Lemma 52 (Efron-Stein Inequality)**.** *Let $X_1, X_1', X_2, X_2', \dots, X_n, X_n'$ be independent random variables, and let $R_n = f(X_1, \dots, X_n)$ be a square-integrable function of $X_1, \dots, X_n$. Let*

$$R_{ni}' = f(X_1, \dots, X_{i-1}, X_i', X_{i+1}, \dots, X_n), \quad i = 1, \dots, n.$$

Then,

$$\operatorname{Var}[R_n] \le \sum_{i=1}^n \mathbb{E}(R_n - R_{ni}')_+^2.$$

With these results in place, we turn to proving the one-sample case of Theorem 19(ii). In view of Lemma 30, it suffices to assume $P \neq Q$, in which case $\operatorname{Var}[\phi_0(X)] > 0$. We abbreviate $\widehat{P}_n = \widehat{P}_n^{(\mathrm{bc})}$, and we begin with the following result.

Proposition 53. *Assume the same conditions as Theorem 19(ii). Define*

$$R_n = W_2^2(\widehat{P}_n, Q) - \int \phi_0 d\widehat{P}_n.$$

Then, as $n \to \infty$, $n \operatorname{Var}(R_n) \to 0$.

L.1 Proof of Proposition 53

Let $X_1' \sim P$ denote a random variable independent of $X_1, \dots, X_n$, and let

$$P_n' = \frac{1}{n}\delta_{X_1'} + \frac{1}{n}\sum_{i=2}^n \delta_{X_i}$$

denote the corresponding empirical measure. Let $\widehat{P}_n'$ be the distribution with density

$$\widehat{p}_n' = \sum_{\zeta\in\Phi} \widehat{\beta}_\zeta' \zeta + \sum_{j=j_0}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi' \xi = \sum_{j=j_0-1}^{J_n} \sum_{\xi\in\Psi_j} \widehat{\beta}_\xi' \xi, \quad \text{where } \widehat{\beta}_\xi' = \int \xi d\widehat{P}_n', \ \xi \in \Psi,$$

where we write $\Psi_{J_0-1} = \Phi$ for ease of notation. Set

$$R_n' = W_2^2(\widehat{P}_n', Q) - \int \phi_0 d\widehat{P}_n'.$$

By Lemma 52, it will suffice to prove that $n^2\mathbb{E}(R_n - R_n')_+^2 = o(1)$. Let $(\widehat{\phi}_n, \widehat{\psi}_n)$ be a pair of Kantorovich potentials between $\widehat{P}_n$ and Q. Without loss of generality, we may assume that $\int \widehat{\phi}_n d\mathcal{L} = \int \phi_0 d\mathcal{L}$ for all $n \geq 1$. By the Kantorovich duality, we have

$$\begin{aligned}
W_2^2(\widehat{P}_n, Q) &= \int \widehat{\phi}_n d\widehat{P}_n + \int \widehat{\psi}_n dQ, \\
W_2^2(\widehat{P}_n', Q) &= \sup_{(\phi,\psi)\in\mathcal{K}} \int \phi d\widehat{P}_n' + \int \psi dQ \\
&\geq \int \widehat{\phi}_n d\widehat{P}_n' + \int \widehat{\psi}_n dQ = W_2^2(\widehat{P}_n, Q) + \int \widehat{\phi}_n d(\widehat{P}_n' - \widehat{P}_n).
\end{aligned}$$

It follows that, on the event E_n,

$$R_n - R_n' \leq \int (\widehat{\phi}_n - \phi_0) d(\widehat{P}_n - \widehat{P}_n').$$

In view of Lemma 52, the claim will follow if we are able to show that $n^2\mathbb{E}(R_n - R_n')_+ = o(1)$. Arguing similarly as in the proof of, for instance, Lemma 11, it holds that $\mathbb{P}(\widehat{p}_n = \widetilde{p}_n) \lesssim n^{-3}$. Using this fact and the above inequality, it will suffice to prove that the quantity

$$\Delta_n := n^2\mathbb{E}\left(\int (\widehat{\phi}_n - \phi_0)(\widetilde{p}_n - \widetilde{p}_n') d\mathcal{L}\right)_+$$

vanishes as $n \to \infty$. To this end, notice that

$$\int (\widehat{\phi}_n - \phi_0)(\widetilde{p}_n - \widetilde{p}_n') d\mathcal{L} = \int (\widehat{\phi}_n - \phi_0) \left(\sum_{j=j_0-1}^{J_n} \sum_{\xi\in\Psi_j} (\widehat{\beta}_\xi - \widehat{\beta}_\xi')\xi\right)$$

$$= \frac{1}{n} \sum_{j=j_0-1}^{J_n} \sum_{\xi \in \Psi_j} (\xi(X_1) - \xi(X_1')) \int (\widehat{\phi}_n - \phi_0)\xi.$$

