Minimax Lower Bounds for Estimating Distributions on Low-dimensional Spaces

Anonymous authors
Paper under double-blind review

Abstract

Recent statistical analyses of Generative Adversarial Networks (GAN) suggest that the error in estimating the target distribution in terms of the β-Hölder Integral Probability Metric (IPM) scales as $\mathcal{O}\left(n^{-\frac{\beta}{\overline{d}_{\mathbb{M}}+\delta}} \vee n^{-1/2}\log n\right)$. Here $\overline{d}_{\mathbb{M}}$ is the upper Minkowski dimension of the corresponding support $\mathbb{M}$ of the data distribution and δ is a positive constant. It is, however, unknown as to whether this rate is minimax optimal, i.e. whether there are estimators that achieve a better test-error rate. In this paper, we show that the minimax rate for estimating unknown distributions in the β-Hölder IPM on $\mathbb{M}$ scales as $\Omega\left(n^{-\frac{\beta}{\underline{d}_{\mathbb{M}}-\delta}} \vee n^{-1/2}\right)$, where $\underline{d}_{\mathbb{M}}$ is the lower Minkowski dimension of $\mathbb{M}$. Thus if the low-dimensional structure $\mathbb{M}$ is regular in the Minkowski sense, i.e. $\overline{d}_{\mathbb{M}} = \underline{d}_{\mathbb{M}}$, GANs are roughly minimax optimal in estimating distributions on $\mathbb{M}$. We also show that the minimax estimation rate in the p-Wasserstein metric scales as $\Omega\left(n^{-\frac{1}{\underline{d}_{\mathbb{M}}-\delta}} \vee n^{-1/(2p)}\right)$.

Nonparametric density estimation, aimed at approximating a probability distribution from a finite collection of identically and independently distributed (i.i.d.) samples, holds extensive application in the realms of statistics and machine learning. Nonparametric density estimation finds application in various fields such as mode estimation (Parzen, 1962), nonparametric classification (Rigollet, 2007; Chaudhuri et al., 2008), Monte Carlo computational methods (Doucet et al., 2001), and clustering (Chaudhuri & Dasgupta, 2010; Chakraborty et al., 2021; Rinaldo & Wasserman, 2010), among others. Typical techniques for nonparametric density estimation encompass the histogram method, kernel method, k-Nearest Neighbor (kNN) method (Devroye & Wagner, 1977; Bhattacharya & Mack, 1987; Zhao & Lai, 2022), wavelet-based methods (Donoho et al., 1996) and more. Notably, the recent advancements in deep learning have led to the groundbreaking concept of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which has revolutionised the field of nonparametric density estimation to obtain superhuman performance, especially for handling vision data.

The empirical successes of GANs have motivated researchers to study their theoretical guarantees. Biau et al. (2020) analyzed the asymptotic properties of vanilla GANs along with parametric rates. Biau et al. (2021) also analyzed the asymptotic properties of WGANs. Liang (2021) explored the min-max rates for WGANs for different non-parametric density classes and under a sampling scheme from a kernel density estimate of the data distribution; while Schreuder et al. (2021) studied the finite-sample rates under adversarial noise. Uppal et al. (2019) derived the convergence rates for Besov discriminator classes for WGANs. Luise et al. (2020) conducted a theoretical analysis of WGANs under an optimal transport-based paradigm. Recently, Asatryan et al. (2023) and Belomestny et al. (2021) improved upon the works of Biau et al. (2020) to understand the behaviour of GANs for Hölder class of density functions. Arora et al. (2017) showed that generalisation might not hold in standard metrics. However, they show that under a restricted "neural-net distance", the GAN is indeed guaranteed to generalize well. Recently, Arora et al. (2018) showed that GANs and their variants might not be well-equipped against mode collapse.

Although significant progress has been made in our theoretical understanding of GAN, some limitations of the existing results are yet to be addressed. For instance, the generalisation bounds frequently suffer from the curse of dimensionality. In practical applications, data distributions tend to have high dimensionality,

making the convergence rates that have been proven exceedingly slow. However, high-dimensional data, such as images, texts, and natural languages, often possess latent low-dimensional structures that reduce the complexity of the problem. For example, it is hypothesised that natural images lie on a low-dimensional structure, in spite of its high-dimensional pixel-wise representation (Pope et al., 2020). Though in classical statistics there have been various approaches, especially using kernel tricks and Gaussian process regression that achieve a fast rate of convergence that depends only on their low intrinsic dimensionality (Bickel & Li, 2007; Kim et al., 2019), such results are largely unexplored in the context of GANs. Recently, Huang et al. (2022) expressed the generalisation rates for GAN when the data has low-dimensional support in the Minkowski sense and the latent space is one-dimensional; while Dahal et al. (2022) derived the convergence rates under the Wasserstein-1 distance in terms of the manifold dimension. It is important to note that the compact Riemannian manifold assumption of the support of the target distribution and the assumption of a bounded density of the target distribution on this manifold by Dahal et al. (2022) is a very strong assumption that might not hold in practice.

