
Nonparametric Density Estimation and Convergence of GANs under Besov IPM Losses

Ananya Uppal

Department of Mathematical Sciences
Carnegie Mellon University
auppal@andrew.cmu.edu

Shashank Singh* Barnabás Póczos

Machine Learning Department
Carnegie Mellon University
{ssss1,bapoczos}@cs.cmu.edu

Abstract

We study the problem of estimating a nonparametric probability density under a large family of losses called Besov IPMs, which include, for example, L^p distances, total variation distance, and generalizations of both Wasserstein and Kolmogorov-Smirnov distances. For a wide variety of settings, we provide both lower and upper bounds, identifying precisely how the choice of loss function and assumptions on the data interact to determine the minimax optimal convergence rate. We also show that linear distribution estimates, such as the empirical distribution or kernel density estimator, often fail to converge at the optimal rate. Our bounds generalize, unify, or improve several recent and classical results. Moreover, IPMs can be used to formalize a statistical model of generative adversarial networks (GANs). Thus, we show how our results imply bounds on the statistical error of a GAN, showing, for example, that GANs can strictly outperform the best linear estimator.

1 Introduction

This paper studies the problem of estimating a nonparametric probability density, using an integral probability metric as a loss. That is, given a sample space $\mathcal{X} \subseteq \mathbb{R}^D$, suppose we observe n IID samples $X_1, \dots, X_n \stackrel{\text{iid}}{\sim} p$ from a probability density p over \mathcal{X} that is unknown but assumed to lie in a regularity class \mathcal{P} . We seek an estimator $\hat{p} : \mathcal{X}^n \rightarrow \mathcal{P}$ of p , with the goal of minimizing a loss

$$d_{\mathcal{F}}(p, \hat{p}(X_1, \dots, X_n)) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim p} [f(X)] - \mathbb{E}_{X \sim \hat{p}(X_1, \dots, X_n)} [f(X)] \right|, \quad (*)$$

where \mathcal{F} , called the *discriminator class*, is some class of bounded, measurable functions on \mathcal{X} .

Metrics of the form $(*)$ are called *integral probability metrics* (IPMs), or \mathcal{F} -IPMs², and can capture a wide variety of metrics on probability distributions by choosing \mathcal{F} appropriately [39]. This paper studies the case where both \mathcal{F} and \mathcal{P} belong to the family of Besov spaces, a large family of nonparametric smoothness spaces that include, as examples, L^p , Lipschitz/Hölder, and Hilbert-Sobolev spaces. The resulting IPMs include, as examples, L^p , total variation, Kolmogorov-Smirnov, and Wasserstein distances. We have two main motivations for studying this problem:

1. This problem unifies nonparametric density estimation with the central problem of empirical process theory, namely bounding quantities of the form $d_{\mathcal{F}}(P, \hat{P})$ when \hat{P} is the empirical distribution $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ of the data [43]. Whereas empirical process theory typically avoids restricting \mathcal{P} and fixes the estimator $\hat{P} = P_n$, focusing on the discriminator class \mathcal{F} , nonparametric density estimation typically fixes the loss to be an L^p distance, and seeks a good estimator \hat{P} for a given

^{*}Now at Google.

²While the name IPM seems most widely used [39, 50, 7, 60], many other names have been used for these quantities, including *adversarial loss* [48, 13], *MMD* [17], and \mathcal{F} -*distance* or *neural net distance* [5].

distribution class \mathcal{P} . In contrast, we study how constraints on \mathcal{F} and \mathcal{P} *jointly* determine convergence rates of a number of estimates \hat{P} of P . In particular, since Besov spaces comprise perhaps the largest commonly-studied family of nonparametric function spaces, this perspective allows us to unify, generalize, and extend several classical and recent results in distribution estimation (see Section 3).

2. This problem is a theoretical framework for analyzing generative adversarial networks (GANs). Specifically, given a GAN whose discriminator and generator networks encode functions in \mathcal{F} and \mathcal{P} , respectively, recent work [32, 28, 29, 48] showed that a GAN can be seen as a distribution estimate³

$$\hat{P} = \operatorname{argmin}_{Q \in \mathcal{P}} \sup_{f \in \mathcal{F}} \left| \mathbb{E}_{X \sim Q} [f(X)] - \mathbb{E}_{X \sim \tilde{P}_n} [f(X)] \right| = \operatorname{argmin}_{Q \in \mathcal{P}} d_{\mathcal{F}}(Q, \tilde{P}_n), \quad (1)$$

i.e., an estimate which directly minimizes empirical IPM risk with respect to a (regularized) empirical distribution \tilde{P}_n . While, in the original GAN model [21], \tilde{P}_n was the empirical distribution $P_n = \frac{1}{n} \sum_{i=1}^n \delta_{X_i}$ of the data, Liang [28] showed that, under smoothness assumptions on the population distribution, performance is improved by replacing P_n with a regularized version \tilde{P}_n , equivalent to the instance noise trick that has become standard in GAN training [49, 35]. We show, in particular, that, when \tilde{P}_n is a wavelet-thresholding estimate, a GAN based on sufficiently large fully-connected neural networks with ReLU activations learns Besov probability distributions at the optimal rate.

2 Set up and Notation

For non-negative real sequences $\{a_n\}_{n \in \mathbb{N}}$, $\{b_n\}_{n \in \mathbb{N}}$, $a_n \lesssim b_n$ indicates $\limsup_{n \rightarrow \infty} \frac{a_n}{b_n} < \infty$, and $a_n \asymp b_n$ indicates $a_n \lesssim b_n \lesssim a_n$. For $p \in [1, \infty]$, $p' := \frac{p}{p-1}$ denotes the Hölder conjugate of p (with $1' = \infty$, $\infty' = 1$). $L^p(\mathbb{R}^D)$ (resp. l^p) denotes the set of functions f (resp. sequences a) with $\|f\|_p := (\int |f(x)|^p dx)^{1/p} < \infty$ (resp. $\|a\|_{l^p} := (\sum_{n \in \mathbb{N}} |a_n|^p)^{1/p} < \infty$).

2.1 Multiresolution Approximation and Besov Spaces

We now provide some notation that is necessary to define the family of Besov spaces studied in this paper. Since the statements and formal justifications behind these definitions are a bit complex, some technical details are relegated to the Appendix, and several well-known examples from the rich class of resulting spaces are given in Section 3. The diversity of Besov spaces arises from the fact that, unlike the Hölder or Sobolev spaces that they generalize, Besov spaces model functions simultaneously across multiple spatial scales. In particular, they rely on the following notion:

Definition 1. A *multiresolution approximation (MRA)* of $L^2(\mathbb{R}^D)$ is an increasing sequence $\{V_j\}_{j \in \mathbb{Z}}$ of closed linear subspaces of $L^2(\mathbb{R}^D)$ with the following properties:

1. $\bigcap_{j=-\infty}^{\infty} V_j = \{0\}$, and the closure of $\bigcup_{j=-\infty}^{\infty} V_j = L^2(\mathbb{R}^D)$.
2. For $f \in L^2(\mathbb{R}^D)$, $k \in \mathbb{Z}^D$, $j \in \mathbb{Z}$, $f(x) \in V_0 \Leftrightarrow f(x-k) \in V_0$ & $f(x) \in V_j \Leftrightarrow f(2x) \in V_{j+1}$.
3. For some “father wavelet” $\phi \in V_0$, $\{\phi(x-k) : k \in \mathbb{Z}^D\}$ is an orthonormal basis of $V_0 \subset L^2(\mathbb{R}^D)$.

For intuition, consider the best-known MRA of $L^2(\mathbb{R})$, namely the Haar wavelet basis. Let $\phi(x) = 1_{\{[0,1)\}}$ be the Haar father wavelet, let $V_0 = \text{Span}\{\phi(x-k) : k \in \mathbb{Z}\}$ be the span of translations of ϕ by an integer, and let V_j defined recursively for all $j \in \mathbb{Z}$ by $V_j = \{f(2x) : f(x) \in V_{j-1}\}$ be the set of horizontal scalings of functions in V_{j-1} by $1/2$. Then, $\{V_j\}_{j \in \mathbb{Z}}$ is an MRA of $L^2(\mathbb{R})$.

The importance of an MRA is that it generates an orthonormal basis of $L^2(\mathbb{R}^D)$, via the following:

Lemma 2 ([36], Section 3.9). *Let $\{V_j\}_{j \in \mathbb{Z}}$ be an MRA of $L^2(\mathbb{R}^D)$ with father wavelet ϕ . Then, for $E = \{0, 1\}^D \setminus (0, \dots, 0)$, there exist “mother wavelets” $\{\psi_\epsilon\}_{\epsilon \in E}$ such that $\{2^{Dj/2} \psi_\epsilon(2^j x - k) : \epsilon \in E, k \in \mathbb{Z}^D\} \cup \{2^{Dj/2} \phi(2^j x - k) : k \in \mathbb{Z}^D\}$ is an orthonormal basis of $V_j \subseteq L^2(\mathbb{R}^D)$.*

Let $\Lambda_j = \{2^{-j}k + 2^{-j-1}\epsilon : k \in \mathbb{Z}^D, \epsilon \in E\} \subseteq \mathbb{R}^D$. Then k, ϵ are uniquely determined for any $\lambda \in \Lambda_j$. Thus, for all $\lambda \in \Lambda := \bigcup_{j \in \mathbb{Z}} \Lambda_j$, we can let $\psi_\lambda(x) = 2^{Dj/2} \psi_\epsilon(2^j x - k)$. Equipped with the orthonormal basis $\{\psi_\lambda : \lambda \in \Lambda\}$ of $L^2(\mathbb{R}^D)$, we are almost ready to define Besov spaces.

For technical reasons (see, e.g., [36, Section 3.9]), we need MRAs of smoother functions than Haar wavelets, which are called *r-regular*. Due to space constraints, *r*-regularity is defined precisely in

³We assume a good optimization algorithm for computing (1), although this is also an active area of research.

Appendix A; we note here that standard r -regular MRAs exist, such as the Daubechies wavelet [11]. We assume for the rest of the paper that the wavelets defined above are supported on $[-A, A]$.

Definition 3 (Besov Space). Let $0 \leq \sigma < r$, and let $p, q \in [1, \infty]$. Given an r -regular MRA of $L^2(\mathbb{R}^D)$ with father and mother wavelets ϕ, ψ respectively, the *Besov space* $B_{p,q}^\sigma(\mathbb{R}^D)$ is defined as the set of functions $f : \mathbb{R}^D \rightarrow \mathbb{R}$ such that, the wavelet coefficients

$$\alpha_k := \int_{\mathbb{R}^D} f(x)\phi(x-k)dx \text{ for } k \in \mathbb{Z}^D \quad \text{and} \quad \beta_\lambda := \int_{\mathbb{R}^D} f(x)\psi_\lambda(x)dx \text{ for } \lambda \in \Lambda,$$

satisfy $\|f\|_{B_{p,q}^\sigma} := \|\{\alpha_k\}_{k \in \mathbb{Z}^D}\|_{l^p} + \left\| \left\{ 2^{j(\sigma+D(1/2-1/p))} \|\{\beta_\lambda\}_{\lambda \in \Lambda_j}\|_{l^p} \right\}_{j \in \mathbb{N}} \right\|_{l^q} < \infty$

The quantity $\|f\|_{B_{p,q}^\sigma}$ is called the *Besov norm* of f , and, for any $L > 0$, we write $B_{p,q}^\sigma(L)$ to denote the closed Besov ball $B_{p,q}^\sigma(L) = \{f \in B_{p,q}^\sigma : \|f\|_{B_{p,q}^\sigma} \leq L\}$. When the constant L is unimportant (e.g., for *rates* of convergence), $B_{p,q}^\sigma$ denotes a ball $B_{p,q}^\sigma(L)$ of finite but arbitrary radius L .

2.2 Formal Problem Statement

Having defined Besov spaces, we now formally state the statistical problem we study in this paper. Fix an r -regular MRA. We observe n IID samples $X_1, \dots, X_n \stackrel{IID}{\sim} p$ from an unknown probability density p lying in a Besov ball $B_{p_g, q_g}^{\sigma_g}(L_g)$ with $\sigma_g < r$. We want to estimate p , measuring error with an IPM $d_{B_{p_d, q_d}^{\sigma_d}(L_d)}$. Specifically, for general $\sigma_d, \sigma_g, p_d, p_g, q_d, q_g$, we seek to bound minimax risk

$$M(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g}) := \inf_{\hat{p}} \sup_{p \in B_{p_g, q_g}^{\sigma_g}} \mathbb{E}_{X_{1:n}} \left[d_{B_{p_d, q_d}^{\sigma_d}}(p, \hat{p}(X_1, \dots, X_n)) \right] \quad (2)$$

of estimating densities in $\mathcal{F}_g = B_{p_g, q_g}^{\sigma_g}$, where the infimum is taken over all estimators $\hat{p}(X_1, \dots, X_n)$. In the rest of this paper, we suppress dependence of $\hat{p}(X_1, \dots, X_n)$ on X_1, \dots, X_n , writing simply \hat{p} .

3 Related Work

The current paper unifies, extends, or improves upon a number of recent and classical results in the nonparametric density estimation literature. Two areas of prior work are most relevant:

Nonparametric estimation over inhomogeneous smoothness spaces First is the classical study of estimation over inhomogeneous smoothness spaces under L^p losses. Nemirovski [41] first noticed that, over classes of regression functions with inhomogeneous (i.e., spatially-varying) smoothness, many widely-used regression estimators, called “linear” estimators (defined precisely in Section 4.2), are provably unable to converge at the minimax optimal rate, in L^2 loss. Donoho et al. [14] identified a similar phenomenon for estimating probability densities in a Besov space $B_{p_g, q_g}^{\sigma_g}$ on \mathbb{R} under $L^{p'_d}$ losses with $p'_d > p_g$, corresponding to the case $\sigma_d = 0, D = 1$ in our work. [14] also showed that the wavelet-thresholding estimator we consider in Section 4.1 *does* converge at the minimax optimal rate. We generalize these phenomena to many new loss functions; in many cases, linear estimators continue to be sub-optimal, whereas the wavelet-thresholding estimator continues to be optimal. We also show that sub-optimality of linear estimators is more pronounced in higher dimensions.

