Nonparametric Density Estimation with Adversarial Losses

Shashank Singh[1,2,*] **Ananya Uppal**[3] **Boyue Li**[4]
Chun-Liang Li[1] **Manzil Zaheer**[1] **Barnabás Póczos**[1]
[1]Machine Learning Department [2]Department of Statistics & Data Science
[3]Department of Mathematical Sciences [4]Language Technologies Institute
Carnegie Mellon University
[*]Corresponding Author: `sss1@cs.cmu.edu`

Abstract

We study minimax convergence rates of nonparametric density estimation under a large class of loss functions called "adversarial losses", which, besides classical $\mathcal{L}^p$ losses, includes maximum mean discrepancy (MMD), Wasserstein distance, and total variation distance. These losses are closely related to the losses encoded by discriminator networks in generative adversarial networks (GANs). In a general framework, we study how the choice of loss and the assumed smoothness of the underlying density together determine the minimax rate. We also discuss implications for training GANs based on deep ReLU networks, and more general connections to learning implicit generative models in a minimax statistical sense.

1 Introduction

Generative modeling, that is, modeling the distribution from which data are drawn, is a central task in machine learning and statistics. Often, prior information is insufficient to guess the form of the data distribution. In statistics, generative modeling in these settings is usually studied from the perspective of nonparametric density estimation, in which histogram, kernel, orthogonal series, and nearest-neighbor methods are popular approaches with well-understood statistical properties [64, 61, 19, 9].

Recently, machine learning has made significant empirical progress in generative modeling, using such tools as generative adversarial networks (GANs) and variational autoencoders (VAEs). Computationally, these methods are quite distinct from classical density estimators; they usually rely on deep neural networks, fit by black-box optimization, rather than a mathematically prescribed smoothing operator, such as convolution with a kernel or projection onto a finite-dimensional subspace.

Ignoring the implementation of these models, from the perspective of statistical analysis, these recent methods have at least two main differences from classical density estimators. First, they are *implicit*, rather than *explicit* (or *prescriptive*) generative models [14, 38]; that is, rather than an estimate of the probability of a set or the density at a point, they return novel samples from the data distribution. Second, in many recent models, loss is measured not with $\mathcal{L}^p$ distances (as is conventional in nonparametric statistics [64, 61]), but rather with weaker losses, such as

$$d_{\mathcal{F}_D}(P, Q) = \sup_{f \in \mathcal{F}_D} \left| \mathop{\mathbb{E}}_{X \sim P} [f(X)] - \mathop{\mathbb{E}}_{X \sim Q} [f(X)] \right|, \tag{1}$$

where $\mathcal{F}_D$ is a *discriminator class* of bounded, Borel-measurable functions, and P and Q lie in a *generator class* $\mathcal{F}_G$ of Borel probability measures on a sample space $\mathcal{X}$. Specifically, GANs often use losses of this form because (1) can be approximated by a discriminator neural network.

This paper attempts to help bridge the gap between traditional nonparametric statistics and these recent advances by studying these two differences from a statistical minimax perspective. Specifically,

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

under traditional statistical smoothness assumptions, we identify (i.e., prove matching upper and lower bounds on) minimax convergence rates for density estimation under several losses of the form (1). We also discuss some consequences this has for particular neural network implementations of GANs based on these losses. Finally, we study connections between minimax rates for explicit and implicit generative modeling, under a plausible notion of risk for implicit generative models.

1.1 Adversarial Losses

The quantity (1) has been extensively studied, in the case that $\mathcal{F}_D$ is a reproducing kernel Hilbert space (RKHS) under the name *maximum mean discrepancy* (MMD; [23, 60]), and, in a wider context under the name *integral probability metric* (IPM; [40, 55, 56, 10]). [7] also called (1) the *$\mathcal{F}_D$-distance*, or, when $\mathcal{F}_D$ is a family of functions that can be implemented by a neural network, the *neural network distance*. We settled on the name "adversarial loss" because, without assuming any structure on $\mathcal{F}_D$, this matches the intuition of the expression (1), namely that of an adversary selecting the most distinguishing linear projection $f \in \mathcal{F}_D$ between the true density P and our estimate $\widehat{P}$ (e.g., by the discriminator network in a GAN).

One can check that $d_{\mathcal{F}_D} : \mathcal{F}_G \times \mathcal{F}_G \to [0, \infty]$ is a pseudometric (i.e., it is non-negative and satisfies the triangle inequality, and $d_{\mathcal{F}_D}(P, Q) > 0 \Rightarrow P \neq Q$, although $d_{\mathcal{F}_D}(P, Q) = 0 \not\Rightarrow P = Q$ unless $\mathcal{F}_D$ is sufficiently rich). Many popular (pseudo)metrics between probability distributions, including $\mathcal{L}^p$ [64, 61], Sobolev [31, 39], maximum mean discrepancy (MMD; [60])/energy [59, 47], total variation [63], (1-)Wasserstein/Kantorovich-Rubinstein [25, 63], Kolmogorov-Smirnov [27, 52], and Dudley [17, 1] metrics can be written in this form, for appropriate choices of $\mathcal{F}_D$.

The **main contribution of this paper** is a statistical analysis of the problem of estimating a distribution P from n IID observations using the loss $d_{\mathcal{F}_D}$, in a minimax sense over $P \in \mathcal{F}_G$, for fairly general nonparametric smoothness classes $\mathcal{F}_D$ and $\mathcal{F}_G$. General upper and lower bounds are given in terms of decay rates of coefficients of functions in terms of an (arbitrary) orthonormal basis of $\mathcal{L}^2$ (including, e.g., Fourier or wavelet bases); note that this does *not* require $\mathcal{F}_D$ or $\mathcal{F}_G$ to have any inner product structure, only that $\mathcal{F}_D \subseteq \mathcal{L}^1$. We also discuss some consequences for density estimators based on neural networks (such as GANs), and consequences for the closely related problem of implicit generative modeling (i.e., of generating novel samples from a target distribution, rather than estimating the distribution itself), in terms of which GANs and VAEs are usually cast.

Paper Organization: Section 2 provides our formal problem statement and required notation. Section 3 discusses related work on nonparametric density estimation, with further discussion of the theory of GANs provided in the Appendix. Sections 4 and 5 contain our main theoretical upper and lower bound results, respectively. Section 6 develops our general results from Sections 4 and 5 into concrete minimax convergence rates for some important special cases. Section 7 uses our theoretical results to upper bound the error of perfectly optimized GANs. Section 8 establishes some theoretical relationships between the convergence of optimal density estimators and optimal implicit generative models. The Appendix provides proofs of our theoretical results, further applications, further discussion of related and future work, and experiments on simulated data that support our theoretical results.

2 Problem Statement and Notation

We now provide a formal statement of the problem studied in this paper in a very general setting, and then define notation required for our specific results.

Formal Problem Statement: Let $P \in \mathcal{F}_G$ be an unknown probability measure on a sample space $\mathcal{X}$, from which we observe n IID samples $X_{1:n} = X_1, ..., X_n \overset{IID}{\sim} P$. In this paper, we are interested in using the samples $X_{1:n}$ to estimate the measure P, with error measured using the adversarial loss $d_{\mathcal{F}_D}$. Specifically, for various choices of spaces $\mathcal{F}_D$ and $\mathcal{F}_G$, we seek to bound the minimax rate

$$M(\mathcal{F}_D, \mathcal{F}_G) := \inf_{\widehat{P}} \sup_{P \in \mathcal{F}_G} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{F}_D} \left(P, \widehat{P}(X_{1:n}) \right) \right]$$

of estimating distributions assumed to lie in a class $\mathcal{F}_G$, where the infimum is taken over all estimators $\widehat{P}$ (i.e., all (potentially randomized) functions $\widehat{P} : \mathcal{X}^n \to \mathcal{F}_G$). We will discuss both the case when $\mathcal{F}_G$ is known *a priori* and the *adaptive* case when it is not.

2.1 Notation

For a non-negative integer n, we use $[n] := \{1, 2, ..., n\}$ to denote the set of positive integers at most n. For sequences $\{a_n\}_{n\in\mathbb{N}}$ and $\{b_n\}_{n\in\mathbb{N}}$ of non-negative reals, $a_n \lesssim b_n$ and, similarly $b_n \gtrsim a_n$, indicate the existence of a constant $C > 0$ such that $\limsup_{n\to\infty} \frac{a_n}{b_n} \leq C$. $a_n \asymp b_n$ indicates $a_n \lesssim b_n \lesssim a_n$. For functions $f : \mathbb{R}^d \to \mathbb{R}$, we write

$$\lim_{\|z\|\to\infty} f(z) := \sup_{\{z_n\}_{n\in\mathbb{N}}:\|z_n\|\to\infty} \lim_{n\to\infty} f(z_n),$$

where the supremum is taken over all diverging $\mathbb{R}^d$-valued sequences. Note that, by equivalence of finite-dimensional norms, the exact choice of the norm $\|\cdot\|$ does not matter here. We will also require summations of the form $\sum_{z\in\mathcal{Z}} f(z)$ in cases where $\mathcal{Z}$ is a (potentially infinite) countable index set and $\{f(z)\}_{z\in\mathcal{Z}}$ is summable but not necessarily absolutely summable. Therefore, to ensure that the summation is well-defined, the order of summation will need to be specified, depending on the application (as in, e.g., Section 6).

Fix the sample space $\mathcal{X} = [0, 1]^d$ to be the d-dimensional unit cube, over which λ denotes the usual Lebesgue measure. Given a measurable function $f : \mathcal{X} \to \mathbb{R}$, let, for any Borel measure μ on $\mathcal{X}$, $p \in [1, \infty]$, and $L > 0$,

$$\|f\|_{\mathcal{L}^p_\mu} := \left(\int_{\mathcal{X}} |f|^p \, d\mu\right)^{1/p} \qquad \text{and} \qquad \mathcal{L}^p_\mu(L) := \left\{f : \mathcal{X} \to \mathbb{R} \;\middle|\; \|f\|_{\mathcal{L}^p_\mu} < L\right\}$$

(taking the appropriate limit if $p = \infty$) denote the Lebesgue norm and ball of radius L, respectively.

Fix an orthonormal basis $\mathcal{B} = \{\phi_z\}_{z\in\mathcal{Z}}$ of $\mathcal{L}^2_\lambda$ indexed by a countable family $\mathcal{Z}$. To allow probability measures P without densities (i.e., $P \not\ll \mu$), we assume each basis element $\phi_z : \mathcal{X} \to \mathbb{R}$ is a bounded function, so that $\widetilde{P}_z := \mathbb{E}_{X\sim P}\left[\phi_z(X)\right]$ is well-defined. For constants $L > 0$ and $p \geq 1$ and real-valued net $\{a_z\}_{z\in\mathcal{Z}}$, our results pertain to generalized ellipses of the form

$$\mathcal{H}_{p,a}(L) = \left\{ f \in \mathcal{L}^1(\mathcal{X}) : \left(\sum_{z\in\mathcal{Z}} a_z^p |\widetilde{f}_z|^p\right)^{1/p} \leq L \right\}.$$

(where $\widetilde{f}_z := \int_{\mathcal{X}} f\phi_z \, d\mu$ is the z^{th} coefficient of f in the basis $\mathcal{B}$). We sometimes omit dependence on L (e.g., $\mathcal{H}_{p,a} = \mathcal{H}_{p,a}(L)$) when its value does not matter (e.g., when discussing *rates* of convergence).

A particular case of interest is the scale of the Sobolev spaces defined for $s, L \geq 0$ and $p \geq 1$ by

$$\mathcal{W}^{s,p}(L) = \left\{ f \in \mathcal{L}^1(\mathcal{X}) : \left(\sum_{z\in\mathcal{Z}} |z|^{sp} |\widetilde{f}_z|^p\right)^{1/p} \leq L \right\}.$$

For example, when $\mathcal{B}$ is the standard Fourier basis and s is an integer, for a constant factor c depending only on s and the dimension d,

$$\mathcal{W}^{s,p}(cL) := \left\{ f \in \mathcal{L}^p_\lambda \middle| \left\| f^{(s)} \right\|_{\mathcal{L}^p_\lambda} < L \right\}$$

corresponds to the natural standard smoothness class of $\mathcal{L}^p_\lambda$ functions having s^{th}-order (weak) derivatives $f^{(s)}$ in $\mathcal{L}^p_\lambda(L)$ [31]).

3 Related Work

Our results apply directly to many of the losses that have been used in GANs, including 1-Wasserstein distance [5, 24], MMD [32], Sobolev distances [39], and the Dudley metric [1]. As discussed in the Appendix, slightly different assumptions are required to obtain results for the Jensen-Shannon divergence (used in the original GAN formulation of [22]) and other f-divergences [42].

Given their generality, our results relate to many prior works on distribution estimation, including classical work in nonparametric statistics and empirical process theory, as well as more recent work studying Wasserstein distances and MMD. Here, we briefly survey known results for these problems. There have also been a few other statistical analyses of the GAN framework; due to space constraints, we discuss these works in the Appendix.

$\mathcal{L}^2_\lambda$ distances: Classical work on nonparametric statistics has typically focused on the problem of smooth density estimation under $\mathcal{L}^2_\lambda$ loss, corresponding the adversarial loss $d_{\mathcal{F}_D}$ with $\mathcal{F}_D = \mathcal{L}^2_\lambda(L_D)$ (the Hölder dual) of $\mathcal{L}^2$ [64, 61]. In this case, when $\mathcal{F}_G = \mathcal{W}^{t,2}(L_G)$ is a Sobolev class, then the minimax rate is typically $M(\mathcal{F}_D, \mathcal{F}_G) \asymp n^{-\frac{t}{2t+d}}$, matching the rates given by our main results.