Despite these recent advances, it remains uncertain whether GAN estimates of the target distribution are optimal in the minimax sense for estimating distributions that are supported on a low-dimensional structure. In this paper, we address this gap in the current literature by providing a comprehensive analysis of the minimax lower bound for estimating and demonstrate that when n independent and identically distributed samples are available from any target distribution on a set $\mathbb{M}$, the convergence rate for any estimator is at least $\Omega\left(n^{-\frac{\beta}{\underline{d}_{\mathbb{M}}-\delta}} \vee n^{-1/2}\right)$, where δ is a positive constant in the range of $(0, \underline{d}_{\mathbb{M}})$, and $\underline{d}_{\mathbb{M}}$ is the lower Minkowski dimension of the set $\mathbb{M}$. Thus, when the set $\mathbb{M}$ is Minkowski regular, GANs almost match this rate, when the networks are properly chosen. Additionally, we demonstrate that the minimax estimation rate in the p-Wasserstein metric decreases in proportion to $\Omega\left(n^{-\frac{1}{\underline{d}_{\mathbb{M}}-\delta}} \vee n^{-1/(2p)}\right)$.

1 Background

1.1 Related Work

Recent research has explored the minimax rates under the Wasserstein distances under various settings. Singh & Póczos (2018) demonstrated the minimax convergence rates assuming the distribution is compactly supported. In a related context, Liang (2021) and Uppal et al. (2019) established minimax convergence rates for the Wasserstein-1 distance under a smoothness assumption on the corresponding density. It has been demonstrated that estimating under the Integral Probability Metric (IPM) with smooth functions can lead to enhanced rates of convergence for empirical measures Kloeckner (2020). Niles-Weed & Berthet (2022) established the minimax convergence rates for Besov densities for the Wasserstein-p metric. For smooth densities, the derived minimax rates can be improved (McDonald, 2017; Liang, 2021) in the sense that estimating a smooth density is easier than estimating a non-smooth one. However, all the aforementioned findings primarily consider the minimax rates when the corresponding distribution varies across all probability measures on a compact set, resulting in rates of $\mathcal{O}(n^{-1/d})$ or similar. Recently, Tang & Yang (2023) derived the minimax rates when the data is supported on a smooth manifold and has a smooth density w.r.t. the Lebesgue measure on this manifold. In contrast, when measures are confined to a low-dimensional structure, we demonstrate that the resulting minimax rates can be enhanced, incorporating a dependence on the exponent of n as the Minkowski dimension of this structure. Notably, this not only eliminates the necessity of a manifold structure of the support but also relaxes the assumption of absolute continuity and smoothness of the underlying measure.

1.2 Preliminaries and Notations

Before we go into the details of the theoretical results, we introduce some notation and recall some preliminary concepts.

We use the notation $x \vee y := \max\{x, y\}$. $B_\varrho(x, r)$ denotes the open ball of radius r around x, with respect to (w.r.t.) the metric ϱ. For any measure γ, the support of γ is defined as, $\text{supp}(\gamma) = \{x : \gamma(B_\varrho(x, r)) >$

0, for all $r > 0\}$. For any function $f : S \to \mathbb{R}$, and any measure γ on S, let $\|f\|_{\mathbb{L}_p(\gamma)} := \left(\int_S |f(x)|^p d\gamma(x)\right)^{1/p}$, if $0 < p < \infty$. Also let, $\|f\|_{\mathbb{L}_\infty(\gamma)} := \operatorname{ess\,sup}_{x\in\operatorname{supp}(\gamma)} |f(x)|$. We say $A_n \precsim B_n$ (also written as $A_n = \mathcal{O}(B_n) \iff B_n = \Omega(A_n)$) if there exists $C > 0$, independent of n, such that $A_n \leq CB_n$. We say $A_n \asymp B_n$, if $A_n \precsim B_n$ and $B_n \precsim A_n$. For any $k \in \mathbb{N}$, we let $[k] = \{1, \ldots, k\}$. $\mathbb{1}(\cdot)$ denotes the indicator function. For any measure μ, $\mu^{\otimes_n}$ denotes the n-product measure of μ. We also recall some useful definitions as follows.