Using the locality of the wavelet basis (Lemma 24(ii)) and the Cauchy-Schwarz inequality, we obtain

$$\Delta_n \lesssim J_n \sum_{j=j_0-1}^{J_n} \sum_{\xi \in \Psi_j} \mathbb{E}[\xi^2(X)] \int \big\|\widehat{\phi}_n - \phi_0\big\|^2 |\xi|^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} \int \big\|\widehat{\phi}_n - \phi_0\big\|^2 |\xi|^2.$$

In the final step, we again used Lemma 24(ii) together with the fact that p is bounded over $[0,1]^d$ (since $p \in \mathcal{C}^{\alpha-1}([0,1]^d)$), implying that

$$\mathbb{E}[\xi^2(X)] = \int \xi^2(x)p(x)dx \lesssim \int \xi^2(x)dx = 1.$$

By Lemma 24, for all $\xi \in \Psi_j$ and $j \geq j_0$, we have $\mathrm{supp}(\xi) \subseteq I_\xi$ for a rectangle $I_\xi \subseteq [0,1]^d$ satisfying $\mathrm{diam}(I_\xi) \lesssim 2^{-j}$, and $\|\xi\|_{L^\infty(I_\xi)} \lesssim 2^{dj/2}$. Thus,

$$n^2 \mathbb{E}(R_n - R_n')_+^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{dj} \int_{I_\xi} \big\|\widehat{\phi}_n - \phi_0\big\|^2.$$

Apply the Poincaré inequality in Lemma 35 together with the bound $\mathrm{diam}(I_\xi) \lesssim 2^{-j}$ to deduce

$$n^2 \mathbb{E}(R_n - R_n')_+^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{(d-2)j} \int_{I_\xi} \big\|\nabla(\widehat{\phi}_n - \phi_0)\big\|^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{(d-2)j} \big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2,$$

where the final inequality holds due to the assumption that p has a positive density over $[0,1]^d$, which, due to the continuity of p, implies that there is a constant $\gamma^{-1} > 0$ such that $p \geq \gamma^{-1}$ over $[0,1]^d$. Apply Theorem 10 to deduce that

$$n^2 \mathbb{E}(R_n - R_n')_+^2 \lesssim J_n \sum_{j=j_0-1}^{J_n} 2^{(d-2)j} \big\|\widehat{T}_n - T_0\big\|_{L^2(P)}^2 \lesssim J_n \left(2^{J_n(d-2-2\alpha)} \vee \frac{(\log n)^2}{n} \right).$$

Since $d < 2(\alpha+1)$ and $J_n \asymp \log n$, the above display is of order $o(1)$, thus the claim follows from Lemma 52. □

To prove the claim from here, write

$$\sqrt{n}\Big(W_2^2(\widehat{P}_n, Q) - \mathbb{E}W_2^2(\widehat{P}_n, Q)\Big) = \sqrt{n} \int \phi_0(\widehat{p}_n - p_{J_n}) + \sqrt{n}\Big(R_n - \mathbb{E}[R_n]\Big),$$

where recall that $p_{J_n} = \mathbb{E}[\widehat{p}_n]$. It follows from Proposition 53 that the final term of the above display converges to zero in probability. Furthermore, $\sqrt{n} \int \phi_0(\widehat{p}_n - p_{J_n}) \rightsquigarrow N(0, \mathrm{Var}[\phi_0(X)])$ by Lemma 48, and the claim follows. □

Acknowledgements. TM wishes to thank Aram-Alexandre Pooladian for conversations related to this work, and for his comments on an earlier version of this manuscript. TM was supported in part by the Natural Sciences and Engineering Research Council of Canada, through a PGS D scholarship. TM, SB and LW were supported in part by National Science Foundation grants DMS-1713003 and DMS-2113684. SB was additionally supported by a Google Research Scholar Award and an Amazon Research Award. JNW gratefully acknowledges the support of National Science Foundation grant DMS-2015291.

References

Ajtai, M., Komlós, J., and Tusnády, G. (1984). On optimal matchings. *Combinatorica*, 4(4):259–264.