Maximum Mean Discrepancy (MMD): When $\mathcal{F}_D$ is a reproducing kernel Hilbert space (RKHS), the adversarial loss $d_{\mathcal{F}_D}$ has been widely studied under the name *maximum mean discrepancy (MMD)* [23, 60]. When the RKHS kernel is translation-invariant, one can express $\mathcal{F}_D$ in the form $\mathcal{H}_{2,a}$, where a is determined by the spectrum of the kernel, and so our analysis holds for MMD losses with translation-invariant kernels (see Example 6). To the best of our knowledge, minimax rates for density estimation under MMD loss have not been established in general; our analysis suggests that density estimation under an MMD loss is essentially equivalent to the problem of estimating kernel mean embeddings studied in [60], as both amount to density estimation while ignoring bias, and both typically have a parametric $n^{-1/2}$ minimax rate. Note that the related problems of estimating MMD itself, and of using it in statistical tests for homogeneity and dependence, have received extensive theoretical treatment [23, 46].

Wasserstein Distances: When $\mathcal{F}_D = \mathcal{W}^{1,\infty}(L)$ is the class of 1-Lipschitz functions, $d_{\mathcal{F}_D}$ is equivalent to the *(order-*1*) Wasserstein* (also called *earth-mover's* or *Kantorovich-Rubinstein*) distance. In this case, when $\mathcal{F}_G$ contains all Borel measurable distributions on $\mathcal{X}$, minimax bounds have been established under very general conditions (essentially, when the sample space $\mathcal{X}$ is an arbitrary totally bounded metric space) in terms of covering numbers of $\mathcal{X}$ [65, 50, 30]. In the particular case that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^d$ of full dimension (i.e., having non-empty interior, comparable to the case $\mathcal{X} = [0,1]^d$ that we study here), these results imply a minimax rate of $M(\mathcal{F}_D, \mathcal{F}_G) = n^{-\min\left\{\frac{1}{2}, \frac{1}{d}\right\}}$, matching our rates. Notably, these upper bounds are derived using the empirical distribution, which *cannot* benefit from smoothness of the true distribution (see [65]). At the same time, it is obvious to generalize smoothing estimators to sample spaces that are not sufficiently nice subsets of $\mathbb{R}^d$.

Sobolev IPMs: The closest work to the present is [33], which we believe was the first work to analyze how convergence rates jointly depend on (Sobolev) smoothness restrictions on both $\mathcal{F}_D$ and $\mathcal{F}_G$. Specifically, for Sobolev spaces $\mathcal{F}_D = \mathcal{W}^{s,p}$ and $\mathcal{F}_G = \mathcal{W}^{t,q}$ with $p, q \geq 2$ (compare our Example 4), they showed

$$n^{-\frac{s+t}{2t+d}} \lesssim M(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}) \lesssim n^{-\frac{s+t}{2(s+t)+d}}. \tag{2}$$

Our main results in Sections 4 and 5 improve on this in two main ways. First, our results generalize to and are tight for many spaces besides Sobolev spaces. Examples include when $\mathcal{F}_D$ is a reproducing kernel Hilbert space (RKHS) with translation-invariant kernel, or when $\mathcal{F}_G$ is the class of all Borel probability measures. Our bounds also allow other (e.g., wavelet) estimators, whereas the bounds of [33] are for the (uniformly $\mathcal{L}^\infty_\lambda$-bounded) Fourier basis. Second, the lower and upper bounds in (2) diverge by a factor polynomial in n. We tighten the upper bound to match the lower bound, identifying, for the first time, minimax rates for many problems of this form (e.g., $M(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}) \asymp n^{-\frac{s+t}{2t+d}}$ in the Sobolev case above). Our analysis has several interesting implications:

1. When $s > d/2$, the convergence becomes *parametric*: $M(W^{s,2}, \mathcal{F}_G) \asymp n^{-1/2}$, for *any class of distributions $\mathcal{F}_G$*. This highlights that the loss $d_{\mathcal{F}_D}$ is quite weak for large s, and matches known minimax results for the Wasserstein case $s = 1$ [12, 50].
2. Our upper bounds, as in [33], are for smoothing estimators (namely, the orthogonal series estimator 3). In contrast, previous analyses of Wasserstein loss focused on convergence of the (unsmoothed) empirical distribution $\widehat{P}_E$ to the true distribution, which typically occurs at rate of $\asymp n^{-1/d} + n^{-1/2}$, where d is the intrinsic dimension of the support of P [12, 65, 50]. Moreover, if $\mathcal{F}_G$ includes all Borel probability measures, this rate is minimax optimal [50]. The loose upper bound of [33] left open the questions of whether (when $s < d/2$) a very small amount ($t \in \left(0, \frac{2s^2}{d-2s}\right]$) of smoothness improves the minimax rate and, more importantly, whether smoothed estimators are outperformed by $\widehat{P}_E$ in this regime. Our results imply that, for $s < d/2$, the minimax rate strictly improves with smoothness t, and that, as long as the support of P

has full dimension, the smoothed estimator *always* converges faster than $\widehat{P}_E$. An important open problem is to simultaneously leverage when P is smooth *and* has support of low intrinsic dimension; many data (e.g., images) likely enjoy both these properties.

3. [33] suggested over-smoothing the estimate (the smoothing parameter ζ discussed in Equation (3) below was set to $\zeta \asymp n^{\frac{1}{2(s+t)+d}}$) compared to the case of $\mathcal{L}^2_\lambda$ loss, and hence it was not clear how to design estimators that adapt to unknown smoothness under losses $d_{W^{s,p}}$. We show that the optimal smoothing ($\zeta \asymp n^{\frac{1}{2t+d}}$) under $d_{W^{s,p}}$ loss is identical to that under $\mathcal{L}^2_\lambda$ loss, and we use this to design an adaptive estimator (see Corollary 5).
4. Our bounds imply improved performance bounds for optimized GANs, discussed in Section 7.

4 Upper Bounds for Orthogonal Series Estimators

This section gives upper bounds on the adversarial risk of the following density estimator. For any finite set $Z \subseteq \mathcal{Z}$, let $\widehat{P}_Z$ be the truncated series estimate

$$\widehat{P}_Z := \sum_{z \in Z} \widehat{P}_z \phi_z, \quad \text{where, for any } z \in \mathcal{Z}, \quad \widehat{P}_z := \frac{1}{n} \sum_{i=1}^{n} \phi_z(X_i). \tag{3}$$

Z is a tuning parameter that typically corresponds to a smoothing parameter; for example, when $\mathcal{B}$ is the Fourier basis and $Z = \{z \in \mathbb{Z}^d : \|z\|_\infty \leq \zeta\}$ for some $\zeta > 0$, $\widehat{P}_Z$ is equivalent to a kernel density estimator using a sinc product kernel $K_h(x) = \prod_{j=1}^{d} \frac{2}{h} \frac{\sin(2\pi x/h)}{2\pi x/h}$ with bandwidth $h = 1/\zeta$ [43].

We now present our main upper bound on the minimax rate of density estimation under adversarial losses. The upper bound is given by the orthogonal series estimator given in Equation (3), but we expect kernel and other standard linear density estimators to converge at the same rate.

Theorem 1 (Upper Bound)**.** *Suppose that $\mu(\mathcal{X}) < \infty$ and there exist constants $L_D, L_G > 0$, real-valued nets $\{a_z\}_{z\in\mathcal{Z}}$, $\{b_z\}_{z\in\mathcal{Z}}$ such that $\mathcal{F}_D = \mathcal{H}_{p,a}(\mathcal{X}, L_D)$ and $\mathcal{F}_G = \mathcal{H}_{q,b}(\mathcal{X}, L_G)$, where $p, q \geq 1$. Let $p' = \frac{p}{p-1}$ denote the Hölder conjugate of p. Then, for any $P \in \mathcal{F}_G$,*

$$\mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right) \right] \leq L_D \frac{c_{p'}}{\sqrt{n}} \left\| \left\{ \frac{\|\phi_z\|_{\mathcal{L}^\infty_P}}{a_z} \right\}_{z \in Z} \right\|_{p'} + L_D L_G \left\| \left\{ \frac{1}{a_z b_z} \right\}_{z \in \mathcal{Z}\setminus Z} \right\|_{\frac{1}{1-1/p-1/q}} \tag{4}$$

The two terms in the bound (4) demonstrate a bias-variance tradeoff, in which the first term (*variance*) increases with the truncation set Z and is typically independent of the class $\mathcal{F}_G$ of distributions, while the second term (*bias*) decreases with Z at a rate depending on the complexity of $\mathcal{F}_G$.

Corollary 2 (Sufficient Conditions for Parametric Rate)**.** *Consider the setting of Theorem 1. If*

$$A := \sum_{z \in \mathcal{Z}} \frac{\|\phi_z\|^2_{\mathcal{L}^\infty_P}}{a_z^2} < \infty \quad \textit{and} \quad \max\{a_z, b_z\} \to \infty.$$

whenever $\|z\| \to \infty$, then, the minimax rate is parametric; specifically, $M(\mathcal{F}_D, \mathcal{F}_G) \leq L_D\sqrt{A/n}$. In particular, letting $c_z := \sup_{x\in\mathcal{X}} |\phi_z(x)|$ for each $z \in \mathcal{Z}$, this occurs whenever $\sum_{z\in\mathcal{Z}} \frac{c_z^2}{a_z^2} < \infty$.

In many contexts (e.g., if $P \ll \lambda$ and $\lambda \ll P$), the simpler condition $\sum_{z\in\mathcal{Z}} \frac{c_z^2}{a_z^2} < \infty$ suffices. The first, and slightly weaker condition in terms of $\|\phi_z\|^2_{\mathcal{L}^\infty_P}$ is useful when we restrict $\mathcal{F}_G$; e.g., if $\mathcal{B}$ is the wavelet basis (defined in the Appendix) and $\mathcal{F}_G$ contains only discrete distributions supported on at most k points, then $\|\phi_{i,j}\|^2_{\mathcal{L}^\infty_P} = 0$ for all but k values of $j \in [2^i]$, at each resolution $i \in \mathbb{N}$. The assumption $\max\left\{\lim_{\|z\|\to\infty} a_z, \lim_{\|z\|\to\infty} b_z\right\} = \infty$ is quite mild; for example, the Riemann-Lebesgue lemma and the assumption that $\mathcal{F}_D$ is bounded in $\mathcal{L}^\infty_\lambda \subseteq \mathcal{L}^1_\lambda$ together imply that this condition always holds if $\mathcal{B}$ is the Fourier basis.

5 Minimax Lower Bound

In this section, we lower bound the minimax risk $M(\mathcal{F}_D, \mathcal{F}_G)$ of distribution estimation under $d_{\mathcal{F}_D}$ loss over $\mathcal{F}_G$, for the case when $\mathcal{F}_D = \mathcal{H}_{p,a}$ and $\mathcal{F}_G := \mathcal{H}_{q,b}$ are generalized ellipses. As we show in some examples in Section 6, our lower bound rate matches our upper bound rate in Theorem 1 for many spaces $\mathcal{F}_D$ and $\mathcal{F}_G$ of interest. Our lower bound also suggests that the assumptions in Corollary 2 are typically necessary to guarantee the parametric convergence rate $n^{-1/2}$.

Theorem 3 (Minimax Lower Bound)**.** *Fix $\mathcal{X} = [0,1]^d$, and let p_0 denote the uniform density (with respect to Lebesgue measure) on $\mathcal{X}$. Suppose $\{p_0\} \cup \{\phi_z\}_{z\in\mathcal{Z}}$ is an orthonormal basis in $\mathcal{L}^2_\mu$, and $\{a_z\}_{z\in\mathcal{Z}}$ and $\{b_z\}_{z\in\mathcal{Z}}$ are two real-valued nets. Let $L_D, L_G \geq 0$ and $p, q \geq 2$. For any $Z \subseteq \mathcal{Z}$, let*

$$A_Z := |Z|^{1/2} \sup_{z\in Z} a_z \quad \text{ and } \quad B_Z := |Z|^{1/2} \sup_{z\in Z} b_z.$$

Then, for $\mathcal{F}_D = \mathcal{H}_{p,a}(L_D)$ and $\mathcal{F}_G := \mathcal{H}_{q,b}(L_G)$, for any $Z \subseteq \mathcal{Z}$ satisfying

$$B_Z \geq 16 L_G \sqrt{\frac{n}{\log 2}} \quad \text{ and } \quad 2\frac{L_G}{B_Z} \sum_{z\in Z} \|\phi_z\|_{\mathcal{L}^\infty_\mu} \leq 1, \tag{5}$$

we have $M(\mathcal{F}_D, \mathcal{F}_G) \geq \dfrac{L_G L_D |Z|}{64 A_Z B_Z} = \dfrac{L_G L_D}{64\left(\sup_{z\in Z} a_z\right)\left(\sup_{z\in Z} b_z\right)}$.

As in most minimax lower bounds, our proof relies on constructing a finite set Ω_G of "worst-case" densities in $\mathcal{F}_G$, lower bounding the distance $d_{\mathcal{F}_D}$ over Ω_G, and then letting elements of Ω_G shrink towards the uniform distribution p_0 at a rate such that the average information (here, Kullback-Leibler) divergence between each $p \in \Omega_G$ and p_0 does not grow with n. The first condition in (5) ensures that the information divergence between each $p \in \Omega_G$ and p_0 is sufficiently small, and typically results in tuning of Z identical (in rate) to its optimal tuning in the upper bound (Theorem 1).

The second condition in (5) is needed to ensure that the "worst-case" densities we construct are everywhere non-negative. Hence, this condition is not needed for lower bounds in the Gaussian sequence model, as in Theorem 2.3 of [33]. However, failure of this condition (asymptotically) corresponds to the breakdown point of the asymptotic equivalence between the Gaussian sequence model and the density estimation model in the regime of very low smoothness (e.g., in the Sobolev setting, when $t < d/2$; see [11]), and so finer analysis is needed to establish lower bounds here.