Definition 1 (Covering and Packing Numbers)**.** For a metric space (S, ϱ), the ϵ-covering number w.r.t. ϱ is defined as: $\mathcal{N}(\epsilon; S, \varrho) = \inf\{n \in \mathbb{N} : \exists\, x_1, \ldots x_n \text{ such that } \cup_{i=1}^n B_\varrho(x_i, \epsilon) \supseteq S\}$. Similarly, the ϵ-packing number is defined as: $\mathcal{M}(\epsilon; S, \varrho) = \sup\{m \in \mathbb{N} : \exists\, x_1, \ldots x_m \in S \text{ such that } \varrho(x_i, x_j) \geq \epsilon, \text{ for all } i \neq j\}$.

Definition 2 (Hölder functions)**.** Let $f : S \to \mathbb{R}$ be a function, where $S \subseteq \mathbb{R}^d$. For a multi-index $\boldsymbol{s} = (s_1, \ldots, s_d)$, let, $\partial^{\boldsymbol{s}} f = \frac{\partial^{|\boldsymbol{s}|} f}{\partial x_1^{s_1} \ldots \partial x_d^{s_d}}$, denote the weak partial derivative of f, where, $|\boldsymbol{s}| = \sum_{\ell=1}^d s_\ell$. We say that a function $f : S \to \mathbb{R}$ is β-Hölder (for $\beta > 0$) if

$$\|f\|_{\mathbb{H}^\beta} := \sum_{\boldsymbol{s}:0\leq|\boldsymbol{s}|\leq\lfloor\beta\rfloor} \|\partial^{\boldsymbol{s}} f\|_\infty + \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor\beta\rfloor} \sup_{x\neq y} \frac{\|\partial^{\boldsymbol{s}} f(x) - \partial^{\boldsymbol{s}} f(y)\|}{\|x - y\|^{\beta-\lfloor\beta\rfloor}} < \infty.$$

If $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, then we define $\|f\|_{\mathbb{H}^\beta} = \sum_{j=1}^{d_2} \|f_j\|_{\mathbb{H}^\beta}$. For notational simplicity, let, $\mathbb{H}^\beta(S_1, S_2, C) = \{f : S_1 \to S_2 : \|f\|_{\mathbb{H}^\beta} \leq C\}$. Here, both S_1 and S_2 are both subsets of real vector spaces. If $S_1 = [0, 1]^d$, $S_2 = \mathbb{R}$ and $C = 1$, we write $\mathbb{H}^\beta$ in stead of $\mathbb{H}^\beta(S_1, S_2, C)$.

Next, we recall the definitions of Total Variation and Wasserstein-p distances as well as Integral Probability Metrics (IPMs).

Definition 3 (Total Variation Distance)**.** Let Ω be a Polish space and suppose that μ and ν be two probability measures defined on Ω. Then, the total variation distance between μ and ν is defined as,

$$\mathrm{TV}(\mu, \nu) = \sup_{B\in\mathscr{B}(\Omega)} |\mu(B) - \nu(B)| = \inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{P}_{(X,Y)\sim\gamma}(X \neq Y). \tag{1}$$

Here, $\mathscr{B}(\Omega)$ denotes the Borel σ-algebra on Ω and $\Gamma(\mu, \nu)$ denotes the set of all measure couples between μ and ν. The reader is referred to Proposition 4.7 of Levin & Peres (2017) for a proof of the second equality in (1).

Definition 4 (Wasserstein p-distance)**.** *Let $(\Omega, dist)$ be a Polish space and let μ and ν be two probability measures on the same with finite p-moments. Then the p-Wasserstein distance between μ and ν is defined as:*

$$\mathbb{W}_p(\mu, \nu) = \left(\inf_{\gamma\in\Gamma(\mu,\nu)} \mathbb{E}_{(X,Y)\sim\gamma} \left(\mathrm{dist}(X, Y)\right)^p\right)^{1/p}.$$

In what follows, we take dist to be the ℓ_2-norm on $\mathbb{R}^d$.

