arXiv:1902.01778v2 [math.ST] 6 Feb 2019

Estimation of smooth densities in Wasserstein distance

Jonathan Weed* and Quentin Berthet†

Massachusetts Institute of Technology
Cambridge University

Abstract. The Wasserstein distances are a set of metrics on probability distributions supported on $\mathbb{R}^d$ with applications throughout statistics and machine learning. Often, such distances are used in the context of variational problems, in which the statistician employs in place of an unknown measure a proxy constructed on the basis of independent samples. This raises the basic question of how well measures can be approximated in Wasserstein distance. While it is known that an empirical measure comprising i.i.d. samples is rate-optimal for general measures, no improved results were known for measures possessing smooth densities. We prove the first minimax rates for estimation of smooth densities for general Wasserstein distances, thereby showing how the curse of dimensionality can be alleviated for sufficiently regular measures. We also show how to construct discretely supported measures, suitable for computational purposes, which enjoy improved rates. Our approach is based on novel bounds between the Wasserstein distances and suitable Besov norms, which may be of independent interest.

AMS 2000 subject classifications: Primary 62G07.
Key words and phrases: Wasserstein distance, nonparametric density estimation, optimal transport.

1. INTRODUCTION

Wasserstein distances are an increasingly common tool in statistics and machine learning. Their popularity can be traced back to their empirical success on a wide range of practical problems (see, e.g., Peyré and Cuturi, 2017, for a survey) and a line of recent computational advances leading to much faster algorithms (Altschuler et al., 2017; Cuturi, 2013).

Wasserstein distances are a special case of the problem of *optimal transport*, one of the foundational problems of optimization (Kantorovitch, 1942; Monge, 1781), and a very important topic in analysis (Villani, 2008). This problem asks how one can transport mass with distribution μ to have another distribution ν, with minimal global transport cost. This problem also has the probabilistic interpretation, known as the Monge–Kantorovich formulation, of finding a coupling

*This work was supported in part by a Josephine de Kármán Fellowship.
†This work was supported in part by The Alan Turing Institute under the EPSRC grant EP/N510129/1.

minimizing a cost between random variables X and Y with given marginal distributions. The Wasserstein distance emerges as the minimum value of this problem, and creates a natural tool to compare distributions, with W_p corresponding to the $\|\cdot\|^p$ transport cost:

$$W_p^p(\mu,\nu) = \inf_{\pi \in \mathcal{M}(\mu,\nu)} \int \|x-y\|^p \mathrm{d}\pi(x,y)\,,$$

where the set $\mathcal{M}(\mu,\nu)$ denotes the set of joint measures with marginals μ and ν, respectively.

In many modern applications, a Wasserstein distance is used as a loss function in an optimization problem over measures. Solving such problems involves optimizing functionals of the form $\nu \mapsto W_p(\nu,\mu)$ where μ is unknown. Given n i.i.d. samples from μ, much of the statistics literature adopts the plug-in approach and focuses on using the empirical distribution $\hat{\mu}_n$ to obtain the estimated functional $\nu \mapsto W_p(\nu,\hat{\mu}_n)$. In this case, the rates of convergence are of order $n^{-1/d}$, and the sample size required for a particular precision is exponential in the dimension, a phenomenon known as the *curse of dimensionality*. Moreover, it is known that this exponential dependence is tight, in the sense that no better estimate is available in general (Singh and Póczos, 2018).

Our work adopts a different approach to show that the plug-in estimator is suboptimal for measures possessing a smooth density. Estimating the density of a distribution, based on independent samples, is one of the fundamental problems of statistics. The usual goal in these problems is to produce an estimate $\tilde{f}$ which is as close as possible to the unknown density f, measured either at one point of the sample space, or in L_p norm. In this line of work, f is usually assumed to belong to a large, nonparametric class defined via smoothness or regularity conditions, and typically the rates obtained in this setting show that sufficient smoothness can substantially mitigate the curse of dimensionality. This is the subject of a wide literature on *nonparametric density estimation*. In this work, we follow the same philosophy and derive similar rates for W_p distances, over Besov classes of densities $\mathcal{B}^s_{p,q}$. We likewise show that the smoothness parameter s improves the optimal exponent of n in the Wasserstein setting.

Algorithmic aspects are an important part of optimal transport problems. For practical applications, the proposed estimates must therefore also be computationally tractable. We describe a method to produce computationally tractable atomic estimators from any estimator that outperforms the empirical distribution, under minimal assumptions. We study the computational cost of this method, compared to the cost of using the empirical distribution with n atoms, and exhibit a trade-off between computational cost and statistical precision.

1.1 Prior work

The question of establishing minimax rates for estimation in Wasserstein distances has been examined in several recent works. Singh and Póczos (2018) established that, in the absence of smoothness assumptions, the empirical distribution $\hat{\mu}$ is rate optimal in a variety of examples. Their proof relies on a dyadic partitioning argument (see, e.g. Weed and Bach, 2018), and does not appear to extend to the smooth case. Closer to our setting, under a smoothness assumption on the density of μ, Liang (2017) and Singh et al. (2018) showed minimax rates of convergence for the Wasserstein-1 distance. To obtain these rates, these works focus on the dual form of W_1:

$$W_1(\mu, \nu) = \sup_{f \in \mathrm{Lip}} \int f(\mathrm{d}\mu - \mathrm{d}\nu)\,,$$

where the supremum is taken over all 1-Lipschitz functions. This dual formulation puts the Wasserstein-1 distance into the category of *integral probability metrics* (Müller, 1997), for which both Liang (2017) and Singh et al. (2018) obtain general results. It has been shown that choosing functions which are smoother than Lipschitz in this definition can result in improved rates of convergence for empirical measures (Kloeckner, 2018). Crucially, the metric W_p for $p > 1$ is *not* an integral probability metric. Establishing sharp rates for general Wasserstein distances therefore requires different techniques. A separate line of work has focused instead on modifying the definition of the Wasserstein distance to include a regularizing term based on the mutual information of the coupling. It has been shown that this definition enjoys improved convergence rates relative to the unregularized version (Genevay et al., 2018).

Our proofs rely on establishing control of Wassserstein distances by Besov norms of negative smoothness. Similar results have been obtained elsewhere under different conditions. Shirdhonkar and Jacobs (2008) showed that the optimal transportation distance with cost $\|\cdot\|^p$ for $0 < p < 1$ can be characterized explicitly via an expression involving wavelet coefficients, which implies that these distances agree with a particular Besov norm (see 2.1.2). Loeper (2006) (see also Maury et al., 2010) showed that the Wasserstein-2 distance between measures with densities bounded above dominates a negative Sobolev norm, and Peyre (2018) extended this result to show that W_2 is in fact *equivalent* to such a norm when the densities are in addition bounded below. To our knowledge, ours is the first result to establish a connection to Besov norms of negative smoothness and general Wasserstein distances.

The use of wavelet estimators for density estimation has a long history in nonparametric statistics (Donoho et al., 1996; Doukhan and León, 1990; Härdle et al., 1998; Kerkyacharian and Picard, 1992; Walter, 1992). However, while wavelets have been used for computational purposes in the optimal transport community (Chen et al., 2012; Dominitz et al., 2008; Rabin et al., 2011; Shirdhonkar and Jacobs, 2008), the statistical properties of wavelet estimators with respect to Wasserstein distances have remained largely unexplored.

2. MAIN RESULTS

2.1 Problem description and preliminaries

2.1.1 Nonparametric density estimation in Wasserstein distance Our observation consists of an i.i.d. sample of size n drawn from a probability measure μ_f on $\mathbb{R}^d$ with smooth density f. Our goal is to compute an estimator $\tilde{\mu}_n$ that is close to μ_f in expected Wasserstein distance. As noted above, such an estimator can serve as a proxy for μ_f in statistical and computational applications. While estimation of the density f in norms such as L_p is a well studied problem in nonparametric statistics (Tsybakov, 2009), such estimates do not readily lend themselves to guarantees in Wasserstein distance.

For technical reasons, we restrict ourselves to the case of measures supported on a compact set $\Omega \subseteq \mathbb{R}^d$. **We focus throughout on $\Omega := [0,1]^d$.** The extension to other rectangular sets is straightforward; however, non-rectangular sets present nontrivial challenges, which we do not explore here.

2.1.2 Wavelets and Besov spaces We direct the reader to Härdle et al. (1998) and Meyer (1990) for an introduction to the theory of wavelets. In brief, we assume the existence of sets Φ and Ψ_j for $j \geq 0$ of functions in $L_2(\Omega)$ satisfying the standard requirements of a wavelet basis. (See Appendix C for our precise assumptions.)

Wavelets can be used to characterize the Besov spaces $\mathcal{B}^s_{p,q}(\Omega)$. We follow the approach of Cohen (2003) for defining such spaces on bounded domains. Suppose $s > 0$ and $p, q \geq 1$, and let $n > s$ be an integer. Given $h \in \mathbb{R}^d$, set

$$\Delta^1_h f(x) := f(x+h) - f(x)$$
$$\Delta^k_h f(x) := \Delta^1_h(\Delta^{k-1}_h) f(x) \quad \forall 1 < k \leq n\,,$$

where these functions are defined on $\Omega_{h,n} := \{x \in \Omega : x + nh \in \Omega\}$. For $t > 0$, we then define

$$\omega_n(f,t)_p = \sup_{\|h\| \leq t} \|\Delta^n_h f\|_{L_p(\Omega_{h,n})}\,.$$

The function ω_n measures the order-n smoothness of f in L_p. Finally, we define the space $\mathcal{B}^s_{p,q}(\Omega)$ to be the set of functions for which the quantity

$$\|f\|'_{\mathcal{B}^s_{p,q}} := \|f\|_{L_p} + \left\|(2^{sj}\omega_n(f, 2^{-j})_p)_{j \geq 0}\right\|_{\ell_q}$$

is finite.

Assuming that the elements of Φ and Ψ_j have r continuous derivatives for $r > s$ and that polynomials of degree up to $\lfloor s \rfloor$ lie in the span of Φ, the norm $\|\cdot\|'_{\mathcal{B}^s_{p,q}}$ is equivalent to a sequence norm based on wavelet coefficients. Given $f \in L_p(\Omega)$, denote by $\alpha = \{\alpha_\phi\}_{\phi \in \Phi}$ the vector defined by $\alpha_\phi := \int f(x)\phi(x)\,\mathrm{d}x$ and for $j \geq 0$ denote by $\beta_j = \{\beta_\psi\}_{\psi \in \Psi_j}$ the vector whose entries are given by $\beta_\psi := \int f(x)\psi(x)\,\mathrm{d}x$. Then $\|\cdot\|'_{\mathcal{B}^s_{p,q}}$ is equivalent to $\|\cdot\|_{\mathcal{B}^s_{p,q}}$ defined by

$$\|f\|_{\mathcal{B}^s_{p,q}} := \|\alpha\|_{\ell_p} + \left\|2^{js}2^{dj(\frac{1}{2}-\frac{1}{p})}\|\beta_j\|_{\ell_p}\right\|_{\ell_q}\,. \tag{1}$$

This expression can then be used directly to define a norm when $s < 0$ (see Cohen, 2003, Theorem 3.8.1), as long as the elements of Φ and Ψ_j have r continuous derivatives for $r > |s|$ and polynomials of degree up to $\lfloor |s| \rfloor$ lie in the span of Φ. In what follows, we therefore adopt (1) as our primary definition and assume throughout that the wavelet system has sufficient regularity that the equivalence of $\|\cdot\|'_{\mathcal{B}^s_{p,q}}$ and $\|\cdot\|_{\mathcal{B}^s_{p,q}}$ holds.