6 Examples

In this section, we apply our bounds from Sections 4 and 5 to compute concrete minimax convergence rates for two examples choices of $\mathcal{F}_D$ and $\mathcal{F}_G$, namely Sobolev spaces and reproducing kernel Hilbert spaces. Due to space constraints, we consider only the Fourier basis here, but, in the Appendix, we also discuss an estimator in the Sobolev case using the Haar wavelet basis.

For the purpose of this section, suppose that $\mathcal{X} = [0, 2\pi]^d$, $\mathcal{Z} = \mathbb{Z}^d$, and, for each $z \in \mathcal{Z}$, ϕ_z is the z^{th} standard Fourier basis element given by $\phi_z(x) = e^{i\langle z,x\rangle}$ for all $x \in \mathcal{X}$. In this case, we will always choose the truncation set Z to be of the form $Z := \{z \in \mathcal{Z} : \|z\|_\infty \leq \zeta\}$, for some $\zeta > 0$, so that $|Z| \leq \zeta^d$. Moreover, for every $z \in Z$, $\|\phi_z\|_{\mathcal{L}^\infty_\mu} = 1$, and hence $C_Z \leq 1$.

Example 4 (Sobolev Spaces)**.** Suppose that, for some $s, t \geq 0$, $a_z = \|z\|_\infty^s$ and $b_z = \|z\|_\infty^t$. Then, setting $\zeta = n^{\frac{1}{2t+d}}$ in Theorems 1 and 3 gives that there exist constants $C > c > 0$ such that

$$cn^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}} \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}\right) \leq Cn^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}. \tag{6}$$

Combining the observation that the s-Hölder space $\mathcal{W}^{s,\infty} \subseteq \mathcal{W}^{s,2}$ with the lower bound (over $\mathcal{W}^{s,\infty}$) in Theorem 3.1 of [33], we have that (6) also holds when $\mathcal{W}^{s,2}$ is replaced with $\mathcal{W}^{s,p}$ for any $p \in [2, \infty]$ (e.g., in the case of the Wasserstein metric $d_{\mathcal{W}^{1,\infty}}$).

So far, we have assumed the smoothness t of the true distribution P is known, and used that to tune the parameter ζ of the estimator. However, in reality, t is not known. In the next result, we leverage the fact that the rate-optimal choice $\zeta = n^{\frac{1}{2t+d}}$ above does not rely on the loss parameters s, together with Theorem 1 to construct an *adaptively minimax estimator*, i.e., one that is minimax and fully-data dependent. There is a large literature on adaptive nonparametric density estimation under $\mathcal{L}^2_\mu$ loss; see [19] for accessible high-level discussion and [21] for a technical but comprehensive review.

Corollary 5 (Adaptive Upper Bound for Sobolev Spaces)**.** *There exists an adaptive choice* $\widehat{\zeta} : \mathcal{X}^n \to \mathbb{N}$ *of the hyperparameter* ζ *(independent of* s, t*), such that, for any* $s, t \geq 0$*, there exists a constant* $C > 0$ *(independent of* n*), such that*

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P}_{Z_{\widehat{\zeta}(X_{1:n})}} \right) \right] \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2} \right) \tag{7}$$

Due to space constraints, we present the actual construction of the adaptive $\widehat{\zeta}$ in the Appendix, but, in brief, it is a standard construction based on leave-one-out cross-validation under $\mathcal{L}^2_\mu$ loss which is known (e.g., see Sections 7.2.1 and 7.5.2 of [36]) to be adaptively minimax under $\mathcal{L}^2_\mu$ loss. Using the fact that our upper bound Theorem 1 uses a choice of ζ is independent of the loss parameter s, we show that the $d_{\mathcal{W}^{s,\infty}}$ risk of $\widehat{P}_\zeta$ can be factored into its $\mathcal{L}^2_\mu$ risk and a component (ζ^{-s}) that is independent of t. Since $\mathcal{L}^2_\mu$ risk can be rate-minimized in independently of t, it follows that the $d_{\mathcal{W}^{s,\infty}}$ risk can be rate-minimized independently of t. Adaptive minimaxity then follows from Theorem 3.

Example 6 (Reproducing Kernel Hilbert Space/MMD Loss)**.** Suppose $\mathcal{H}_k$ is a reproducing kernel Hilbert space (RKHS) with reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ [6, 8]. If k is translation invariant (i.e., there exists $\kappa \in \mathcal{L}^2_\mu$ such that, for all $x, y \in \mathcal{X}$, $k(x, y) = \kappa(x - y)$), then Bochner's theorem (see, e.g., Theorem 6.6 of [66]) implies that, up to constant factors,

$$\mathcal{H}_k(L) := \{ f \in \mathcal{H}_k : \|f\|_{\mathcal{H}_k} \leq L \} = \left\{ f \in \mathcal{H}_k : \sum_{z \in \mathcal{Z}} |\widetilde{\kappa}_z|^2 |\widetilde{f}_z|^2 < L^2 \right\}.$$

Thus, in the setting of Theorem 1, we have $\mathcal{H}_k = \mathcal{H}_{2,a}$, where $a_z = |\widetilde{\kappa}_z|$ satisfies $\sum_{z \in \mathcal{Z}} a_z^{-2} = \|\kappa\|^2_{\mathcal{L}^2_\mu} < \infty$. Corollary 2 then gives $M(\mathcal{H}_k(L_D), \mathcal{F}_G) \leq L_D \|\kappa\|_{\mathcal{L}^2_\mu} n^{-1/2}$ for *any class* $\mathcal{F}_G$. It is well-known known that MMD can always be *estimated* at the parametric rate $n^{-1/2}$ [23]; however, to the best of our knowledge, only recently has it been shown that any probability distribution can be estimated at the rate $n^{-1/2}$ under MMD loss[54], emphasizing the fact that MMD is a very weak metric. This has important implications for applications such as two-sample testing [46].

7 Consequences for Generative Adversarial Neural Networks (GANs)

This section discusses implications of our minimax bounds for GANs. Neural networks in this section are assumed to be fully-connected, with rectified linear unit (ReLU) activations. [33] used their upper bound result (2) to prove a similar theorem, but, since their upper bound was loose, the resulting theorem was also loose. The following results are immediate consequences of our improvement (Theorem 1) over the upper bound (2) of [33], and so we refer to that paper for the proof. Key ingredients are an oracle inequality proven in [33], an upper bound such as Theorem 1, and bounds of [67] on the size of a neural network needed to approximate functions in a Sobolev class.

In the following, $\mathcal{F}_D$ denotes the set of functions that can be encoded by the discriminator network and $\mathcal{F}_G$ denotes the set of distributions that can be encoded by the generator network. $P_n := \frac{1}{n} \sum_{i=1}^n 1_{\{X_i\}}$ denotes the empirical distribution of the observed data $X_{1:n} \overset{IID}{\sim} P$.

Theorem 7 (Improvement of Theorem 3.1 in Liang [33])**.** *Let* $s, t > 0$*, and fix a desired approximation accuracy* $\epsilon > 0$*. Then, there exists a GAN architecture, in which*

1. *the discriminator* $\mathcal{F}_D$ *has at most* $O(\log(1/\epsilon))$ *layers and* $O(\epsilon^{-d/s} \log(1/\epsilon))$ *parameters,*
2. *and the generator* $\mathcal{F}_G$ *has at most* $O(\log(1/\epsilon))$ *layers and* $O(\epsilon^{-d/t} \log(1/\epsilon))$ *parameters,*

$$\textit{such that, if } \widehat{P}_*(X_{1:n}) := \mathop{\mathrm{argmin}}_{\widehat{P} \in \mathcal{F}_G} d_{\mathcal{F}_D}\left(P_n, \widehat{P} \right), \textit{ is the optimized GAN estimate of } P,$$

$$\textit{then} \quad \sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P}_*(X_{1:n}) \right) \right] \leq C \left(\epsilon + n^{-\min\left\{ \frac{1}{2}, \frac{s+t}{2t+d} \right\}} \right).$$

The discriminator and generator in the above theorem can be implemented as described in [67]. The assumption that the GAN is perfectly optimized may be strong; see [41, 34] for discussion of this.

Though we do not present this result due to space constraints, we can similarly improve the upper bound of [33] (their Theorem 3.2) for very deep neural networks, further improving on the previous state-of-the-art bounds of [4] (which did not leverage smoothness assumptions on P).

8 Minimax Comparison of Explicit and Implicit Generative Models

In this section, we draw formal connections between our work on density estimation (explicit generative modeling) and the problem of implicit generative modeling under an appropriate measure of risk. In the sequel, we fix a class $\mathcal{F}_G$ of probability measures on a sample space $\mathcal{X}$ and a loss function $\ell : \mathcal{F}_G \times \mathcal{F}_G \to [0, \infty]$ measuring the distance of an estimate $\widehat{P}$ from the true distribution P. ℓ need not be an adversarial loss $d_{\mathcal{F}_D}$, but our discussion does apply to all ℓ of this form.

8.1 A Minimax Framework for Implicit Generative Models

Thus far, we have analyzed the *minimax risk of density estimation*, namely

$$M_D(\mathcal{F}_G, \ell, n) = \inf_{\widehat{P}} \sup_{P \in \mathcal{F}_G} R_D(P, \widehat{P}), \text{ where } R_D(P, \widehat{P}) = \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\ell(P, \widehat{P}(X_{1:n})) \right] \tag{8}$$

denotes the *density estimation risk of* $\widehat{P}$ *at* P and the infimum is taken over all estimators (i.e., (potentially randomized) functions $\widehat{P} : \mathcal{X}^n \to \mathcal{F}_G$). Whereas density estimation is a classical statistical problem to which we have already contributed novel results, our motivations for studying this problem arose from a desire to better understand recent work on implicit generative modeling.

Implicit generative models, such as GANs [5, 22] and VAEs [26, 48], address the problem of *sampling*, in which we seek to construct a *generator* that produces novel samples from the distribution P [38]. In our context, a generator is a function $\widehat{X} : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ that takes in n IID samples $X_{1:n} \sim P$ and a source of randomness (a.k.a., *latent variable*) $Z \sim Q_Z$ with known distribution Q_Z (independent of $X_{1:n}$) on a space $\mathcal{Z}$, and returns a novel sample $\widehat{X}(X_{1:n}, Z) \in \mathcal{X}$.

The evaluating the performance of implicit generative models, both in theory and in practice, is difficult, with solutions continuing to be proposed [57], some of which have proven controversial. Some of this controversy stems from the fact that many of the most straightforward evaluation objectives are optimized by a trivial generator that 'memorizes' the training data (e.g., $\widehat{X}(X_{1:n}, Z) = X_Z$, where Z is uniformly distributed on $[n]$). One objective that can avoid this problem is as follows. For simplicity, fix the distribution Q_Z of the latent random variable $Z \sim Q_Z$ (e.g., $Q_Z = \mathcal{N}(0, I)$). For a fixed training set $X_{1:n} \overset{IID}{\sim} P$ and latent distribution $Z \sim Q_Z$, we define the *implicit distribution of a generator* $\widehat{X}$ as the conditional distribution $P_{\widehat{X}(X_{1:n},Z)|X_{1:n}}$ over $\mathcal{X}$ of the random variable $\widehat{X}(X_{1:n}, Z)$ given the training data. Then, for any $P \in \mathcal{F}_G$, we define the *implicit risk of* $\widehat{X}$ *at* P by

$$R_I(P, \widehat{X}) := \mathop{\mathbb{E}}_{X_{1:n} \sim P} \left[\ell(P, P_{\widehat{X}(X_{1:n},Z)|X_{1:n}}) \right].$$

We can then study the *minimax risk of sampling*, $M_I(\mathcal{F}_G, \ell, n) := \inf_{\widehat{X}} \sup_{P \in \mathcal{F}_G} R_I(P, \widehat{X})$. A few remarks about $M_I(\mathcal{F}, \ell, n)$: First, we implicitly assumed $\ell(P, P_{\widehat{X}(X_{1:n},Z)|X_{1:n}})$ is well-defined, which is not obvious unless $P_{\widehat{X}(X_{1:n},Z)} \in \mathcal{F}_G$. We discuss this assumption further below. Second, since the risk $R_I(P, \widehat{X})$ depends on the unknown true distribution P, we cannot calculate it in practice. Third, for the same reason (because $R_P(P, \widehat{X})$ depends directly on P rather than particular data $X_{1:n}$), it detect lack-of-diversity issues such as mode collapse. As we discuss in the Appendix, these latter two points are distinctions from the recent work of [7] on generalization in GANs.

8.2 Comparison of Explicit and Implicit Generative Models

Algorithmically, sampling is a very distinct problem from density estimation; for example, many computationally efficient Monte Carlo samplers rely on the fact that a function *proportional* to the density of interest can be computed much more quickly than the exact (normalized) density function [13]. In this section, we show that, given unlimited computational resources, the problems of density estimation and sampling are equivalent in a minimax statistical sense. Since exactly minimax estimators ($\operatorname{argmin}_{\widehat{P}} \sup_{P \in \mathcal{F}_G} R_D(P, \widehat{P})$) often need not exist, the following weaker notion is useful for stating our results:

Definition 8 (Nearly Minimax Sequence)**.** A sequence $\{\widehat{P}_k\}_{k \in \mathbb{N}}$ of density estimators (resp., $\{\widehat{X}_k\}_{k \in \mathbb{N}}$ of generators) is called *nearly minimax over* $\mathcal{F}_G$ if $\lim_{k \to \infty} \sup_{P \in \mathcal{F}_G} R_{P,D}(\widehat{P}_k) = M_D(\mathcal{F}_G, \ell, n)$ (resp., $\lim_{k \to \infty} \sup_{P \in \mathcal{F}_G} R_{P,I}(\widehat{X}_k) = M_I(\mathcal{F}_G, \ell, n)$).