Definition 5 (Integral Probability Metric)**.** *For a function class $\mathcal{F}$, the $\mathcal{F}$-Integral Probability Metric (IPM) between two probabiloty measures μ and ν is defined as,*

$$\|\mu - \nu\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\int f d\mu - \int f d\nu\right|.$$

1.3 Minkowski Dimension

Often, real data is hypothesized to lie on a lower-dimensional structure within the high-dimensional representative feature space. To characterize this low-dimensionality of the data, researchers have defined various notions of the effective dimension of the underlying measure from which the data is assumed to be generated. Among these approaches, the most popular ones use some sort of rate of increase of the covering number, in the log-scale, of most of the support of this data distribution. Let (S, ϱ) be a compact Polish space and let μ be a probability measure defined on it. Throughout the remainder of the paper, we take ϱ to be the ℓ_∞-norm. We characterize this low-dimensional nature of the data, through the Minkwoski dimension of the support of μ. We recall the definition of Minkowski dimensions (Falconer, 2004),

Definition 6 (Minkowski dimension)**.** *For a bounded metric space (S, ϱ), the upper Minkwoski dimension of S is defined as*

$$\overline{d}_S = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}(\epsilon;\, S,\, \varrho)}{\log(1/\epsilon)}.$$

Similarly, the lower Minkowski dimension of S is given by,

$$\underline{d}_S = \liminf_{\epsilon \downarrow 0} \frac{\log \mathcal{N}(\epsilon;\, S,\, \varrho)}{\log(1/\epsilon)}.$$

If $\underline{d}_S = \overline{d}_S$, we say that S is Minkowski regular and a has Minkowski dimension of $d_S = \lim_{\epsilon \downarrow 0} \frac{\log \mathcal{N}(\epsilon; S, \varrho)}{\log(1/\epsilon)}$.

The Minkowski dimension essentially measures the how the covering number of S is affected by the radius of the covering balls. Since this notion of dimensionality depends only on the covering numbers and does not assume the existence of a smooth correspondence to a smaller dimensional Euclidean space, this notion not only incorporates smooth manifolds but also covers highly non-smooth sets such a fractals. In the literature, Kolmogorov & Tikhomirov (1961) provided a comprehensive study on the dependence of the covering number of different function classes on the underlying Minkowski dimension of the support. Nakada & Imaizumi (2020) showed how deep learners can incorporate this low-dimensionality of the data that is also reflected in their convergence rates. Recently, Huang et al. (2022) showed that WGANs can also adapt to this low-dimensionality of the data. In particular they showed that when the data is independent and identically distributed from a distribution μ, for the GAN estimate for the density (denoted as $\hat{\mu}^{\mathrm{GAN}}$), $\|\mu - \hat{\mu}^{\mathrm{GAN}}\|_{\mathbb{H}^\beta}$ decays at a rate of $\mathcal{O}\left(n^{-\frac{\beta}{\overline{d}_{\mathbb{M}}+\delta}} \vee n^{-1/2} \log n\right)$, where $\mathbb{M}$ is the support of μ and $\delta > 0$ is a pre-fixed constant. In the following section, we attempt to understand whether this rate is optimal or not.

2 Theoretical Analysis

Suppose that $\mathbb{M} \subseteq [0,1]^d$ and let $\Pi_{\mathbb{M}}$ denote the set of all probability distributions on $\mathbb{M}$. We assume that one has access to n samples, $X_1, \ldots, X_n$, generated independently from $\mu \in \Pi_{\mathbb{M}}$. The goal is to understand how good any estimate of μ, based on the data, performs. We characterise this performance in terms of the β-Hölder IPM or the p-Wasserstein distance, i.e. for an estimate $\hat{\mu}$, its performance is measured as $\|\mu - \hat{\mu}\|_{\mathbb{H}^\beta}$ or $\mathbb{W}_p(\hat{\mu}, \mu)$. To characterise this notion of best performing estimator, researchers use the concept of minimax risk i.e. the risk of the best performing estimator that achieves the maximum risk with respect to all members in $\Pi_{\mathbb{M}}$. Formally, the minimax risk for the problem is given by,

$$\mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \Pi_{\mathbb{M}}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathbb{H}^\beta} \quad \text{or} \quad \mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \Pi_{\mathbb{M}}} \mathbb{E}_\mu \mathbb{W}_p(\hat{\mu}, \mu),$$

where the infimum is taken over all measurable estimates of μ, i.e. on $\{\hat{\mu} : (X_1, \ldots, X_n) \to \Pi_{[0,1]^d} : \hat{\mu} \text{ is measurable}\}$. Here, we write $\mathbb{E}_\mu$ to denote that the expectation is taken with respect to the joint distribution of $X_1, \ldots, X_n$, which are independently and identically distributed as μ. Theorem 7 states the main lower bound of this paper, which lower bounds $\mathfrak{M}_n$ in terms of the lower-Minkowski dimension of $\mathbb{M}$ and the number of samples n, when n is large. The proof of this result is given in Section 3.