2.1.3 Notation The quantities C and c will refer to constants whose value may change from line to line. **All constants throughout may depend on the choice of wavelet system and the dimension.** Since we are interested in establishing optimal rates of decay with respect to the exponent (i.e., finding γ such that the rate $n^{-\gamma}$ holds), we leave finer control on dimension-dependent constants to future work. We freely use the notation $a \lesssim b$ to indicate that there exists a constant C for which $a \leq Cb$ holds. Again, such constants may depend on the multiresolution and dimension. The notation $a \asymp b$ indicates that $a \lesssim b$ and $b \lesssim a$.

We set denote by $\mathcal{D}(\Omega)$ the set of probability density functions on Ω, and by $\mathcal{P}$ the set of all probability measures on $\mathbb{R}^d$. Given a density f, we denote by μ_f the associated measure. We write $a \wedge b := \min\{a, b\}$ and $a \vee b := \max\{a, b\}$ for real numbers a and b.

2.2 Minimax estimation of smooth densities

In this section, we give our main statistical results on the problem of estimating densities in Wasserstein distance. These results reveal several striking phenomena: (i) the minimax rate of estimation can improve significantly for smooth densities, and (ii) the optimal rates depend strongly on whether the density in question is bounded away from 0. Indeed, we show that the optimal rate for general densities is strictly worse than the corresponding rate for densities bounded below, no matter the smoothness. While the first phenomenon is well known in non-parametric statistics, the second phenomenon does not occur in classical density estimation problems. As we explore further below, this behavior is fundamental to the Wasserstein distances.

We define two classes of probability densities on Ω. Given $m, L > 0$, set

$$\begin{aligned} \mathcal{B}^s_{p,q}(L) &:= \{f \in L_p(\Omega) : \|f\|_{\mathcal{B}^s_{p,q}} \leq L, \int f(x)\,\mathrm{d}x = 1, f \geq 0\} \\ \mathcal{B}^s_{p,q}(L; m) &:= \mathcal{B}^s_{p,q}(L) \cap \{f : f \geq m\}\,. \end{aligned}$$

We note that if s is sufficiently large and L is sufficiently small then in fact $\mathcal{B}^s_{p,q}(L) \subseteq \mathcal{B}^s_{p,q}(L; m)$ for m a constant. We assume throughout that $m < 1$, since when $m \geq 1$, the class $\mathcal{B}^s_{p,q}(L; m)$ is trivial.

2.2.1 Bounded densities Our first result gives an upper bound on the rate of estimation for functions in $\mathcal{B}^s_{p,q}(L; m)$.

THEOREM 1. *For any $m > 0$, $s \geq 0$, and $p \in [1, \infty)$, there exists an estimator $\hat{f}$ such that for any $p' \geq p$ and $q \geq 1$, the estimator satisfies*

$$\sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(\mu_f, \mu_{\hat{f}}) \lesssim \begin{cases} n^{-\frac{1+s}{d+2s}} & d \geq 3 \\ n^{-1/2} \log n & d = 2 \\ n^{-1/2} & d = 1\,. \end{cases}$$

The upper bound in Theorem 1 is achieved by a wavelet estimator. As s ranges between 0 and ∞, the upper bound interpolates between the dimension-dependent rate $n^{-1/d}$ and the fully parametric rate $n^{-1/2}$.

Our lower bounds nearly match the upper bounds proved in Theorem 1, up to a logarithmic factor in the $d = 2$ case.

THEOREM 2. *For any $p, p', q \geq 1$, and $s \geq 0$,*

$$\inf_{\tilde{\mu} \in \mathcal{P}} \sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_p(\mu_f, \tilde{\mu}) \gtrsim \begin{cases} n^{-\frac{1+s}{d+2s}} & d \geq 3 \\ n^{-1/2} & d \leq 2\,, \end{cases}$$

where the infimum is taken over all estimators $\tilde{\mu}$ based on n observations.

The rates Theorems 1 and 2 evince two phenomena not present in L_p density estimation. First, in low dimension ($d \le 2$), the rates are independent of s, so that there is no benefit to smoothness. Second, even in the case when $s = 0$, nontrivial estimation is possible at the rate $n^{-1/d}$ when $d \ge 3$.

Our bounds are obtained via the following technical result, which establishes a connection between Wasserstein distances and Besov norms of negative smoothness.

THEOREM 3. *Let $p \in [1, \infty)$. If f, g are two densities in $L_p([0,1]^d)$ satisfying $m \le f, g \le M$ for $m, M > 0$, then*

$$M^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,\infty}} \lesssim W_p(\mu_f, \mu_g) \lesssim m^{-1/p'} \|f - g\|_{\mathcal{B}^{-1}_{p,1}},$$

where $\frac{1}{p} + \frac{1}{p'} = 1$.

Theorem 3 can be viewed as a partial extension of the dual formulation of W_1 to W_p for $p > 1$. Indeed, the inclusions $\mathcal{B}^1_{\infty,1} \subseteq \mathrm{Lip} \subseteq \mathcal{B}^1_{\infty,\infty}$, where Lip is the space of bounded Lipschitz functions, imply $\|f - g\|_{\mathcal{B}^{-1}_{1,\infty}} \lesssim W_1(\mu_f, \mu_g) \lesssim \|f - g\|_{\mathcal{B}^{-1}_{1,1}}$. Theorem 3 establishes the analogous result when $p > 1$, but only when the densities f and g are bounded. A proof of this theorem appears in Section 3.

We prove Theorems 1 and 2 in Section 4.

2.2.2 Unbounded densities Surprisingly, the density estimation problem over the class $\mathcal{B}^s_{p,q}(L)$ is strictly harder than the corresponding problem over $\mathcal{B}^s_{p,q}(L; m)$, even under a smoothness assumption. We prove the following lower bound.

THEOREM 4. *For any $p, p', q \ge 1$, and $s \ge 0$, if L is a sufficiently large constant, then*

$$\inf_{\tilde{\mu} \in \mathcal{P}} \sup_{f \in \mathcal{B}^s_{p',q}(L)} W_p(\mu_f, \tilde{\mu}) \gtrsim n^{-\frac{1+s/p}{d+s}} \vee n^{-1/2p},$$

where the infimum is taken over all estimators $\tilde{\mu}$ based on n observations.

Note that, when $p \ge 2$, this rate is worse than the upper bound given in Theorem 1 for all $s > 0$ and $d \ge 1$. This establishes that the class of densities bounded from below is strictly easier to estimate than the class of all densities, for all nontrivial smoothness parameters.

When $s \in [0, 1)$, we can also prove an upper bound. While this bound does not match the lower bound above, it nevertheless verifies qualitatively the behavior present in Theorem 4. Moreover, the estimator we construct is a *histogram*. This property enables the use of such an estimator in practical applications. We take up this point in Section 6.

THEOREM 5. *For any $s \in [0, 1)$, there exists a histogram estimator $\hat{f}$ such that for any $1 \le p \le p' < \infty$ and $1 \le q \le \infty$, the estimator satisfies*

$$\sup_{f \in \mathcal{B}^s_{p',q}(L)} \mathbb{E} W_p(\mu_f, \mu_{\hat{f}}) \lesssim \begin{cases} n^{-\frac{1+s/p}{d+2s}} & d > 2p \\ n^{-\frac{1}{2p}} \log n & d = 2p \\ n^{-\frac{1}{2p}} & d < 2p\,. \end{cases}$$

The proofs of both Theorems 5 and 4 appear in Section 5.

2.3 Computational aspects of smooth density estimation

In many computational applications, it is significantly simpler to work with discrete measures supported on a finite number of points, since in general there is no closed form expression for the Wasserstein distance between continuous measures. Unfortunately, the estimators presented in Section 2.2 are not of this form, so it is unclear whether smoothness of the underlying measure can be exploited in applications. However, a simple argument shows that optimal rates can be achieved by *resampling* from the smooth estimator we construct to obtain a discrete distribution supported on $M \geq n$ points which achieves an accelerated rate for $s \in [0, 1)$. We extract one simple result in this direction.

THEOREM 6. *For any $s \in [0, 1)$, there exists an estimator $\bar{\mu}_{n,M}$, supported on $M = o(n^2)$ points, such that for any $1 \leq p \leq p' < \infty$ and $1 \leq q \leq \infty$, the estimator enjoys the same rate as in Theorem 5. Moreover, $\bar{\mu}_{n,M}$ can be computed in time $O(M)$.*

Additional computational considerations along with a proof of Theorem 6 appear in Section 6.

3. CONTROLLING THE WASSERSTEIN DISTANCE BY BESOV NORMS

The main goal of this section is a proof of Theorem 3, which establishes that the Wasserstein distance between two measures on $\Omega = [0, 1]^d$ can be controlled by a Besov norm of the difference in their densities *as long as their densities are bounded above and below*. We also establish that no analogous result can hold for arbitrary densities. While we give upper and lower bounds, the Besov norms appearing in the two bounds do not agree. We do not know whether under some conditions the W_p distance is in fact *equivalent* to a particular Besov norm $\|\cdot\|_{\mathcal{B}^{-1}_{p,q}}$ for some $q \in [1, +\infty]$.

The results of this section are closely results to results of Shirdhonkar and Jacobs (2008) and Peyre (2018), who established similar results for $p < 1$ and $p = 2$, respectively. In Section 3.1, we show the upper bound of Theorem 3, and in Section 3.2 we show that no similar bound can exist once the assumption that the density is bounded away from zero is relaxed. The lower bound is proved in Section 3.3.

3.1 Upper bound

Let f and g be probability densities in $L_p(\Omega)$ for $p \in [1, \infty)$ with the following wavelet expansions.

$$
\begin{aligned}
f &= \sum_{\phi \in \Phi} \alpha_\phi \phi + \sum_{j \geq 0} \sum_{\psi \in \Psi_j} \beta_\psi \psi \\
g &= \sum_{\phi \in \Phi} \alpha'_\phi \phi + \sum_{j \geq 0} \sum_{\psi \in \Psi_j} \beta'_\psi \psi \,,
\end{aligned}
\tag{2}
$$

where we assume (see Assumption 2 in Appendix C) that constant functions lie in the span of Φ. For the upper bound, we do not need to assume any additional regularity—in particular, Proposition 1 holds for the Haar wavelet basis (see

Triebel, 2010). By definition, the expansions in (2) hold in L_2, but in fact convergence also holds in L_p assuming that $f, g \in L_p(\Omega)$ (Härdle et al., 1998, Remark 8.4).

We prove the following proposition in Appendix A.

PROPOSITION 1. *Let $1 \le p < \infty$. If $f, g \ge m$ on $[0,1]^d$, then*

$$W_p(\mu_f, \mu_g) \lesssim m^{-1/p'} \left(\|\alpha - \alpha'\|_{\ell_p} + \left\| 2^{js} 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta_j\|_{\ell_p} \right\|_{\ell_1} \right) ,$$

where $\frac{1}{p} + \frac{1}{p'} = 1$.

3.2 Densities not bounded below

We now show that no statement like Proposition 1 can hold for densities not bounded below. Indeed, in this case, under mild assumptions, it is impossible to control $W_p(\mu_f, \mu_g)$ by any function norm when $p > 1$. This stands in sharp contrast to the fact that, when $p = 1$, the dual formulation of W_1 implies that the Wasserstein distance is such a norm.