The following theorem identifies sufficient conditions under which, in the statistical minimax framework described above, density estimation is no harder than sampling. The idea behind the proof is as follows: If we have a good sampler $\widehat{X}$ (i.e., with $R_I(\widehat{X})$ small), then we can draw m 'fake' samples from $\widehat{X}$. We can use these 'fake' samples to construct a density estimate $\widehat{P}$ of the implicit distribution of $\widehat{X}$ such that, under the technical assumptions below, $R_D(\widehat{P}) - R_I(\widehat{X}) \to 0$ as $m \to \infty$.

Theorem 9 (Conditions under which Density Estimation is Statistically no harder than Sampling)**.** *Let $\mathcal{F}_G$ be a family of probability distributions on a sample space $\mathcal{X}$. Suppose*

- ***(A1)*** $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ *is non-negative, and there exists $C_\triangle > 0$ such that, for all $P_1, P_2, P_3 \in \mathcal{F}_G$, $\ell(P_1, P_3) \leq C_\triangle \left(\ell(P_1, P_2) + \ell(P_2, P_3) \right)$.*
- ***(A2)*** $M_D(\mathcal{F}_G, \ell, m) \to 0$ *as $m \to \infty$.*
- ***(A3)*** *For all $m \in \mathbb{N}$, we can draw m IID samples $Z_1, ..., Z_m \overset{IID}{\sim} Q_Z$ of the latent variable Z.*
- ***(A4)*** *there exists a nearly minimax sequence of samplers $\widehat{X}_k : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ such that, for each $k \in \mathbb{N}$, almost surely over $X_{1:n}$, $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}} \in \mathcal{F}_G$.*

Then, $M_D(\mathcal{F}_G, \ell, n) \leq C_\triangle M_I(\mathcal{F}_G, \ell, n)$.

Assumption (A1) is a generalization of the triangle inequality (and reduces to the triangle inequality when $C_\triangle = 1$). This weaker assumption applies, for example, when ℓ is the Jensen-Shannon divergence (with $C_\triangle = 2$) used in the original GAN formulation of [22], even though this does not satisfy the triangle inequality [20]). Assumption (A2) is equivalent to the existence of a uniformly ℓ-risk-consistent estimator over $\mathcal{F}_G$, a standard property of most distribution classes $\mathcal{F}_G$ over which density estimation is studied (e.g., our Theorem 1). Assumption (A3) is a natural design criterion of implicit generative models; usually, Q_Z is a simple parametric distribution such as a standard normal.

Finally, Assumption (A4) is the most mysterious, because, currently, little is known about the minimax theory of samplers when $\mathcal{F}_G$ is a large space. On one hand, since $M_I(\mathcal{F}_G, \ell, n)$ is an infimum over $\widehat{X}$, Theorem 9 continues to hold if we restrict the class of samplers (e.g., to those satisfying Assumption (A4) or those we can compute). On the other hand, even without restricting $\widehat{X}$, this assumption may not be too restrictive, because nearly minimax samplers are necessarily close to $P \in \mathcal{F}_G$. For example, if $\mathcal{F}_G$ contains only smooth distributions but $\widehat{X}$ is the trivial empirical sampler described above, then $\ell(P, P_{\widehat{X}})$ should be large and $\widehat{X}$ is unlikely to be minimax optimal.

Finally, in practice, we often do not know estimators that are nearly minimax for finite samples, but may have estimators that are rate-optimal (e.g., as given by Theorem 1), i.e., that satisfy

$$C := \limsup_{n \to \infty} \frac{\sup_{P \in \mathcal{F}_G} R_I(P, \widehat{X})}{M_I(\mathcal{F}_G, \ell, n)} < \infty.$$

Under this weaker assumption, it is straightforward to modify our proof to conclude that

$$\limsup_{n \to \infty} \frac{M_D(\mathcal{F}_G, \ell, n)}{M_I(\mathcal{F}_G, \ell, n)} \leq C_\triangle C.$$

The converse result ($M_D(\mathcal{F}_G, \ell, n) \geq M_I(\mathcal{F}_G, \ell, n)$) is simple to prove in many cases, and is related to the well-studied problem of Monte Carlo sampling [49]; we discuss this briefly in the Appendix.

9 Conclusions

Given the recent popularity of implicit generative models in many applications, it is important to theoretically understand why these models appear to outperform classical methods for similar problems. This paper provided new minimax bounds for density estimation under adversarial losses, both with and without adaptivity to smoothness, and gave several applications, including both traditional statistical settings and perfectly optimized GANs. We also gave simple conditions under which minimax bounds for density estimation imply bounds for the problem of implicit generative modeling, suggesting that sampling is typically not *statistically* easier than density estimation. Thus, for example, the strong curse of dimensionality that is known to afflict to nonparametric density estimation Wasserman [64] should also limit the performance of implicit generative models such as GANs. The Appendix describes several specific avenues for further investigation, including whether the curse of dimensionality can be avoided when data lie on a low-dimensional manifold.

Acknowledgments

This work was partly supported by NSF grant IIS1563887, the Darpa D3M program, AFRL FA8750-17-2-0212, and the NSF Graduate Research Fellowship DGE-1252522.

References

[1] Ehsan Abbasnejad, Javen Shi, and Anton van den Hengel. Deep lipschitz networks and dudley GANs, 2018. URL `https://openreview.net/forum?id=rkw-jlb0W`.

[2] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

[3] Niall H Anderson, Peter Hall, and D Michael Titterington. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *Journal of Multivariate Analysis*, 50(1):41–54, 1994.

[4] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pages 214–223, 2017.

[6] Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[7] Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *International Conference on Machine Learning*, pages 224–232, 2017.

[8] Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

[9] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer, 2015.

[10] Leon Bottou, Martin Arjovsky, David Lopez-Paz, and Maxime Oquab. Geometrical insights for implicit generative modeling. *arXiv preprint arXiv:1712.07822*, 2017.

[11] Lawrence D Brown, Cun-Hui Zhang, et al. Asymptotic nonequivalence of nonparametric experiments when the smoothness index is $1/2$. *The Annals of Statistics*, 26(1):279–287, 1998.

[12] Guillermo Canas and Lorenzo Rosasco. Learning probability measures with respect to optimal transport metrics. In *Advances in Neural Information Processing Systems*, pages 2492–2500, 2012.

[13] Siddhartha Chib and Edward Greenberg. Understanding the Metropolis-Hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[14] Peter J Diggle and Richard J Gratton. Monte Carlo methods of inference for implicit statistical models. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 193–227, 1984.

[15] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of Statistics*, pages 508–539, 1996.

[16] Simon S Du, Jayanth Koushik, Aarti Singh, and Barnabás Póczos. Hypothesis transfer learning via transformation functions. In *Advances in Neural Information Processing Systems*, pages 574–584, 2017.

[17] RM Dudley. Speeds of metric probability convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 22(4):323–332, 1972.

[18] Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. *arXiv preprint arXiv:1505.03906*, 2015.

[19] Sam Efromovich. Orthogonal series density estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):467–476, 2010.

[20] Dominik Maria Endres and Johannes E Schindelin. A new metric for probability distributions. *IEEE Transactions on Information theory*, 49(7):1858–1860, 2003.

[21] Alexander Goldenshluger and Oleg Lepski. On adaptive minimax density estimation on R^d. *Probability Theory and Related Fields*, 159(3-4):479–543, 2014.

[22] 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.

[23] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[24] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, pages 5769–5779, 2017.

[25] Leonid Vasilevich Kantorovich and Gennady S Rubinstein. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[27] Andrey Kolmogorov. Sulla determinazione empirica di una lgge di distribuzione. *Inst. Ital. Attuari, Giorn.*, 4:83–91, 1933.

[28] Samory Kpotufe and Vikas Garg. Adaptivity to local smoothness and dimension in kernel regression. In *Advances in neural information processing systems*, pages 3075–3083, 2013.

[29] Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabas Poczos, and Larry A Wasserman. On estimating L_2^2 divergence. In *AISTATS*, 2015.

[30] Jing Lei. Convergence and concentration of empirical measures under wasserstein distance in unbounded functional spaces. *arXiv preprint arXiv:1804.10556*, 2018.

[31] Giovanni Leoni. *A first course in Sobolev spaces*. American Mathematical Soc., 2017.

[32] 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 2200–2210, 2017.

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

[34] Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. *arXiv preprint arXiv:1802.06132*, 2018.

[35] 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.

[36] Pascal Massart. *Concentration inequalities and model selection*, volume 6. Springer, 2007.

[37] Shahar Mendelson. Learnability in hilbert spaces with reproducing kernels. *journal of complexity*, 18(1): 152–170, 2002.

[38] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

[39] Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev gan. *arXiv preprint arXiv:1711.04894*, 2017.

[40] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[41] Vaishnavh Nagarajan and J Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems*, pages 5591–5600, 2017.

[42] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-GAN: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.

[43] Mark Owen. *Practical signal processing*. Cambridge university press, 2007.

[44] David Pollard. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*, pages i–86. JSTOR, 1990.

[45] Novi Quadrianto, James Petterson, and Alex J Smola. Distribution matching for transduction. In *Advances in Neural Information Processing Systems*, 2009.

[46] Aaditya Ramdas, Sashank Jakkam Reddi, Barnabás Póczos, Aarti Singh, and Larry A Wasserman. On the decreasing power of kernel and distance based nonparametric hypothesis tests in high dimensions. In *AAAI*, pages 3571–3577, 2015.

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

[48] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

[49] Christian P Robert. *Monte Carlo methods*. Wiley Online Library, 2004.

[50] Shashank Singh and Barnabás Póczos. Minimax distribution estimation in Wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

[51] Shashank Singh, Bharath K Sriperumbudur, and Barnabás Póczos. Minimax estimation of quadratic Fourier functionals. *arXiv preprint arXiv:1803.11451*, 2018.

[52] Nickolay Smirnov. Table for estimating the goodness of fit of empirical distributions. *The annals of mathematical statistics*, 19(2):279–281, 1948.

[53] Bharath Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Aapo Hyvärinen, and Revant Kumar. Density estimation in infinite dimensional exponential families. *The Journal of Machine Learning Research*, 18(1): 1830–1888, 2017.

[54] Bharath Sriperumbudur et al. On the optimal estimation of probability measures in weak and strong topologies. *Bernoulli*, 22(3):1839–1893, 2016.

[55] 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.

[56] Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, Gert RG Lanckriet, et al. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

[57] D. Sutherland, H-Y Tung, H. Strathmann, S. De, A. Ramdas, A. Smola, and A. Gretton. Generative models and model criticism via optimized maximum mean discrepancy. In *ICLR*, 2017. URL `https://arxiv.org/abs/1611.04488`.

[58] Dougal J Sutherland. *Scalable, Flexible and Active Learning on Distributions*. PhD thesis, Carnegie Mellon University, 2016.

[59] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The annals of statistics*, 35(6):2769–2794, 2007.

[60] Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

[61] Alexandre B Tsybakov. *Introduction to nonparametric estimation*. Springer Series in Statistics. Springer, New York, 2009.

[62] Vladimir N Vapnik and A Ya Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[63] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[64] Larry Wasserman. *All of Nonparametric Statistics*. Springer Science & Business Media, 2006.

[65] 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.

[66] Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

[67] Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

10 Further Related Work

As noted in the main paper, our problem setting is quite general, and thus overlaps with several previous settings that have been studied. First, we note the analysis of [35], which also studied convergence of distribution estimation under adversarial losses. Considering a somewhat broader class of non-metric losses (including, e.g., Jensen-Shannon divergence), which they call *adversarial divergences*, [35] provided consistency results (in distribution) for a number of GAN formulations, assuming convergence of the min-max GAN optimization problem to a generator-optimal equilibrium. However, they did not study rates of convergence.

Our results can also be viewed as a refinement of several results from empirical process and learning theory, especially the wealth of literature on the case where $\mathcal{F}_D$ is a Glivenko-Cantelli (GC, a.k.a., Vapnik-Chervonenkis (VC)) class [44]. Corollary 2 can be interpreted as showing that spaces $\mathcal{F}_D$ that are sufficiently small in terms of orthonormal basis expansions are $n^{-1/2}$-uniformly GC/VC classes [2, 62]. In particular, this gives a simple functional-analytic proof of this property for the general case when $\mathcal{F}_D$ is a ball in a translation-invariant RKHS. On the other hand, some related results, cast in terms of fat-shattering dimensions [37, 18], appear to lead to slower rates for RKHSs.

Glivenko-Cantelli classes are defined without regards to the class $\mathcal{F}_G$ of possible distributions. However, the more interesting consequences of our results are for the case that $\mathcal{F}_G$ is restricted, as in Theorem 1. In Example 4 this allowed us to characterize the interaction between smoothness constraints on the discriminator class $\mathcal{F}_D$ and the generator class $\mathcal{F}_G$, showing in particular, that, when $\mathcal{F}_D$ is large, restricting $\mathcal{F}_G$ improves convergence rates. Aside for the results of [33] and many results for the specific case $\mathcal{F}_D = \mathcal{L}^2_\lambda$, we do not know of any results that show this.

Several prior works have studied the closely related problem of estimating certain adversarial metrics, including $\mathcal{L}^2$ distance [29], MMD [23], Sobolev distances [51], and others [56]. In some cases, these metrics can themselves be estimated far more efficiently than the underlying distribution under that loss, and these estimators have various applications including two-sample/homogeneity and independence testing [3, 23, 46], and distributional [58], transfer [16], and transductive [45] learning.

There has also been some work studying the min-max optimization problem in terms of which GANs are typically cast [41, 34]. However, in this work, as in [35, 33], we implicitly assume the optimization procedure has converged to a generator-optimal equilibrium. Another work that studies adversarial losses is [10], which focuses on a comparison of Wasserstein distance and MMD in the context of implicit generative modeling.

10.1 Other statistical analyses of GANs

Our results are closely related to some previous work studying the *generalization error* of GANs under MMD [18] or Jensen-Shannon divergence, Wasserstein, or other adversarial losses [7].