Distribution estimation under IPMs The second, more recent body of results [28, 48, 29] concerns nonparametric distribution estimation under IPM losses. Prior work focused on the case where \mathcal{F} and \mathcal{P} are both Sobolev ellipsoids, corresponding to the case $p_d = q_d = p_g = q_g = 2$ in our work. Notably, over these smaller spaces (of homogeneous smoothness), the linear estimators mentioned above are minimax rate-optimal. Perhaps the most important finding of these works is that the curse of dimensionality pervading classical nonparametric statistics is significantly diminished under weaker loss functions than L^p losses (namely, many IPMs). For example, Singh et al. [48] showed that, when $\sigma_d > D/2$, one can estimate P at the parametric rate $n^{-1/2}$ in the loss $d_{B_{2,2}^{\sigma_d}}$, without *any* regularity assumptions whatsoever on the probability distribution P . We generalize this to other losses $d_{B_{p_d, q_d}^{\sigma_d}}$.

These papers were motivated in part by a desire to understand theoretical properties of GANs, and, in particular, Liang [28] and Singh et al. [48] helped establish (1) as a valid statistical model of GANs. In particular, we note that Singh et al. [48] showed that the implicit generative modeling problem (“sampling”) in terms of which GANs are usually framed, is equivalent, in terms of minimax

convergence rates, to nonparametric density estimation, justifying our focus on the latter problem in this paper. We show, in Section 4.3, that, given a sufficiently good optimization algorithm, GANs based on appropriately constructed deep neural networks can learn Besov densities at the minimax optimal rate. In this context, our results are among the first to suggest theoretically that GANs can outperform classical density estimators (namely, linear estimators mentioned above).

Liu et al. [32] provided general sufficient conditions for weak consistency of GANs in a generalization of the model (1). Since many IPMs, such as Wasserstein distances, metrize weak convergence of probability measures under mild additional assumptions Villani [54], this implies consistency under these IPMs. However, Liu et al. [32] did not study *rates* of convergence.

We end this section with a brief survey of known results for estimating distributions under specific Besov IPM losses, noting that our results (Equations (3) and (4) below) generalize all these rates:

1. L^p Distances: If $\mathcal{F}_d = L^{p'} = B_{p', p'}^0$, then, for distributions P, Q with densities $p, q \in L^p$, $d_{\mathcal{F}_d}(P, Q) = \|p - q\|_{L^p}$. These are the most well-studied losses in nonparametric statistics, especially for $p \in \{1, 2, \infty\}$ [42, 55, 53]. [14] studied the minimax rate of convergence of density estimation over Besov spaces under L^p losses, obtaining minimax rates $n^{-\frac{\sigma_g}{2\sigma_g+D}} + n^{-\frac{\sigma_g+D(1-1/p_g-1/p_d)}{2\sigma_g+D(1-2/p_g)}}$ over general estimators, and $n^{-\frac{\sigma_g-D/p_g+D/p'_d}{2\sigma_g+D}} + n^{-\frac{\sigma_g-D/p_g+D/p'_d}{2\sigma_g+D-2D/p_g+2D/p'_d}}$ when restricted to linear estimators.

2. Wasserstein Distance: If $\mathcal{F}_d = C^1(1) \asymp B_{\infty, \infty}^1$ is the space of 1-Lipschitz functions, then $d_{\mathcal{F}_d}$ is the 1-Wasserstein or Earth mover's distance (via the Kantorovich dual formulation [24, 54]). A long line of work has established convergence rates of the empirical distribution to the true distribution in spaces as general as unbounded metric spaces [56, 26, 47]). In the Euclidean setting, this is well understood [15, 2, 19], although, to the best of our knowledge, minimax lower bounds have been proven only recently [47]; this setting intersects with our work in the case $\sigma_d = 1, \sigma_g = 0, p_d = \infty$, matching our minimax rate of $n^{-1/D} + n^{-1/2}$. More general p -Wasserstein distances W_p ($p \geq 1$) cannot be expressed exactly as IPMs, but, our results complement recent results of Weed and Berthet [57], who showed that, for densities p and q that are bounded above and below (i.e., $0 < m \leq p, q \leq M < \infty$), the bounds $M^{-1/p'} d_{B_{p', \infty}^1}(p, q) \leq W_p(p, q) \leq m^{-1/p'} d_{B_{p', 1}^1}(p, q)^{\frac{1+\sigma_g}{1+\sigma_g}}$ hold; for such densities, our rates match theirs ($n^{-\frac{\sigma_g}{2\sigma_g+D}} + n^{-1/2}$) up to polylogarithmic factors. Weed and Berthet [57] showed that, without the lower-boundedness assumption ($m > 0$), minimax rates under W_p are strictly slower (by a polynomial factor in n).

In machine learning applications, Arora et al. [5] recently used this rate to argue that, for data from a continuous distribution, Wasserstein GANs [4] cannot generalize at a rate faster than $n^{-1/D}$ (at least without additional regularization, as we use in Theorem 9). A variant in which $\mathcal{F}_d \subset C^1 \cap L^\infty$ is both uniformly bounded and 1-Lipschitz gives rise to the Dudley metric [16], which has also been suggested for use in GANs [1]. Finally, we note that the more general distances induced by $\mathcal{F}_d = B_{\infty, \infty}^{\sigma_d}$ have been useful for deriving central limit theorems [8, Section 4.8].

3. Kolmogorov-Smirnov Distance: If $\mathcal{F}_d = \text{BV} \asymp B_{1, \cdot}^1$ is the set of functions of bounded variation, then, in the 1-dimensional case, $d_{\mathcal{F}_d}$ is the well-known Kolmogorov-Smirnov metric [10], and so the famous Dvoretzky-Kiefer-Wolfowitz inequality [34] gives a parametric convergence rate of $n^{-1/2}$.

4. Sobolev Distances: If $\mathcal{F}_d = \mathcal{W}^{\sigma_d, 2} = B_{2, 2}^\sigma$ is a Hilbert-Sobolev space, for $\sigma \in \mathbb{R}$, then $d_{\mathcal{F}_d} = \|\cdot - \cdot\|_{\mathcal{W}^{-\sigma_d, 2}}$ is the corresponding negative Sobolev pseudometric [59]. Recent work [28, 48, 29] established a minimax rate of $n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+1}} + n^{-1/2}$ when $\mathcal{F}_g = \mathcal{W}^{\sigma_g, 2}$ is also a Hilbert-Sobolev space.

4 Main Results

The three **main technical contributions** of this paper are as follows:

1. We prove lower and upper bounds (Theorems 4 and 5, respectively) on minimax convergence rates of distribution estimation under IPM losses when the distribution class $\mathcal{P} = B_{p_g, q_g}^{\sigma_g}$ and the discriminator class $\mathcal{F} = B_{p_d, q_d}^{\sigma_d}$ are Besov spaces; these rates match up to polylogarithmic factors in the sample size n . Our upper bounds use the wavelet-thresholding estimator proposed in Donoho et al. [14], which we show converges at the optimal rate for a much wider range of losses than previously known. Specifically, if $M(\mathcal{F}, \mathcal{P})$ denotes minimax risk (2), we show that for $p'_d \geq p_g, \sigma_g \geq D/p_g$,

$$M(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g}) \asymp \max \left\{ n^{-1/2}, n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}, n^{-\frac{\sigma_g + \sigma_d + D(1-1/p_g - 1/p_d)}{2\sigma_g + D(1-2/p_g)}} \right\}. \quad (3)$$

2. We show (Theorem 7) that, for $p'_d \geq p_g$ and $\sigma_g \geq D/p_g$, no estimator in a large class of distribution estimators, called “linear estimators”, can converge at a rate faster than

$$M_{\text{lin}}(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g}) \gtrsim n^{-\frac{\sigma_g + \sigma_d - D/p_g + D/p'_d}{2\sigma_g + D(1-2/p_g) + 2D/p'_d}}. \quad (4)$$

“Linear estimators” include the empirical distribution, kernel density estimates with uniform bandwidth, and the orthogonal series estimators recently used in Liang [28] and Singh et al. [48]). The lower bound (4) implies that, in many settings (discussed in Section 5), linear estimators converge at sub-optimal rates. This effect is especially pronounced when the data dimension D is large and the distribution P has relatively sparse support (e.g., if P is supported near a low-dimensional manifold).

3. We show that the minimax convergence rate can be achieved by a GAN with generator and discriminator networks of bounded size, after some regularization. As one of the first theoretical results separating performance of GANs from that of classic nonparametric tools such as kernel methods, this may help explain GANs’ successes with high-dimensional data such as images.

4.1 Minimax Rates over Besov Spaces

We now present our main lower and upper bounds for estimating densities that live in a Besov space under a Besov IPM loss. Then, we have the following lower bound on the convergence rate:

Theorem 4. (Lower Bound) Let $r > \sigma_g \geq D/p_g$, then,

$$M(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g}) \gtrsim \max \left(n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}, \left(\frac{\log n}{n} \right)^{\frac{\sigma_g + \sigma_d + D - D/p_g - D/p_d}{2\sigma_g + D - 2D/p_g}} \right) \quad (5)$$

Before giving a corresponding upper bound, we describe the estimator on which it depends.

Wavelet-Thresholding: Our upper bound uses the wavelet-thresholding estimator proposed by [14]:

$$\hat{p}_n = \sum_{k \in \mathbb{Z}} \hat{\alpha}_k \phi_k + \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \hat{\beta}_\lambda \psi_\lambda + \sum_{j=j_0}^{j_1} \sum_{\lambda \in \Lambda_j} \tilde{\beta}_\lambda \psi_\lambda. \quad (6)$$

\hat{p}_n estimates p via its truncated wavelet expansion, with $\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n \phi_k(X_i)$, $\hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^n \psi_\lambda(X_i)$, and $\tilde{\beta}_\lambda = \hat{\beta}_\lambda \mathbf{1}_{\{\hat{\beta}_\lambda > \sqrt{j/n}\}}$ are empirical estimates of respective coefficient of the wavelet expansion of p . As [14] first showed, attaining optimality over Besov spaces requires truncating high-resolution terms (of order $j \in [j_0, j_1]$) when their empirical estimates are too small; this “nonlinear” part of the estimator distinguishes it from the “linear” estimators we study in the next section. The hyperparameters j_0 and j_1 are set to $j_0 = \frac{1}{2\sigma_g + D} \log_2 n$, $j_1 = \frac{1}{2\sigma_g + D - 2D/p_g} \log_2 n$.

Theorem 5. (Upper Bound) Let $r > \sigma_g \geq D/p_g$ and $p'_d > p_g$. Then, for a constant C depending only on p'_d , σ_g , p_g , q_g , D , L_g , L_d and $\|\psi_\epsilon\|_{p'_d}$,

$$M(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g}) \leq C \left(\sqrt{\log n} \left(n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}} + n^{-\frac{\sigma_g + \sigma_d - D/p_g + D/p'_d}{2\sigma_g + D - 2D/p_g}} \right) + n^{-1/2} \right) \quad (7)$$

We will comment only briefly on Theorems 4 and 5 here, leaving extended discussion for Section 5. First, note that the lower bound (5) and upper bound (7) are essentially tight; they differ only by a polylogarithmic factor in n . Second, both bounds contain two main terms of interest. The simpler term, $n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}$, matches the rate observed in the Sobolev case by Singh et al. [48]. The other term is unique to more general Besov spaces. Depending on the values of D , σ_d , σ_g , p_d , and p_g , one of these two terms dominates, leading to two main regimes of convergence rates, which we call the “Sparse” regime and the “Dense” regime. Section 5 discusses these and other interesting phenomena in detail.

4.2 Minimax Rates of Linear Estimators over Besov Spaces

We now show that, for many Besov densities and IPM losses, many widely-used nonparametric density estimators cannot converge at the optimal rate (5). These estimators are as follows:

Definition 6 (Linear Estimator). Let (Ω, \mathcal{F}, P) be a probability space. An estimate \hat{P} of P is said to be *linear* if there exist functions $T_i(X_i, \cdot) : \mathcal{F} \rightarrow \mathbb{R}$ such that for all measurable $A \in \mathcal{F}$,

$$\hat{P}(A) = \sum_{i=1}^n T_i(X_i, A). \quad (8)$$

Classic examples of linear estimators include the empirical distribution $(T_i(X_i, A) = \frac{1}{n} \mathbf{1}_{\{X_i \in A\}})$, the kernel density estimate $(T_i(X_i, A) = \frac{1}{n} \int_A K(X_i, \cdot))$ for some bandwidth $h > 0$ and smoothing kernel $K : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$) and the orthogonal series estimate $(T_i(X_i, A) = \frac{1}{n} \sum_{j=1}^J g_j(X_i) \int_A g_j)$ for some cutoff J and orthonormal basis $\{g_j\}_{j=1}^\infty$ (e.g., Fourier, wavelet, or polynomial) of $L^2(\Omega)$).

Theorem 7 (Minimax rate for Linear Estimators). Suppose $r > \sigma_g \geq D/p_g$,

$$M_{lin} \left(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g} \right) := \inf_{\hat{P}_{lin}} \sup_{p \in \mathcal{F}_g} \mathbb{E}_{X_{1:n}} \left[d_{\mathcal{F}_d} \left(\mu_p, \hat{P} \right) \right] \asymp n^{-\frac{1}{2}} + n^{-\frac{\sigma_g + \sigma_d - D/p_g + D/p'_d}{2\sigma_g + D - 2D/p_g + 2D/p'_d}} + n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}$$

where the inf is over all linear estimates of $p \in \mathcal{F}_g$, and μ_p is the distribution with density p .