THEOREM 7. *Let $\|\cdot\|$ be any norm on functions on Ω, and suppose that there exists a function h in $L_1(\Omega)$, not identically zero, satisfying*

- $\int_\Omega h \, \mathrm{d}x = 0$
- $\|h\| < \infty$
- *The sets $\overline{\{h > 0\}}$ and $\overline{\{h < 0\}}$ are disjoint.*

Then for any $p > 1$,

$$\sup_{f,g \in \mathcal{D}(\Omega)} \frac{W_p(\mu_f, \mu_g)}{\|f - g\|} = \infty .$$

A proof appears in Appendix A.

3.3 Lower bound

We can prove a lower bound similar to Proposition 1 when f and g are bounded above. Unlike the assumption that the densities are bounded below required for Proposition 1, this assumption is relatively benign, insofar as it holds automatically for continuous densities on $[0,1]^d$. For Proposition 2, we require the wavelets in (2) to possess at least one continuous derivative (see Assumption 2 in Appendix C). A proof appears in Appendix A.

PROPOSITION 2. *Let $1 \le p < \infty$. If $f, g \le M$ on $[0,1]^d$, then*

$$W_p(\mu_f, \mu_g) \gtrsim M^{-1/p'} \left(\|\alpha - \alpha'\|_{\ell_p} + \left\| 2^{js} 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta_j\|_{\ell_p} \right\|_{\ell_\infty} \right) ,$$

where $\frac{1}{p} + \frac{1}{p'} = 1$.

4. WAVELET ESTIMATION IN FOR BOUNDED DENSITIES

In this section, we employ the results of Section 3 to prove Theorems 1 and 2. We show that the minimax rate over $\mathcal{B}^s_{p,q}(L; m)$ can be achieved by a wavelet estimator. We do not address the issue of adaptivity (to m or to the smoothness s) here, but note that it can be handled by known techniques in wavelet density estimation (Donoho and Johnstone, 1995).

4.1 Upper bound

To prove Theorem 1, we introduce the following estimator based on a wavelet expansion of regularity $r > \max\{s, 1\}$ (see Assumption 2 in Appendix C) truncated to level J, for some $J \geq 0$ to be chosen. Set

$$\tilde{\alpha}_\phi := \frac{1}{n}\sum_{i=1}^n \phi(X_i) \quad \phi \in \Phi$$
$$\tilde{\beta}_\psi := \frac{1}{n}\sum_{i=1}^n \psi(X_i) \quad \psi \in \Psi_j, 0 \leq j \leq J$$

and let $\tilde{f} := \sum_{\phi\in\Phi} \tilde{\alpha}_\phi \phi + \sum_{0\leq j\leq J}\sum_{\psi\in\Psi_j} \tilde{\beta}_\psi \psi$. While such an estimator can already yield optimal rates in L_p (Kerkyacharian and Picard, 1992), $\tilde{f}$ may fail to be a probability density, in which case the quantity $W_p(\mu_f, \mu_{\tilde{f}})$ is undefined. We therefore focus on the estimator

$$\hat{f} := \min_{g\in\mathcal{D}(m)} \|g - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}},$$

where $\mathcal{D}(m)$ is the set of probability densities on Ω bounded below by m. By construction, $\hat{f}$ is a density, so that $W_p(\mu_f, \mu_{\hat{f}})$ is meaningful. The proof of Theorem 1 now follows from standard facts in wavelet density estimation. It appears in Appendix A

4.2 Lower bound

Our lower bound follows almost directly from the bound proved by Kerkyacharian and Picard (1992) to establish minimax rates for density estimation in L_p over Besov spaces. We defer the proof to Appendix A.

5. GENERAL SMOOTH DENSITIES

In this section, we give results for general smooth densities (Theorems 4 and 5). Our main result is a lower bound showing that the rate of estimation over the class $\mathcal{B}^s_{p',q}(L)$ is strictly worse than the rate over the class $\mathcal{B}^s_{p',q}(L; m)$ when L is large enough that $\mathcal{B}^s_{p',q}(L) \not\subseteq \mathcal{B}^s_{p',q}(L; m)$. We also give an upper bound when the smoothness parameter s is less than 1, which nearly matches our lower bounds.

5.1 Lower bounds

We assume that L is large enough that $\mathcal{B}^s_{p',q}(L)$ contains a function g_0 whose support lies entirely inside $(0, 1/3)^d$. It is easy to see that this goal is indeed achievable by choosing g_0 to be suitable compactly supported smooth bump functions, as long as L is a large enough constant.

The lower bound is based on the following fundamental lemma, which gives a lower bound on the Wasserstein distances for a pair of measures with disconnected support.

LEMMA 1. *Let μ and ν be measures on $\mathbb{R}^d$. Suppose there exist two compact sets S and T such that $d(S, T) \geq c$ and such that the supports of μ and ν lie in $S \cup T$. Then*

$$W_p(\mu, \nu) \geq c|\mu(S) - \nu(S)|^{1/p}$$

PROOF. Assume without loss of generality that $\mu(S) \geq \nu(S)$. Then any coupling between μ and ν must assign mass at least $\mu(S) - \nu(S)$ to $S \times T$, so that $W_p^P(\mu, \nu) \geq c^p \lambda$. $\square$

The proof of Theorem 4 boils down to applying Lemma 1 to appropriately chosen measures. We defer the proof to Appendix A.

5.2 Upper bounds

We now show how to prove an upper bound for general densities that achieves the rate $n^{-\frac{1+s/p}{d+2s}}$ for $s < 1$. The construction is based on the following observation. For $j \geq 0$, let $\mathcal{Q} := \bigcup_{j \geq 0} \mathcal{Q}_j$ be the dyadic decomposition of $[0,1]^d$, where $\mathcal{Q}_j$ consists of a partition of $[0,1]^d$ into cubes with sides of length 2^{-j}. If μ and ν are two measures on $[0,1]^d$, then such a decomposition can be used to obtain an upper bound on the Wasserstein distance between μ and ν (see, e.g., Weed and Bach, 2018, Proposition 1):

$$W_p^p(\mu, \nu) \lesssim \sum_{j \geq 0} 2^{-jp} \sum_{Q \in \mathcal{Q}_j} |\mu(Q) - \nu(Q)| \,. \tag{3}$$

When μ and ν possess densities f and g, respectively, the expression on the right side of the above inequality is an expansion of $f - g$ with respect to the Haar wavelet basis.

We can therefore again employ a wavelet estimator using the Haar wavelet, as in the proof of Theorem 1. The definition of the Haar wavelet implies that such an estimator is in fact a histogram, that is, its density is constant on each cube in $\mathcal{Q}_J$, where J represents the level at which the wavelet expansion is truncated. A full proof appears in Appendix A.

6. COMPUTATIONAL ASPECTS

One of the motivations for this line of work is found in applications of optimal transport techniques for data analysis and machine learning, with unknown distributions and access to an independent sample of size n. Many so-called *variational Wasserstein problems* involve the problem of minimizing a functional $F : \nu \mapsto W_p(\nu, \mu)$ with unknown μ. These problems, such as minimum Kantorovich estimators (Bassetti et al., 2006) and Wasserstein barycenters (Agueh and Carlier, 2011), are increasingly common in practical applications (Peyré and Cuturi, 2017), especially when the minimization is taken over a parametric class, with $\nu = \nu_\theta$ for $\theta \in \Theta$.

Solving variational Wasserstein problems in practice requires first obtaining an empirical estimate of the functional F on the basis of data drawn from μ, and then writing the resulting optimization problem in a computationally tractable form. The first issue is typically addressed by obtaining an estimator $\tilde{\mu}_n$ of μ and then estimating the functional via the plug-in principle. Indeed, the triangle inequality implies that

$$\sup_{\nu \in \mathcal{P}} |W_p(\nu, \mu) - W_p(\nu, \tilde{\mu}_n)| = W_p(\mu, \tilde{\mu}_n) \,,$$

where equality is achieved at $\mu = \nu$. Following this approach, guarantees in Wasserstein distance between μ and the estimator $\tilde{\mu}_n$ therefore yield uniform deviation bounds for these functionals over the set of all probability measures on $\mathbb{R}^d$.

To solve the resulting optimization problem, finite discretizations are often taken for ν and $\tilde{\mu}_n$ to render the resulting problem amenable to discrete optimization techniques (Altschuler et al., 2017; Cuturi, 2013). For this reason, the estimator $\tilde{\mu}_n$ is often taken to be the empirical distribution $\hat{\mu}_n$, since this measure is a finitely supported measure and enjoys the rate

$$\mathbb{E}W_p(\mu, \hat{\mu}_n) \lesssim n^{-1/d},$$

as long as $d > 2p$, which is minimax optimal over the class of compactly supported probability measures (Singh and Póczos, 2018).

However, Sections 4 and 5 establish that under natural regularity assumptions for μ, estimators based on density estimation statistically outperform the empirical distribution. Focusing on the regime $d > 2p$, Theorems 1 and 5 yield guarantees of the form

$$\mathbb{E}W_p(\mu, \tilde{\mu}_n) \lesssim n^{-\gamma^*(s)/d},$$

where $\gamma^*(s) \geq 1$ increases as the smoothness of μ increases.

These results are summarized in the following table, highlighting that the optimal exponent $\gamma^*(s)/d$ interpolates between $1/d$ (for $s = 0$) and $1/2p$ (for s going to ∞). The value $s = 1$ is of special interest, as it corresponds to the maximum smoothness which can be exploited by a histogram estimator, which is most relevant for computational aspects.

Nonparametric class	optimal $\gamma^*(s)$	$\gamma^*(0)/d$	$\gamma^*(\infty)/d$	$\gamma^*(1)/d$
$\mathcal{B}^s_{p,q}(L)$	$\frac{1+s/p}{1+2s/d}$	$1/d$	$1/2p$	$\frac{1+1/p}{d+2}$

Note that for $d > 2p$, the exponent $\gamma^*(s)$ is greater than 1. In the light of these results, there is an apparent tension between two objectives: statistical precision and computational efficiency. On the one hand, using the empirical measure as an estimator of the unknown distribution permits efficient computation of Wasserstein distances: the optimal transport problem reduces to a linear program in finite dimension. The statistical performance of this approximation can however be suboptimal: for smooth densities the lower bound in $n^{-1/d}$ applies to all n-atomic distributions (Dudley, 1969), which is strictly worse than the rate appearing in the table above. On the other hand, wavelet estimators can attain minimax-optimal statistical precision, but even in the simple case of histograms (piecewise-constant densities), there is no explicit or simple way to solve optimal transport problems involving such measures.

We therefore propose a procedure to leverage the regularity of the distribution, and to handle our proposed estimators in a computationally efficient manner. The idea is to exploit the best of both worlds, by creating an atomic measure with $M \geq n$ atoms, based on a density estimator. Such measures statistically outperform the empirical measure, and optimal transport problems can be explicitly solved on these measures. If it is possible to efficiently sample from one of these estimators, it is always possible to extract an atomic distribution out of it, as described in the following

Definition 1. Let $\tilde{\mu}_n$ be a probability measure from which one can efficiently sample points, and let $Z_1, \dots, Z_M$ be an i.i.d. sample from $\tilde{\mu}_n$. The *estimator resample distribution* is $\bar{\mu}_{n,M} := \frac{1}{M}\sum_{i=1}^M \delta_{Z_i}$.