Assume, for simplicity, that ℓ satisfies a weak triangle inequality (Assumption (A1) above), and let P denote the true distribution from which the data are drawn IID. Then, we can bound the true loss $\ell(P, \widehat{P})$ of an estimator $\widehat{P}$ in terms of the approximation error $\ell(P, P_*)$ (corresponding to bias) and generalization error $\ell(P_*, \widehat{P})$ (i.e., corresponding to variance):

$$\ell(P, \widehat{P}) \leq C_\triangle \left(\ell(P, P_*) + \ell(P_*, \widehat{P}) \right),$$

where $P_* := \operatorname{argmin}_{Q \in \widehat{\mathcal{F}}} \ell(P, Q)$ denotes the optimal approximation of P in some restricted class $\widehat{\mathcal{F}} \subseteq \mathcal{F}_G$ of estimators in which $\widehat{P}$ lies.

Bounding the approximation error $\ell(P, P_*)$ typically requires restricting the space $\mathcal{F}_G$ in which P lies. Theorem 1 of [18] and Theorem 3.1 of [7] focus on bounding the generalization error $\ell(P_*, \widehat{P})$, and thus avoid making such assumptions on P. However, our Theorem 1 shows that, when $\mathcal{F}_D$ is sufficiently small (e.g., an RKHS, as in [18]), $\ell = d_{\mathcal{F}_D}$ is so weak that $\ell(P, P_*)$ can be bounded even when $\mathcal{F}_G$ includes *all* probability measures. In particular, while [18] gave only high-probability bounds of order $n^{-1/2}$ on the *generalization error* $\ell(P_*, \widehat{P})$ in terms of the fat-shattering dimension of the RKHS, we show that, for any RKHS with a translation-invariant kernel, the *total* risk $\mathbb{E}[\ell(P, \widehat{P})]$ can be bounded at the parametric rate of $n^{-1/2}$.

[7] also showed that, if $\widehat{\mathcal{F}}$ is too large (specifically, if $\widehat{\mathcal{F}}$ contains the empirical distribution), then the generalization error $\ell(P_*, \widehat{P})$ (or, specifically, an empirical estimate thereof) need not vanish as the sample size increases, or, in the case of Wasserstein distance, if the dimension d grows faster than logarithmically with the sample size n. Our Theorem 1 showed that, if $\widehat{F}$ contains only (e.g., orthogonal series) estimates of a fixed smoothness (e.g., orthogonal series estimates with a fixed ζ), then the generalization error decays at the rate $\asymp \zeta^{d/2} n^{-1/2}$ (the first term on the right-hand side of 4), so that $d \in o(\log n)$ is still necessary[1]. Our minimax lower bound 3 suggests that, without making significantly stronger assumptions, we cannot hope to avoid this curse of dimensionality, at least without sacrificing approximation error (bias).

11 Proof of Upper Bound

In this section, we prove our main upper bound, Theorem 1. We begin with a simple lemma showing that, under mild assumptions, we can write an adversarial loss in terms of an $\mathcal{L}^2_\lambda$ basis expansion.

Lemma 10 (Basis Expansion of Adversarial Loss)**.** *Consider a class $\mathcal{F}_D$ of discriminator functions, two probability distributions P and Q, and an orthonormal basis $\{\phi_z\}_{z \in \mathcal{Z}}$ of $\mathcal{L}^2_\lambda(\mathcal{X})$. Moreover, suppose that either of the following conditions holds:*

1. *$P, Q \ll \lambda$ have densities $p, q \in \mathcal{L}^2_\lambda$.*
2. *For every $f \in \mathcal{F}_D$, the expansion of f in the basis $\mathcal{B}$ converges uniformly (over $\mathcal{X}$) to f. That is,*

$$\lim_{Z \uparrow \mathcal{Z}} \sup_{x \in \mathcal{X}} \left| f(x) - \sum_{z \in Z} \widetilde{f}_z(x) \phi_z(x) \right| \to 0.$$

Then, we can expand the adversarial loss $d_{\mathcal{F}_D}$ over $\mathcal{P}$ as

$$d_{\mathcal{F}_D}(P, Q) = \sup_{f \in \mathcal{F}_D} \sum_{z \in \mathcal{Z}} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z \right).$$

Condition 1 above is quite straightforward, and would be taken for granted in most classical nonparametric analysis. When $\mathcal{B}$ is the Fourier basis, the assumption that $p, q \in \mathcal{L}^r_\mu$ for $r = 2$ can be weakened to any $r > 1$ using Hölder's inequality together with the facts that $f \in \mathcal{L}^{r'}$ and that Fourier series converge in $\mathcal{L}^{r'}$ (where $r' = \frac{r}{r-1}$ denote the Hölder conjugate of r).

Since we are also interested in probability distributions that lack density functions, we provide the fairly mild Condition 2 as an alternative. As an example of this condition in the Fourier case, suppose $\mathcal{F}_D$ is uniformly equi-continuous, say, with modulus of continuity $\omega : [0, \infty) \to [0, \infty)$ satisfying $\omega(\epsilon) \in o\left(\frac{1}{\log 1/\epsilon} \right)$. Then, there exists a constant $C > 0$ such that

$$\sup_{x \in \mathcal{X}} \left| f(x) - \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \right| \leq K(\log \zeta) \omega \left(\frac{2\pi}{\zeta} \right). \tag{9}$$

As a concrete example of this, it suffices if every f is α_f-Hölder continuous for some $\alpha_f > 0$. Finally, we note that, if P and Q are allowed to be arbitrary, then the above uniform convergence assumption is essentially also necessary.

Proof. First note that it suffices to show that, for all $f \in \mathcal{F}_D$,

$$\mathop{\mathbb{E}}_{X \sim P} [f(X)] - \mathop{\mathbb{E}}_{X \sim Q} [f(X)] = \sum_{z \in \mathcal{Z}} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z \right).$$

We show this separately for the two sets of assumptions considered:

[1] The case of Jensen-Shannon divergence requires an additional uniform lower boundedness assumption and is discussed in the Appendix.

1. **Case 1: P, Q have a densities $p, q \in \mathcal{L}^2_\mu$.** Then $\widetilde{P}_z = \langle p, \phi_z \rangle_{\mathcal{L}^2}$, and so, by the Plancherel Theorem, since $f \in \mathcal{L}^\infty_\mu(\mathcal{X}) \subseteq \mathcal{L}^2_\mu(\mathcal{X})$,
$$\mathop{\mathbb{E}}_{X \sim P}\left[f(X)\right] = \int_{\mathcal{X}} f p \, d\mu = \langle f, p \rangle_{\mathcal{L}^2_\mu} = \sum_{z \in \mathcal{Z}} \widetilde{f}_z \widetilde{P}_z < \infty.$$
Similarly, $\mathbb{E}_{X \sim Q}\left[f(X)\right] = \sum_{z \in \mathcal{Z}} \widetilde{f}_z \widetilde{Q}_z < \infty$. Since these quantities are finite, we can split the sum of differences
$$\sum_{z \in \mathcal{Z}} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z\right) = \sum_{z \in \mathcal{Z}} \widetilde{f}_z \widetilde{P}_z - \sum_{z \in \mathcal{Z}} \widetilde{f}_z \widetilde{Q}_z = \mathop{\mathbb{E}}_{X \sim P}\left[f(X)\right] - \mathop{\mathbb{E}}_{X \sim Q}\left[f(X)\right].$$

2. **Case 2: For every $f \in \mathcal{F}_D$, the basis expansion of f in $\mathcal{B}$ converges uniformly (over $\mathcal{X}$) to f.** Then,
$$\begin{aligned}
&\left| \mathop{\mathbb{E}}_{X \sim P}\left[f(X)\right] - \mathop{\mathbb{E}}_{X \sim Q}\left[f(X)\right] - \sum_{|z| \leq \zeta} \widetilde{f}_z \left(\widetilde{P}_z - \widetilde{Q}_z\right) \right| \\
&= \left| \int_{\mathcal{X}} f(x) \, dP - \int_{\mathcal{X}} f(x) \, dQ - \sum_{|z| \leq \zeta} \widetilde{f}_z \left(\int_{\mathcal{X}} \phi_z(x) \, dP - \int_{\mathcal{X}} \phi_z(x) \, dQ \right) \right| \\
&= \left| \int_{\mathcal{X}} f(x) \, dP - \int_{\mathcal{X}} f(x) \, dQ - \int_{\mathcal{X}} \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \, dP - \int_{\mathcal{X}} \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \, dQ \right| \\
&= \left| \int_{\mathcal{X}} f(x) - \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \, dP + \int_{\mathcal{X}} f(x) - \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \, dQ \right| \\
&\leq \int_{\mathcal{X}} \left| f(x) - \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \right| \, dP + \int_{\mathcal{X}} \left| f(x) - \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \right| \, dQ \\
&\leq 2 \sup_{x \in \mathcal{X}} \left| f(x) - \sum_{|z| \leq \zeta} \widetilde{f}_z \phi_z(x) \right| \to 0 \quad \text{as } \zeta \to \infty.
\end{aligned}$$

□

Theorem 1. Suppose that $\mu(\mathcal{X}) < \infty$ and there exist constants $L_D, L_G > 0$, real-valued nets $\{a_z\}_{z \in \mathcal{Z}}$, $\{b_z\}_{z \in \mathcal{Z}}$ such that $\mathcal{F}_D = \mathcal{H}_{p,a}(\mathcal{X}, L_D)$ and $\mathcal{F}_G = \mathcal{H}_{q,b}(\mathcal{X}, L_G)$, where $p, q \geq 1$. Let $p' = \frac{p}{p-1}$ denote the Hölder conjugate of p. Then, for any $P \in \mathcal{F}_G$,
$$\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right)\right] \leq L_D \frac{c_{p'}}{\sqrt{n}} \left\| \left\{ \frac{\|\phi_z\|_{\mathcal{L}^\infty_P}}{a_z} \right\}_{z \in Z} \right\|_{p'} + L_D L_G \left\| \left\{ \frac{1}{a_z b_z} \right\}_{z \in \mathcal{Z} \setminus Z} \right\|_{1/(1-1/p-1/q)}.$$

Proof. By Lemma 10,
$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n}}\left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right)\right] &= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f \in \mathcal{F}_D} \sum_{z \in \mathcal{Z}} |\widetilde{f}_z \left(\widetilde{P}_z - \widehat{P}_z\right)|\right] \\
&= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f \in \mathcal{F}_D} \sum_{z \in Z} |\widetilde{f}_z \left(\widetilde{P}_z - \widehat{P}_z\right)| + \sum_{z \in \mathcal{Z} \setminus Z} |\widetilde{f}_z \left(\widetilde{P}_z - \widehat{P}_z\right)|\right] \\
&= \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f \in \mathcal{F}_D} \sum_{z \in Z} |\widetilde{f}_z \left(\widetilde{P}_z - \widehat{P}_z\right)| + \sum_{z \in \mathcal{Z} \setminus Z} |\widetilde{f}_z \widetilde{P}_z|\right] \\
&\leq \mathop{\mathbb{E}}_{X_{1:n}}\left[\sup_{f \in \mathcal{F}_D} \sum_{z \in Z} |\widetilde{f}_z \left(\widetilde{P}_z - \widehat{P}_z\right)|\right] + \sup_{f \in \mathcal{F}_D} \sum_{z \in \mathcal{Z} \setminus Z} |\widetilde{f}_z \widetilde{P}_z|.
\end{aligned}$$

Note that we have decomposed the risk into two terms, the first comprising estimation error (variance) and the second comprising approximation error (bias). Indeed, in the case that $\mathcal{F}_D = \mathcal{L}^2(\mathcal{X})$, the above becomes precisely the usual bias-variance decomposition of mean squared error.

To bound the first term, applying the Holder's inequality, the fact that $f \in \mathcal{F}_D$, and Jensen's inequality (in that order), we have

$$\begin{aligned}
\underset{X_{1:n}}{\mathbb{E}} \left[\sup_{f \in \mathcal{F}_D} \sum_{z \in Z} |\widetilde{f}_z \left(\widetilde{P}_z - \widehat{P}_z \right) | \right] &= \underset{X_{1:n}}{\mathbb{E}} \left[\sup_{f \in \mathcal{F}_D} \sum_{z \in Z} a_z |\widetilde{f}_z| \frac{|\widetilde{P}_z - \widehat{P}_z|}{a_z} \right] \\
&\leq \underset{X_{1:n}}{\mathbb{E}} \left[\sup_{f \in \mathcal{F}_D} \left(\sum_{z \in Z} a_z^p |\widetilde{f}_z|^p \right)^{\frac{1}{p}} \left(\sum_{z \in Z} \left(\frac{|\widetilde{P}_z - \widehat{P}_z|}{a_z} \right)^{p'} \right)^{\frac{1}{p'}} \right] \\
&\leq L_D \underset{X_{1:n}}{\mathbb{E}} \left[\left(\sum_{z \in Z} \left(\frac{|\widetilde{P}_z - \widehat{P}_z|}{a_z} \right)^{p'} \right)^{\frac{1}{p'}} \right] \\
&\leq L_D \left(\sum_{z \in Z} \frac{\mathbb{E}_{X_{1:n}} \left[\left| \widetilde{P}_z - \widehat{P}_z \right|^{p'} \right]}{a_z^{p'}} \right)^{\frac{1}{p'}} \leq \frac{L_D}{\sqrt{n}} \left(\sum_{z \in Z} \frac{\|\phi_z\|_{\mathcal{L}_P^\infty}^{p'}}{a_z^{p'}} \right)^{\frac{1}{p'}},
\end{aligned}$$

where $p' = \frac{p}{p-1}$ is the Hölder conjugate of p. In the last inequality we have used Rosenthal's inequality i.e.,

$$\underset{X_{1:n}}{\mathbb{E}} \left[\left| \widetilde{P}_z - \widehat{P}_z \right|^{p'} \right] \leq c_{p'} \frac{\|\phi_z\|_{\mathcal{L}_P^\infty}^{p'}}{n^{p'/2}}.$$