One can check that the above error decays no faster than $n^{-\frac{\sigma_g + \sigma_d + D - D/p_g - D/p_d}{2\sigma_g + D - 2D/p_g}}$. Comparing with the rate in Theorem 5, this implies that, in certain cases, convergence the rate for linear estimators is strictly slower than that for general estimators; i.e., linear estimators fail to achieve the minimax optimal rate over certain Besov space. We defer detailed discussion of this phenomenon to Section 5.

4.3 Upper Bounds on a Generative Adversarial Network

Pioneered by Goodfellow et al. [21] as a mechanism for applying deep neural networks to the problem of unsupervised image generation, Generative adversarial networks (GANs) have since been widely applied not only to computer vision [61, 25], but also to such diverse problems and data as machine translation using natural language data [58], discovering drugs [23] and designing materials [46] using molecular structure data, inferring expression levels using gene expression data [12], and sharing patient data under privacy constraints using electronic health records [9]. Besides the Jensen-Shannon divergence used by [21], many GAN formulations have been proposed based on minimizing other losses, including the Wasserstein metric [4, 22], total variation distance [31], χ^2 divergence [33], MMD [27], Dudley metric [1], and Sobolev metric [38]. The diversity of data types and losses with which GANs have been used motivates studying GANs in a very general (nonparametric) setting. In particular, Besov spaces likely comprise the largest widely-studied family of nonparametric smoothness class; indeed, most of the losses listed above are Besov IPMs.

GANs are typically described as a two-player minimax game between a generator network N_g and a discriminator network N_d ; we denote by \mathcal{F}_d the class of functions that can be implemented by N_d and by \mathcal{F}_g the class of distributions that can be implemented by N_g . A recent line of work has argued that a natural statistical model for a GAN as a distribution estimator is

$$\hat{P} := \operatorname{argmin}_{Q \in \mathcal{F}_g} \sup_{f \in \mathcal{F}_d} \mathbb{E}_{X \sim Q} [f(X)] - \mathbb{E}_{X \sim \tilde{P}_n} [f(X)], \quad (9)$$

where \tilde{P}_n is an (appropriately regularized) empirical distribution, and that, when \mathcal{F}_d and \mathcal{F}_g respectively approximate classes \mathcal{F} and \mathcal{P} well, one can bound the risk, under \mathcal{F} -IPM loss, of estimating distributions in \mathcal{P} by (9) [32, 28, 48, 29]. We emphasize, that, as Singh et al. [48] showed, the minimax risk in this framework is identical to that under the “sampling” (or “implicit generative modeling” [37]) framework in terms of which GANs are usually cast.⁴

In this section, we show such a result for Besov spaces; namely, we show the existence of a particular GAN (specifically, a sequence of GANs, necessarily growing with the sample size n), that estimates distributions in a Besov space at the minimax optimal rate (7) under Besov IPM losses. This

⁴As in these previous works, we assume implicitly that the optimum (9) can be computed; this complex saddle-point problem is itself the subject of a related but distinct and highly active area of work [40, 3, 30, 20].

construction uses a standard neural network architecture (a fully-connected neural network with rectified linear unit (ReLU) activations), and a simple data regularizer \tilde{P}_n , namely the wavelet-thresholding estimator described in Section 4.1. Our results extend those of Liang [28] and Singh et al. [48], for Wasserstein loss over Sobolev spaces, to general Besov IPM losses over Besov spaces. We begin with a formal definition of the network architectures that we consider:

Definition 8. A *fully-connected ReLU network* $f_{(A_1, \dots, A_H), (b_1, \dots, b_H)} : \mathbb{R}^W \rightarrow \mathbb{R}$ has the form

$$A_H \eta(A_{H-1} \eta(\cdots \eta(A_1 x + b_1) \cdots) + b_{H-1}) + b_H,$$

where, for each $\ell \in [H-1]$, $A_\ell \in \mathbb{R}^{W \times W}$, and $A_H \in \mathbb{R}^{1 \times W}$ and the ReLU operation $\eta(x) = \max\{x, 0\}$ is applied element-wise to vectors in \mathbb{R}^W .

The size of $f_{(A_1, \dots, A_H), (b_1, \dots, b_H)}(x)$ can be measured in terms of the following four (hyper)parameters: the *depth* H , the *width* W , the *sparsity* $S := \sum_{\ell \in [H]} \|A_\ell\|_{0,0} + \|b_\ell\|_0$ (i.e., the total number of non-zero weights), and the *maximum weight* $B := \max\{\|A_\ell\|_{\infty, \infty}, \|b_\ell\|_\infty : \ell \in [H]\}$. For given size parameters H, W, S, B we write $\Phi(H, W, S, B)$ to denote the set of functions satisfying the corresponding size constraints.

Our results rely on a recent construction (Lemma 17 in the Appendix), by [51], of a fully-connected ReLU network that approximates Besov functions. [51] used this approximation to bound the risk of a neural network for nonparametric regression over Besov spaces, under L^r loss. Here, we use this approximation result Lemma 17 to bound the risk of a GAN for nonparametric distribution estimation over Besov spaces, under the much larger class of Besov IPM losses. Our precise result is as follows:

Theorem 9 (Convergence Rate of a Well-Optimized GAN). *Fix a Besov density class $B_{p_g, q_g}^{\sigma_g}$ with $\sigma_g > D/p_g$ and discriminator class $B_{p_d, q_d}^{\sigma_d}$ with $\sigma_d > D/p_d$. Then, for any desired approximation error $\epsilon > 0$, one can construct a GAN \hat{p} of the form (9) (with \tilde{p}_n) with discriminator network $N_d \in \Phi(H_d, W_d, S_d, B_d)$ and generator network $N_g \in \Phi(H_g, W_g, S_g, B_g)$, s.t. for all $p \in B_{p_g, q_g}^{\sigma_g}$*

$$\mathbb{E} \left[d_{B_{p_d, q_d}^{\sigma_d}}(\hat{p}, p) \right] \lesssim \epsilon + \mathbb{E} d_{B_{p_d, q_d}^{\sigma_d}}(\tilde{p}_n, p)$$

where H_d, H_g grow logarithmically with $1/\epsilon$, $W_d, S_d, B_d, W_g, S_g, B_g$ grow polynomially with $1/\epsilon$ and $C > 0$ is a constant that depends only on $B_{p_d, q_d}^{\sigma_d}$ and $B_{p_g, q_g}^{\sigma_g}$.

This theorem implies that the rate of convergence of the GAN estimate \hat{p} of the form 9 is the same as the convergence rate of the estimator \tilde{p}_n with which the GAN estimate is generated (Here we assume that all distributions have densities). Therefore, given our upper bound from theorem 5 we have the following direct consequence.

Corollary 10. *For a Besov density class $B_{p_g, q_g}^{\sigma_g}$ with $\sigma_g > D/p_g$ and discriminator class $B_{p_d, q_d}^{\sigma_d}$ with $\sigma_d > D/p_d$ there exists an appropriately constructed GAN estimate \hat{p} s.t.*

$$d_{\mathcal{F}_d}(\hat{p}, p) \leq \left(n^{-\eta(D, \sigma_d, p_d, \sigma_g, p_g)} \sqrt{\log n} \right)$$

where $\eta(D, \sigma_d, p_d, \sigma_g, p_g) = \min \left\{ \frac{1}{2}, \frac{\sigma_g + \sigma_d}{2\sigma_g + D}, \frac{\sigma_g + \sigma_d + D - D/p_g - D/p'_d}{2\sigma_g + D(1 - 2/p_g)} \right\}$ is the exponent from (7).

In other words there is a GAN estimate that is minimax rate optimal for a smooth class of densities over an IPM generated by a smooth class of discriminator functions.

5 Discussion of Results

In this section, we discuss some general phenomena that can be gleaned from our technical results.

First, we note that, perhaps surprisingly, q_d and q_g do not appear in our bounds. Tao [52] suggests that q_d and q_g may have only logarithmic effects (contrasted with the polynomial effects of σ_d, p_d, σ_g , and p_g). Thus, a more fine-grained analysis to close the polylogarithmic gap between our lower and upper bounds for general estimators (Theorems 4 and 5) might require incorporating q_d and q_g .

On the other hand, the parameters σ_d, p_d, σ_g , and p_g each play a significant role in determining minimax convergence rates, in both the linear and general cases. We first discuss each of these parameters independently, and then discuss some interactions between them.

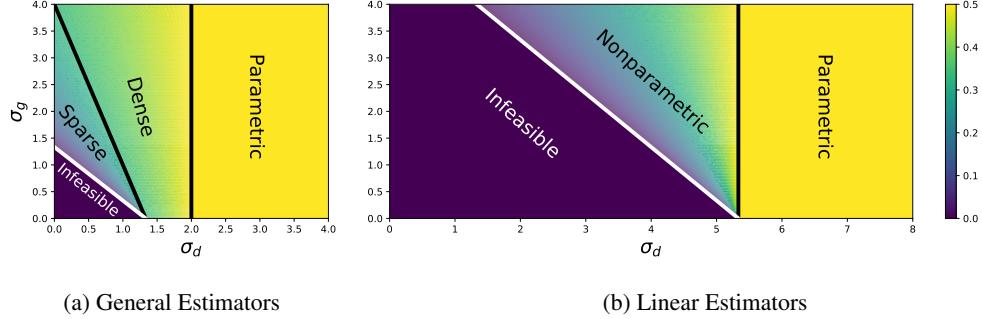


Figure 1: Minimax convergence rates as functions of discriminator smoothness σ_d and distribution function smoothness σ_g , for (a) general and (b) linear estimators, in the case $D = 4$, $p_d = 1.2$, $p_g = 2$. Color shows exponent of minimax convergence rate (i.e., $\alpha(\sigma_d, \sigma_g)$) such that $M(B_{1.2,q_d}^{\sigma_d}(\mathbb{R}^D), B_{2,q_g}^{\sigma_g}(\mathbb{R}^D)) \asymp n^{-\alpha(\sigma_d, \sigma_g)}$), ignoring polylogarithmic factors.

Roles of the smoothness orders σ_d and σ_g As a visual aid for understanding our results, Figure 1 show phase diagrams of minimax convergence rates, as functions of discriminator smoothness σ_d and distribution smoothness σ_g , in the illustrative case $D = 4$, $p_d = 1.2$, $p_g = 2$. When $1/p_g + 1/p_d > 1$, a minimum total smoothness $\sigma_d + \sigma_g \geq D(1/p_d + 1/p_g - 1)$ is needed for consistent estimation to be possible – this fails in the “Infeasible” region of the phase diagrams. Intuitively, this occurs because \mathcal{F}_d is not contained in the topological dual \mathcal{F}'_g of \mathcal{F}_g . For linear estimators, even greater smoothness $\sigma_d + \sigma_g \geq D(1/p_d + 1/p_g)$ is needed. At the other extreme, for highly smooth discriminator functions, both linear and nonlinear estimators converge at the parametric rate $O(n^{-1/2})$, corresponding to the “Parametric” region. In between, rates for linear estimators vary smoothly with σ_d and σ_g , while rates for nonlinear estimators exhibit another phase transition on the line $\sigma_g + 3\sigma_d = D$; to the left lies the “Sparse” case, in which estimation error is dominated by a small number of large errors at locations where the distribution exhibits high local variation; to the right lies the “Dense” case, where error is relatively uniform on the sample space.

The left boundary $\sigma_d = 0$ corresponds to the classical results of Donoho et al. [14], who consequently identified the “Infeasible”, “Sparse”, and “Dense” phases, but not the “Parametric” phase. When restricting to linear estimators, the “Infeasible” region grows and the “Parametric” region shrinks.

Role of the powers p_d and p_g At one extreme ($p_d = \infty$) lie L^1 or total variation loss ($\sigma_d = 0$), Wasserstein loss ($\sigma_d = 1$), and its higher-order generalizations, for which we showed the rate

$$M(B_{\infty,q_d}^{\sigma_d}, B_{p_g,q_g}^{\sigma_g}) \asymp n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + n^{-1/2},$$

generalizing the rate first shown by Singh et al. [48] for Hilbert-Sobolev classes to other distribution classes, such as $\mathcal{F}_g = \text{BV}$. Because discriminator functions in this class exhibit homogeneous smoothness, these losses effectively weight the sample space relatively uniformly in importance, the “Sparse” region in Figure (1a) vanishes, and linear estimators can perform optimally.

At the other extreme ($p_d = 1$) lie L^∞ loss ($\sigma_d = 0$), Kolmogorov-Smirnov loss ($\sigma_d = 1$), and its higher-order generalizations, for which we have shown that the rate is always

$$M(B_{1,q_d}^{\sigma_d}, B_{p_g,q_g}^{\sigma_g}) \asymp n^{-\frac{\sigma_g+\sigma_d+D(1-1/p_d-1/p_g)}{2\sigma_g+D(1-2/p_g)}} + n^{-1/2};$$

except in the parametric regime ($D \leq 2\sigma_d$), this rate differs from that of Singh et al. [48]. Because discriminator functions can have inhomogeneous smoothness, and hence weight some portions of the sample space much more heavily than others, the “Dense” region in Figure 1a vanishes, and linear estimators are always sub-optimal. We note that Sadhanala et al. [45] recently proposed using these higher-order distances (integer $\sigma_d > 1$) in a fast two-sample test that generalizes the well-known Kolmogorov-Smirnov test, improving sensitivity to the tails of distributions; our results may provide a step towards understanding theoretical properties of this test.