The distribution $\bar{\mu}_{n,M}$ is "simply" the empirical distribution of a sample of size M from a distribution. However, we retain n in the notation, to highlight that the Z_i's are themselves drawn from an estimator based on a sample of size n from an unknown μ. We recall the following result for compactly supported distributions.

Proposition 3 (Fournier and Guillin, 2015). *For $d > 2p$, the estimator resample distribution $\bar{\mu}_{n,M}$ satisfies*

$$\mathbb{E}W_p(\tilde{\mu}_n, \bar{\mu}_{n,M}) \lesssim M^{-1/d}\,.$$

As a consequence of this result, resampling from the estimated distribution yields an atomic measure as close in Wasserstein distance to the original estimator as desired, since M can be chosen by the statistician. This approach shares some similarities with the concept of the parametric bootstrap (Wasserman, 2004), where a sample of the same size is drawn from an estimator. Conceptually, this is however quite different: our focus is not on inference and we do not aim to to create a proxy of our original sample sharing similar probabilistic properties. The resample of size M is created to approximate, up to statistical precision, the estimate of μ by an atomic measure, as in finite element methods in numerical analysis. It is naturally only useful to chose an approximation error $M^{-1/d}$ of the same order as the estimation error, as seen in the following.

Corollary 1. *Assume $d > 2p$. Let μ be in a nonparametric class such that there exists an estimator $\tilde{\mu}_n$ from which one can efficiently sample, and such that $\mathbb{E}W_p(\tilde{\mu}_n, \mu) \lesssim n^{-\gamma^*/d}$. For any $\gamma \in [1, \gamma^*]$, the estimator resample distribution $\bar{\mu}_{n,M}$ with $M = n^\gamma$ satisfies*

$$\mathbb{E}W_p(\mu, \bar{\mu}_{n,M}) \lesssim n^{-\gamma/d}\,.$$

The proofs of this corollary and Theorem 6, which follows directly, appear in Appendix A.

In some examples, the choice of γ^* can be left to the practitioner. For our estimators, this corresponds to choosing the depth of the wavelet decomposition. Taking piecewise constant estimators (histograms) limits the exponent γ to $\gamma^* = \gamma^*(1)$. In any case, it is also possible to chose $M = n^\gamma$ for $\gamma \in (1, \gamma^*(1)]$, and let the approximation error $n^{-\gamma/d}$ dominate the statistical error $n^{-\gamma^*(1)/d}$.

Using the estimator resample distribution $\bar{\mu}_{n,M}$ instead of $\hat{\mu}_n$ requires solving optimal transport problems of size $M = n^\gamma$ instead of n. This naturally increases the computational cost. This motivates the question of quantifying the statistical and computational tradeoffs of our proposal. The dependency of the algorithmic cost of solving optimal transportation problems on the size of the distribution is the subject of a large literature (see Peyré and Cuturi, 2017), from which we extract a simple bound.

PROPOSITION 4 (Altschuler et al., 2017; Dvurechensky et al., 2018). *Given two distributions α and β supported on at most M atoms on a set of diameter 1, an additive approximation to $W_p^p(\alpha,\beta)$ of accuracy ε can be computed in time $O(M^2\log(M)/\varepsilon^2)$.*

The following describes the interplay between statistical precision and computational efficiency for estimating the Wasserstein distance between distributions. We prove this theorem in Appendix A.

THEOREM 8. *Let μ be in a nonparametric class such that there exists an estimator $\tilde{\mu}_n$ from which one can efficiently sample, and such that*

$$\mathbb{E}W_p(\tilde{\mu}_n,\mu) \lesssim n^{-\gamma^*/d}$$

Given a sample of size n from μ and known ν, for any $\gamma \in [1,\gamma^*]$ an estimate $\tilde{W}_{p,n}$ of $W_p(\nu,\mu)$ satisfying

$$\mathbb{E}|\tilde{W}_{p,n} - W_p(\nu,\mu)| \lesssim n^{-\gamma/d}$$

can be computed in time $O\big(n^{\gamma(2+2p/d)}\log(n)\big)$.

Taking $\gamma = 1$, $\tilde{\mu}_n = \hat{\mu}_n$ and $M = n$ with $\bar{\mu}_{n,M} = \hat{\mu}_n$ (without resampling) is always possible. It yields an algorithm that outputs a $n^{-1/d}$ approximation in time $\tilde{O}\big(n^{2+2p/d}\big)$. However, whenever another estimator $\tilde{\mu}_n$ with precision $n^{-\gamma/d}$ exists for $\gamma \in (1,\gamma^*(s)]$, it is possible to obtain a better approximation with error $n^{-\gamma/d}$ in time of order $\tilde{O}(n^{\gamma(2+2p/d)}$. This quantifies the computational cost for added statistical precision. We summarize these results in the following table, for $\gamma^* = \gamma^*(1)$ for histogram estimators (from which it is easy to sample). The parameter γ can be taken in the full range from 1 to $\gamma^*(1)$.

γ	precision $n^{-\gamma/d}$	$M = n^\gamma$	time $n^{\gamma(2+2/d)}\log(n)$
$\gamma = 1$	$n^{-\frac{1}{d}}$	n	$n^{2+2p/d}\log(n)$
$\gamma^*(1) = \frac{1+1/p}{1+2/d}$	$n^{-\frac{1+1/p}{d+2}}$	$n^{\frac{1+1/p}{1+2/d}}$	$n^{2(1+1/p)(1+\frac{p-2}{d+2})}\log(n)$

In the high-dimensional limit, we obtain that a histogram estimator can improve the exponent in the precision by a factor γ of nearly $1+1/p$ at the price of increasing the exponent in the running time by nearly the same factor. The choice of M, which can be left to the statistician, determines the value of γ.

REFERENCES

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

J. Altschuler, J. Weed, and P. Rigollet. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 1961–1971, 2017.

F. Bassetti, A. Bodini, and E. Regazzini. On minimum Kantorovich distance estimators. *Statist. Probab. Lett.*, 76(12):1298–1302, 2006.

J.-D. Benamou and Y. Brenier. A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem. *Numer. Math.*, 84(3):375–393, 2000.

Y. Brenier. Extended Monge-Kantorovich theory. In *Optimal transportation and applications (Martina Franca, 2001)*, volume 1813 of *Lecture Notes in Math.*, pages 91–121. Springer, Berlin, 2003.

G. Chen, M. A. Iwen, S. Chin, and M. Maggioni. A fast multiscale framework for data in high-dimensions: Measure estimation, anomaly detection, and compressive measurements. In *2012 Visual Communications and Image Processing, VCIP 2012, San Diego, CA, USA, November 27-30, 2012*, pages 1–6, 2012.

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

A. Cohen, I. Daubechies, and P. Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1):54–81, 1993.

M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 2292–2300, 2013.

A. Dominitz, S. Angenent, and A. Tannenbaum. On the computation of optimal transport maps using gradient flows and multiresolution analysis. In *Recent advances in learning and control*, volume 371 of *Lect. Notes Control Inf. Sci.*, pages 65–78. Springer, London, 2008.

D. L. Donoho and I. M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.*, 90(432):1200–1224, 1995.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

P. Doukhan and J. R. León. Déviation quadratique d'estimateurs de densité par projections orthogonales. *C. R. Acad. Sci. Paris Sér. I Math.*, 310(6):425–430, 1990.

R. M. Dudley. The speed of mean Glivenko-Cantelli convergence. *Ann. Math. Statist*, 40:40–50, 1969.

P. Dvurechensky, A. Gasnikov, and A. Kroshnin. Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn's algorithm. *arXiv preprint arXiv:1802.04367*, 2018.

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

A. Genevay, L. Chizat, F. Bach, M. Cuturi, and G. Peyré. Sample complexity of Sinkhorn divergences. *arXiv preprint arXiv:1810.02733*, 2018.

W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelets, approximation, and statistical applications*, volume 129 of *Lecture Notes in Statistics*. Springer-Verlag, New York, 1998.

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

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

B. Kloeckner. Empirical measures: regularity is a counter-curse to dimensionality. *arXiv preprint arXiv:1802.04038*, 2018.

R. A. Kronmal and A. V. Peterson, Jr. On the alias method for generating random variables from a discrete distribution. *Amer. Statist.*, 33(4):214–218, 1979.

L. LeCam. Convergence of estimates under dimensionality restrictions. *Ann. Statist.*, 1:38–53, 1973.

T. Liang. How well can generative adversarial networks (GAN) learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.

G. Loeper. Uniqueness of the solution to the Vlasov-Poisson system with bounded density. *J. Math. Pures Appl. (9)*, 86(1):68–79, 2006.

B. Maury, A. Roudneff-Chupin, and F. Santambrogio. A macroscopic crowd motion model of gradient flow type. *Math. Models Methods Appl. Sci.*, 20(10):1787–1821, 2010.

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

G. Monge. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie royale des sciences*, 1:666–704, 1781.

J. Moser. On the volume elements on a manifold. *Trans. Amer. Math. Soc.*, 120:286–294, 1965.

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

G. Peyré and M. Cuturi. Computational optimal transport. Technical report, 2017.
R. Peyre. Comparison between W_2 distance and $\dot{H}^1$ norm, and localization of Wasserstein distance. *ESAIM: Control, Optimisation and Calculus of Variations*, 2018.
J. Rabin, G. Peyré, J. Delon, and M. Bernot. Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer, 2011.
H. P. Rosenthal. On the span in L^p of sequences of independent random variables. II. *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. II: Probability theory*, pages 149–167, 1972.
S. Shirdhonkar and D. W. Jacobs. Approximate earth mover's distance in linear time. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008), 24-26 June 2008, Anchorage, Alaska, USA*, 2008.
S. Singh and B. Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.
S. Singh, A. Uppal, B. Li, C.-L. Li, M. Zaheer, and B. Póczos. Nonparametric density estimation under adversarial losses. *arXiv preprint arXiv:1805.08836*, 2018.
H. Triebel. *Bases in function spaces, sampling, discrepancy, numerical integration*, volume 11 of *EMS Tracts in Mathematics*. European Mathematical Society (EMS), Zürich, 2010.
A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.
C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
G. G. Walter. Approximation of the delta function by wavelets. *J. Approx. Theory*, 71(3): 329–343, 1992.
L. Wasserman. *All of statistics*. Springer Texts in Statistics. Springer-Verlag, New York, 2004.
J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 2018. To appear.

APPENDIX A: OMITTED PROOFS

A.1 Proof of Theorem 1

The following lemma shows that the quality of the estimator can be controlled by the distance between f and $\tilde{f}$ in Besov norm.

LEMMA 2. *For any $f \in \mathcal{B}^s_{p,q}(L; m)$,*

$$W_p(\mu_f, \mu_{\hat{f}}) \lesssim \|f - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} .$$

PROOF. By assumption, both f and $\hat{f}$ are bounded below by m. Theorem 3 therefore implies

$$W_p(\mu_f, \mu_{\hat{f}}) \lesssim \|f - \hat{f}\|_{\mathcal{B}^{-1}_{p,1}} \leq \|f - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} + \|\hat{f} - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} \lesssim \|f - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} ,$$

where the final inequality uses the fact that $f \in \mathcal{D}(m)$. ☐

We require the following proposition, whose proof appears in Section A.11.

PROPOSITION 5. *Let f have wavelet expansion as in (2), and let $0 \leq j \leq J$. If $n \geq 2^{dJ}$ and $p \geq 1$, then*

$$\mathbb{E}\|\alpha - \tilde{\alpha}\|_{\ell_p} \lesssim \frac{1}{n^{1/2}}$$
$$\mathbb{E}\|\beta_j - \tilde{\beta}_j\|_{\ell_p} \lesssim \frac{2^{dj/p}}{n^{1/2}} .$$

We are now in a position to prove Theorem 1.