For the second term, by Holder's inequality,

$$\begin{aligned}
\sup_{f \in \mathcal{F}_D} \sum_{z \in \mathcal{Z} \setminus Z} |\widetilde{f}_z \widetilde{P}_z| &\leq \sup_{f \in \mathcal{F}_D} \left(\sum_{z \in \mathcal{Z} \setminus Z} \left(a_z |\widetilde{f}_z| \right)^p \right)^{1/p} \left(\sum_{z \in \mathcal{Z} \setminus Z} \left(\frac{|\widetilde{P}_z|}{a_z} \right)^{p'} \right)^{1/p'} \\
&\leq L_D \left\| \left\{ \frac{b_z \widetilde{P}_z}{b_z a_z} \right\}_{z \in \mathcal{Z} \setminus Z} \right\|_{p'} \\
&\leq L_D \left\| \{ b_z \widetilde{P}_z \}_{z \in \mathcal{Z} \setminus Z} \right\|_q \left\| \left\{ \frac{1}{b_z a_z} \right\}_{z \in \mathcal{Z} \setminus Z} \right\|_{\frac{p'q}{q-p'}} \qquad \text{by Holder} \\
&= L_D L_G \left\| \left\{ \frac{1}{a_z b_z} \right\}_{z \in \mathcal{Z} \setminus Z} \right\|_{\frac{1}{1-(1/p+1/q)}}
\end{aligned}$$

□

12 Proof of Lower Bound

Theorem 3 (Minimax Lower Bound). Let $\lambda(\mathcal{X}) = 1$, and let p_0 denote the uniform density (with respect to Lebesgue measure) on $\mathcal{X}$. Suppose $\{p_0\} \cup \{\phi_z\}_{z \in \mathcal{Z}}$ is an orthonormal basis in $\mathcal{L}_\lambda^2$, suppose $\{a_z\}_{z \in \mathcal{Z}}$ and $\{b_z\}_{z \in \mathcal{Z}}$ are two real-valued nets, and let $L_D, L_G \geq 0$. For any $Z \subseteq \mathcal{Z}$, define

$$A_Z := |Z|^{1/p} \sup_{z \in Z} a_z \quad \text{and} \quad B_Z := |Z|^{1/q} \sup_{z \in Z} b_z.$$

Then, for $\mathcal{H}_D = \mathcal{H}_{p,a}(L_D)$ and $\mathcal{H}_G := \mathcal{H}_{b,q}(L_G)$, for any $Z \subseteq \mathcal{Z}$ satisfying

$$B_Z \geq 16 L_G \sqrt{\frac{n}{\log 2}} \tag{10}$$

and

$$2\frac{L_G}{B_Z}\sum_{z\in Z}\|\phi_z\|_{\mathcal{L}^\infty_\mu}\leq 1, \tag{11}$$

we have

$$M(\mathcal{H}_D,\mathcal{H}_G)\geq\frac{L_G L_D|Z|}{64A_Z B_Z}=\frac{L_G L_D|Z|^{1-1/p-1/q}}{64\left(\sup_{z\in Z}a_z\right)\left(\sup_{z\in Z}b_z\right)}.$$

Proof. We will follow a standard procedure for proving minimax lower bounds based on the Varshamov-Gilbert bound and Fano's lemma (as outlined, e.g., Chapter 2 of Tsybakov [61]). The proof is quite similar to a standard proof for the case of $\mathcal{L}^2_\lambda$-loss, based on constructing a finite "worst-case" subset $\Omega_G\subseteq\mathcal{F}_G$ of densities over which estimation is difficult. The main difference is that we also construct a similar finite "worst-case" subset $\Omega_D\subseteq\mathcal{F}_D$ of the discriminator class $\mathcal{F}_D$, which we use to lower bound $d_{\mathcal{F}_D}\geq d_{\Omega_D}$ over Ω_G. Specifically, we will use the following result:

Lemma 11 (Simplified Form of Theorem 2.5 of Tsybakov [61])**.** *Fix a family $\mathcal{P}$ of distributions over a sample space $\mathcal{X}$ and fix a pseudo-metric $\rho:\mathcal{P}\times\mathcal{P}\to[0,\infty]$ over $\mathcal{P}$. Suppose there exists a set $T\subseteq\mathcal{P}$ such that*

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

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

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

where the inf *is taken over all estimators $\widehat{p}$ (i.e., (potentially randomized) functions of $\widehat{p}:\mathcal{X}\to\mathcal{P}$).*

Note that, compared to Theorem 2.5 of Tsybakov [61], we have loosened some of the constants in order to provide a simpler finite-sample statement.

Suppose $Z\subseteq\mathcal{Z}$ satisfies condition (10) and (11). For each $\tau\in\{-1,1\}^Z$ define

$$p_\tau:=p_0+c_G\sum_{z\in Z}\tau_z\phi_z,$$

where $c_G=\frac{L_G}{B_Z}$, and let $\Omega_G:=\left\{p_\tau:\tau\in\{-1,1\}^Z\right\}$.

Since each ϕ_z is orthogonal to p_0, each $p\in\Omega_G$ has unit mass $\int_{\mathcal{X}}p\,d\lambda=1$, and, by assumption (11),

$$\|p_\tau-p_0\|_{\mathcal{L}^\infty_\lambda}=\left\|\frac{L_G}{B_Z}\sum_{z\in Z}\tau_z\phi_z\right\|_{\mathcal{L}^\infty_\lambda}\leq\frac{L_G}{B_Z}\sum_{z\in Z}\|\phi_z\|_{\mathcal{L}^\infty_\lambda}\leq 0.5,$$

which implies that each $p\in\Omega_G$ is lower bounded on $\mathcal{X}$ by 0.5. Thus, each $p\in\Omega_G$ is a probability density. Note that, if we had worked with Gaussian sequences, as in Liang [33], we would not need to check this, and could hence omit assumption (11). Finally, by construction, for each $p\in\Omega_G$,

$$\|p\|_b^q=\sum_{z\in Z}b_z^q|p_z|^q=c^q\sum_{z\in Z}b_z^q\leq c^q|Z|\sup_{z\in Z}b_z^q=L_G^q$$

so that $\Omega_G\subseteq\mathcal{H}_{b,q}(L_G)$. Also, for $c_D:=\frac{L_D}{A_Z}$ and for each $\tau\in\{-1,1\}^Z$, let

$$f_\tau:=\frac{L_D}{A_Z}\sum_{z\in Z}\tau_z\phi_z,$$

and define $\Omega_D:=\left\{f_\tau:\tau\in\{-1,1\}^Z\right\}$. By construction, for each $f_\tau\in\Omega_D$,

$$\|f_\tau\|_a^p=\frac{L_D^p}{A_Z^p}\sum_{z\in Z}a_z^p\leq\frac{L_D^p}{A_Z^p}|Z|\sup_{z\in Z}a_z^p=L_D^p,$$

so that $\Omega_D \subseteq \mathcal{H}_{p,a}(L_D)$. Then, for any $\tau, \tau' \in \{-1,1\}^Z$,

$$d_{\mathcal{F}_D}\left(p_\tau, p_{\tau'}\right) \geq d_{\Omega_D}\left(p_\tau, p_{\tau'}\right) = \sup_{\tau'' \in \{-1,1\}^Z} \sum_{z \in Z} f_{\tau'',z} c_G(\tau_z - \tau'_z) = 2 c_G c_D \omega\left(\tau, \tau'\right),$$

where $\omega\left(\tau, \tau'\right) := \sum_{z \in Z} 1_{\{\tau_z \neq \tau'_z\}}$ denotes the Hamming distance between τ and τ'. By the Varshamov-Gilbert bound (Lemma 2.9 of Tsybakov [61]), we can select $T \subseteq \{-1,1\}^Z$ such that $\log |T| \geq \frac{|Z| \log 2}{8}$ and, for each $\tau, \tau' \in T$,

$$\omega\left(\tau, \tau'\right) \geq \frac{|Z|}{8}, \quad \text{so that} \quad d_{\mathcal{F}}\left(\theta_\tau, \theta_{\tau'}\right) \geq \frac{c_G c_D |Z|}{4}.$$

Moreover, for any $\tau \in \{-1,1\}^Z$, using the facts that $-\log(1+x) \leq x^2 - x$ for all $x \geq -0.5$ and that $\int_{\mathcal{X}} p_\tau \, dx = 1 = \int_{\mathcal{X}} p_0 \, dx$,

$$\begin{aligned}
D_{KL}(p_\tau^n, p_0^n) &= n D_{KL}(p_\tau, p_0) \\
&= n \int_{\mathcal{X}} p_\tau(x) \log \frac{p_\tau(x)}{p_0(x)} \, dx \\
&= -n \int_{\mathcal{X}} p_\tau(x) \log \left(1 + \frac{p_0(x) - p_\tau(x)}{p_\tau(x)}\right) dx \\
&\leq n \int_{\mathcal{X}} p_\tau(x) \left(\left(\frac{p_0(x) - p_\tau(x)}{p_\tau(x)} \right)^2 - \frac{p_0(x) - p_\tau(x)}{p_\tau(x)} \right) dx \\
&= n \int_{\mathcal{X}} \frac{\left(p_0(x) - p_\tau(x)\right)^2}{p_\tau(x)} \, dx \\
&\leq 2n \int_{\mathcal{X}} \left(p_0(x) - p_\tau(x)\right)^2 \, dx \\
&= 2n \|p_0 - p_\tau\|_{\mathcal{L}^2_\lambda} = 2n \frac{L_G^2}{B_Z^2} |Z| \leq n \frac{L_G^2}{B_Z^2} \frac{16}{\log 2} \log |T| \leq \frac{\log |T|}{16},
\end{aligned}$$

where the last two inequalities follow from the Varshamov-Gilbert bound and assumption (10), respectively. Combining the above results, Lemma 11 gives a minimax lower bound of

$$M(\mathcal{F}_D, \mathcal{F}_G) \geq \frac{c_G c_D |Z|}{64} = \frac{L_G L_D |Z|}{64 A_Z B_Z}.$$

□

13 Proofs and Further Discussion of Applications in Section 6

Example 4 (Sobolev Spaces, Oracle and Adaptive estimators in Fourier basis). Suppose that, for some $s, t \geq 0$, $a_z = \left(1 + \|z\|_\infty^2\right)^{s/2}$ and $b_z = \left(1 + \|z\|_\infty^2\right)^{t/2}$. Then, one can check that, for $c = \frac{2^{d-2s} d}{d-2s}$,

$$\sum_{z \in Z} a_z^{-2} \leq 1 + c\left(\zeta^{d-2s} - 1\right), \quad \sup_{z \in \mathcal{Z} \setminus Z} a_z^{-1} \leq \zeta^{-s}, \quad \text{and} \quad \sup_{z \in \mathcal{Z} \setminus Z} b_z^{-1} \leq \zeta^{-t},$$

so that Theorem 1 gives

$$\mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right) \right] \leq \frac{L_D}{\sqrt{n}} \left(1 + c \zeta^{d/2 - s}\right) + L_D L_G \zeta^{-(s+t)}. \tag{12}$$

Setting $\zeta = n^{\frac{1}{2t+d}}$ gives

$$\mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right) \right] \leq C n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}, \quad \text{where} \quad C := L_D\left(2\sqrt{c} + L_G\right).$$

On the other hand, as long as $t > d/2$, setting

$$\zeta = \left(256 L_G^2 \frac{n}{\log 2}\right)^{\frac{1}{2t+d}}$$

satisfies the conditions of Theorem 3, giving the minimax lower bound

$$M(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}) \geq \frac{L_G L_D}{64 \zeta^{s+t}} = c_1 n^{-\frac{s+t}{2t+d}} \quad \text{where} \quad c_1 = \frac{L_G L_D}{64} \left(\frac{\log 2}{256 L_G^2} \right)^{\frac{t+s}{2t+d}}.$$

Classical methods can also be used to show that, for all values of s and t, $M(\mathcal{H}_{s,2}, \mathcal{H}_{t,2}) \geq c_2 n^{-1/2}$. Thus, we conclude, there exist constants $C, c > 0$ such that

$$c n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}} \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2}\right) \leq C n^{-\min\left\{\frac{1}{2}, \frac{s+t}{2t+d}\right\}}. \tag{13}$$

Combining the observation that the s-Hölder space $\mathcal{W}^{s,\infty} \subseteq \mathcal{W}^{s,2}$ with the lower bound in Theorem 3.1 of Liang [33], we have that (13) also holds when $\mathcal{H}_{s,2}$ is replaced with $\mathcal{W}^{s,\infty}$ (e.g., in the case of the Wasserstein metric $d_{\mathcal{W}^{1,\infty}}$), or indeed $\mathcal{W}^{s,q}$ for any $q \geq 2$.