Comparison of linear and general rates Letting $\sigma'_g := \sigma_g - D(1/p_g + 1/p_d)$, one can write the sparse term of the linear minimax rate in the same form as the Dense rate, replacing σ_g with σ'_g :

$$M_{\text{lin}} \left(B_{p_d, q_d}^{\sigma_d}, B_{p_g, q_g}^{\sigma_g} \right) \asymp n^{-\frac{\sigma'_g + \sigma_d}{2\sigma'_g + D}}. \quad (10)$$

This is not a coincidence; Morrey's inequality [18, Section 5.6.2] in functional analysis tells us that for general $\sigma_g > D(1/p_g + 1/p_d)$, $\sigma'_g := \sigma_g - D(1/p_g + 1/p_d)$ is largest possible value such that the embedding $B_{p_g, p_g}^{\sigma_g} \subseteq B_{p_d, p_d}^{\sigma'_g}$ holds. In the extreme case $p_d = \infty$ (corresponding to generalizations of total variation loss), one can interpret the rate (10) as saying that linear estimators benefit only from homogeneous (e.g., Hölder) smoothness, and not from weaker inhomogeneous (e.g., Besov) smoothness. For general p_d , linear estimator can still benefit from inhomogeneous smoothness, but to a lesser extent than general minimax optimal estimators.

Conclusions We have shown, up to log factors, unified minimax convergence rates for a large class of pairs of \mathcal{F}_d -IPM losses and distribution classes \mathcal{F}_g . By doing so, we have generalized several phenomena that had observed in special cases previously. First, under sufficiently weak loss functions, distribution estimation is possible at the parametric rate $O(n^{-1/2})$ even over very large nonparametric distribution classes. Second, in many cases, optimal estimation requires estimators that adapt to inhomogeneous smoothness conditions; many commonly used distribution estimators fail to do this, and hence converge at sub-optimal rates, or even fail to converge. Finally, GANs with sufficiently large fully-connected ReLU neural networks using wavelet-thresholding regularization perform statistically minimax rate-optimal distribution estimation over inhomogeneous nonparametric smoothness classes (assuming the GAN optimization problem can be solved accurately). Importantly, since GANs optimize IPM losses much weaker than traditional L^p losses, they may be able to learn reasonable approximations of even high-dimensional distributions with tractable sample complexity, perhaps explaining why they excel in the case of image data. Thus, our results suggest that the curse of dimensionality may be less severe than indicated by classical nonparametric lower bounds.

A Technical Definitions and Notation

As noted in the main text, we need a multiresolution approximation (MRA) satisfying an r -regularity condition, defined as follows:

Definition 11. Given a non-negative integer r , an MRA is called *r-regular* if the function ϕ can be chosen in such a way that, for every $m \in \mathbb{N}$ and multi-index $\alpha = (\alpha_1, \dots, \alpha_D) \in \mathbb{N}^D$ satisfying $|\alpha| \leq r$, for some constant $C_{\alpha, m}$, $|\partial^\alpha \phi(x)| \leq C_{\alpha, m}(1 + |x|)^{-m}$. Here, $\partial^\alpha = (\partial/\partial x_1)^{\alpha_1} \cdots (\partial/\partial x_D)^{\alpha_D}$ is the mixed derivative of index α , $|\alpha| = \sum_{j=1}^D \alpha_j$ and $|x|$ is any of the equivalent norms on a finite dimensional Euclidean space. That is, all derivatives of ϕ of order up to r are bounded and decay at a rate faster than any polynomial.

While constructing an r -regular MRA is nontrivial, it suffices for our purpose to note that r -regular MRAs exist; the most famous example is the Daubechies wavelet [11, 36].

We also note the following result showing that for any function in V_j (i.e., at a certain “level” in the MRA) its L^p norm is equivalent to the l^p sequence norm of its coefficients in the wavelet basis; this helps motivate the sequence-based definition of the Besov norm.

Proposition 12 (Meyer [36], Section 6.10, Proposition 7). *There exist positive constants C, C' s.t. for every $1 \leq p \leq \infty$, $j \in \mathbb{Z}$ and $\{\alpha_k\} \in l^p$, $f(x) = \sum a_k 2^{Dj/2} \psi_\epsilon(2^j x - k)$, $\epsilon \in E$, $k \in \mathbb{Z}^D$,*

$$C \|f\|_p \leq 2^{Dj(1/2 - 1/p)} \left(\sum |a_k|^p \right)^{1/p} \leq C' \|f\|_p.$$

Appendix A.1 of Donoho et al. [14] offers a more extended background of Besov spaces, including how the sequence-based definition corresponds to more conventional smoothness measures (moduli of continuity), as well as some direct connections between Besov spaces and minimax theory for linear estimators.

B Upper Bound - Linear Case

For any density function p let

$$\begin{aligned}\alpha_k^p &= \int \phi_k(x)p(x)dx \\ \beta_\lambda^p &= \int \psi_\lambda(x)p(x)dx\end{aligned}$$

We first show that Besov IPMs essentially measure the distance in co-efficient space between compactly supported densities.

Lemma 13. *For any compactly supported probability densities $p, q \in L_{p'_d}$ where $\mathcal{F}_d = B_{p_d, q_d}^{\sigma_d}$*

$$d_{\mathcal{F}_d}(p, q) = \sup_{f \in \mathcal{F}_d} \left| \sum_{k \in \mathbb{Z}} \alpha_k^f (\alpha_k^p - \alpha_k^q) + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \beta_\lambda^q) \right|$$

where for $f \in \mathcal{F}_d$

$$f = \sum_{k \in \mathbb{Z}} \alpha_k^f \phi_k + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f \psi_\lambda$$

Proof. We notice that the convergence to f above is in the L_∞ norm. So for probability measures P, Q we have,

$$\begin{aligned}d_{\mathcal{F}_d}(p, q) &= \sup_{f \in \mathcal{F}_d} |\mathbb{E}_{X \sim p}[f(X)] - \mathbb{E}_{X \sim q}[f(X)]| \\ &= \sup_{f \in \mathcal{F}_d} \left| \int_{\mathcal{X}} f(x)p(x)dx - \int_{\mathcal{X}} f(x)q(x)dx \right| \\ &= \sup_{f \in \mathcal{F}_d} \left| \int_{\mathcal{X}} \left(\sum_{k \in \mathbb{Z}} \alpha_k^f \phi_k(x) + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f \psi_\lambda(x) \right) (p(x) - q(x)) dx \right|\end{aligned}$$

If p, q are compactly supported on $[-B, B]$ then we can assume WLOG that f is compactly supported on $[-B, B]$ so convergence of f_n to f in L^∞ norm implies convergence in L^1 norm. Therefore,

$$\begin{aligned}d_{\mathcal{F}_d}(P, Q) &= \sup_{f \in \mathcal{F}_d} \left| \sum_{k \in \mathbb{Z}} \int_{\mathcal{X}} \alpha_k^f \phi_k (dP(x) - dQ(x)) + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \int_{\mathcal{X}} \beta_\lambda^f \psi_\lambda (dP(x) - dQ(x)) \right| \\ &= \sup_{f \in \mathcal{F}_d} \left| \sum_{k \in \mathbb{Z}} \alpha_k^f (\alpha_k^p - \alpha_k^q) + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \beta_\lambda^q) \right|\end{aligned}$$

□

We will need the following inequalities to estimate the error of the wavelet estimator under the IPM loss.

The first lemma is the standard upper bound on the m th moment of a sum of IID random variables with bounded variance. The second is a standard concentration inequality used to bound large deviations in our error estimate.

Lemma 14. (Rosenthal's Inequality ([44])) *Let $m \in \mathbb{R}$ and Y_1, \dots, Y_n be IID random variables with $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] \leq \sigma^2$. Then there is a constant c_m that depends only on m s.t.*

$$\begin{aligned}\mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n Y_i \right|^m \right] &\leq c_m \left(\frac{\sigma^m}{n^{m/2}} + \frac{\mathbb{E}|Y_1|^m}{n^{m-1}} \right) && \text{for } 2 < m < \infty, \\ \mathbb{E} \left[\left| \frac{1}{n} \sum_{i=1}^n Y_i \right|^m \right] &\leq \sigma^m n^{-m/2} && \text{for } 1 \leq m \leq 2.\end{aligned}$$

Lemma 15. (Bernstein's Inequality ([6])) If Y_1, \dots, Y_n are IID random variables such that $\mathbb{E}[Y_i] = 0$, $\mathbb{E}[Y_i^2] = \sigma^2$ and $|Y_i| \leq \|Y\|_\infty < \infty$, then

$$\Pr\left(\left|\frac{1}{n} \sum_{i=1}^n Y_i\right| > \lambda\right) \leq 2 \exp\left(-\frac{n\lambda^2}{2(\sigma^2 + \|Y\|_\infty \lambda/3)}\right)$$

where $\|Y\|_\infty = \text{ess sup } Y$.

Given discriminator and generator classes as

$$\begin{aligned}\mathcal{F}_d &= \{f : \|f\|_{p_d, q_d}^{\sigma_d} \leq L_d\} \\ \mathcal{F}_g &= \{p : \|p\|_{p_g, q_g}^{\sigma_g} \leq L_g\} \cap \mathcal{P} \\ \mathcal{P} &= \{p : p \geq 0, \|p\|_{L^1} = 1, \text{supp}(p) \subseteq [-T, T]\},\end{aligned}$$

we decompose $f \in \mathcal{F}_d$ as

$$f = \sum_{k \in \mathbb{Z}} \alpha_k \phi_k + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda \psi_\lambda.$$

We use the linear wavelet estimator to demonstrate the upper bound. Let X_1, \dots, X_n be IID with density $p \in \mathcal{F}_g$ and consider the wavelet estimator of p i.e.

$$\begin{aligned}p &= \sum_{k \in \mathbb{Z}} \alpha_k^p \phi_k + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^p \psi_\lambda \\ \hat{p}_n &= \sum_{k \in \mathbb{Z}} \hat{\alpha}_k \phi_k + \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \hat{\beta}_\lambda \psi_\lambda\end{aligned}$$

where

$$\begin{aligned}\alpha_k^p &= \mathbb{E}_{X \sim p} [\phi_k(X)] & \hat{\alpha}_k &= \frac{1}{n} \sum_{i=1}^n \phi_k(X_i) \\ \beta_\lambda^p &= \mathbb{E}_{X \sim p} [\psi_\lambda(X)] & \hat{\beta}_\lambda &= \frac{1}{n} \sum_{i=1}^n \psi_\lambda(X_i)\end{aligned}$$

Then applying lemma 13, we bound

$$\begin{aligned}d_{\mathcal{F}_d}(p, \hat{p}_n) &\leq \sup_{f \in \mathcal{F}_d} \sum_{k \in \mathbb{Z}} \alpha_k (\alpha_k^p - \hat{\alpha}_k) \\ &\quad + \sup_{f \in \mathcal{F}_d} \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \beta_\lambda (\beta_\lambda^p - \hat{\beta}_\lambda) \\ &\quad + \sup_{f \in \mathcal{F}_d} \sum_{j \geq j_1} \sum_{\lambda \in \Lambda_j} \beta_\lambda \beta_\lambda^p\end{aligned}$$

where the first two terms constitute the stochastic error and the last term is the bias. We bound these separately below. We first prove a few lemmas that will be used repeatedly to upper bound the different terms.

Lemma 16. Let $n_1, n_2 \in \mathbb{N} \cup \{\infty\}$ and η be any sequence of numbers. Then

$$\mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{j=n_1}^{n_2} \sum_{\lambda \in \Lambda_j} \gamma_\lambda \eta_\lambda \leq L_D \sum_{j=n_1}^{n_2} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(\mathbb{E}_{X_1, \dots, X_n} \sum_{\lambda \in \Lambda_j} |\eta_\lambda|^{p'_d} \right)^{1/p'_d}$$

Note that if the above is true also if $\gamma = \alpha^f$ and $n_1 = n_2 = 0$.

Proof. Since $f \in \mathcal{F}_d$, applying Hölder's inequality twice we get,

$$\begin{aligned}
& \mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{j=n_1}^{n_2} \sum_{\lambda \in \Lambda_j} \gamma_\lambda \eta_\lambda \leq \mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{j=n_1}^{n_2} \|\gamma\|_{p_d} \|\eta\|_{p'_d} \\
& \leq \mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \left(\sum_{j=n_1}^{n_2} \left(2^{j(\sigma_d + D/2 - D/p_d)} \|\gamma\|_{p_d} \right)^{q_d} \right)^{1/q_d} \\
& \quad \times \sum_{j=n_1}^{n_2} 2^{-j(\sigma_d + D/2 - D/p_d)} \|\eta\|_{p'_d} \quad (l^1 \subseteq l^{q'_d}) \\
& \leq L_D \sum_{j=n_1}^{n_2} 2^{-j(\sigma_d + D/2 - D/p_d)} \mathbb{E}_{X_1, \dots, X_n} \|\eta\|_{p'_d} \\
& \leq L_D \sum_{j=n_1}^{n_2} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(\mathbb{E}_{X_1, \dots, X_n} \sum_{\lambda \in \Lambda_j} |\eta_\lambda|^{p'_d} \right)^{1/p'_d}
\end{aligned}$$

where p'_d is the conjugate of p_d i.e. $\frac{1}{p_d} + \frac{1}{p'_d} = 1$ and we applied Jensen's to get the last inequality.

□

Lemma 17. Let $f \in B_{p_g, q_g}^{\sigma_g}$ where $\sigma_g > D/p_g$ then

$$\|f\|_\infty \leq 4A \|\psi\|_\infty L_g (1 - 2^{(\sigma_g - D/p_g)q'_g})^{-1/q'_g}$$

This implies that sufficiently smooth Besov spaces $B_{p_g, q_g}^{\sigma_g}$ are uniformly bounded.