Proof of Theorem 1. Denote by f_J the projection of f to the span of $\Phi \cup \{\bigcup_{0 \le j \le J} \Psi_j\}$, i.e..

$$f_J = \sum_{\phi \in \Phi} \alpha_\phi \phi + \sum_{0 \le j \le J} \sum_{\psi \in \Psi} \beta_\psi \psi \,.$$

The assumption that $f \in \mathcal{B}^s_{p',q}(L)$ implies by the definition of $\|\cdot\|_{\mathcal{B}^s_{p',q}}$ and Lemma 9 that

$$2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j\|_{\ell_p} \lesssim 2^{dj(\frac{1}{2}-\frac{1}{p'})} \|\beta_j\|_{\ell_{p'}} \lesssim 2^{-js}$$

for all $j \ge 0$; hence

$$\|f - f_J\|_{\mathcal{B}^{-1}_{p,1}} = \sum_{j > J} 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j\|_{\ell_p} \lesssim 2^{-J(s+1)} \,.$$

Lemma 2 implies

$$\begin{aligned}
\mathbb{E} W_p(\mu_f, \mu_{\hat{f}}) &\lesssim \mathbb{E}\|f_J - \tilde{f}\|_{\mathcal{B}^{-1}_{p,1}} + \|f - f_J\|_{\mathcal{B}^{-1}_{p,1}} \\
&\lesssim \mathbb{E}\|\alpha - \tilde{\alpha}\|_{\ell_p} + \sum_{0 \le j \le J} \sum_{\psi \in \Psi_j} 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \mathbb{E}\|\beta_j - \tilde{\beta}_j\|_{\ell_p} + 2^{-J(s+1)} \\
&\lesssim \sum_{0 \le j \le J} 2^{-j} \left(\frac{2^{dj}}{n}\right)^{1/2} + 2^{-J(s+1)}
\end{aligned}$$

by Proposition 5, as long as $2^J \le n^{1/d}$. If $d \ge 3$, the last term in the sum dominates and choosing J such that $2^J \asymp n^{\frac{1}{d+2s}}$ yields the claim. If $d \le 2$, we choose J so that $2^J \asymp n^{\frac{1}{2+2s}}$. The sum is then of order $n^{-1/2}$ if $d = 1$, or $n^{-1/2} \log n$ if $d = 2$. ☐

A.2 Proof of Theorem 2

By the monotonicity of the Wasserstein-p distances in p, it suffices to prove the lower bound for $p = 1$. Given an index J to be specified and a vector $\varepsilon \in \{\pm 1\}^{|\Psi_J|}$, we write

$$f_\varepsilon := 1 + \frac{1}{2} \sum_{\psi \in \Psi_J} n^{-1/2} \varepsilon_\psi \psi \,.$$

As long as $n^{-1/2} 2^{J(s+d/2)} \lesssim 1$, the functions f_ε all lie in $\mathcal{B}^s_{p',q}(L; m)$.

Moreover, note that for any $\varepsilon, \varepsilon' \in \{\pm 1\}^{|\Psi_J|}$, Proposition 2 implies

$$W_1(\mu_{f_\varepsilon}, \mu_{f_{\varepsilon'}}) \gtrsim \|f_\varepsilon - f_{\varepsilon'}\|_{\mathcal{B}^{-1}_{1,\infty}} = 2^{-J(1+d/2)} n^{-1/2} \rho(\varepsilon, \varepsilon') \,.$$

where $\rho(\varepsilon, \varepsilon')$ is the Hamming distance between ε and ε'.

Moreover, when $\rho(\varepsilon, \varepsilon') = 1$, the fact that $f_\varepsilon, f_{\varepsilon'} \ge m$ implies that the Hellinger distance satisfies

$$\int (\sqrt{f_\varepsilon} - \sqrt{f_{\varepsilon'}})^2 \,\mathrm{d}x \lesssim \int (f_\varepsilon - f_{\varepsilon'})^2 \,\mathrm{d}x = n^{-1}$$

Therefore, since W_1 is a metric, a standard application of Assouad's lemma (Tsybakov, 2009, Theorem 2.12) implies that

$$\begin{aligned}\inf_{\tilde{\mu}} \sup_{f \in \mathcal{B}^s_{p',q}(L;m)} \mathbb{E} W_1(\hat{\mu}, \mu_f) &\gtrsim \inf_{\hat{\varepsilon}} \sup_{\varepsilon \in \{\pm 1\}^{|\Psi_J|}} \mathbb{E} W_1(\mu_{f_{\hat{\varepsilon}}}, \mu_{f_\varepsilon}) \\ &\gtrsim 2^{-J(1+d/2)} |\Psi_J| n^{-1/2} \\ &\gtrsim 2^{-J} 2^{dJ/2} n^{-1/2}\,,\end{aligned}$$

where the infimum is taken over all estimators $\tilde{\mu}$ constructed from n samples and where the final inequality is a consequence of Lemma 9. Choosing J such that $2^J \asymp n^{\frac{1}{d+2s}}$ when $d \geq 2$ and $J = 0$ when $d = 1$ yields the claim. □

A.3 Proof of Theorem 4

We first prove the $n^{-1/2p}$ bound. Let $g_0 \in \mathcal{B}^s_{p',q}(L)$ be supported in $[0, 1/3]^d$ and let g_1 be a translation of g_0 supported on $[2/3, 1]^d$. For $\lambda \in [-1, 1]$, define $f_\lambda := \frac{1}{2}((1+\lambda)g_0 + (1-\lambda)g_1)$. Then for any $\lambda \in [-1, 1]$, the densities $f_{-\lambda}$ and f_λ satisfy

$$\begin{aligned}\int (\sqrt{f_\lambda} - \sqrt{f_{-\lambda}})^2 \,\mathrm{d}x &= \frac{1}{2} \int (\sqrt{1+\lambda} - \sqrt{1-\lambda})^2 (g_0 + g_1) \,\mathrm{d}x \\ &\lesssim \lambda^2\,.\end{aligned}$$

On the other hand, by Lemma 1, choosing $S = [0, 1/3]^d$ yields

$$W_p(\mu_f, \mu_{f'}) \gtrsim \lambda^{1/p}\,.$$

Therefore, if we choose $\lambda \asymp n^{-1/2}$, then the claim follows from the method of LeCam (1973).

We now prove the $n^{-\frac{1+s/p}{d+s}}$ bound. We proceed via Assouad's lemma. Let g_0 be as above. For $M > 0$ to be specified, the characterization of $\mathcal{B}^s_{p',q}$ by $\|\cdot\|'_{\mathcal{B}^s_{p',q}}$ implies that there exists a universal constant $c > 0$ such that $h(x) := cM^{-s} g_0(Mx)$ also lies in $\mathcal{B}^s_{p',q}(L)$. We denote by Γ a set of vectors in $\mathbb{R}^d$ such that for any $\gamma_1, \gamma_2 \in \Gamma$, the supports of the functions $h(x - \gamma_1)$, $h(x - \gamma_2)$, and g_0 are all separated by at least cM^{-1} for c a small constant. By a volume argument, we can choose Γ such that $|\Gamma| \asymp M^d$. We assume that $|\Gamma|$ is even.

We divide the elements of $|\Gamma|$ into pairs and label them $\{(\gamma_i^+, \gamma_i^-)\}_{i=1}^{|\Gamma|/2}$. For any $\varepsilon \in \{\pm 1\}^{|\Gamma|/2}$, define

$$f_\varepsilon := \sum_{i \in |\Gamma|/2} h(x - \gamma_i^{\varepsilon_i}) + \kappa g_0\,,$$

where $\kappa := 1 - \frac{|\Gamma|}{2} \int h \,\mathrm{d}x$ is chosen to ensure that f_ε integrates to 1.

Given $\varepsilon, \varepsilon' \in \{\pm 1\}^{|\Gamma|/2}$, define

$$\Delta(\varepsilon, \varepsilon') := \{\gamma_i^+ \in \Gamma : \varepsilon_i = +1, \varepsilon'_i = -1\}\,.$$

In other words, $\Delta(\varepsilon, \varepsilon')$ is a set of $\gamma \in \Gamma$ such that $h(x - \gamma)$ appears in the density f_ε but not in $f_{\varepsilon'}$. We set

$$S := \bigcup_{\gamma \in \Delta(\varepsilon, \varepsilon')} \mathrm{supp}(h(x - \gamma))\,.$$

If we denote by ρ the Hamming distance, then the density f_ε assigns mass $|\Delta(\varepsilon,\varepsilon')|cM^{-s-d} \gtrsim \rho(\varepsilon,\varepsilon')M^{-s-d}$ to S, and $f_{\varepsilon'}$ assigns zero mass to this set. By construction, the rest of the support of f_ε and $f_{\varepsilon'}$ lies at distance at least cM^{-1} from S. Therefore, by Lemma 1,

$$W_p(\mu_{f_{\varepsilon'}}, \mu_{f_\varepsilon}) \gtrsim M^{-1}(\rho(\varepsilon,\varepsilon')M^{-s-d})^{1/p} \gtrsim \rho(\varepsilon,\varepsilon')M^{-\frac{s}{p}-1-d},$$

where the last inequality follows from the fact that $\rho(\varepsilon,\varepsilon') \le |\Gamma|/2 \lesssim M^d$. Moreover, if $\rho(\varepsilon,\varepsilon') = 1$, then

$$\int (\sqrt{f_\varepsilon(x)} - \sqrt{f_{\varepsilon'}(x)})^2 \,\mathrm{d}x \le \int |f_\varepsilon(x) - f_{\varepsilon'}(x)| \,\mathrm{d}x \lesssim M^{-s-d}.$$

Therefore, if we choose $M \asymp n^{\frac{1}{s+d}}$, then Assouad's lemma (Tsybakov, 2009, Theorem 2.12) and the fact that W_p satisfies the triangle inequality imply

$$\begin{aligned}
\inf_{\hat\mu} \sup_{f \in f \in \mathcal{B}^s_{p',q}(L)} \mathbb{E}W_p(\hat\mu, \mu_{f_\varepsilon}) &\gtrsim \inf_{\hat\varepsilon} \sup_{\varepsilon \in \{\pm 1\}^{|\Gamma|/2}} \mathbb{E}W_p(\mu_{f_{\hat\varepsilon}}, \mu_{f_\varepsilon}) \\
&\gtrsim M^{-\frac{s}{p}-1} \\
&\gtrsim n^{-\frac{1+s/p}{d+s}},
\end{aligned}$$

as claimed. □

A.4 Proof of Theorem 5

For $j \ge 0$, denote by Ψ_j the elements of the d-dimensional Haar wavelet basis at scale 2^{-j} (see Triebel, 2010, Section 2.3). Note that in the case of the Haar wavelet basis over the cube, the set Φ of scaling functions contains only the function which is identically 1 on $[0,1]^d$. We have the expansion $f = 1 + \sum_{j\ge0}\sum_{\psi\in\Psi_j}\beta_\psi\psi$ in L_2.

Fix some $J \ge 0$ to be chosen. Set

$$\hat\beta_\psi := \frac{1}{n}\sum_{i=1}^n \psi(X_i) \quad \psi \in \Psi_j, 0 \le j \le J,$$

and define $\hat f := 1 + \sum_{0\le j\le J}\sum_{\psi\in\Psi_j}\hat\beta_\psi\psi$.