Ambrosio, L., Colombo, M., De Philippis, G., and Figalli, A. (2012). Existence of Eulerian solutions to the semigeostrophic equations in physical space: The 2-dimensional periodic case. *Communications in Partial Differential Equations*, 37:2209–2227.

Bahouri, H., Chemin, J.-Y., and Danchin, R. (2011). *Fourier Analysis and Nonlinear Partial Differential Equations*, volume 343. Springer Science & Business Media.

Berrett, T. B. and Samworth, R. J. (2019). Efficient two-sample functional estimation and the super-oracle phenomenon. *arXiv preprint arXiv:1904.09347*.

Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley & Sons.

Black, E., Yeom, S., and Fredrikson, M. (2020). Fliptest: Fairness testing via optimal transport. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 111–121.

Bobkov, S. and Ledoux, M. (2019). One-dimensional empirical measures, order statistics, and Kantorovich transport distances. *Memoirs of the American Mathematical Society*, 261.

Boissard, E. and Le Gouic, T. (2014). On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. In *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, volume 50, pages 539–563.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press.

Bousquet, O., Boucheron, S., and Lugosi, G. (2003). Introduction to statistical learning theory. In *Summer School on Machine Learning*, pages 169–207. Springer.

Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on Pure and Applied Mathematics*, 44:375–417.

Caffarelli, L. A. (1991). Some regularity properties of solutions of Monge Ampère equation. *Communications on Pure and Applied Mathematics*, 44:965–969.

Caffarelli, L. A. (1992a). Boundary regularity of maps with convex potentials. *Communications on Pure and Applied Mathematics*, 45:1141–1151.

Caffarelli, L. A. (1992b). The Regularity of Mappings with a Convex Potential. *Journal of the American Mathematical Society*, 5:99–104.

Caffarelli, L. A. (1996). Boundary Regularity of Maps with Convex Potentials–II. *Annals of Mathematics*, 144:453–496.

Chaudhuri, K. and Dasgupta, S. (2010). Rates of convergence for the cluster tree. In *Advances in Neural Information Processing Systems 24*, pages 343–351.

Chernozhukov, V., Galichon, A., Hallin, M., and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics*, 45:223–256.

Chizat, L., Roussillon, P., Léger, F., Vialard, F.-X., and Peyré, G. (2020). Faster Wasserstein Distance Estimation with the Sinkhorn Divergence. *Advances in Neural Information Processing Systems 33*, pages 2257–2269.

Cohen, A. (2003). *Numerical Analysis of Wavelet Methods*, volume 32 of *Studies in Mathematics and Its Applications*. North-Holland Publishing Co., Amsterdam.

Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the Interval and Fast Wavelet Transforms. *Applied and Computational Harmonic Analysis*.

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

Courty, N., Flamary, R., Tuia, D., and Rakotomamonjy, A. (2016). Optimal transport for domain adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1853–1865.

Cover, T. (1968). Estimation by the nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:50–55.

Cuevas, A. (2009). Set estimation: Another bridge between statistics and geometry. *Bol. Estad. Investig. Oper*, 25:71–85.

Cuevas, A. and Fraiman, R. (1997). A Plug-in Approach to Support Estimation. *The Annals of Statistics*, 25:2300–2312.

Daubechies, I. (1988). Orthonormal Bases of Compactly Supported Wavelets. *Communications on Pure and Applied Mathematics*, 41:909–996.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.

de Lara, L., González-Sanz, A., and Loubes, J.-M. (2021). A Consistent Extension of Discrete Optimal Transport Maps for Machine Learning Applications. *arXiv preprint arXiv:2102.08644*.

De Philippis, G. and Figalli, A. (2014). The Monge–Ampère equation and its link to optimal transportation. *Bulletin of the American Mathematical Society*, 51:527–580.

Deb, N., Bhattacharya, B. B., and Sen, B. (2021a). Efficiency Lower Bounds for Distribution-Free Hotelling-Type Two-Sample Tests Based on Optimal Transport. *arXiv preprint arXiv:2104.01986*.

Deb, N., Ghosal, P., and Sen, B. (2021b). Rates of Estimation of Optimal Transport Maps using Plug-in Estimators via Barycentric Projections. *Advances in Neural Information Processing Systems 34*.

Deb, N. and Sen, B. (2021). Multivariate Rank-Based Distribution-Free Nonparametric Testing Using Measure Transportation. *Journal of the American Statistical Association*, 0:1–16.

del Barrio, E., Cuesta-Albertos, J. A., Hallin, M., and Matrán, C. (2020). Center-Outward Distribution Functions, Quantiles, Ranks, and Signs in $\mathbb{R}^d$. *arXiv preprint arXiv:1806.01238*.

del Barrio, E. and Loubes, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *The Annals of Probability*, 47:926–951.