Proof. We have that $\sum_{k \in \mathbb{Z}^D} \alpha_k \phi_k + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda \psi_\lambda$ converges to f in L_∞ . So, using the fact that $l^{p_d} \subseteq l^\infty$ and proposition 12,

$$\|f\|_\infty \leq 2A \|\psi\|_\infty \left(\|\{\alpha_k\}_{k \in \mathbb{Z}^D}\|_\infty + \sum_{j \geq 0} 2^{Dj/2} \|\{\beta_\lambda\}_{\lambda \in \Lambda_j}\|_\infty \right).$$

We can upper bound, by Hölder's inequality,

$$\begin{aligned}
\sum_{j \geq 0} 2^{Dj/2} \|\{\beta_\lambda\}_{\lambda \in \Lambda_j}\|_\infty & \leq \sum_{j \geq 0} \frac{1}{2^{j(\sigma_g - D/p_g)}} \times 2^{j(\sigma_g + D/2 - D/p_g)} \|\{\beta_\lambda\}_{\lambda \in \Lambda_j}\|_\infty \\
& \leq \left(\sum_{j \geq 0} \frac{1}{2^{j(\sigma_g - D/p_g)q'_g}} \right)^{1/q'_g} \left(\sum_{j \geq 0} 2^{jq_g(\sigma_g + D/2 - D/p_g)} \|\{\beta_\lambda\}_{\lambda \in \Lambda_j}\|_\infty^{q_g} \right)^{1/q_g} \\
& \leq \left(\frac{1}{1 - 2^{-(\sigma_g - D/p_g)q'_g}} \right)^{1/q'_g} \left(\sum_{j \geq 0} 2^{jq_g(\sigma_g + D/2 - D/p_g)} \|\{\beta_\lambda\}_{\lambda \in \Lambda_j}\|_{p_g}^{q_g} \right)^{1/q_g} \\
& \leq \left(1 - 2^{-(\sigma_g - D/p_g)q'_g} \right)^{-1/q'_g} \|f\|_{p_g, q_g}^{\sigma_g} \\
& \leq \left(1 - 2^{-(\sigma_g - D/p_g)q'_g} \right)^{-1/q'_g} L_g.
\end{aligned}$$

Putting the above together we obtain the required upper bound.

□

We also need a few preliminary results namely, the moments of error of linear estimates of the wavelet coefficients are essentially bounded by $1/\sqrt{n}$ and the probability that this error is large is negligibly small. In particular,

Lemma 18. (Moment Bounds) Let $X_1, \dots, X_n \sim p$, $m \geq 1$ s.t. there is a constant c with $\mathbb{E}_p |\psi_\lambda(X)|^m \leq c 2^{Dj(m/2-1)}$. Let

$$\begin{aligned}\gamma_\lambda^p &= \mathbb{E}[\psi_\lambda(X)], \\ \widehat{\gamma}_\lambda &= \frac{1}{n} \sum_{i=1}^n \psi_\lambda(X_i),\end{aligned}$$

Then for all j s.t. $2^{Dj} \in \mathcal{O}(n)$,

$$\mathbb{E}[|\widehat{\gamma}_{jk} - \gamma_{jk}|^m] \leq cn^{-m/2}.$$

where $c = c_m (\mathbb{E}_p |\psi_\lambda(X)|^2)^{m/2}$ is a constant.

Proof. Since ψ_λ is bounded for every λ , let

$$Y_i = \psi_\lambda(X_i) - \mathbb{E}[\psi_\lambda(X)]$$

then for all $m \geq 1$, applying Jensen's inequality repeatedly we get

$$\begin{aligned}\mathbb{E}[|Y_i|^m] &\leq \mathbb{E}[(|\psi_\lambda(X_i)| + |\mathbb{E}[\psi_\lambda(X_i)]|)^m] && \text{(triangle inequality)} \\ &\leq 2^{m-1} (\mathbb{E}[|\psi_\lambda(X_i)|^m] + |\mathbb{E}[\psi_\lambda(X_i)]|^m) && \text{(Jensen's)} \\ &\leq 2^m \mathbb{E}[|\psi_\lambda(X_i)|^m]. && \text{(Jensen's)}\end{aligned}$$

Therefore, by Rosenthal's inequality we have,

$$\mathbb{E}[|\gamma_\lambda^p - \widehat{\gamma}_\lambda|^m] \leq c_m \left(\left(\frac{1}{p} \mathbb{E}[\psi_\lambda(X)]^2 \right)^{m/2} + c \left(\frac{2^{Dj}}{n} \right)^{(m/2-1)_+} \right) n^{-m/2}$$

where c_m is a constant that only depends on m . Therefore,

$$\mathbb{E}[|\gamma_\lambda^p - \widehat{\gamma}_\lambda|^m] \leq c_m \left(\frac{1}{p} \mathbb{E}[\psi_\lambda(X)]^2 \right)^{m/2} n^{-m/2}$$

□

Note that we have from above $2^{Dj_1} \leq n$ so this bound holds for any $j \leq j_1$.

Lemma 19. (Large Deviations) Let $X_1, \dots, X_n \sim p$ such that for a constant c , $\mathbb{E}_p |\psi_\lambda(X)|^2 \leq c$. Let

$$\begin{aligned}\gamma_\lambda^p &= \mathbb{E}[\psi_\lambda(X)], \\ \widehat{\gamma}_\lambda &= \frac{1}{n} \sum_{i=1}^n \psi_\lambda(X_i),\end{aligned}$$

Let $l = \sqrt{j/n}$ and $\gamma > 0$, then, for all j s.t. $2^{Dj} \in o(n)$, we have,

$$\Pr(|\widehat{\gamma}_\lambda - \gamma_\lambda| > (K/2)l) \leq 2 \times 2^{-\gamma nl^2}$$

where K large enough such that

$$\frac{K^2}{8(c + \|\psi_\epsilon\|_\infty (K/3))} > \log 2\gamma$$

Proof. Applying Bernstein's inequality we have

$$\begin{aligned}\Pr(|\widehat{\gamma}_\lambda - \gamma_\lambda| > (K/2)l) &\leq 2 \exp \left(-\frac{n(K/2)^2 l^2}{2(c + 2^{Dj/2} \|\psi_\epsilon\|_\infty (K/3)l)} \right) \\ &\leq 2 \exp \left(-\frac{K^2 nl^2}{8(L_g + \|\psi_\epsilon\|_\infty (K/3))} \right)\end{aligned}$$

This implies for K satisfying the above condition,

$$\Pr(|\widehat{\gamma}_\lambda - \gamma_\lambda| > (K/2)l) \leq 2 \times 2^{(-\gamma nl^2)}$$

□

Now for every $j \leq j_1$, l satisfies the requirements of the above lemma. So if $nl^2 (= j) \rightarrow \infty$ as $n \rightarrow \infty$ the probability of large deviation goes to zero.

Lemma 20. (Variance) Let $X_1, \dots, X_n \sim p$ where p is compactly supported, such that for a constant c , $\mathbb{E}_p |\psi_\lambda(X)|^m \leq c 2^{Dj(m/2-1)}$. Let $\mathcal{F}_d = B_{p_d, q_d}^{\sigma_d}$, then the variance of a linear wavelet estimator \hat{p} with j_0 terms i.e.

$$\hat{p}_n = \sum_{k \in \mathbb{Z}} \hat{\alpha}_k \phi_k + \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \hat{\beta}_\lambda \psi_\lambda$$

is bounded by

$$d_{\mathcal{F}_d}(\hat{p}_n, \mathbb{E}[\hat{p}_n]) \leq c \left(\frac{1}{\sqrt{n}} + \frac{2^{j_0(D/2-\sigma_d)}}{\sqrt{n}} \right)$$

where $c = c_{p'_d} (\mathbb{E}_p |\psi_\lambda(X)|^2)^{1/2}$ is a constant.

Proof. Since $\mathcal{F}_d = B_{p_d, q_d}^{\sigma_d}$ and p is compactly supported we can, by lemma 13 upper bound

$$\mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{k \in \mathbb{Z}} \alpha_k^f (\alpha_k^p - \hat{\alpha}_k) + \mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \hat{\beta}_\lambda)$$

Since, for a constant c , $\mathbb{E}_p |\psi_\lambda(X)|^m \leq c 2^{Dj(m/2-1)}$ we can apply the moment bound below. For the first term we have, (taking $\gamma = \alpha$ and $n_1 = n_2 = 0$ in lemma 16 above)

$$\begin{aligned} & \mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{k \in \mathbb{Z}} \alpha_k^f (\alpha_k^p - \hat{\alpha}_k) \\ & \leq L_D \left(\sum_k \mathbb{E}_{X_1, \dots, X_n} |\alpha_k^p - \hat{\alpha}_k|^{p'_d} \right)^{1/p'_d} \quad (\text{finitely many terms}) \\ & \leq c L_D \|p\|_\infty \left((T + A) n^{-p'_d/2} \right)^{1/p'_d} \quad (\text{moment bound}) \\ & \leq c n^{-1/2} \end{aligned}$$

where we use the fact only finitely many of the α s are non-zero because of the compactness of the support of the densities we consider and the compactness of the wavelets. Similarly taking $\gamma = \beta$, $n_1 = 0$, $n_2 = j_0$ in lemmas 16 we have, using the moment bound as above,

$$\begin{aligned} & \mathbb{E}_{X_1, \dots, X_n} \sup_{f \in \mathcal{F}_d} \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \hat{\beta}_\lambda) \\ & \leq c \|p\|_\infty L_D \sum_{j=0}^{j_0} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(2^{Dj} (T + A) n^{-p'_d/2} \right)^{1/p'_d} \\ & \leq L_D \sum_{j=0}^{j_0} 2^{-j(\sigma_d + D/2 - D/p_d)} 2^{Dj/p'_d} n^{-1/2} \\ & \leq c L_D \|p\|_\infty \sum_{j=0}^{j_0} 2^{j(D/2 - \sigma_d)} n^{-1/2} \\ & \leq c \|p\|_\infty \begin{cases} 2^{j_0(D/2 - \sigma_d)} n^{-1/2} & \sigma_d \leq D/2 \\ n^{-1/2} & \sigma_d > D/2 \end{cases} \end{aligned}$$

□

Lemma 21. (Bias) Let $X_1, \dots, X_n \sim p$ where $p \in B_{p_g, q_g}^{\sigma_g}$ is compactly supported and $\sigma_g \geq D/p_g$, $\mathcal{F}_d = B_{p_d, q_d}^{\sigma_d}$. Then the bias of a linear wavelet estimator \hat{p} with j_0 terms is bounded by

$$d_{\mathcal{F}_d}(p, \mathbb{E}[\hat{p}_n]) \leq c 2^{-j_0(\sigma_d + \sigma_g - (D/p_g - D/p'_d)_+)} \mathbb{E}_p [\hat{p}_n]$$

where c is a constant that depends on p_d and $\|\psi\|_m$.

Proof. Since p is compactly supported, by lemma 13 we need to upper bound

$$\sup_{\beta \in \mathcal{F}_d} \sum_{j \geq j_1} \sum_{\lambda \in \Lambda} \beta_\lambda^f \beta_\lambda^p$$

Using lemma 16 and the fact that $\sigma_g \geq D/p_g$

$$\begin{aligned} & \sup_{\beta \in \mathcal{F}_d} \sum_{j \geq j_0} \sum_{\lambda \in \Lambda} \beta_\lambda^f \beta_\lambda^p \\ & \leq L_D \sum_{j \geq j_0} 2^{-j(\sigma_d + D/2 - D/p_d)} \|\beta^p\|_{p'_d} \\ & = L_D \sum_{j \geq j_0} \frac{2^{j(\sigma_g + D/2 - D/p_g)}}{2^{j(\sigma_d + \sigma_g + D - D/p_d - D/p_g)}} 2^{j(D/p'_d - D/p_g) +} \|\beta^p\|_{p_g} \\ & \leq L_D \sum_{j \geq j_0} \frac{2^{j(D/p'_d - D/p_g) +}}{2^{j(\sigma_d + \sigma_g + D/p'_d - D/p_g)}} \sup_{j \geq j_0} 2^{j(\sigma_g + D/2 - D/p_g)} \|\beta^p\|_{p_g} \\ & \leq 2^{-j_0(\sigma_d + \sigma_g - (D/p_g - D/p'_d) +)} \|p\|_{p_g q_g}^{\sigma_g} \quad (\sigma_g \geq D/p_g) \\ & \leq c 2^{-j_0(\sigma_d + \sigma_g - (D/p_g - D/p'_d) +)} \end{aligned}$$

□

Using lemmas 21 and 20 we get the following upper bound on the bias and variance of the linear wavelet estimator.

$$c \left(n^{-1/2} + n^{-1/2} 2^{j_0(D/2 - \sigma_d)} + 2^{-j_0(\sigma_g + \sigma_d - D/p_g + D - D/p_d)} \right)$$

which when minimized for j_0 gives,

$$2^{j_0} = n^{1/(2\sigma_g + D + 2D/p'_d - 2D/p_g)}$$

which implies an upper bound of

$$\lesssim n^{-1/2} + n^{-\frac{\sigma_g + \sigma_d - D/p_g + D - D/p_d}{2\sigma_g + D + 2D/p'_d - 2D/p_g}}$$

as desired.

C Proof of the Lower Bound

In this section we prove our main lower bound i.e. Theorem 4 using Fano's lemma and the Varshamov Gilbert bound as summarized below.

Lemma 22. (*Fano's Lemma; Simplified Form of Theorem 2.5 of [53]*)

Fix a family \mathcal{P} of distributions over a sample space \mathcal{X} and fix a pseudo-metric $\rho : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty]$ over \mathcal{P} . Suppose there exists a set $T \subseteq \mathcal{P}$ such that there is a $p_0 \in T$ with $p \ll p_0 \forall p \in T$ and

$$s := \inf_{p, p' \in T} \rho(p, p') > 0 \quad , \quad \sup_{p \in T} D_{KL}(p, p_0) \leq \frac{\log |T|}{16},$$

where $D_{KL} : \mathcal{P} \times \mathcal{P} \rightarrow [0, \infty]$ denotes Kullback-Leibler divergence. Then,

$$\inf_{\hat{p}} \sup_{p \in \mathcal{P}} \mathbb{E} [\rho(p, \hat{p})] \geq \frac{s}{16}$$

where the inf is taken over all estimators \hat{p} .