If we denote by V_j the space of functions spanned by $\{1, (\Psi_k)_{0\le k<j}\}$ and let K_j be the L_2 projection onto V_j, then the bound of (3) reads

$$W_p^p(\mu_f, \mu_{\hat f}) \lesssim \sum_{j\ge0} 2^{-jp}\|K_j(f-\hat f)\|_{L_1}.$$

Expanding each term as a sum of wavelets and applying the triangle inequality term-by-term yields

$$W_p^p(\mu_f, \mu_{\hat f}) \lesssim \sum_{0\le j\le J} 2^{-jp}2^{-dj/2}\|\beta_j - \hat\beta_j\|_{\ell_1} + \sum_{j>J} 2^{-jp}2^{-dj/2}\|\beta_j\|_{\ell_1}.$$

The assumption that $f \in \mathcal{B}^s_{p',q}(L)$ with $s \in [0,1)$, $p', q \ge 1$ implies $2^{-dj/2}\|\beta_j\|_{\ell_1} \le 2^{dj(\frac12 - \frac{1}{p'})}\|\beta_j\|_{\ell_{p'}} \lesssim 2^{-js}$ for all $j \ge 0$ (Triebel, 2010, Proposition 2.20), which yields

$$W_p^p(f, \hat f) \lesssim \sum_{0\le j\le J} 2^{-jp}2^{-dj/2}\|\beta_j - \hat\beta_j\|_{\ell_1} + 2^{-J(s+p)},$$

and hence, applying Proposition 5,

$$\mathbb{E}W_p^p(\mu_f, \mu_{\hat{f}}) \lesssim \sum_{0 \leq j \leq J} 2^{-jp} \left(\frac{2^{dj}}{n}\right)^{1/2} + 2^{-J(s+p)}$$

Choosing J such that $2^J \asymp n^{\frac{1}{d+2s}}$ when $d \geq 2p$ and $2^J \asymp n^{\frac{1}{2p+2s}}$ otherwise yields the claim. □

A.5 Proof of Theorem 6

We consider the estimator resample distribution constructed from the histogram estimator give by Theorem 5. If $d \leq 2p$, then Fournier and Guillin (2015, Theorem 1) implies that the empirical distribution $\hat{\mu}_n$ already achieves the rate appearing in Theorem 5; therefore, choosing $M = n$ suffices. If $d > 2p$, we apply Corollary 1. Theorem 5 implies that the histogram estimator $\mu_{\hat{f}}$ achieves the rate $n^{-\gamma^*/d}$ with $\gamma^* = \frac{1+\frac{s}{p}}{1+\frac{2s}{d}}$. Choosing $\gamma = \gamma^*$ and noting that $\gamma^* < 2$ yields the claim.

Finally, to show that constructing $\bar{\mu}_{n,M}$ takes time $O(M)$, we note that the histogram estimator constructed in Theorem 5 is piecewise constant with $O(n)$ pieces. Constructing the histogram estimator takes $O(n)$ time, and sampling M points from a histogram distribution can be done in time $O(n + M) = O(M)$ by the alias method (Kronmal and Peterson, 1979). □

A.6 Proof of Proposition 1

We will follow a strategy originally developed by Moser (1965) for the purpose of showing that all volume forms on a smooth, compact manifold are equivalent up to automorphism. We define a vector field V on Ω satisfying

$$\nabla \cdot V = f - g \tag{4}$$

$$\|V\|_{L^p} \lesssim \|\alpha - \alpha'\|_{\ell_p} + \sum_{j \geq 0} 2^{-j} 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta_j - \beta_j'\|_{\ell_p}\,, \tag{5}$$

where the first condition is intended in the distributional sense that

$$-\int_\Omega \nabla h \cdot V \, \mathrm{d}x = \int_\Omega h(f - g) \, \mathrm{d}x\,.$$

for all $h \in C^1(\Omega)$. In particular, we require the boundary condition $V \cdot \mathbf{n} = 0$ on $\partial\Omega$, where $\mathbf{n}$ is an outward-pointing normal.

In Appendix B, we establish the following proposition.

Proposition 6. *There exists a vector field V satisfying (4) and (5).*

To show the theorem, we appeal to the following characterization of the Wasserstein distance. Denote by K_Ω the set of pairs of measures (ρ, E) on $\Omega \times [0, 1]$ where ρ is scalar valued and E is vector valued.

Theorem 9 (Benamou and Brenier, 2000; Brenier, 2003). *For any measures μ and ν on Ω and $p \in [1, \infty)$,*

$$W_p^p(\mu, \nu) = \inf_{(\rho, E) \in K_\Omega} \left\{ \mathcal{B}_p(\rho, E) : \rho(\cdot, 0) = \mu, \rho(\cdot, 1) = \nu, \partial_t \rho + \nabla_x \cdot E = 0 \right\}\,,$$

where

$$\mathcal{B}_p(\rho, E) := \begin{cases} \int_{\Omega\times[0,1]} \left\| \frac{dE}{d\rho}(x,t) \right\|^p \mathrm{d}\rho(x,t) & \text{if } E \ll \rho, \\ +\infty & \text{otherwise.} \end{cases}$$

Let us show how to prove the theorem. We choose ρ and E absolutely continuous with respect to the Lebesgue measure on $\Omega \times [0,1]$, and consequently identify them with their density. First, set $\rho(x,t) = (1-t)f(x) + tg(x)$. Note that $\rho(x,t) \geq m$ for all $x \in \Omega, t \in [0,1]$, and clearly $\rho(\cdot, 0) = f$ and $\rho(\cdot, 1) = g$. We then choose E to be constant in time, setting

$$E(x,t) = V(x) \quad \text{ for } t \in [0,1].$$

Since $\nabla_x \cdot E = f - g$, the pair (ρ, E) defined in this way satisfies the continuity equation $\partial_t \rho + \nabla_x \cdot E = 0$.

For all $t \in [0,1]$, we have the bound

$$\left\| \frac{dE}{d\rho}(x,t) \right\|^p \mathrm{d}\rho(x,t) \leq \|V(x)\|^p m^{-p+1} \,\mathrm{d}x \,.$$

We obtain

$$\begin{aligned} W_p(\mu_f, \mu_g) &\leq \left(\int_{\Omega\times[0,1]} \left\| \frac{dE}{d\rho}(x,t) \right\|^p \mathrm{d}\rho(x,t) \right)^{1/p} \\ &\leq m^{1/p-1} \|V\|_{L_p} \\ &\lesssim m^{-1/p'} \left(\|\alpha - \alpha'\|_{\ell_p} + \sum_{j\geq 0} 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j - \beta_j'\|_{\ell_p} \right), \end{aligned}$$

as claimed. □

A.7 Proof of Theorem 7

Let $h_+ := h \vee 0$ and $h_- := -(h \wedge 0)$, so that $h = h_+ - h_-$, and note that by assumption $\int_\Omega h_+ \,\mathrm{d}x = \int_\Omega h_- \,\mathrm{d}x > 0$. Let $\rho := \int_\Omega h_- \,\mathrm{d}x$. For any $\lambda \in [0,1]$, set

$$\begin{aligned} f_\lambda &= \frac{1}{2\rho}((1+\lambda)h_+ + (1-\lambda)h_-) \\ g_\lambda &= \frac{1}{2\rho}((1-\lambda)h_+ + (1+\lambda)h_-)\,. \end{aligned}$$

Note that $f_\lambda, g_\lambda \in \mathcal{D}(\Omega)$, and $\|f - g\| = \frac{\lambda}{\rho}\|h\|$. On the other hand, since the compact sets $\overline{\{h > 0\}}$ and $\overline{\{h < 0\}}$ are disjoint, there exist two sets S and T and $c > 0$ such that $\mathrm{supp}(h_+) \subseteq S$ and $\mathrm{supp}(h_-) \subseteq T$ and $\|x - y\| \geq c$ for any $x \in S, y \in T$. Lemma 1, below, therefore implies $W_p(\mu_{f_\lambda}, \mu_{g_\lambda}) \geq c|\int_S (f - g)\,\mathrm{d}x|^{1/p} = c\lambda^{1/p}$.

We obtain

$$\sup_{f,g\in\mathcal{D}(\Omega)} \frac{W_p(\mu_f, \mu_g)}{\|f-g\|} \geq \sup_{\lambda\in(0,1)} \frac{W_p(\mu_{f_\lambda}, \mu_{g_\lambda})}{\|f_\lambda - g_\lambda\|} \gtrsim \sup_{\lambda\in(0,1)} \lambda^{1/p-1} = \infty\,.$$

□

A.8 Proof of Proposition 2

We use the following fact, due to Maury et al. (2010, Lemma 3.4 and subsequent remark):

LEMMA 3 (Maury et al., 2010). *For all $h \in C^1(\Omega)$,*

$$\int_\Omega h(f-g)\,\mathrm{d}x \leq M^{1/p'}\|\nabla h\|_{L_{p'}(\Omega)} W_p(\mu,\nu)\,.$$

Fix an index $j \geq 0$. Let h be a function of the form

$$h = \sum_{\phi\in\Phi} \kappa_\phi \phi + \sum_{\psi \in \Psi_j} \lambda_\psi \psi\,,$$

for some vectors κ and λ satisfying $\|\kappa\|_{\ell_{p'}} \leq 1$ and $\|\lambda\|_{\ell_{p'}} \leq 2^{-j+dj(\frac{1}{2}-\frac{1}{p})}$.

We require the following bound, whose proof appears in Section A.12.

LEMMA 4. *If $\|\kappa\|_{\ell_{p'}} \leq 1$ and $\|\lambda\|_{\ell_{p'}} \leq 2^{-j+dj(\frac{1}{2}-\frac{1}{p})}$, then*

$$\|\nabla h\|_{L_{p'}(\Omega)} \lesssim 1$$

Applying Lemmas 3 and 4, we obtain

$$W_p(f,g) \gtrsim M^{-1/p'} \int_\Omega h(f-g)\,\mathrm{d}x = M^{-1/p'}\left(\sum_{\phi\in\Phi}\kappa_\phi(\alpha_\phi - \alpha'_\phi) + \sum_{\psi\in\Psi_j}\lambda_\psi(\beta_\psi - \beta'_\psi)\right)\,.$$