Delalande, A. and Mérigot, Q. (2021). Quantitative Stability of Optimal Transport Maps under Variations of the Target Measure. *arXiv preprint arXiv:2103.05934*.

Divol, V. (2021). A short proof on the rate of convergence of the empirical measure for the Wasserstein distance. *arXiv preprint arXiv:2101.08126*.

Dudley, R. M. (1969). The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40:40–50.

Efromovich, S. (1999). *Nonparametric Curve Estimation: Methods, Theory, and Applications*. Springer Series in Statistics. Springer-Verlag, New York.

Encinas, L. H. and Masque, J. M. (2003). A short proof of the generalized Faà di Bruno's formula. *Applied Mathematics Letters*, 16:975–979.

Fan, J. and Hu, T.-C. (1992). Bias correction and higher order kernel functions. *Statistics & Probability Letters*, 13:235–243.

Figalli, A. (2017). *The Monge–Ampère Equation and Its Applications*. European Math. Soc., Zürich.

Finlay, C., Gerolin, A., Oberman, A. M., and Pooladian, A.-A. (2020). Learning normalizing flows from Entropy-Kantorovich potentials. *arXiv preprint arXiv:2006.06033*.

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738.

Freitag, G. and Munk, A. (2005). On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships. *Journal of Multivariate Analysis*, 94:123–158.

Ghodrati, L. and Panaretos, V. M. (2021). Distribution-on-Distribution Regression via Optimal Transport Maps. *arXiv preprint arXiv:2104.09418*.

Ghosal, P. and Sen, B. (2022). Multivariate Ranks and Quantiles using Optimal Transport: Consistency, Rates, and Nonparametric Testing. *The Annals of Statistics*, 50:1012–4037.

Gigli, N. (2011). On Hölder continuity-in-time of the optimal transport map towards measures along a curve. *Proceedings of the Edinburgh Mathematical Society*, 54:401–409.

Gilbarg, D. and Trudinger, N. S. (2001). *Elliptic Partial Differential Equations of Second Order*. Classics in Mathematics. Springer-Verlag, Berlin Heidelberg, 2 edition.

Giné, E. and Guillou, A. (2002). Rates of strong uniform consistency for multivariate kernel density estimators. In *Annales de l'Institut Henri Poincare (B) Probability and Statistics*, volume 38, pages 907–921. Elsevier.

Giné, E. and Nickl, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli*, 14.

Giné, E. and Nickl, R. (2009). Uniform limit theorems for wavelet density estimators. *The Annals of Probability*, 37:1605–1646.

Giné, E. and Nickl, R. (2016). *Mathematical Foundations of Infinite-Dimensional Statistical Models*, volume 40. Cambridge University Press.

Gordaliza, P., Del Barrio, E., Fabrice, G., and Loubes, J.-M. (2019). Obtaining Fairness using Optimal Transport Theory. In *International Conference on Machine Learning*, pages 2357–2365.

Grafakos, L. (2009). *Modern Fourier Analysis*, volume 250. Springer.

Gunsilius, F. (2021). On the Convergence Rate of Potentials of Brenier Maps. *To Appear, Econometric Theory*.

Guo, H. and Kou, J. (2019). Strong Uniform Convergence Rates of Wavelet Density Estimators with Size-Biased Data. *Journal of Function Spaces*, 2019.

Györfi, L., Kohler, M., Krzyzak, A., and Walk, H. (2006). *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.

Hallin, M., del Barrio, E., Cuesta-Albertos, J., and Matrán, C. (2021). Distribution and quantile functions, ranks and signs in dimension d: A measure transportation approach. *The Annals of Statistics*, 49:1139–1165.

Han, Y., Jiao, J., Weissman, T., and Wu, Y. (2020). Optimal rates of entropy estimation over Lipschitz balls. *The Annals of Statistics*, 48:3228–3250.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (2012). *Wavelets, Approximation, and Statistical Applications*, volume 129. Springer Science & Business Media.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2004). *Fundamentals of Convex Analysis*. Springer Science & Business Media.

Hütter, J.-C. and Rigollet, P. (2021). Minimax rates of estimation for smooth optimal transport maps. *The Annals of Statistics*, 49:1166–1194.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Kantorovich, L. V. (1948). On a problem of Monge. In *CR (Doklady) Acad. Sci. URSS (NS)*, volume 3, pages 225–226.