Corollary 12 (Adaptive Upper Bound for Sobolev Spaces). *For any $t, \zeta \geq 0$ and $s \in (0, d/2)$,*

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P}_{Z_\zeta} \right) \right] \leq C \zeta^{-s} \sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{L}^2_\mu} \left(P, \widehat{P}_{Z_\zeta} \right) \right], \tag{14}$$

where $C := \sqrt{2} \left(1 + \frac{2^{d-2s} d}{d-2s} \right)$ does not depend on n or ζ. Hence, if $\widehat{\zeta}(X_{1:n})$ is any adaptive scheme for choosing ζ (i.e., if computing $\widehat{\zeta}$ does not require knowledge of t), then $\widehat{P}_{\widehat{\zeta}}$ is adaptively minimax under the loss $d_{\mathcal{W}^{s,2}}$; that is, for all $t > 0$, there exists $C > 0$ such that

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P}_{Z_{\widehat{\zeta}}} \right) \right] \leq M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2} \right).$$

One common scheme for choosing $\widehat{\zeta}$ is to use a leave-one-out cross-validation scheme. Specifically, for

$$\widehat{J}(\zeta) := \|\widehat{P}_\zeta\|_2^2 - \frac{2}{n} \sum_{i=1}^n \widehat{P}_{\zeta,-i}(X_i), \quad \textit{where} \quad \widehat{P}_{\zeta,-i} := \sum_{z \in Z_\zeta} \left(\frac{1}{n-1} \sum_{j \in [n] \setminus \{i\}} \phi_z(X_j) \right) \phi_z$$

is a computation of the estimate $\widehat{P}_\zeta$ omitting the i^{th} sample X_i, one can show that $\mathbb{E}_{X_{1:n} \overset{IID}{\sim} P} \left[\widehat{J}(\zeta) \right] = \mathbb{E}_{X_{1:n} \overset{IID}{\sim} P} \left[d^2_{\mathcal{L}^2_\mu} \left(P, \widehat{P}_\zeta \right) \right] - \|P\|^2_{\mathcal{L}^2_\mu}$, so that, up to an additive constant independent of ζ, $\widehat{J}(\zeta)$ is an unbiased estimate of the squared $\mathcal{L}^2_\mu$-risk using the parameter ζ. Based on this, setting

$$\widehat{\zeta} := \mathop{\mathrm{argmin}}_{\zeta \in [0, n^{-1/d}]} J(\zeta),$$

one can show that $\widehat{P}_{\widehat{\zeta}}$ is adaptively minimax over all Sobolev spaces $\mathcal{W}^{t,2}$ with $t > 0$; that is, for all $t > 0$,

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{L}^2_\mu} \left(P, \widehat{P}_{\widehat{\zeta}} \right) \right] \asymp M\left(\mathcal{L}^2_\mu, \mathcal{W}^{t,2} \right). \tag{15}$$

This equivalence (14) *implies that we can generalize the adaptive minimaxity bound* (15) *to*

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P}_{\widehat{\zeta}} \right) \right] \asymp M\left(\mathcal{W}^{s,2}, \mathcal{W}^{t,2} \right). \tag{16}$$

for all $s \in [0, d/2]$.

Proof. A proof of the adaptive minimaxity of the cross-validation estimator in $d_{\mathcal{L}^2_\mu}$ can be found in Sections 7.2.1 and 7.5.1 of Massart [36]. Therefore, we prove only Inequality (14) here. To do this, we combine Theorem 1 with a lower bound on the worst-case performance of the orthogonal series estimator under $\mathcal{L}^2_\mu$ loss, which we establish by explicitly constructing a worst-case true distribution as follows.

Define $P_\zeta := 1 + L_G \zeta^{-t} \phi_\zeta$ (where ϕ_ζ is any ϕ_z satisfying $\|z\|_\infty = \zeta$), one can easily check that $P_\zeta \in \mathcal{W}^{t,2}$, and that, for any z with $\|z\| < \zeta$,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P_\zeta} \left[\left(\widetilde{(P_\zeta)}_z - \widehat{P}_z \right)^2 \right] &= \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P_\zeta} \left[\left(\frac{1}{n} \sum_{i=1}^n \phi_z(X_i) \right)^2 \right] \\
&= \frac{1}{n} \mathop{\mathbb{E}}_{X \sim P_\zeta} \left[\phi_z^2(X) \right] \\
&= \frac{1}{n} \int_{\mathcal{X}} \phi_z^2(x) \left(1 + L_G \zeta^{-t} \phi_\zeta(x) \right) \, dx \\
&\geq \frac{1}{n} \int_{\mathcal{X}} \phi_z^2(x) \, dx = \frac{1}{n}
\end{aligned}$$

(with equality if $\zeta \neq 2z$). Also, let

$$f := \frac{L_D}{\sqrt{2}} \sum_{\|z\| < \zeta} \frac{\left(\widetilde{P_\zeta}_z - \widehat{P}_z \right)}{\sqrt{|Z_\zeta|}} \phi_z + \frac{L_D}{\sqrt{2}} \phi_\zeta$$

so that

$$\|f\|_2^2 = \frac{L_D^2}{2} \sum_{\|z\| < \zeta} \frac{\left(\widetilde{P_\zeta}_z - \widehat{P}_z \right)^2}{|Z_\zeta|} + \frac{L_D^2}{2} \leq \frac{L_D^2}{2} \sum_{\|z\| < \zeta} |Z_\zeta|^{-1} + \frac{L_D^2}{2} \leq L_D^2,$$

and hence $f \in \mathcal{L}_\mu^2(1)$. Then,

$$\begin{aligned}
\mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P_\zeta} \left[d_{\mathcal{L}_\mu^2} \left(P_\zeta, \widehat{P}_{Z_\zeta} \right) \right] &\geq \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\sum_{\|z\| < \zeta} \widetilde{f}_z \left(\widetilde{P_\zeta}_z - \widehat{P}_z \right)^2 + \widetilde{f}_\zeta \widetilde{P_\zeta}_z \right] \\
&= \frac{L_D}{\sqrt{2|Z_\zeta|}} \sum_{\|z\| < \zeta} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\left(\widetilde{P_\zeta}_z - \widehat{P}_z \right)^2 \right] + \frac{L_D L_G}{\sqrt{2}} \zeta^{-t} \\
&\geq \frac{L_D}{\sqrt{2|Z_\zeta|}} \sum_{\|z\| < \zeta} \frac{1}{\sqrt{n}} + \frac{L_D L_G}{\sqrt{2}} \zeta^{-t} = \frac{L_D}{\sqrt{2}} \left(\sqrt{\frac{\zeta^d}{n}} + L_G \zeta^{-t} \right)
\end{aligned}$$

It follows that

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{L}_\mu^2} \left(P, \widehat{P}_{Z_\zeta} \right) \right] \geq \frac{L_D}{\sqrt{2}} \left(\sqrt{\frac{\zeta^d}{n}} + \zeta^{-t} \right).$$

On the other hand, as we already saw, Theorem 1 gives

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P} \right) \right] \leq \left(1 + \frac{2^{d-2s} d}{d - 2s} \right) L_D \left(\sqrt{\frac{\zeta^d}{n}} + L_G \zeta^{-t} \right) \zeta^{-s}.$$

Combining these two inequalities gives

$$\sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{W}^{s,2}} \left(P, \widehat{P} \right) \right] \leq C \zeta^{-s} \sup_{P \in \mathcal{W}^{t,2}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[d_{\mathcal{L}_\mu^2} \left(P, \widehat{P}_{Z_\zeta} \right) \right].$$

□

13.1 Wavelet Basis

Our previous applications were given in terms of the Fourier basis. In this section, we demonstrate that our upper and lower bounds can give tight minimax results using other bases (in this case, the Haar wavelet basis).

Suppose that $\mathcal{X} = [0,1]^D$, and suppose that a function $f : \mathcal{X} \to \mathbb{R}$ has Haar wavelet basis coefficients $\widetilde{f}_{i,j}$, indexed by $z \in \mathcal{Z} := \{(i,j) \in \mathbb{N} \times \mathbb{N} : j \in [2^i]\}$, where $i \in \mathbb{N}$ is the order and $j \in [2^i]$ is the index within that order.

One can show (see, e.g., Donoho et al. [15]) that the Besov seminorm $\|\cdot\|_{\mathcal{B}^q_{p,q}}$ satisfies

$$\|f\|^q_{\mathcal{B}^r_{p,q}} = \sum_{i\in\mathbb{N}} 2^{iqs} \left(\sum_{j\in[2^i]} |\widetilde{f}_{i,j}|^p \right)^{q/p} = \sum_{i\in\mathbb{N}} 2^{iqs} \|\widetilde{f}_i\|^q_p,$$

where $s = r + \frac{1}{2} - \frac{1}{p}$. In particular, when $p = q = 2$, $s = r$, and one can show that $\mathcal{B}^r_{p,q} = \mathcal{W}^r_2$, and

$$\|f\|^q_{\mathcal{B}^r_{p,q}} = \sum_{(i,j)\in\mathcal{Z}} 2^{2is} |\widetilde{f}_{i,j}|^2,$$

For some $\zeta > 0$, we will choose the truncation set Z to be of the form

$$Z = \{(i,j) \in \mathcal{Z} : i \leq \zeta\}.$$

Note that, for each $i \in \mathbb{N}$, since $\phi_{i,1}, ..., \phi_{i,2^i}$ have disjoint supports

$$\sup_{x\in\mathcal{X}} \sum_{j\in[2^i]} |\phi_{i,j}(x)| = \sup_{x\in\mathcal{X}} \sup_{j\in[2^i]} |\phi_{i,j}(x)| = 2^{i/2}.$$

Thus,

$$\sum_{j\in[2^i]} \|\phi_{i,j}\|^2_{\mathcal{L}^2_P} = \sum_{j\in[2^i]} \int_{\mathcal{X}} \phi^2_{i,j}(x) dP \leq \int_{\mathcal{X}} \left(\sum_{j\in[2^i]} \phi_{i,j}(x) \right)^2 dP = 2^i.$$

Example 13 (Sobolev Space, Wavelet Basis)**.** Suppose that, for some $s, t \geq 0$, $a_{i,j} = 2^{is}$ and $b_{i,j} = 2^{it}$. Then, one can check that, for some $c > 0$

$$\sum_{z\in\mathcal{Z}} \frac{\|\phi_z\|^2_{\mathcal{L}^2_P}}{a^2_z} = \sum_{i\leq\zeta} \sum_{j\in[2^i]} \frac{\|\phi_{i,j}\|^2_{\mathcal{L}^2_P}}{2^{2is}} = \sum_{i\leq\zeta} \frac{2^i}{2^{2is}} = \frac{2^{(\zeta+1)(1-2s)} - 1}{2^{1-2s} - 1} \asymp 2^{\zeta(1-2s)}.$$

Also, $\sup_{z\in\mathcal{Z}\setminus Z} a^{-1}_z \leq 2^{-s\zeta}$ and $\sup_{z\in\mathcal{Z}\setminus Z} b^{-1}_z \leq 2^{-t\zeta}$. Thus, Theorem 1 gives

$$\mathop{\mathbb{E}}_{X_{1:n}} \left[d_{\mathcal{F}_D}\left(P, \widehat{P}\right) \right] \lesssim L_D \left(\sqrt{\frac{c}{n}} 2^{(d/2-s)\zeta} + L_G 2^{-(s+t)\zeta} \right).$$

By letting $\zeta = \log_2 \xi$, we can easily see that this is identical, up to constants, to the bound for the Sobolev case. In contrast to Fourier basis, a larger variety of function spaces (such as inhomogeneous Besov spaces) can be expressed in terms of wavelet basis. The classical work of Donoho et al. [15] showed that, under $\mathcal{L}^p_\mu$ losses, linear estimators, such as that analyzed in our Theorem 1 are sub-optimal in these spaces, but that relatively simple thresholding estimators can recover the minimax rate. We leave it to future work to understand how this phenomenon extends to more general adversarial losses.

14 Proofs and Applications of Explicit & Implicit Generative Modeling Results (Section 8 of Main Paper)

Here, we prove Theorem 9 from the main text, provide some discussion of when the converse direction $M_I(\mathcal{P}, \ell, n) \leq M_D(\mathcal{P}, \ell, n)$ holds, and also provide some concrete applications.

14.1 Proofs of Theorem 9 and Converse

Theorem 9 (Conditions under which Density Estimation is Statistically no harder than Sampling). Let $\mathcal{F}_G$ be a family of probability distributions on a sample space $\mathcal{X}$. Assume the following:

(A1) $\ell : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is non-negative, and there exists $C_\triangle > 0$ such that, for all $P_1, P_2, P_3 \in \mathcal{F}_G$,

$$\ell(P_1, P_3) \leq C_\triangle \left(\ell(P_1, P_2) + \ell(P_2, P_3)\right).$$

(A2) $M_D(\mathcal{F}_G, \ell, m) \to 0$ as $m \to \infty$.

(A3) For all $m \in \mathbb{N}$, we can draw m IID samples $Z_{1:m} = Z_1, ..., Z_m \overset{IID}{\sim} Q_Z$ of the latent variable Z.

(A4) there exists a nearly minimax sequence of samplers $\widehat{X}_k : \mathcal{X}^n \times \mathcal{Z} \to \mathcal{X}$ such that, for each $k \in \mathbb{N}$, almost surely over $X_{1:n}$, $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}} \in \mathcal{F}_G$.

Then, $M_D(\mathcal{F}_G, \ell, n) \leq C_\triangle M_I(\mathcal{F}_G, \ell, n)$.