Lemma 23. (*Varshamov-Gilbert bound ([53])*) Let $\Omega = \{0, 1\}^m$ where $m \geq 8$. Then there exists a subset $\{w^0, \dots, w^M\}$ of Ω such that $w^0 = (0, \dots, 0)$ and

$$\omega(w^j, w^k) \geq \frac{m}{8} \quad \forall 0 \leq j, k \leq M$$

where $M \geq 2^{m/8}$, where $\omega(w^j, w^k) = \sum_{i=1}^m 1_{\{w_i^j \neq w_i^k\}}$ is the Hamming distance.

Proof. (of Theorem 4) We follow the method in Donoho et. al. [14] and separate our proof into “sparse” and “dense” cases. As is standard procedure, for both cases we pick a finite subset of densities from \mathcal{F}_g over which estimation is difficult. Since any function in a Besov space can be defined by its wavelet coefficients we pick a set of densities by an appropriate choice of wavelet coefficients.

Here we also need to pick a subset of functions from \mathcal{F}_d so as to estimate $d_{\mathcal{F}_d}$. Following the method in [47] we pick from \mathcal{F}_d , functions that are analogous to the ones we pick from \mathcal{F}_g so that we measure the difference in the densities along the chosen perturbations.

We now fill in the details. We first let g_0 be a density function supported on an interval that contains $[-A, A]^D$ such that $\|g_0\|_{\sigma_g p_g q_g} \leq L_G/2$ and $g_0 = c > 0$ on $[-A, A]^D$.

At a particular resolution j , we choose 2^{Dj} wavelets with disjoint supports; pick $\psi_\lambda = 2^{Dj/2}\psi_{\epsilon_1}(2^{Dj}x - k)$ indexed by $\lambda = 2^{-j}k + 2^{-(j+1)}\epsilon_1$ s.t. $k \in K_j$ where

$$K_j = \{-(2^j - 1)A + 2lA, l = 0, \dots, (2^j - 1)\}^D$$

and $\epsilon_1 = (1, 0, \dots, 0)$ (i.e. we pick the first wavelet). Note here that if $\lambda \neq \lambda'$ then ψ_λ and $\psi_{\lambda'}$ have disjoint support.

We now describe our choice of densities based on the set of coefficients $\zeta \subseteq \{\tau \in \mathbb{Z}^{|K_j|} : |\tau_\lambda| \leq 1\}$ i.e.

$$\Omega_g := \{g_0 + c_g \sum_{\lambda} \tau_{\lambda} \psi_{\lambda} : \tau \in \zeta, \lambda = 2^{-j}k + 2^{-j-1}\epsilon_1, k \in K_j\}.$$

If we pick c_g to be small enough, every p in Ω_g is a density function and is lower bounded on $[-A, A]^D$. Specifically if c_g s.t.

$$c_g \leq \frac{c}{2\|\psi\|_{\infty}} 2^{-Dj/2}$$

then $\int g_0 + c_g \sum_{\lambda} \tau_{\lambda} \psi_{\lambda} = 1$ (since $\int \psi_{\lambda} = 0$) and,

$$\|g_0 - p\|_{\infty} = c_g 2^{Dj/2} \|\psi\|_{\infty} \leq c/2$$

so that p is lower bounded on the domain of ψ_{λ} by $c/2$ for every λ . This also implies that p is always positive.

Now the following lemma states that if you have a small perturbation of a density s.t. the density is lower bounded on the support of the perturbation then the KL divergence between the perturbed and the original density is upper bounded by the L^2 norm of the perturbation.

Lemma 24. *Let $g = g_0 + h$, g_0 be density functions such that $h \leq g_0$. If $S = \text{supp}(h) \subseteq \text{supp}(g)$ and $c \leq g$ on S , where c is a constant. Then*

$$D_{KL}(g^n, g_0^n) \leq cn \|g_0 - g\|_{L^2}^2$$

Proof. Since $g \leq 2g_0$ we have,

$$\frac{g_0 - g}{g} \geq -\frac{1}{2}$$

so using the fact that $-\log(1+x) \leq x^2 - x$ for all $x \geq -1/2$ we get

$$\begin{aligned} D_{KL}(g^n, g_0^n) &= n D_{KL}(g, g_0) \\ &= n \int_S g(x) \log \frac{g(x)}{g_0(x)} dx \\ &= -n \int_S g(x) \log \left(1 + \frac{g_0(x) - g(x)}{g(x)}\right) dx \\ &\leq n \int_S g(x) \left(\left(\frac{g_0(x) - g(x)}{g(x)}\right)^2 - \frac{g_0(x) - g(x)}{g(x)}\right) dx \\ &= n \int_S \frac{(g_0(x) - g(x))^2}{g(x)} dx \end{aligned}$$

which, since $g \geq c$ on S , is smaller than $cn \int_S (g_0(x) - g(x))^2$ as desired. \square

Using this fact we conclude that for any $p_\tau \in \Omega_g$,

$$KL(p_\tau, g_0) \leq nc_g^2 c \left\| \sum_{\lambda} \tau_\lambda \psi_\lambda \right\|_{L^2}^2 = cn c_g^2 \|\tau\|_2^2$$

Following the technique in [48] we also pick an analogous set of functions that live in \mathcal{F}_d so that we can lower bound $d_{\mathcal{F}_D}$. In particular let

$$\Omega_d := \left\{ c_d \sum_{\lambda} \tau_\lambda \psi_\lambda : \tau \in \zeta, \lambda = 2^{-j}k + 2^{-j-1}\epsilon_1, k \in K_j \right\}$$

It now, only remains to choose appropriate sets ζ for the wavelet coefficients in each of the sparse and dense cases. In the remainder let c be a constant not necessarily the same.

Sparse or low-smoothness case:

For the sparse/lower smoothness case we choose worst case densities to be perturbations along only a specific scaling of the wavelet at a time. In particular, let

$$\zeta = \left\{ \tau : \tau_\lambda = 1, \tau_{\lambda'} = 0, \lambda' \neq \lambda = 2^{-j}k + 2^{-(j+1)}\epsilon_1, k \in K_j \right\}$$

We know from above that for any $c_g \leq c2^{-Dj/2}$, every $p \in \Omega_g$ is a density such that $D_{KL}(p^n, g_0^n) \leq cn c_g^2 \|\tau\|_2$. Now, we need

$$\|g_0 + c_g \psi_\lambda\|_{p_g q_g}^{\sigma_g} \leq \|g_0\|_{p_g q_g}^{\sigma_g} + 2^{j(\sigma_g + D/2 - D/p_g)} c_g \leq L_g$$

so that $\Omega_g \subseteq \mathcal{F}_g$. Since $\sigma_g \geq D/p_g$ the choice of $c_g = c2^{-j(\sigma_g + D/2 - D/p_g)}$ suffices. Similarly, $c_d = L_d 2^{-j(\sigma_d + D/2 - D/p_d)}$ implies $\Omega_d \subseteq \mathcal{F}_d$.

Then we pick j large enough such that the KL divergence between any p_τ and g_0 is small. This enables us to apply Fano's lemma from above and get a lower bound.

So we need $cn c_g^2 \leq \frac{\log |\zeta|}{16} = \frac{\log |K_j|}{16}$ i.e.

$$n \leq cj/c_g^2 \iff n \leq 2^{2j(\sigma_g + D/2 - D/p_g)} j$$

for the KL divergence to be small. Given such a j we have,

$$d_{\mathcal{F}_d}(p_\lambda, p_{\lambda'}) \geq \sup_{f \in \Omega_d} \left| \int c_g(f(x)(\psi_\lambda - \psi_{\lambda'})) dx \right| = \|\psi\|_{L^2}^2 c_g c_d$$

(since, $\|\psi_\lambda\|_{L^2}^2 = \|\psi\|_{L^2}^2$). So, if $2^j = (n/\log n)^{\frac{1}{2\sigma_g + D - 2D/p_g}}$ we have,

$$M(\mathcal{F}_g, \mathcal{F}_d) \gtrsim \left(\frac{\log n}{n} \right)^{\frac{\sigma_g + \sigma_d + D - D/p_g - D/p_d}{2\sigma_g + D - 2D/p_g}}$$

Dense or higher smoothness case:

In the dense case, we choose our set of densities by perturbing g_0 along every scaling of the wavelet simultaneously i.e. let

$$\zeta = \left\{ \tau : \tau_\lambda \in \{-1, +1\} \right\}$$

Now, we need

$$\left\| g_0 + c_g \sum_{\lambda} \tau_\lambda \psi_\lambda \right\|_{p_g q_g}^{\sigma_g} \leq \|g_0\|_{p_g q_g}^{\sigma_g} + 2^{j(\sigma_g + D/2)} c_g \leq L_g$$

so that $\Omega_g \subseteq \mathcal{F}_g$. The choice of $c_g = c2^{-j(\sigma_g + D/2)}$ suffices. Similarly, $c_d = L_d 2^{-j(\sigma_d + D/2)}$ implies $\Omega_d \subseteq \mathcal{F}_d$.

Now the Varshamov-Gilbert bound from above implies we can pick a subset of Ω_G with size at least $2^{|K_j|/8}$ such that $\omega(\tau_\lambda, \tau_{\lambda'}) \geq |K_j|/8$ which gives,

$$\begin{aligned} d_{\mathcal{F}_d}(p_\lambda, p_{\lambda'}) &= \sup_{f \in \Omega_d} \left| \int c_g(f(x)(\psi_\lambda - \psi_{\lambda'}) dx \right| \\ &= c_g c_d \omega(\tau_\lambda, \tau_{\lambda'}) \geq c_g c_d \frac{2^{Dj}}{4} \end{aligned}$$

We pick j large enough such that the KL divergence between any p_τ and g_0 is small. This enables us to apply Fano's lemma from above and get a lower bound. In particular we need, for any $p_\tau \in \Omega_g$, $D_{KL}(p_\tau^n, g_0^n) \leq cnc_g^2 \|\tau\|_2 = cnc_g^2 |K_j|$ to be at most $\frac{\log |\zeta|}{16} = \frac{|K_j|}{16}$ which is equivalent to $n \leq 2^{j(2\sigma_g + D)}$. Then by Fano's lemma the lower bound in the dense case is

$$n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}$$

We combine the above two cases to get the following lower bound on the rate

$$\gtrsim \max(n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}, n^{-\frac{\sigma_g + \sigma_d + D - D/p_g - D/p_d}{2\sigma_g + D - 2D/p_g}})$$

□

D Proof of the Upper Bound

We use the wavelet thresholding estimate as introduced in [14] to get an upper bound on our minimax rate.

Proof. (of theorem 5) We first upper bound our error by three terms namely, the stochastic error, the bias and the non-linear terms. The stochastic error is bounded above as usual by the above moment bound. The bias is bounded above by virtue of our density belonging to the besov space $B_{p_g, q_g}^{\sigma_g}$. The non-linear terms are more delicate. We follow the procedure in [14] and split them into four groups the first two of which are shown to be negligible as the probability of large deviations falls exponentially rapidly from Bernstein's inequality above. We simplify the upper bounds on the other two terms considerably by paying a penalty on the rate by the factor that is logarithmic in the sample size. We now fill in the details of the proof.

We first let our discriminator and generator classes be

$$\begin{aligned} \mathcal{F}_d &= \{f : \|f\|_{p_d, q_d}^{\sigma_d} \leq L_d\} \\ \mathcal{F}_g &= \{p : \|p\|_{p_g, q_g}^{\sigma_g} \leq L_g\} \cap \mathcal{P} \\ \mathcal{P} &= \{p : p \geq 0, \|p\|_{L^1} = 1, \text{supp}(p) \subseteq [-T, T]\} \end{aligned}$$

Given X_1, \dots, X_n be IID with density $p \in \mathcal{F}_g$ and the thresholded wavelet estimator of p i.e.

$$\begin{aligned} p &= \sum_{k \in \mathbb{Z}} \alpha_k^p \phi_k + \sum_{j \geq 0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^p \psi_\lambda \\ \widehat{p}_n &= \sum_{k \in \mathbb{Z}} \widehat{\alpha}_k \phi_k + \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \widehat{\beta}_\lambda \psi_\lambda + \sum_{j=j_0}^{j_1} \sum_{\lambda \in \Lambda_j} \widetilde{\beta}_\lambda \psi_\lambda \end{aligned}$$

where

$$\begin{aligned} \alpha_k^p &= \mathbb{E}_{X \sim p} [\phi_k(X)] & \widehat{\alpha}_k &= \frac{1}{n} \sum_{i=1}^n \phi_k(X_i) \\ \beta_\lambda^p &= \mathbb{E}_{X \sim p} [\psi_\lambda(X)] & \widehat{\beta}_\lambda &= \frac{1}{n} \sum_{i=1}^n \psi_\lambda(X_i) \\ & & & \widetilde{\beta}_\lambda = \widehat{\beta}_\lambda \mathbf{1}_{\{\widehat{\beta}_\lambda > t\}} \end{aligned}$$

with $t = K\sqrt{j/n}$, where K is a constant to be specified later, and

$$2^{j_0} = n^{\frac{1}{2\sigma_g + D}}$$

$$2^{j_1} = n^{\frac{1}{2\sigma_g + D - 2D/p_g}}$$

we can upper bound the error as,

$$\begin{aligned} d_{\mathcal{F}_d}(p, \hat{p}_n) &\leq \sup_{f \in \mathcal{F}_d} \sum_{k \in \mathbb{Z}} \alpha_k^f (\alpha_k^p - \hat{\alpha}_k) &+ \sup_{f \in \mathcal{F}_d} \sum_{j=0}^{j_0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \hat{\beta}_\lambda) \\ &+ \sup_{f \in \mathcal{F}_d} \sum_{j \geq j_0} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \tilde{\beta}_\lambda) &+ \sup_{f \in \mathcal{F}_d} \sum_{j \geq j_1} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f \beta_\lambda^p \end{aligned}$$

where the first three terms constitute the stochastic error (the non-linear terms or thresholded terms are also called ‘detail’ terms [14]) and the last term is the bias. In particular:

1. The first term in our upper bound of the risk is the stochastic error or the variance of a linear wavelet estimator with j_0 terms. Note that since $\sigma_g \geq D/p_g$ $p \in \mathcal{F}_g$ implies by lemma 17 that $\|p\|_\infty < \infty$. Then by substitution

$$\mathbb{E}_p |\psi_\lambda(X)|^{p'_d} \leq 2^{-Dj(p'_d/2-1)}$$

Therefore by lemma 20 we have an upper bound here of

$$cn^{-1/2} (2^{j_0(D/2-\sigma_d)} + 1) \lesssim n^{-\frac{\sigma_g+\sigma_d}{2\sigma_g+D}} + n^{-1/2}$$

2. The third term is the bias of a linear wavelet estimator with j_1 terms which by lemma 21 for $p'_d \geq p_g$ is bounded above by

$$c2^{-j_1(\sigma_d+\sigma_g-D/p_g+D/p'_d)} \lesssim n^{-\frac{\sigma_g+\sigma_d+D-D/p_g-D/p_d}{2\sigma_g+D-2D/p_g}}$$

3. For the second term we have, by lemmas 13 and 16

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_d} \sum_{j \geq j_0} \sum_{\lambda \in \Lambda} \beta_\lambda^f (\beta_\lambda^p - \tilde{\beta}_\lambda) &\leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left(\mathbb{E} \sum_{\lambda \in \Lambda_j} |\beta_\lambda^p - \tilde{\beta}_\lambda|^{p'_d} 1_A \right)^{1/p'_d} \\ &\leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left(\sum_{\lambda \in \Lambda_j} \mathbb{E} |\beta_\lambda^p - \tilde{\beta}_\lambda|^{p'_d} 1_A \right)^{1/p'_d} \end{aligned}$$

where we are only summing over finitely many terms. The set A is given by the following cases:

(For the upper bounds of the first two cases we have chosen γ (which in turn determines the value of K) to be large enough so that the exponent of 2^j is negative and thus we can upper bound the geometric series by a constant multiple of the first term.)

- a) Let A be the set of k s.t. $\hat{\beta}_\lambda > t$ and $\beta_\lambda^p < t/2$ and $r \geq 1/p'_d$ then

$$\begin{aligned} L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left(\sum_{\lambda \in \Lambda_j} \mathbb{E} |\beta_\lambda^p - \tilde{\beta}_\lambda|^{p'_d} 1_A \right)^{1/p'_d} \\ \leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d+D/2-D/p_d)} \left(\sum_{\lambda \in \Lambda_j} (\mathbb{E} |\beta_\lambda^p - \tilde{\beta}_\lambda|^{p'_d r})^{1/r} \Pr(A)^{1/r'} \right)^{1/p'_d} \end{aligned}$$

Using the large deviation and moment bound

$$\Pr(A) \leq \Pr(|\hat{\beta}_\lambda - \beta_\lambda^p| \geq t/2) \leq c2^{-\gamma j}$$

we get,

$$\begin{aligned} &\leq c \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(2^{Dj} n^{-p'_d/2} 2^{-j\gamma/r'} \right)^{1/p'_d} \\ &\leq c \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d - D/p'_d)} n^{-1/2} 2^{-\gamma j/p'_d r'} \\ &\leq cn^{-1/2} \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d - D/2 + \gamma/p'_d r')} \\ &\leq cn^{-1/2} 2^{-j_0(\sigma_d - D/2 + \gamma/p'_d r')} \\ &\lesssim n^{-\frac{\sigma_g + \sigma_d + \gamma/p'_d r'}{2\sigma_g + D}}, \end{aligned}$$

which is negligible compared to the linear term.

(b) Let B be the set of k s.t. $\hat{\beta}_\lambda < t$ and $\beta_\lambda^p > 2t$ then same as above

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}_d} \sum_{j=j_0}^{j_1} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f \beta_\lambda^p 1_B &\leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \|\beta_\lambda^p\|_{p'_d} (\Pr(B))^{1/p'_d} \\ &\leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \|\beta_\lambda^p\|_{p'_d} 2^{-\gamma j/p'_d} \\ &\leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + \sigma'_g + \gamma/p'_d)} \sup_{j_0 \leq j \leq j_1} 2^{j(\sigma'_g + D/2 - D/p'_d)} \|\beta_\lambda\|_{p'_d} \\ &\leq L_D L_G \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + \sigma'_g + \gamma/p'_d)} \\ &\leq L_D L_G C 2^{-j_0(\sigma_d + \sigma'_g + \gamma/p'_d)} \\ &\lesssim n^{-\frac{\sigma_d + \sigma'_g + \gamma}{2\sigma_g + D}} \end{aligned}$$

which is negligible compared to the bias term.

(c) Let C be the set of k s.t. $\hat{\beta}_\lambda > t$ and $\beta_\lambda^p > t/2$ then:

$$\begin{aligned}
& \mathbb{E} \sup_{f \in \mathcal{F}_d} \sum_{j=j_0}^{j_1} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f (\beta_\lambda^p - \tilde{\beta}_\lambda) 1_C \\
& \leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(\sum_{k \in C} \mathbb{E} |\beta_\lambda^p - \tilde{\beta}_\lambda|^{p'_d} \right)^{1/p'_d} \\
& \leq L_D \sum_{j=j_0}^{j_1} C n^{-1/2} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(\sum_{k \in C} \left(\frac{2\beta_\lambda^p \sqrt{n/j}}{K} \right)^{p_g} \right)^{1/p'_d} \\
& \leq L_D \sum_{j=j_0}^{j_1} C n^{-1/2} (\sqrt{n/j})^{p_g/p'_d} 2^{-j(\sigma_d + D/2 - D/p_d)} \|\beta^p\|_{p_g}^{p_g/p'_d} \\
& \leq L_D \sum_{j=j_0}^{j_1} C n^{-1/2} (\sqrt{n/j})^{p_g/p'_d} 2^{-j(\sigma_d + D/2 - D/p_d)} 2^{-j(\sigma_g + D/2 - D/p_g)p_g/p'_d} \\
& \quad \sup_{j_0 \leq j \leq j_1} \|\beta\|_{p_g} 2^{j(\sigma_g + D/2 - D/p_g)} \\
& \leq C L_D L_G n^{1/2(p_g/p'_d - 1)} \sum_{j=j_0}^{j_1} 2^{-j((\sigma_g + D/2)p_g/p'_d + \sigma_d - D/2)} j^{-p_g/2p'_d} \\
& \leq C L_D L_G n^{1/2(p_g/p'_d - 1)} 2^{-j_m((\sigma_g + D/2)p_g/p'_d + \sigma_d - D/2)}
\end{aligned}$$

where

$$j_m = \begin{cases} j_0 & (2\sigma_g + D)p_g \geq (D - 2\sigma_d)p'_d \\ j_1 & (2\sigma_g + D)p_g \leq (D - 2\sigma_d)p'_d \end{cases}$$

In the first case we have an upper bound of

$$\lesssim n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + D}}$$

and in the second case we have an upper bound of

$$\lesssim n^{-\frac{\sigma_g + \sigma_d + D - D/p_d - D/p_g}{2\sigma_g + D - 2D/p_g}}$$

(d) Let E be the set of k s.t. $\hat{\beta}_\lambda < t$ and $\beta_\lambda^p < 2t$ then:

$$\begin{aligned}
& \mathbb{E} \sup_{f \in \mathcal{F}_d} \sum_{j=j_0}^{j_1} \sum_{\lambda \in \Lambda_j} \beta_\lambda^f \beta_\lambda^p 1_D \\
& \leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(\sum_{\lambda \in \Lambda_j} |\beta_\lambda^p|^{p'_d} \right)^{1/p'_d} \\
& \leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} \left(\sum_{\lambda \in \Lambda_j} |\beta_\lambda^p|^{p_g} (2t)^{p'_d - p_g} \right)^{1/p'_d} \quad p'_d \geq p_g \\
& = L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} (2t)^{1-p_g/p'_d} \|\beta\|_{p_g}^{p_g/p'_d} \\
& \leq L_D \sum_{j=j_0}^{j_1} 2^{-j(\sigma_d + D/2 - D/p_d)} (2\sqrt{j/n})^{1-p_g/p'_d} 2^{-j(\sigma_g + D/2 - D/p_g)p_g/p'_d} L_g \\
& \leq c \sqrt{j_1} n^{1/2(p_g/p'_d - 1)} \sum_{j=j_0}^{j_1} 2^{-j((\sigma_g + D/2)p_g/p'_d + \sigma_d - D/2)} j^{-p_g/2p'_d} \\
& \lesssim \left(n^{-\frac{\sigma_g + \sigma_d}{2\sigma_g + 1}} + n^{-\frac{\sigma_g + \sigma_d + D - D/p_d - D/p_g}{2\sigma_g + D - 2D/p_g}} \right) \sqrt{\log n}
\end{aligned}$$

□

E Proof of Theorem 7

Lower Bound

Proof. Just as in the proof of the lower bound above we let $j \geq 0$ and

$$\Omega_g := \{g_0 \pm c_g \psi_\lambda : \lambda = 2^{-j}k + 2^{-j-1}\epsilon_1, k \in K_j\}$$

where $\epsilon_1 = (1, 0, \dots, 0)$. Here we let $g_0 = 2^{Dj}c$ on at least $[-A, A]^D$ and

$$c_g = \min \left(\frac{c}{2 \|\psi\|_\infty} 2^{-Dj/2}, \frac{L_g}{2} 2^{-j(\sigma_g + D/2 - D/p_g)} \right)$$

such that $\Omega_g \subseteq \mathcal{F}_g$. We also let

$$\Omega_d := \{c_d \sum_{\lambda} \tau_{\lambda} \psi_{\lambda} : \lambda = 2^{-j}k + 2^{-j-1}\epsilon_1, k \in K_j, \|\tau\| \leq L_d\}$$

s.t.

$$c_d \leq L_d 2^{-j(\sigma_d + D/2 - 1/p_d)}$$

i.e. $\Omega_d \subseteq \mathcal{F}_d$.

Then for any linear estimate \widehat{P} with $\widehat{\alpha}_{\lambda} = \int \psi_{\lambda}(x) d\widehat{P}(x)$,

$$\begin{aligned} & \sup_{P \in \mathcal{F}_g} \mathbb{E} \sup_{f \in \mathcal{F}_d} \left| \int f(x) (dP(x) - d\widehat{P}(x)) \right| \\ & \geq \sup_{p \in \Omega_g} \mathbb{E} \sup_{f \in \Omega_d} \left| \int f(x) (p(x) dx - d\widehat{P}(x)) \right| \\ & = \sup_{\lambda: k \in K_j} \frac{c_d}{2} \mathbb{E}_{g_0 + c_g \psi_{\lambda}} \left(\sup_{\tau: \|\tau\|_{p_d} \leq L_d} \sum_{\lambda' \neq \lambda} |\tau_{\lambda'} \widehat{\alpha}_{\lambda'}| + |\tau_{\lambda}| |c_g - \widehat{\alpha}_{\lambda}| \right) \\ & \quad + \mathbb{E}_{g_0 - c_g \psi_{\lambda}} \left(\sup_{\tau: \|\tau\|_{p_d} \leq L_d} \sum_{\lambda' \neq \lambda} |\tau_{\lambda'} \widehat{\alpha}_{\lambda'}| + |\tau_{\lambda}| |c_g - \widehat{\alpha}_{\lambda}| \right) \\ & \geq \sup_{\lambda: k \in K_j} \frac{c_d}{2} \\ & \quad \sup_{\tau: \|\tau\|_{p_d} \leq L_d} \left(\sum_{\lambda' \neq \lambda} \mathbb{E}_{g_0 + c_g \psi_{\lambda}} |\tau_{\lambda'}| |\widehat{\alpha}_{\lambda'}| + \mathbb{E}_{g_0 - c_g \psi_{\lambda}} |\tau_{\lambda'}| |\widehat{\alpha}_{\lambda'}| + \mathbb{E}_{g_0 + c_g \psi_{\lambda}} |\tau_{\lambda}| |c_g - \widehat{\alpha}_{\lambda}| + \mathbb{E}_{g_0 - c_g \psi_{\lambda}} |\tau_{\lambda}| |c_g - \widehat{\alpha}_{\lambda}| \right) \\ & = \sup_{\lambda: k \in K_j} \frac{c_d}{2} \\ & \quad \left(\sum_{\lambda' \neq \lambda} \left(\mathbb{E}_{g_0 + c_g \psi_{\lambda}} |\widehat{\alpha}_{\lambda'}| \right)^{p'_d} + \left(\mathbb{E}_{g_0 - c_g \psi_{\lambda}} |\widehat{\alpha}_{\lambda'}| \right)^{p'_d} + \left(\mathbb{E}_{g_0 + c_g \psi_{\lambda}} |c_g - \widehat{\alpha}_{\lambda}| \right)^{p'_d} + \left(\mathbb{E}_{g_0 - c_g \psi_{\lambda}} |c_g - \widehat{\alpha}_{\lambda}| \right)^{p'_d} \right)^{1/p'_d} \\ & \geq c_d \left(\frac{1}{2^{Dj}} \sum_{\lambda' \neq \lambda} \left(\mathbb{E}_{g_0 + c_g \psi_{\lambda}} |\widehat{\alpha}_{\lambda'}| \right)^{p'_d} + \left(\mathbb{E}_{g_0 - c_g \psi_{\lambda}} |\widehat{\alpha}_{\lambda'}| \right)^{p'_d} + \left(\mathbb{E}_{g_0 + c_g \psi_{\lambda}} |c_g - \widehat{\alpha}_{\lambda}| \right)^{p'_d} + \left(\mathbb{E}_{g_0 - c_g \psi_{\lambda}} |c_g - \widehat{\alpha}_{\lambda}| \right)^{p'_d} \right)^{1/p'_d} \end{aligned}$$