Theorem 7 (Main Result)**.** *Suppose that $X_1, \ldots, X_n$ are i.i.d. μ and let $\delta \in (0, \underline{d}_{\mathbb{M}})$. Then there exists and $n_0 \in \mathbb{N}$ such that if $n \geq n_0$,*

$$\inf_{\hat{\mu}} \sup_{\mu \in \Pi_{\mathbb{M}}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathbb{H}^\beta} \gtrsim n^{-\frac{1}{\underline{d}_{\mathbb{M}}-\delta}} \vee n^{-1/2}, \tag{2}$$

where the infimum is taken over all measurable estimates of μ, based on the data, $X_1, \ldots, X_n$. Furthermore,

$$\inf_{\hat{\mu}} \sup_{\mu \in \Pi_{\mathbb{M}}} \mathbb{E}_\mu \mathbb{W}_p(\hat{\mu}, \mu) \gtrsim n^{-\frac{1}{\underline{d}_{\mathbb{M}}-\delta}} \vee n^{-\frac{1}{2p}}. \tag{3}$$

If $\mathbb{M}$ is Minkowski regular, from Theorem 7, we note that for any $\delta \in (0, d_{\mathbb{M}})$, we observe that,

$$\inf_{\hat{\mu}} \sup_{\mu \in \Pi_{\mathbb{M}}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathbb{H}^\beta} \gtrsim n^{-\frac{\beta}{d_{\mathbb{M}}-\delta}} \vee n^{-1/2}.$$

From the results derived by Huang et al. (2022), GANs can achieve a rate of convergence of $\mathcal{O}\left(n^{-\frac{\beta}{\bar{d}_{\mathbb{M}}+\delta}} \vee n^{-1/2}\log n\right)$, implying that GANs are almost optimal in learning distributions when the data is low-dimensional in the Minkowski sense, barring poly-log factors in the sample size.

It is important to note that the lower bound in (3) closely resembles the ones derived by Niles-Weed & Berthet (2022) for distributions with Besov densities. Furthermore, from Theorem 1 of Weed & Bach (2019), we note that the empirical distribution $\hat{\mu}_n$ scales as $\mathbb{E}W_p(\hat{\mu}_n, \mu) \precsim n^{-1/(d_p^*(\mu)+\delta)}$, where $d_p^*(\mu)$ denotes the p-upper Wasserstein dimension of μ. Since μ is supported on $\mathbb{M}$, by Proposition 2 of Weed & Bach (2019), $d_p^*(\mu) \leq \bar{d}_{\mathbb{M}}$, when $\underline{d}_{\mathbb{M}} \geq 2p$. Thus, $\mathbb{E}W_p(\hat{\mu}_n, \mu) \precsim n^{-1/(\bar{d}_{\mathbb{M}}+\delta)}$. Hence, when $\underline{d}_{\mathbb{M}} > 2p$, we can choose $\delta > 0$, such that, $\inf_{\hat{\mu}} \sup_{\mu\in\Pi_{\mathbb{M}}} \mathbb{E}_\mu \mathbb{W}_p(\hat{\mu}, \mu) \succsim n^{-1/(\underline{d}_{\mathbb{M}}-\delta)}$. Hence, in this case, the sample mean almost achieves this minimax optimal rate when $\underline{d}_{\mathbb{M}} > 2p$ and $\mathbb{M}$ is Minkowski regular.

We observe that the δ-term is an artifact of the definition of the Minkowski dimension. If $\mathcal{N}(\epsilon, \mathbb{M}, \ell_\infty) \gtrsim \epsilon^{-\underline{d}}$, for some $\underline{d} \in (0, d]$, i.e. when the limit in the computation of the lower Minkowski dimension can be achieved exactly, following the proof of Theorem 7 (see Section 3), we note that under the same assumptions $\mathfrak{M}_n \gtrsim n^{-\beta/\underline{d}} \vee n^{-1/2}$, for n large. The δ-term in the lower bound is only an artefact of the definition of the lower Minkowski dimension and can be removed by assuming the lower bound for the covering number.