Proof. The assumption (A2) implies that there exists a sequence $\{\widehat{P}_m\}_{m \in \mathbb{N}}$ of density estimators $\widehat{P}_m : \mathcal{X}^m \to \mathcal{P}$ that is uniformly consistent in ℓ over $\mathcal{P}$; that is,

$$\lim_{m \to \infty} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{Y_{1:m} \overset{IID}{\sim} P} \left[\ell\left(P, \widehat{P}_m(Y_{1:m})\right)\right]. \tag{17}$$

For brevity, we use the abbreviation $P_{\widehat{X}_k} = P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}}$ in the rest of this proof to denote the conditional distribution of the 'fake data' generated by $\widehat{X}_k$ given the true data. Recalling that the minimax risk is at most the risk of any particular sampler, we have

$$\begin{aligned} M_D(\mathcal{P}, \ell, n) &:= \inf_{\widehat{P}} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P, \widehat{P}(X_{1:n})\right)\right] \\ &\leq \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P, \widehat{P}_m(X_{n+1:n+m})\right)\right]. \end{aligned}$$

Taking $\lim_{m \to \infty}$ gives, by Tonelli's theorem and non-negativity of ℓ,

$$\begin{aligned} & M_D(\mathcal{P}, \ell, n) \\ & \leq \lim_{m \to \infty} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P, \widehat{P}_m(X_{n+1:n+m})\right)\right] \\ & \leq C_\triangle \lim_{m \to \infty} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P, P_{\widehat{X}_k}\right) + \ell\left(P_{\widehat{X}_k}, \widehat{P}_m(X_{n+1:n+m})\right)\right] \\ & \leq C_\triangle \lim_{m \to \infty} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P, P_{\widehat{X}_k}\right) + \ell\left(P_{\widehat{X}_k}, \widehat{P}_m(X_{n+1:n+m})\right)\right] \\ & \leq C_\triangle \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\ell\left(P, P_{\widehat{X}_k}\right)\right] \qquad (18) \\ & + C_\triangle \lim_{m \to \infty} \sup_{P \in \mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P_{\widehat{X}_k}, \widehat{P}_m(X_{n+1:n+m})\right)\right]. \qquad (19) \end{aligned}$$

In the above, we upper bounded $M_D(\mathcal{P}, \ell, n)$ by the sum of two terms, (18) and (19). Since the sequence $\{\widehat{X}_k\}_{k \in \mathbb{N}}$ is nearly minimax, if we were to take an infimum over $k \in \mathbb{N}$ on both sides, the term (18) would become precisely $C_\triangle M_I(\mathcal{P}, \ell, n)$. Therefore, it suffices to observe that the second term (19) is 0. Indeed, by the assumption that $P_{\widehat{X}_k} \in \mathcal{P}$ for all $X_{1:n} \in \mathcal{X}$ and the uniform consistency assumption (17),

$$\begin{aligned}
&\lim_{m\to\infty} \sup_{P\in\mathcal{P}} \mathop{\mathbb{E}}_{\substack{X_{1:n} \overset{IID}{\sim} P \\ Z_{1:m} \overset{IID}{\sim} Q_Z}} \left[\ell\left(P_{\widehat{X}_k}, \widehat{P}_m(X_{n+1:n+m})\right)\right] \\
&\leq \lim_{m\to\infty} \sup_{P\in\mathcal{P}, X_{1:n} \overset{IID}{\sim} P} \mathop{\mathbb{E}}_{Z_{1:m} \overset{IID}{\sim} Q_Z} \left[\ell\left(P_{\widehat{X}_k}, \widehat{P}_m(X_{n+1:n+m})\right)\right] \\
&\leq \lim_{m\to\infty} \sup_{P'\in\mathcal{P}} \mathop{\mathbb{E}}_{X_{n+1:n+m} \overset{IID}{\sim} P'} \left[\ell\left(P, \widehat{P}_m(X_{n+1:n+m})\right)\right] = 0.
\end{aligned}$$

□

For completeness, we provide a very simple result on the converse of Theorem 9:

Theorem 14 (Conditions under which Sampling is Statistically no harder than Density Estimation)**.** *Suppose that, there exists as nearly minimax sequence $\{\widehat{P}_k\}_{k\in\mathbb{N}}$ such that, for any $k \in \mathbb{N}$, we can draw a random sample $\widehat{X}$ from $\widehat{P}_k(X_{1:n})$. Then,*

$$M_D(\mathcal{F}_G, \ell, n) \geq M_I(\mathcal{F}_G, \ell, n).$$

The assumption above that we can draw samples from a nearly minimax sequence of estimators if not particularly insightful, but techniques for drawing such samples have been widely studied in the vast literature of Monte Carlo sampling [49]. As an example, if $\widehat{P}$ is a kernel density estimator with kernel K, then, recalling that K is itself a probability density, of which $\widehat{P}$ is a mixture, we can sample from $\widehat{P}$ simply by choosing a sample uniformly from $X_{1:n}$ and adding noise $\epsilon \sim K$. Alternatively, if $\widehat{P}$ is bounded and has bounded support, then one can perform rejection sampling.

Proof. Since, by definition of the implicit distribution of $\widehat{X}$,

$$P_{\widehat{X}(X_{1:n},Z)|X_{1:n}} = \widehat{P}(X_{1:n})$$

is precisely the implicit distribution of $\widehat{X}$, we trivially have

$$M_I(\mathcal{F}_G, \ell, n) \leq \sup_{P\in\mathcal{F}_G} \mathop{\mathbb{E}}_{X_{1:n} \overset{IID}{\sim} P} \left[\ell\left(P, P_{\widehat{X}(X_{1:n},Z)|X_{1:n}}\right)\right]$$

□

14.2 Applications

Example 15 (Density Estimation and Sampling in Sobolev families under Dual-Sobolev Loss)**.** There exist constants $C > c > 0$ such that, for all $n \in \mathbb{N}$,

$$cn^{-\min\left\{\frac{s+t}{2s+d}, \frac{1}{2}\right\}} \leq M_I\left(\mathcal{W}^{t,2}, d_{\mathcal{W}^{s,2}}, n\right) \leq Cn^{-\min\left\{\frac{s+t}{2s+d}, \frac{1}{2}\right\}}.$$

Proof. Since adversarial losses always satisfy the triangle inequality, the first inequality follows Theorems 9 and the discussion in Example 4. For the second inequality, since we have already established that the orthogonal series estimator $\widehat{P}_Z$ is nearly minimax, by Theorem 14 it suffices to give a scheme for sampling from the distribution $\widehat{P}_Z(X_{1:n})$. Since the sample space $\mathcal{X} = [0,1]^d$ is bounded and the estimator $\widehat{P}_Z(X_{1:n})$ has a bounded density $p : \mathcal{X} \to [0,\infty)$, we can simply perform rejection sampling; that is, repeatedly sample $Z \times Y$ uniformly from $\mathcal{X} \times [0, \sup_{x\in\mathcal{X}} p(x)]$. Let Z^* denote the first Z sample satisfying $Y < p(Z)$. Then, we Z^* will necessarily have the density p. □

Example 16 (Density Estimation and Sampling in Exponential Families under Jensen-Shannon, $\mathcal{L}^q$, Hellinger, and RKHS losses)**.** Let $\mathcal{H}$ be an RKHS over a compact sample space $\mathcal{X} \subseteq \mathbb{R}^d$, and let

$$\mathcal{F}_G := \left\{ p_f : \mathcal{X} \to [0, \infty) \middle| p_f(x) = e^{f(x) - A(f)} \text{ for all } x \in \mathcal{X}, f \in \mathcal{H} \right\},$$

in which $A(f) := \log \int_{\mathcal{X}} e^{f(x)}\, d\mu$ denotes the log-partition function.

The Jensen-Shannon divergence $J : \mathcal{P} \times \mathcal{P} \to [0, \infty]$ is defined by

$$J(P, Q) := \frac{1}{2} \left(D_{KL} \left(P, \frac{P+Q}{2} \right) + D_{KL} \left(Q, \frac{P+Q}{2} \right) \right),$$

where $\frac{P+Q}{2}$ denotes the uniform mixture of P and Q, and, noting that we always have $P \ll \frac{P+Q}{2}$ and $Q \ll \frac{P+Q}{2}$,

$$D_{KL}(P, Q) := \int_{\mathcal{X}} \log \left(\frac{dP}{dQ} \right) dP$$

denotes the Kullback-Leibler divergence. Although J does not satisfy the triangle inequality, one can show that $\sqrt{J}$ is a metric on $\mathcal{P}$ [20], and hence, for all $P, Q \in \mathcal{P}$, by Cauchy-Schwarz,

$$J(P, Q) = \left(\sqrt{J(P, Q)} \right)^2 \leq \left(\sqrt{J(P, R)} + \sqrt{J(R, Q)} \right)^2 \leq 2J(P, R) + 2J(R, Q). \tag{20}$$

Also, under mild regularity conditions on $\mathcal{H}$, Sriperumbudur et al. [53] (in their Theorem 7) provides uniform convergence guarantees for a particular density estimator over $\mathcal{P}$. Combining this the inequality (20), our Theorem 9 implies

$$M_D(\mathcal{P}, J, n) \leq 2M_I(\mathcal{P}, J, n).$$

For the same class $\mathcal{P}$, the convergence results of Sriperumbudur et al. [53] (their Theorems 6 and 7) also imply similar guarantees under several other losses, including the parameter estimation loss $\|f_P - f_{\widehat{P}}\|_H$ in the RKHS metric, as well as the $\mathcal{L}^q_\mu$ and Hellinger metrics H (on the density), so that we have $M_D(\mathcal{P}, \rho, n) \leq M_I(\mathcal{P}, \rho, n)$ when ρ is any of these metrics.

Perhaps more interestingly, in the case of Jensen-Shannon divergence, under certain regularity conditions, we can altogether drop the assumption that $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}} \in \mathcal{P}$ using uniform convergence bounds shown in Section 5 of Sriperumbudur et al. [53] for the mis-specified case; the density estimator described therein converges (uniformly over P_*) to the projection P_* of $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}}$ onto $\mathcal{P}$ even when samples are drawn from $P_{\widehat{X}_k(X_{1:n}, Z)|X_{1:n}}$.

It is also worth pointing out that, when densities in $\mathcal{F}_G$ are additionally assumed to be lower bounded by a positive constant $\kappa > 0$ (i.e.,

$$\kappa := \inf_{p \in \mathcal{F}_G} \inf_{x \in \mathcal{X}} p(x) > 0,$$

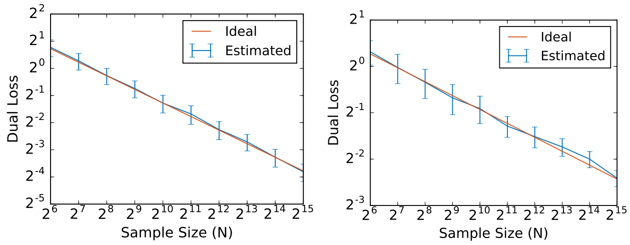
then, by the inequality $-\log(1+x) \leq x^2 - x$ that holds for all $x \geq -0.5$, for all densities $p, q \in \mathcal{F}_G$,

$$\begin{aligned} \int_{\mathcal{X}} p(x) \log \left(\frac{2p(x)}{p(x) + q(x)} \right) dx &= -\int_{\mathcal{X}} p(x) \log \left(1 + \frac{q(x) - p(x)}{2p(x)} \right) dx \\ &\leq \int_{\mathcal{X}} p(x) \left(\left(\frac{q(x) - p(x)}{2p(x)} \right)^2 - \left(\frac{q(x) - p(x)}{2p(x)} \right) \right) dx \\ &= \int_{\mathcal{X}} \frac{(q(x) - p(x))^2}{2p(x)} dx \leq \frac{1}{2\kappa} \|P - Q\|^2_{\mathcal{L}^2_\mu}, \end{aligned}$$

and, therefore, $J(P, Q) \leq \frac{1}{2\kappa} \|P - Q\|_{\mathcal{L}^2_\mu}$. Thus, under this additional assumption of uniform lower-boundedness, standard results for density estimation under $\mathcal{L}^2_\mu$ apply [61].

15 Experimental Results

This section presents some empirical results supporting the theoretical bounds above. First, we consider an example with a finite basis, which should yield the parametric $n^{-1/2}$ rate. In particular, we construct the true distribution P to consist of 6 randomly chosen basis functions in the Fourier basis. We employ the truncated series estimator $\widehat{P}$ of (3) in the same basis using different number of samples n and compute the distance $d_{\mathcal{F}_D}\left(P, \widehat{P}\right)$. Under this setting, the maximization problem of (1) needed to evaluate this distance can be solved in closed form. The risk empirically appears to closely follow our derived minimax rate of $n^{-1/2}$, as shown in Figure 1a. Next, we consider a non-parametric case, in which the number of active basis elements increases as function of n, weighted such that Inequality (6) predicts a rate of $n^{-1/3}$. As expected, the estimated risk, shown in Figure 1b, closely resembles the rate of $n^{-1/3}$.



(a) Parametric Regime

(b) Nonparametric Regime

Figure 1: Simple synthetic experiment to showcase the tightness of the bound.

16 Future Work

In this paper, we showed that minimax convergence rates for distribution estimation under certain adversarial losses can improve when the probability distributions are assumed to be smooth, using an orthogonal series estimator that smooths the observed empirical distribution. On the other hand, recent work has also shown that, at least under Wasserstein losses, minimax convergence rates improve when the distribution is assumed to have support of low intrinsic dimension, even within a high-dimensional ambient space [50]. In any case, further work is needed to understand whether minimax rates further improve when distributions are simultaneously smooth and supported on a set of low intrinsic dimension. It is easy to see that the empirical distribution does *not* benefit from assumed smoothness (see, e.g., Proposition 6 of Weed and Bach [65]). Whether an orthogonal series estimate benefits from low intrinsic dimension may depend on the basis used; the Fourier basis is not likely to benefit, but a wavelet basis, which is spatially localized, may. Nearest neighbor methods have also been shown to benefit from both smoothness and low intrinsic dimensionality, under $\mathcal{L}^2_\mu$ loss, and may therefore be promising [28].

The results in this paper should also be generalized to larger classes of spaces, such as inhomogeneous Besov spaces. Over these spaces, the classic work of Donoho et al. [15] suggests that simple linear density estimators such as the orthogonal series estimator studied in this paper cease to be minimax rate-optimal, but simple non-linear estimators such as wavelet thresholding estimators may continue to be (adaptively) minimax optimal.

The results of Yarotsky [67], on uniform approximation of smooth functions (over Sobolev spaces) by neural networks, we crucial to the result Theorem 7 bounding the error of perfectly optimized GANs. If these approximation-theoretic results can be generalized to other spaces (e.g., RKHSs), then our Theorem 1 can be used to derive performance bounds for perfectly optimized GANs over these spaces.

Finally, it has been widely observed that, in practice, optimization of GANs can be quite difficult [41, 34, 7]. This limits the practical implications of our performance bounds on GANs, which assumed perfect optimization (i.e., convergence to a generator-optimal equilibrium). Conversely, most work studying the optimization landscape of GANs is specific to the noiseless (i.e., "infinite sample size") case, whereas our lower bounds suggest that the sample complexity of training GANs may be substantial. Hence, it is important to generalize these statistical results to the case of imperfect optimization, and, conversely, to understand the effects of statistical noise on the optimization procedure.