Now the expression inside the brackets is bounded below in [14] appendix A.3 by $n^{-1/2} 2^{jD/p'_d}$ where $2^j = n^{\frac{1}{2\sigma_g - 2D/p_g + 2D/p'_d + D}}$ which implies a lower bound in our case of

$$\begin{aligned} & c 2^{-j(\sigma_d + D/2 - D/p_d)} n^{-1/2} 2^{Dj/p'_d} \\ & = c 2^{j(D/2 - \sigma_d)} n^{-1/2} \end{aligned}$$

which gives us a lower bound of

$$\gtrsim n^{-\frac{\sigma_d + \sigma_g - D/p_g + D/p'_d}{2\sigma_g - 2D/p_g + 2D/p'_d + D}}$$

as desired. \square

F Proof of Theorem 9

Here, we prove the following theorem, which upper bounds the risk of an appropriately constructed GAN for learning Besov distributions:

Theorem 25 (Convergence Rate of a Well-Optimized GAN). *Fix a Besov density class $B_{p_g, q_g}^{\sigma_g}$ with $\sigma_g > D/p_g$ and discriminator class $B_{p_d, q_d}^{\sigma_d}$ with $\sigma_d > D/p_d$. Then, for any desired approximation error $\epsilon > 0$, one can construct a GAN \hat{p} of the form (9) (with \tilde{p}_n) with discriminator network $N_d \in \Phi(H_d, W_d, S_d, B_d)$ and generator network $N_g \in \Phi(H_g, W_g, S_g, B_g)$, s.t. for all $p \in B_{p_g, q_g}^{\sigma_g}$*

$$\mathbb{E} \left[d_{B_{p_d, q_d}^{\sigma_d}}(\hat{p}, p) \right] \lesssim \epsilon + \mathbb{E} d_{B_{p_d, q_d}^{\sigma_d}}(\tilde{p}_n, p)$$

where H_d, H_g grow logarithmically with $1/\epsilon$, $W_d, S_d, B_d, W_g, S_g, B_g$ grow polynomially with $1/\epsilon$ and $C > 0$ is a constant that depends only on $B_{p_d, q_d}^{\sigma_d}$ and $B_{p_g, q_g}^{\sigma_g}$.

Our statistical guarantees rely on a recent construction, by Suzuki [51], of a fully-connected ReLU network that approximates Besov functions. Specifically, we leverage the following result:

Lemma 26 (Proposition 1 of Suzuki [51]). *Suppose that $p, q, r \in (0, \infty]$ and $\sigma > \delta := D(1/p - 1/r)_+$ and let $\nu = (\sigma - \delta)/(2\delta)$. Then, for sufficiently small $\epsilon \in (0, 1)$, there exists a constant $C > 0$, depending only on D, p, q, r, σ , such that, for some*

$$H \leq C \log(1/\epsilon), \quad W \leq C \epsilon^{-D/\sigma}, \quad S \leq C \epsilon^{-D/\sigma} \log(1/\epsilon), \quad B \leq C \epsilon^{-(D/\nu+1)(1 \vee (D/p - \sigma)_+)/\sigma},$$

$$\Phi(H, W, S, B) \subseteq B_{p, q}^\sigma(1) \text{ and } \Phi(H, W, S, B) \text{ approximates } B_{p, q}^\sigma(1) \text{ to accuracy } \epsilon \text{ in } L^r; \text{ i.e.,}$$

$$\sup_{f \in B_{p, q}^\sigma(1)} \inf_{f \in \Phi(H, W, S, B)} \|f - \tilde{f}\|_{L^r} \leq C\epsilon.$$

Proof. Liang [28, Inequality 2.2] showed that we can decompose the error, for densities \hat{p}, p ,

$$\begin{aligned} d_{\mathcal{F}_d}(\hat{p}, p) &\leq \inf_{q \in \Phi(H_g, W_g, S_g, B_g)} d_{\mathcal{F}_d}(p, q) \\ &\quad + 2 \sup_{f \in \mathcal{F}_d} \inf_{g \in \Phi(H_d, W_d, S_d, B_d)} \|f - g\|_\infty \\ &\quad + d_{\Phi(H_d, W_d, S_d, B_d)}(p, \tilde{p}_n) + d_{\mathcal{F}_d}(p, \tilde{p}_n), \end{aligned}$$

where the 3 summands above correspond respectively the error of approximating \mathcal{F}_g by $\Phi(L_g, W_g, S_g, B_g)$ (generator approximation error), the error of approximating \mathcal{F}_d by $\Phi(L_d, W_d, S_d, B_d)$ (discriminator approximation error), and statistical error.

To bound the first term, note also that, since we assumed $\sigma_d > D/p_d$, we have the embedding $B_{p_d, q_d}^{\sigma_d} \subseteq L^\infty$, and, in particular, $M := \sup_{f \in B_{p_d, q_d}^{\sigma_d}} \|f\|_{L^\infty} < \infty$. Thus, by Hölder's inequality, the assumption that densities in \mathcal{P} are supported only on $[-T, T]$, and Lemma 26 (with $r = \infty$),

$$\inf_{q \in \mathcal{F}_g} d_{\mathcal{F}_d}(p, q) \leq \inf_{q \in \mathcal{F}_g} (p, q) \sup_{f \in \mathcal{F}_D} \|f\|_{L^1([-T, T])} \|p - q\|_{L^\infty} \leq 2MT\epsilon.$$

To bound the second term, simply observe that, by Lemma 26 (with $r = \infty$),

$$\sup_{f \in \mathcal{F}_d} \inf_{g \in \Phi(L_g, W_g, S_g, B_g)} \|f - g\|_\infty \leq \epsilon.$$

Since, by Lemma 26, $\Phi(L_d, W_d, S_d, B_d) \subseteq B_{p_d, q_d}^{\sigma_d}$, the last term is immediately bounded (in expectation) by $d_{\mathcal{F}_d}(\tilde{p}_n, p)$. Combining the bounds on these three terms gives

$$d_{\mathcal{F}_d}(\hat{p}, p) \leq 2(MT + 1)\epsilon + 2d_{\mathcal{F}_d}(\tilde{p}_n, p).$$

\square

References

- [1] Ehsan Abbasnejad, Javen Shi, and Anton van den Hengel. Deep Lipschitz networks and Dudley GANs, 2018. URL <https://openreview.net/pdf?id=rkw-j1b0W>.
- [2] Miklós Ajtai, János Komlós, and Gábor Tusnády. On optimal matchings. *Combinatorica*, 4(4):259–264, 1984.
- [3] Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.
- [4] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.
- [5] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). *arXiv preprint arXiv:1703.00573*, 2017.
- [6] S.N. Bernstein. On a modification of Chebyshev’s inequality and on the error in Laplace formula. *Collected Works, Izd-vo ’Nauka’, Moscow (in Russian)*, 4:71–80, 1964.
- [7] Leon Bottou, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab. Geometrical insights for implicit generative modeling. In *Braverman Readings in Machine Learning. Key Ideas from Inception to Current State*, pages 229–268. Springer, 2018.
- [8] Louis HY Chen, Larry Goldstein, and Qi-Man Shao. *Normal approximation by Stein’s method*. Springer Science & Business Media, 2010.
- [9] Edward Choi, Siddharth Biswal, Bradley Malin, Jon Duke, Walter F Stewart, and Jimeng Sun. Generating multi-label discrete patient records using generative adversarial networks. *arXiv preprint arXiv:1703.06490*, 2017.
- [10] Wayne W Daniel et al. *Applied nonparametric statistics*. Houghton Mifflin, 1978.
- [11] Ingrid Daubechies. *Ten lectures on wavelets*, volume 61. Siam, 1992.
- [12] Kamran Ghasedi Dizaji, Xiaoqian Wang, and Heng Huang. Semi-supervised generative adversarial network for gene expression inference. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1435–1444. ACM, 2018.
- [13] Hao-Wen Dong and Yi-Hsuan Yang. Towards a deeper understanding of adversarial losses. *arXiv preprint arXiv:1901.08753*, 2019.
- [14] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.
- [15] RM Dudley. The speed of mean Glivenko-Cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.
- [16] RM Dudley. Speeds of metric probability convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 22(4):323–332, 1972.
- [17] GK Dziugaite, DM Roy, and Z Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Uncertainty in Artificial Intelligence-Proceedings of the 31st Conference, UAI 2015*, pages 258–267, 2015.
- [18] Lawrence C Evans. *Partial differential equations*. American Mathematical Society, 2010.
- [19] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.
- [20] Gauthier Gidel, Reyhane Askari Hemmat, Mohammad Pezeshki, Gabriel Huang, Remi Lepriol, Simon Lacoste-Julien, and Ioannis Mitliagkas. Negative momentum for improved game dynamics. *arXiv preprint arXiv:1807.04740*, 2018.
- [21] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [22] Ishaaq Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *Advances in Neural Information Processing Systems*, pages 5767–5777, 2017.

- [23] Artur Kadurin, Sergey Nikolenko, Kuzma Khrabrov, Alex Aliper, and Alex Zhavoronkov. drugan: an advanced generative adversarial autoencoder model for de novo generation of new molecules with desired molecular properties in silico. *Molecular pharmaceutics*, 14(9):3098–3104, 2017.
- [24] Leonid Vasilevich Kantorovich and Gennady S Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ.*, 13(7):52–59, 1958.
- [25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. *arXiv preprint*, 2017.
- [26] Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *arXiv preprint arXiv:1804.10556*, 2018.
- [27] Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. In *Advances in Neural Information Processing Systems*, pages 2203–2213, 2017.
- [28] Tengyuan Liang. How well can generative adversarial networks (GAN) learn densities: A nonparametric view. *arXiv preprint arXiv:1712.08244*, 2017.
- [29] Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.
- [30] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint arXiv:1802.06132*, 2018.
- [31] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. Pacgan: The power of two samples in generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 1505–1514, 2018.
- [32] Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In *Advances in Neural Information Processing Systems*, pages 5551–5559, 2017.
- [33] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2794–2802, 2017.
- [34] Pascal Massart. The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The annals of Probability*, pages 1269–1283, 1990.
- [35] Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? *arXiv preprint arXiv:1801.04406*, 2018.
- [36] Yves Meyer. *Wavelets and operators*, volume 1. Cambridge university press, 1992.
- [37] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.
- [38] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.
- [39] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.
- [40] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent gan optimization is locally stable. In *Advances in Neural Information Processing Systems*, pages 5585–5595, 2017.
- [41] Arkadi S Nemirovski. Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk. SSR Teckhn. Kibernet.*, 3:50–60, 1985.
- [42] Arkadi S Nemirovski. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.
- [43] David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.
- [44] Haskell P. Rosenthal. On the subspaces of $l^p(p > 2)$ spanned by sequences of independent random variables. *Israel Journal of Mathematics*, 8(3):273–303, 1970.

- [45] Veeranjaneyulu Sadhanala, Aaditya Ramdas, Yu-Xiang Wang, and Ryan Tibshirani. A higher-order kolmogorov-smirnov test. In *International Conference on Artificial Intelligence and Statistics*, 2019.
- [46] Benjamin Sanchez-Lengeling, Carlos Outeiral, Gabriel L Guimaraes, and Alan Aspuru-Guzik. Optimizing distributions over molecular space. an objective-reinforced generative adversarial network for inverse-design chemistry (organic). *ChemRxiv Preprint*, 2017.
- [47] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.
- [48] Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabas Poczos. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems 31*, pages 10246–10257, 2018. URL <http://papers.nips.cc/paper/8225-nonparametric-density-estimation-under-adversarial-losses.pdf>.
- [49] Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.
- [50] Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert RG Lanckriet. Non-parametric estimation of integral probability metrics. In *Information Theory Proceedings (ISIT), 2010 IEEE International Symposium on*, pages 1428–1432. IEEE, 2010.
- [51] Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. *arXiv preprint arXiv:1810.08033*, 2018.
- [52] Terence Tao. A type diagram for function spaces. <https://terrytao.wordpress.com/tag/besov-spaces/>, 2011.
- [53] Alexandre B Tsybakov. *Introduction to nonparametric estimation. Revised and extended from the 2004 French original. Translated by Vladimir Zaiats*. Springer Series in Statistics. Springer, New York, 2009.
- [54] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
- [55] Larry Wassermann. All of nonparametric statistics. *New York*, 2006.
- [56] Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.
- [57] Jonathan Weed and Quentin Berthet. Estimation of smooth densities in wasserstein distance. *arXiv preprint arXiv:1902.01778*, 2019.
- [58] Zhen Yang, Wei Chen, Feng Wang, and Bo Xu. Improving neural machine translation with conditional sequence generative adversarial nets. *arXiv preprint arXiv:1703.04887*, 2017.
- [59] Kosaku Yosida. Functional analysis. reprint of the sixth (1980) edition. classics in mathematics. *Springer-Verlag, Berlin*, 11:14, 1995.
- [60] Werner Zellinger, Bernhard A Moser, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Robust unsupervised domain adaptation for neural networks via moment alignment. *Information Sciences*, 2019.
- [61] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017.