Inference for distributions supported on a Manifold When the support is regular, one can say that the minimax rate for estimating distributions decay at a rate whose exponent is inversely proportional to its regularity dimension. In particular, it is well known that We recall that we call a set $\mathbb{M}$ is $\tilde{d}$-regular w.r.t. the $\tilde{d}$-dimensional Hausdorff measure $\mathscr{H}^{\tilde{d}}$ if

$$\mathscr{H}^{\tilde{d}}(B_\varrho(x, r)) \asymp r^{\tilde{d}},$$

for all $x \in \mathbb{M}$ (see Definition 6 of Weed & Bach (2019)). Recall that the d-Hausdroff measure of a set S is defined as,

$$\mathscr{H}^d(S) := \liminf_{\epsilon\downarrow 0} \left\{ \sum_{k=1}^{\infty} r_k^d : S \subseteq \sum_{k=1}^{\infty} B_\varrho(x_k, r_k), r_k \leq \epsilon, \forall k \right\}.$$

It is known (Mattila, 1999) that if $\mathbb{M}$ is $\tilde{d}$-regular, then $d_{\mathbb{M}} = \tilde{d}$. Thus, when $\mathbb{M}$ is $\tilde{d}$-regular, the minimax rate roughly scales at $\Omega(n^{-\beta/\tilde{d}})$. Since, compact $\tilde{d}$-dimensional differentiable manifolds are $\tilde{d}$-regular (Weed & Bach, 2019, Proposition 9), this implies that for when $\mathbb{M}$ is a compact differentiable $\tilde{d}$-dimensional manifold, the error rates scale as $\Omega(n^{-\beta/\tilde{d}})$. This result underscores that GANs are nearly minimax optimal, given that the corresponding upper bounds derived by Dahal et al. (2022) match this minimax rate, albeit under some additional assumptions regarding the smoothness of the manifold. A similar result holds when $\mathbb{M}$ is a nonempty, compact convex set spanned by an affine space of dimension $\tilde{d}$; the relative boundary of a nonempty, compact convex set of dimension $\tilde{d}+1$; or a self-similar set with similarity dimension $\tilde{d}$ as all these sets are $\tilde{d}$-regular by (Weed & Bach, 2019, Proposition 9).

3 Proof of the Main Result (Theorem 7)

As a first step for deriving a minimax bound, we first show that the Hölder IPM can be lower bounded by the total variation distance and the minimum separation of the support of the distributions. For any finite set, we use the notation, $\text{sep}(\Xi) = \inf_{\xi,\xi'\in\Xi:\xi\neq\xi'} \|\xi - \xi'\|_\infty$.

Lemma 8. *Let Ξ be a finite subset of $\mathbb{R}^p$ and let, $P, Q \in \Pi_\Xi$. Then, we can find a constant π_1 (that might depend on β) such that, $\|P - Q\|_{\mathbb{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \geq \pi_1(sep(\Xi))^\beta \|P - Q\|_{TV}$.*

Proof. Let $b(x) = \exp\left(\frac{1}{x^2-1}\right) \mathbb{1}\{|x| \leq 1\}$ be the standard bump function on $\mathbb{R}$. For any $x \in \mathbb{R}^d$ and $\delta \in (0, 1]$, we let, $h_\delta(x) = a\delta^\beta \prod_{j=1}^d b(x_j/\delta)$. Here a is such that $ab(x) \in \mathbb{H}^\beta(\mathbb{R}, \mathbb{R}, C)$. It is easy to observe that $h_\delta \in \mathbb{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$. Let P and Q be two distributions on $\Xi = \{\xi_1, \ldots, \xi_k\}$. Let $\delta = \frac{1}{3}\min_{i\neq j} \|\xi_i - \xi_j\|_\infty$. We define $h^\star(x) = \sum_{i=1}^k \alpha_i h_\delta(x - \xi_i)$, with $\alpha_i \in \{-1, +1\}$, to be chosen later. Since the individual terms in $h^\star$ are

members of $\mathbb{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$ and have disjoint supports, $h^\star \in \mathbb{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$. We take $\alpha_i = 2\mathbb{1}\left(P(\xi) \geq Q(\xi)\right) - 1$. Thus,

$$\begin{aligned}\|P - Q\|_{\mathbb{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \geq \int h^* dP - \int h^* dQ = \sum_{i=1}^k a\delta^\beta \alpha_i (P(\xi_i) - Q(\xi_i)) =& a\delta^\beta \sum_{i=1}^k |P(\xi_i) - Q(\xi_i)| \\ =& 2a\delta^\beta \|P - Q\|_{\mathrm{TV}} \\ =& 2a(\mathrm{sep}(\Xi))^\beta \|P - Q\|_{\mathrm{TV}}.\end{aligned}$$

Taking $\pi_1 = 2a$ gives us the desired result. □

Similar to Lemma 8, we also show that on a discrete space, the p-Wasserstein metric is lower bounded by the total variation distance.