Taking the supremum over κ and λ subject to the constraints $\|\kappa\|_{\ell_{p'}} \leq 1$ and $\|\lambda\|_{\ell_{p'}} \leq 2^{-j+dj(\frac{1}{2}-\frac{1}{p})}$ implies

$$W_p(f,g) \gtrsim M^{-1/p'}\left(\|\alpha-\alpha'\|_{\ell_p} + 2^{-j}2^{dj(\frac{1}{2}-\frac{1}{p})}\|\beta_j - \beta'_j\|_{\ell_p}\right)\,,$$

and taking the supremum over $j \geq 0$ yields the claim. □

A.9 Proof of Corollary 1

We have by the triangle inequality that

$$\begin{aligned}\mathbb{E}W_p(\mu,\bar{\mu}_{n,M}) &\leq \mathbb{E}W_p(\mu,\tilde{\mu}_n) + \mathbb{E}W_p(\tilde{\mu}_n,\bar{\mu}_{n,M}) \\ &\lesssim n^{-\gamma^*/d} + M^{-1/d} \\ &\lesssim n^{-\gamma/d}\,,\end{aligned}$$

having taken M of order n^γ. □

A.10 Proof of Theorem 8

Taking $\tilde{\mu}_n$ such that $\mathbb{E}W_p(\mu,\tilde{\mu}_n) \leq n^{-\gamma^*/d}$, $\bar{\mu}_{n,M}$ the empirical resample distribution of $\tilde{\mu}_n$, and $\bar{\nu}_M$ an M-atomic version of ν (obtained by sampling from or discretizing ν), we have

$$\begin{aligned}\mathbb{E}|W_p(\bar{\nu}_M,\bar{\mu}_{n,M}) - W_p(\nu,\mu)| &\leq \mathbb{E}W_p(\mu,\bar{\mu}_{n,M}) + \mathbb{E}W_p(\nu,\bar{\nu}_M) \\ &\lesssim n^{-\gamma^*/d} + M^{-1/d} \lesssim n^{-\gamma/d}\,,\end{aligned}$$

by taking $M = n^\gamma$. Taking $\varepsilon = n^{-p\gamma/d}$, computing an ε-approximation $\tilde{W}^p_{p,n}$ to $W_p^p(\bar{\nu}_M, \bar{\mu}_{n,M})$ given $\bar{\mu}_{n,M}$ and $\bar{\nu}_M$ takes time $O\big(n^{\gamma(2+2p/d)}\log(n)\big)$ by Proposition 4. Finally, the inequality

$$|\tilde{W}_{p,n} - W_p(\bar{\nu}_M, \bar{\mu}_{n,M})| \leq |\tilde{W}^p_{p,n} - W_p^p(\bar{\nu}_M, \bar{\mu}_{n,M})|^{1/p} \leq \varepsilon^{1/p} = n^{-\gamma/d}$$

yields the claim. □

A.11 Proof of Proposition 5

We first show the claim for $\|\beta_j - \hat{\beta}_j\|_{\ell_p}$, $j \geq 0$. The inequalities of Rosenthal (1972) imply that there exists a constant c_p such that for any $\psi \in \Psi_j$,

$$\mathbb{E}|\beta_\psi - \hat{\beta}_\psi|^p \leq c_p \left(\frac{\sigma^p}{n^{p/2}} + \frac{\mathbb{E}\,|\psi(X) - \mathbb{E}\psi(X)|^p}{n^{p-1}}\mathbb{I}\{p \geq 2\}\right),$$

where $\sigma^2 := \mathbb{E}\,|\psi(X) - \mathbb{E}\psi(X)|^2$ and $X \sim \mu_f$.

Assumption 5 implies that $\|\psi\|_{L_\infty} \lesssim 2^{dj/2}$, so

$$\frac{\mathbb{E}\,|\psi(X) - \mathbb{E}\psi(X)|^p}{n^{p-1}} \leq \frac{2^{dj(p-2)/2}\sigma^2}{n^{p-1}} \leq \frac{\sigma^2}{n^{p/2}},$$

where the last inequality follows from the assumption that $n \geq 2^{dj}$.

In order to establish that $\mathbb{E}\|\beta_j - \hat{\beta}_j\|^p_{\ell_p} \leq \frac{2^{dj}}{n^{p/2}}$, it therefore suffices to show that

$$\sum_{\psi \in \Psi_j} (\mathbb{E}\,|\psi(X) - \mathbb{E}\psi(X)|^2)^{p/2} \lesssim 2^{dj}.$$

We have

$$(\mathbb{E}\,|\psi(X) - \mathbb{E}\psi(X)|^2)^{p/2} \leq (\mathbb{E}\psi(X)^2)^{p/2} = \left(\int_\Omega \psi(x)^2 f(x)\,\mathrm{d}x\right)^{p/2} \leq \int_\Omega \psi(x)^2 f(x)^{p/2}\,\mathrm{d}x,$$

where in the last step we use Jensen's inequality combined with the fact that $\int \psi^2\,\mathrm{d}x = 1$ for all $\psi \in \Psi_j$ (see Assumption 1). Finally, we obtain

$$\begin{aligned}\sum_{\psi \in \Psi_j} (\mathbb{E}\,|\psi(X) - \mathbb{E}\psi(X)|^2)^{p/2} &\leq \int_\Omega \sum_{\psi \in \Psi} \psi(x)^2 f(x)^{p/2}\,\mathrm{d}x \\ &\leq \left\|\sum_{\psi \in \Psi_j} \psi(x)^2\right\|_{L_\infty} \|f\|^{p/2}_{L_{p/2}}\end{aligned}$$

The fact that $f \in L_p(\Omega)$ where Ω is compact implies that $\|f\|_{L_{p/2}} \lesssim \|f\|_{L_p} \lesssim 1$, and Assumptions 4 and 5 imply $\left\|\sum_{\psi\in\Psi_j} \psi(x)^2\right\|_{L_\infty} \lesssim 2^{dj}$. This proves the claim.

To show the analogous bound on $\|\alpha - \hat{\alpha}\|_{\ell_p}$, we again write

$$\mathbb{E}|\alpha_\phi - \hat{\alpha}_\phi|^p \leq c_p \left(\frac{\sigma^p}{n^{p/2}} + \frac{\mathbb{E}\,|\phi(X) - \mathbb{E}\phi(X)|^p}{n^{p-1}}\mathbb{I}\{p \geq 2\}\right)$$

for any $\phi \in \Phi$, where $\sigma^2 := |\phi(X) - \mathbb{E}\phi(X)|^2$. Since $\|\phi\|_{L_\infty} \lesssim 1$ for any $\phi \in \Phi$, this bound simplifies to

$$\mathbb{E}|\alpha_\phi - \hat{\alpha}_\phi|^p \lesssim \frac{1}{n^{p/2}},$$

and hence

$$\mathbb{E}\|\alpha - \hat{\alpha}\|^p_{\ell_p} \leq \frac{|\Phi|}{n^{p/2}} \lesssim \frac{1}{n^{p/2}}$$

by Lemma 9. □

A.12 Proof of Lemma 4

This follows directly from the assumptions on the multiresolution analysis:

$$\begin{aligned}\|\nabla h\|_{L_{p'}(\Omega)} &\leq \left\|\nabla \sum_{\phi\in\Phi} \kappa_\phi \phi\right\|_{L_{p'}(\Omega)} + \left\|\nabla \sum_{\psi\in\Psi_j} \lambda_\psi \psi\right\|_{L_{p'}(\Omega)} \\ &\lesssim \left\|\sum_{\phi\in\Phi} \kappa_\phi \phi\right\|_{L_{p'}(\Omega)} + 2^j \left\|\sum_{\psi\in\Psi_j} \lambda_\psi \psi\right\|_{L_{p'}(\Omega)} \\ &\lesssim \|\kappa\|_{\ell_{p'}} + 2^{j+dj(\frac{1}{2}-\frac{1}{p'})}\|\lambda\|_{\ell_{p'}} \\ &\leq 1 + 2^{dj(1-\frac{1}{p}-\frac{1}{p'})} = 2\,,\end{aligned}$$

where the inequalities follow, respectively, from the triangle inequality, Assumption 6, Lemma 10, and the assumption on κ and λ. ☐

APPENDIX B: PROOF OF PROPOSITION 6

We will proceed by defining vector fields V_ϕ for each $\phi \in \Phi$ and V_ψ for each $\psi \in \Psi_j$, $j \geq 0$ satisfying $\nabla \cdot V_\phi = \phi$ and $\nabla \cdot V_\psi = \psi$, along with appropriate boundary conditions. The desired vector field V will then be obtained as

$$V = \sum_{\phi\in\Phi}(\alpha_\phi - \alpha'_\phi)V_\phi + \sum_{j\geq 0}\sum_{\psi\in\Psi_j}(\beta_\psi - \beta'_\psi)V_\psi\,. \tag{6}$$

An application of Fubini's theorem immediately yields that this definition satisfies (4) in the distributional sense. We conclude by obtaining the desired estimate for $\|V\|_{L_p}$ to show (5)

Definition of V_ϕ for $\phi \in \Phi$. Given $x \in \mathbb{R}^d$, we write $x^{(i)}$ for the vector consisting of the first i coordinates of x. For each $1 \leq i \leq d$, define $\phi^{(i)} : \mathbb{R}^i \to \mathbb{R}$ by

$$\phi^{(i)}(x^{(i)}) = \int_0^1 \cdots \int_0^1 \phi(x^{(i)}, t_{i+1}, \dots, t_d)\, \mathrm{d}t_{i+1} \dots \mathrm{d}t_d\,,$$

and set $\phi^{(0)} = 0$. We define V_ϕ componentwise as

$$(V_\phi)_i(x) = \int_0^{x_i} \phi^{(i)}(x^{(i-1)}, t_i)\, \mathrm{d}t_i - x_i\phi^{(i-1)}(x^{(i-1)}) \qquad 1 \leq i \leq d\,.$$

We now verify that this definition satisfies the desired identity. The proof appears in Section B.1.

LEMMA 5. *If $\{\alpha_\phi\}_{\phi\in\Phi}$ and $\{\alpha'_\phi\}_{\phi\in\Phi}$ satisfy*

$$\int_{[0,1]^d} \sum_{\phi\in\Phi} \alpha_\phi \phi(x)\, \mathrm{d}x = \int_{[0,1]^d} \sum_{\phi\in\Phi} \alpha'_\phi \phi(x)\, \mathrm{d}x\,,$$

then

$$\nabla \cdot \sum_{\phi\in\Phi}(\alpha_\phi - \alpha'_\phi)V_\phi = \sum_{\phi\in\Phi}(\alpha_\phi - \alpha'_\phi)\phi$$

and $\left(\sum_{\phi\in\Phi}(\alpha_\phi - \alpha'_\phi)V_\phi\right)\cdot \mathbf{n} = 0$ *on the boundary of* $[0,1]^d$, *where* $\mathbf{n}$ *is an outward-pointing normal.*

Definition of V_ψ for $\psi \in \Psi_j$, $j \geq 0$. We adopt essentially the same construction as above. First, by Assumption 3, ψ can be written as $\prod_{i=1}^d \psi_i$, where each ψ_i is a univariate function. Assumptions 1 and 2 imply that $\int_{[0,1]^d} \psi(x)\,\mathrm{d}x = 0$, so there exists an index $k \in [d]$ such that $\int_{[0,1]} \psi_k(x_k)\,\mathrm{d}x_k = 0$. We set

$$(V_\psi)_k(x) = \int_0^{x_k} \psi_k(t)\,\mathrm{d}t \cdot \prod_{i \neq k} \psi_i(x_i)\,,$$

and $(V_\psi)_i = 0$ for $i \neq k$.

Lemma 6. *The field V_ψ satisfies*

$$\nabla \cdot V_\psi = \psi$$

and $V_\psi \cdot \mathbf{n} = 0$ on the boundary of $[0,1]^d$, where $\mathbf{n}$ is an outward-pointing normal.

A proof appears in Section B.2.

Norm estimates. We now obtain an estimate for $\|V\|_{L_p}$. We require two lemmas, the proofs of which appear in Sections B.3 and B.4.