Kerkyacharian, G. and Picard, D. (1992). Density estimation in Besov spaces. *Statistics & probability letters*, 13:15–24.

Knott, M. and Smith, C. S. (1984). On the optimal mapping of distributions. *Journal of Optimization Theory and Applications*, 43:39–49.

Kolouri, S., Park, S. R., Thorpe, M., Slepcev, D., and Rohde, G. K. (2017). Optimal mass transport: Signal processing and machine-learning applications. *IEEE Signal Processing Magazine*, 34:43–59.

Komiske, P. T., Mastandrea, R., Metodiev, E. M., Naik, P., and Thaler, J. (2020). Exploring the space of jets with CMS open data. *Physical Review D*, 101:034009.

Krishnamurthy, A., Kandasamy, K., Poczos, B., and Wasserman, L. (2014). Nonparametric estimation of renyi divergence and friends. In *International Conference on Machine Learning*, pages 919–927.

Ledoux, M. (2019). On optimal matching of Gaussian samples. *Journal of Mathematical Sciences*, 238:495–522.

Lei, J. (2020). Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26:767–798.

Leoni, G. (2017). *A First Course in Sobolev Spaces*. American Mathematical Soc.

Levy, B. and Schwindt, E. (2018). Notions of optimal transport theory and how to implement them on a computer. *Computers & Graphics*, 72:135–148.

Liang, T. (2019). On the Minimax Optimality of Estimating the Wasserstein Metric. *arXiv preprint arXiv:1908.10324*.

Ma, X.-N., Trudinger, N. S., and Wang, X.-J. (2005). Regularity of Potential Functions of the Optimal Transportation Problem. *Archive for Rational Mechanics and Analysis*, 177(2):151–183.

Makkuva, A., Taghvaei, A., Oh, S., and Lee, J. (2020). Optimal transport mapping via input convex neural networks. In *International Conference on Machine Learning*, pages 6672–6681. PMLR.

Manole, T. and Niles-Weed, J. (2021). Sharp Convergence Rates for Empirical Optimal Transport with Smooth Costs. *arXiv preprint arXiv:2106.13181*.

Masry, E. (1997). Multivariate probability density estimation by wavelet methods: Strong consistency and rates for stationary time series. *Stochastic processes and their applications*, 67:177–193.

Mazumder, R., Choudhury, A., Iyengar, G., and Sen, B. (2019). A computational framework for multivariate convex regression and its variants. *Journal of the American Statistical Association*, 114:318–331.

Mérigot, Q. (2011). A multiscale approach to optimal transport. In *Computer Graphics Forum*, volume 30, pages 1583–1592.

Mérigot, Q., Delalande, A., and Chazal, F. (2019). Quantitative stability of optimal transport maps and linearization of the 2-Wasserstein space. *arXiv preprint arXiv:1910.05954*.

Meyer, Y. (1991). Ondelettes sur l'intervalle. *Revista Matematica Iberoamericana*, 7:115–133.

Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Munk, A. and Czado, C. (1998). Nonparametric Validation of Similar Distributions and Assessment of Goodness of Fit. *Journal of the Royal Statistical Society: Series B*, 60:223–241.

Nath, J. S. and Jawanpuria, P. (2020). Statistical Optimal Transport posed as Learning Kernel Embedding. *arXiv preprint arXiv:2002.03179*.

Niles-Weed, J. and Rigollet, P. (2019). Estimation of Wasserstein distances in the Spiked Transport Model. *To appear, Bernoulli*.

Onken, D., Fung, S. W., Li, X., and Ruthotto, L. (2021). OT-Flow: Fast and Accurate Continuous Normalizing Flows via Optimal Transport. *arXiv preprint arXiv:2006.00104*.

Panaretos, V. M. and Zemel, Y. (2019). Statistical Aspects of Wasserstein Distances. *Annual Review of Statistics and Its Application*, 6:405–431.

Perrot, M., Courty, N., Flamary, R., and Habrard, A. (2016). Mapping Estimation for Discrete Optimal Transport. In *Advances in Neural Information Processing Systems 29*, pages 4197–4205.

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

Peyre, R. (2018). Comparison between W_2 distance and $\dot{H}^{-1}$ norm, and localisation of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 24:1489–1501.

Rachev, S. T. and Rüschendorf, L. (1998). *Mass Transportation Problems: Volume 1: Theory*. Probability and Its Applications. Springer-Verlag, New York.