Lemma 9. *Let Ξ be a finite subset of $\mathbb{R}^p$ and let, $P, Q \in \Pi_\Xi$. Then, $\mathbb{W}_p(P, Q) \geq \mathit{sep}(\Xi)\|P - Q\|_{TV}^{1/p}$.*

Proof. Let $X \sim P$ and $Y \sim Q$. We note that $\|X - Y\|_2 \geq \mathbb{1}\{X \neq Y\}\,\mathrm{sep}(\Xi)$. Thus,

$$(\mathbb{E}\|X - Y\|_2^p)^{1/p} \geq (\mathbb{P}(X \neq Y))^{1/p}\,\mathrm{sep}(\Xi).$$

Taking infimum w.r.t. all measure couples between P and Q gives us the desired result. □

With the above two lemmas at our disposal, we are now ready to prove the main result of this paper. Recall that if $P \ll Q$, the KL-divergence between P and Q is given by, $\mathrm{KL}(P\|Q) = \int \log(dP/dQ)dP$. Similarly, the χ^2-divergence is given by, $\chi^2(P\|Q) = \int (dP/dQ)^2 - 1$.

3.1 Proof of Theorem 7

With Lemmas 8 and 9, we are now ready to prove Theorem 7. We use Fano's method to obtain the minimax lower bound. We refer the reader to Chapter 15 of Wainwright (2019) for a detailed exposition. Let, $s = \underline{d}_{\mathbb{M}} - \delta$. Thus, we can find $\epsilon_0 \in (0, 1)$, such that if $\epsilon \in (0, \epsilon_0]$, $\mathcal{N}(\epsilon, \mathbb{M}, \ell_\infty) \geq \epsilon^{-s} \implies \mathcal{M}(\epsilon, \mathbb{M}, \ell_\infty) \geq \epsilon^{-s}$. We take $n \geq n_0 = (128(\epsilon_0)^{-s}) \vee 8192$. Suppose $\epsilon = (n/128)^{-1/s}$. Let $\Theta = \{\theta_1, \ldots, \theta_k\}$ be a ϵ-separated set in $\mathbb{M}$. For the above choices of n and ϵ, we observe that, $k = \epsilon^{-s} = n/128 \geq 64$ and $n \geq 64k$.

Let $\phi_j(x) = \mathbb{1}\{x = \theta_j\} - \mathbb{1}\{x = \theta_{\lfloor k/2 \rfloor + j}\}$, for all $j = 1, \ldots, \lfloor k/2 \rfloor$. Let, $\boldsymbol{\omega} \in \{0, 1\}^k$. We define the probability mass function on Θ,

$$P_{\boldsymbol{\omega}}(x) = \frac{1}{k} + \frac{\delta_k}{k} \sum_{j=1}^{\lfloor k/2 \rfloor} \omega_j \phi_j(x),$$

with $\delta_k \in (0, 1/2]$. By construction, $P_{\boldsymbol{\omega}} \in \Pi_{\mathbb{M}}$.

Furthermore,

$$\|P_{\boldsymbol{\omega}} - P_{\boldsymbol{\omega}'}\|_{\mathrm{TV}} = \frac{\delta_k}{k}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1.$$

By the Varshamov-Gilbert bound (Tsybakov, 2009, Lemma 2.9), let $\Omega \subseteq \{0, 1\}^{\lfloor k/2 \rfloor}$ be such that $|\Omega| \geq 2^{\frac{1}{8}\lfloor k/2 \rfloor}$ and $\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1 \geq \frac{1}{8}\lfloor k/2 \rfloor$, for all $\boldsymbol{\omega} \neq \boldsymbol{\omega}'$ both in Ω. Thus for any $\boldsymbol{\omega} \neq \boldsymbol{\omega}'$, both in Ω,

$$\|P_{\boldsymbol{\omega}} - P_{\boldsymbol{\omega}'}\|_{\mathrm{TV}} \geq \frac{\delta_k \lfloor k/2 \rfloor}{8k}. \tag{4}$$