Lemma 7. *For any sequence $\{\alpha_\phi\}_{\phi \in \Phi}$,*

$$\left\| \sum_{\phi \in \Phi} \alpha_\phi V_\phi \right\|_{L_p} \lesssim \|\alpha\|_{\ell_p}\,.$$

Lemma 8. *For any sequence $\{\beta_\psi\}_{\psi \in \Psi_j}$,*

$$\left\| \sum_{\psi \in \Psi_j} \beta_\psi V_\psi \right\|_{L_p} \lesssim 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta\|_{\ell_p}\,.$$

We obtain, for V defined as in (6),

$$\begin{aligned}\|V\|_{L_p} &\leq \left\| \sum_{\phi \in \Phi} (\alpha_\phi - \alpha'_\phi) V_\phi \right\|_{L_p} + \sum_{j \geq 0} \left\| \sum_{\psi \in \Psi_j} (\beta_\psi - \beta'_\psi) V_\psi \right\|_{L_p} \\ &\lesssim \|\alpha - \alpha'\|_{\ell_p} + \sum_{j \geq 0} 2^{-j} 2^{dj(\frac{1}{2}-\frac{1}{p})} \|\beta_j - \beta'_j\|_{\ell_p}\,,\end{aligned}$$

which gives (5). □

B.1 Proof of Lemma 5

The partial derivatives of this vector field satisfy

$$\frac{\mathrm{d}}{\mathrm{d}x_i}(V_\phi)_i = \phi^{(i)}(x^{(i)}) - \phi^{(i-1)}(x^{(i-1)}) \qquad 1 \leq i \leq d\,,$$

so that $\nabla \cdot V = \phi$ on the interior of $[0,1]^d$.

On the boundary $\{x : x_i = 0\}$ we have $(V_\phi)_i = 0$ for all i, and for $i > 1$ on the boundary $\{x : x_i = 1\}$ we have

$$(V_\phi)_i = \int_0^1 \phi^{(i)}(x^{(i-1)}, t_i)\,\mathrm{d}t_i - \phi^{(i-1)}(x^{(i-1)}) = 0\,.$$

Finally, on $\{x : x_1 = 1\}$, we have $(V_\phi)_1 = \int_0^1 \phi^{(1)}(t_1)\,\mathrm{d}t_1 = \int_{[0,1]^d} \phi(x)\,\mathrm{d}x$. Under the assumption that $\int_{[0,1]^d} \sum_{\phi\in\Phi} \alpha_\phi \phi(x)\,\mathrm{d}x = \int_{[0,1]^d} \sum_{\phi\in\Phi} \alpha'_\phi \phi(x)\,\mathrm{d}x$, we therefore obtain that $\sum_{\phi\in\Phi}(\alpha_\phi - \alpha'_\phi)(V_\phi)_1 = 0$ on the face $\{x : x_1 = 1\}$. Therefore the boundary conditions hold as well. □

B.2 Proof of Lemma 6

By construction, $\nabla \cdot V_\psi = \frac{d(V_\psi)_k}{dx_k} = \psi$ on the interior of $[0,1]^d$. The equality $(V_\phi)_m = 0$ holds on the boundaries $\{x : x_m = 0\}$ and $\{x : x_m = 1\}$ for $m \neq k$, and by construction $(V_\phi)_k = 0$ on $\{x : x_k = 0\}$. Finally, since $\int_0^1 \psi_k(t)\,\mathrm{d}t = 0$, we also have $(V_\phi)_k = 0$ on $\{x : x_1 = 1\}$, and thus the boundary conditions are satisfied. □

B.3 Proof of Lemma 7

We will show that $\left\|\sum_{\phi\in\Phi} \alpha_\phi V_\phi\right\|_{L_p([0,1]^d)} \lesssim \left\|\sum_{\phi\in\Phi} \alpha_\phi \phi\right\|_{L_p([0,1]^d)}$ and conclude by appealing to Lemma 10.

The definition of V_ϕ implies that

$$\begin{aligned}
\left|\sum_{\phi\in\Phi} \alpha_\phi (V_\phi)_i(x)\right|^p &= \left|\int_0^{x_i} \sum_{\phi\in\Phi} \alpha_\phi \phi^{(i)}(x^{(i-1)}, t_i)\,\mathrm{d}t_i - x_i \sum_{\phi\in\Phi} \alpha_\phi \phi^{(i-1)}(x^{(i-1)})\right|^p \\
&\leq 2^{p-1}\left|\int_0^{x_i} \sum_{\phi\in\Phi} \alpha_\phi \phi^{(i)}(x^{(i-1)}, t_i)\,\mathrm{d}t_i\right|^p + 2^{p-1} x_i \left|\sum_{\phi\in\Phi} \alpha_\phi \phi^{(i-1)}(x^{(i-1)})\right|^p \\
&\lesssim \int_0^1 \cdots \int_0^1 \left|\sum_{\phi\in\Phi} \alpha_\phi \phi(x^{(i-1)}, t_i, \dots, t_d)\right|^p \,\mathrm{d}t_i \dots \mathrm{d}t_d\,,
\end{aligned}$$

where in the second inequality we use Jensen's inequality and the fact that $x_i \leq 1$.

We obtain

$$\left\|\sum_{\phi\in\Phi} \alpha_\phi V_\phi\right\|_{L_p}^p \lesssim \max_{i\in[d]} \int_\Omega \left|\sum_{\phi\in\Phi} \alpha_\phi (V_\phi)_i\right|^p \lesssim \int_\Omega \left|\sum_{\phi\in\Phi} \alpha_\phi \phi(x)\right|^p \,\mathrm{d}x = \left\|\sum_{\phi\in\Phi} \alpha_\phi \phi\right\|_{L_p}^p$$

The claim then follows from Lemma 10. □

B.4 Proof of Lemma 8

By Assumptions 3 and 4, for $\psi \in \Psi_j$ there exists an interval I with $|I| \lesssim 2^{-j}$ such that $\mathrm{supp}(\psi_k) \subseteq I$. Since $(V_\psi)_k(x_k) = 0$ if $x \notin I$, Hölder's inequality implies that

$$\begin{aligned}\left|\int_0^{x_k} \psi_k(t)\,\mathrm{d}t\right|^p &= \mathbb{1}\{x_k \in I\}\left|\int_0^{x_k} \psi_k(t)\mathbb{1}\{t \in I\}\,\mathrm{d}t\right|^p \\ &\le |I|^{p-1}\mathbb{1}\{x_k \in I\}\int_0^{x_k} |\psi_k(t)|^p\,\mathrm{d}t \\ &\lesssim 2^{-j(p-1)}\mathbb{1}\{x_k \in I\}\int_0^{1} |\psi_k(t)|^p\,\mathrm{d}t\,.\end{aligned}$$

We therefore obtain by Assumption 5 that

$$\int_{[0,1]^d} \|V_\psi\|^p\,\mathrm{d}x \lesssim 2^{-j(p-1)}2^{-j}\int_{[0,1]^d} |\psi|^p \lesssim 2^{-jp}2^{pdj(\frac{1}{2}-\frac{1}{p})}\,.$$

The construction of V_ψ and Assumption 4 imply the that $V_\psi(x) = 0$ if $x \notin I_\psi$, where the sets I_ψ satisfy $\left\|\sum_{\psi\in\Psi_j}\mathbb{1}\{x \in I_\psi\}\right\|_{L_\infty} \lesssim 1$. Hölder's inequality therefore yields

$$\begin{aligned}\left\|\sum_{\psi\in\Psi_j}\beta_\psi V_\psi\right\|_{L_p}^p &\le \int_{[0,1]^d}\left(\sum_{\psi\in\Psi_j}|\beta_\psi|\|V_\psi\|\mathbb{1}\{x\in I_\psi\}\right)^p\mathrm{d}x \\ &\le \int_{[0,1]^d}\left(\sum_{\psi\in\Psi_j}|\beta_\psi|^p\|V_\psi\|^p\right)\left(\sum_{\psi\in\Psi_j}\mathbb{1}\{x\in I_\psi\}\right)^{p-1}\mathrm{d}x \\ &\le \left\|\sum_{\psi\in\Psi_j}\mathbb{1}\{x\in I_\psi\}\right\|_{L_\infty}^{p-1}\int_{[0,1]^d}\sum_{\psi\in\Psi_j}|\beta_\psi|^p\|V_\psi\|^p\,\mathrm{d}x \\ &\lesssim 2^{-jp}2^{pdj(\frac{1}{2}-\frac{1}{p})}\|\beta_j\|_{\ell_p}^p\,.\end{aligned}$$

APPENDIX C: ASSUMPTIONS ON WAVELETS

For completeness, we have extracted the properties we require of the sets Φ and Ψ_j.

ASSUMPTION 1 (Basis). $\Phi \cup \left\{\bigcup_{j\ge 0}\Psi_j\right\}$ forms an orthonormal basis for $L_2([0,1]^d)$.

ASSUMPTION 2 (Regularity). The functions in Φ and Ψ_j for $j \ge 0$ all lie in $\mathcal{C}^r(\Omega)$, and polynomials of degree at most r on Ω lie in $\mathrm{span}(\Phi)$.

ASSUMPTION 3 (Tensor construction). Each $\psi \in \Psi_j$ can be expressed as $\psi(x) = \prod_{i=1}^d \psi_i(x_i)$ for some univariate functions ψ_i.

ASSUMPTION 4 (Locality). For each $\psi \in \Psi_j$ there exists a rectangle $I_\psi \subseteq [0,1]^d$ such that $\mathrm{supp}(\psi) \subseteq I_\psi$, $\mathrm{diam}(I_\psi) \lesssim 2^{-j}$, and $\left\|\sum_{\psi\in\Psi_j}\mathbb{1}\{x\in I_\psi\}\right\|_{L_\infty} \lesssim 1$.

ASSUMPTION 5 (Norm). $\|\psi\|_{L_p(\Omega)} \asymp 2^{dj(\frac{1}{2}-\frac{1}{p})}$ for all $\psi \in \Psi_j$.

ASSUMPTION 6 (Bernstein estimate). $\|\nabla f\|_{L_p(\Omega)} \lesssim 2^j\|f\|_{L_p(\Omega)}$ for any $f \in \mathrm{span}\left(\Phi \cup \left\{\bigcup_{0\le k<j}\Psi_k\right\}\right)$.

These requirements are achievable for $\Omega = [0,1]^d$ due to a classic construction due to Cohen et al. (1993). (See also Cohen, 2003, Chapter 2.12, for further details.)

We state without proof some straightforward consequences of our assumptions.

LEMMA 9. *The sets* Φ *and* Ψ_j *satisfy* $|\Phi| \lesssim 1$ *and* $|\Psi_j| \lesssim 2^{dj}$ *for* $j \geq 0$.

LEMMA 10. *For any vector* $\{\alpha_\phi\}_{\phi \in \Phi}$, $\|\sum_{\phi \in \Phi} \alpha_\phi \phi\|_{L_p(\Omega)} \asymp \|\alpha\|_{\ell_p}$. *Likewise, for any* $\{\beta_\psi\}_{\psi \in \Psi_j}$, $\|\sum_{\psi \in \Psi_j} \beta_\psi \psi\|_{L_p(\Omega)} \asymp 2^{dj(\frac{1}{2} - \frac{1}{p})} \|\beta\|_{\ell_p}$

JONATHAN WEED
DEPARTMENT OF MATHEMATICS
MASSACHUSETTS INSTITUTE OF TECHNOLOGY
77 MASSACHUSETTS AVENUE
CAMBRIDGE, MA 02139-4307, USA
(jweed@mit.edu)

QUENTIN BERTHET
STATISTICAL LABORATORY
DEPARTMENT OF PURE MATHEMATICS AND MATHEMATICAL STATISTICS
UNIVERSITY OF CAMBRIDGE
WILBEFORCE ROAD
CAMBRIDGE, CB3 0WB, UK
(q.berthet@statslab.cam.ac.uk)