Rakotomamonjy, A., Flamary, R., Gasso, G., Alaya, M. Z., Berar, M., and Courty, N. (2021). Optimal Transport for Conditional Domain Matching and Label Shift. *arXiv preprint arXiv:2006.08161*.

Read, A. L. (1999). Linear interpolation of histograms. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 425:357–360.

Redko, I., Courty, N., Flamary, R., and Tuia, D. (2019). Optimal transport for multi-source domain adaptation under target shift. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 849–858. PMLR.

Roe, J. (1999). *Elliptic Operators, Topology, and Asymptotic Methods*. CRC Press.

Rosenthal, H. (1972). On the span in L^p of sequences of independent random variables (II). In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability: Held at the Statistical Laboratory, University of California, June 21-July 18, 1970. Probability Theory*, pages 149–167. Univ of California Press.

Rosenthal, H. P. (1970). On the subspaces of L^p ($p > 2$) spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303.

Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*, volume 87. Birkhäuser.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., and Lander, E. S. (2019). Optimal-Transport Analysis of Single-Cell Gene Expression Identifies Developmental Trajectories in Reprogramming. *Cell*, 176:928–943.e22.

Seijo, E. and Sen, B. (2011). Nonparametric least squares estimation of a multivariate convex regression function. *The Annals of Statistics*, 39:1633–1657.

Shi, H., Drton, M., and Han, F. (2020). Distribution-Free Consistent Independence Tests via Center-Outward Ranks and Signs. *Journal of the American Statistical Association*, 0:1–16.

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

Staudt, T., Hundrieser, S., and Munk, A. (2022). On the Uniqueness of Kantorovich Potentials. *arXiv preprint arXiv:2201.08316*.

Steinerberger, S. (2016). Directional Poincare inequalities along mixing flows. *Arkiv för Matematik*, 54:555–569.

Talagrand, M. (1992). The Ajtai-Komlos-Tusnady matching theorem for general measures. In Dudley, R. M., Hahn, M. G., and Kuelbs, J., editors, *Probability in Banach Spaces, 8: Proceedings of the Eighth International Conference*, pages 39–54. Birkhäuser Boston, Boston, MA.

Tameling, C., Sommerfeld, M., and Munk, A. (2019). Empirical optimal transport on countable metric spaces: Distributional limits and statistical applications. *The Annals of Applied Probability*, 29:2744–2781.

Triebel, H. (2006). *Theory of Function Spaces III*. Monographs in Mathematics, Theory of Function Spaces. Birkhäuser Basel.

Tsybakov, A. B. (2008). *Introduction to Nonparametric Estimation*. Springer Science & Business Media.

Urbas, J. (1997). On the second boundary value problem for equations of Monge-Ampère type. *Journal für die reine und angewandte Mathematik*, 1997:115–124.

Vacher, A., Muzellec, B., Rudi, A., Bach, F., and Vialard, F.-X. (2021). A Dimension-free Computational Upper-bound for Smooth Optimal Transport Estimation. *arXiv preprint arXiv:2101.05380*.

van der Vaart, A. (1998). *Asymptotic statistics.* Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, UK ; New York, NY, USA.

van der Vaart, A. W. (2002). Semiparametric statistics. In Bernard, P., editor, *Lectures on Probability Theory and Statistics: École dÉté de Probabilités de Saint-Flour XXIX - 1999*, École d'Éé de Probabilités de Saint-Flour. Springer-Verlag, Berlin Heidelberg.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer Series in Statistics. Springer-Verlag, New York.

Vapnik, V. (2013). *The Nature of Statistical Learning Theory.* Springer Science & Business Media.

Villani, C. (2003). *Topics in Optimal Transportation.* American Mathematical Soc.

Villani, C. (2008). *Optimal Transport: Old and New*, volume 338. Springer Science & Business Media.

Wang, W., Ozolek, J. A., Slepčev, D., Lee, A. B., Chen, C., and Rohde, G. K. (2011). An optimal transportation approach for nuclear structure-based pathology. *IEEE Transactions on Medical Imaging*, 30:621–631.

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25:2620–2648.

Weed, J. and Berthet, Q. (2019). Estimation of smooth densities in Wasserstein distance. *To appear, Annals of Statistics.*

Zhu, J., Guha, A., Xu, M., Ma, Y., Lei, R., Loffredo, V., Nguyen, X., and Zhao, D. (2021). Functional Optimal Transport: Mapping Estimation and Domain Adaptation for Functional data. *arXiv preprint arXiv:2102.03895.*