Hence, by Lemma 8, $\|P_{\boldsymbol{\omega}} - P_{\boldsymbol{\omega}'}\|_{\mathbb{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \geq \pi_1 \epsilon^\beta \frac{\delta_k \lfloor k/2 \rfloor}{k}$. Similarly, by Lemma 9, we note that $\mathbb{W}_p(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}'}) \geq \epsilon \left(\frac{\delta_k \lfloor k/2 \rfloor}{k}\right)^{1/p}$. Furthermore, we observe that

$$\mathrm{KL}(P_{\boldsymbol{\omega}}^{\otimes n} \| P_{\boldsymbol{\omega}'}^{\otimes n}) = n\,\mathrm{KL}(P_{\boldsymbol{\omega}} \| P_{\boldsymbol{\omega}'}) \leq n\,\chi^2(P_{\boldsymbol{\omega}} \| P_{\boldsymbol{\omega}'}) = n \sum_{i=1}^k \frac{(P_{\boldsymbol{\omega}}(\xi_i) - P_{\boldsymbol{\omega}'(\xi_i)})^2}{P_{\boldsymbol{\omega}}(\xi_i)} \leq 2nk \sum_{i=1}^k (P_{\boldsymbol{\omega}}(\xi_i) - P_{\boldsymbol{\omega}'}(\xi_i)^2$$

$$\begin{aligned}&\leq 2nk\lfloor k/2\rfloor(2\delta_k/k)^2\\&=8\frac{n\lfloor k/2\rfloor\delta_k^2}{k}.\end{aligned}$$

Thus,

$$\frac{1}{|\Omega|^2}\sum_{\boldsymbol{\omega},\boldsymbol{\omega}'\in\Omega}\mathrm{KL}(P_{\boldsymbol{\omega}}^{\otimes_n}\|P_{\boldsymbol{\omega}'}^{\otimes_n})\leq 8\frac{n\lfloor k/2\rfloor\delta_k^2}{k}.$$

Let $\mathcal{P}=\{P_{\boldsymbol{\omega}}:\boldsymbol{\omega}\in\Omega\}$. Let $J\sim\mathrm{Unif}(\Omega)$ and $Z|J=\boldsymbol{\omega}\sim P_\omega$. By the convexity of KL divergence (see equation 15.34 of Wainwright (2019)), we know that,

$$I(Z;J)\leq\frac{1}{|\Omega|^2}\sum_{\boldsymbol{\omega},\boldsymbol{\omega}'\in\Omega}\mathrm{KL}(P_{\boldsymbol{\omega}}^{\otimes_n}\|P_{\boldsymbol{\omega}'}^{\otimes_n})\leq 8\frac{n\lfloor k/2\rfloor\delta_k^2}{k}.$$

Thus,

$$\frac{I(Z;J)+\log 2}{\log|\Omega|}\leq 8\frac{I(Z;J)+\log 2}{\lfloor k/2\rfloor\log 2}\leq 64\frac{n\delta_k^2}{k\log 2}+\frac{8}{\lfloor k/2\rfloor}\leq 64\frac{n\delta_k^2}{k\log 2}+\frac{1}{4}. \tag{5}$$

The last inequality follows since $k\geq 64$. We take $\delta_k=\frac{1}{16}\sqrt{\frac{k\log 2}{n}}$. Clearly, $\epsilon\leq 1/2$ as $n\geq 64k$. This choice of ϵ makes,

$$\frac{I(Z;J)+\log 2}{\log|\Omega|}\leq\frac{1}{2}.$$

Thus, by Theorem 15.2 of Wainwright (2019),

$$\begin{aligned}\inf_{\hat{\mu}}\sup_{\mu\in\Pi_{\mathbb{M}}}\mathbb{E}_\mu\|\hat{\mu}-\mu\|_{\mathbb{H}^\beta}\geq\pi_1\epsilon^\beta\frac{\delta_k\lfloor k/2\rfloor}{k}=\pi_1\epsilon^\beta\frac{\lfloor k/2\rfloor}{k}\frac{1}{16}\sqrt{\frac{k\log 2}{n}}&\geq\pi\epsilon^\beta\frac{\lfloor k/2\rfloor}{k}\frac{1}{16}\sqrt{\frac{\log 2}{128}}\\&\geq\pi_2\epsilon^\beta\asymp n^{-\beta/s}.\end{aligned} \tag{6}$$

$$\text{Similarly,}\quad\inf_{\hat{\mu}}\sup_{\mu\in\Pi_{\mathbb{M}}}\mathbb{E}_\mu\mathbb{W}_p(\hat{\mu},\mu)\geq\epsilon\left(\frac{\delta_k\lfloor k/2\rfloor}{k}\right)^{1/p}\gtrsim n^{-1/s}. \tag{7}$$

To show that $\inf_{\hat{\mu}}\sup_{\mu\in\Pi_{\mathbb{M}}}\|\hat{\mu}-\mu\|_{\mathbb{H}^\beta}\gtrsim n^{-1/2}$, we use Le Cam's method (Wainwright, 2019, Chapter 15.2). Let $\theta_0,\theta_1\in\mathbb{M}$ be such that $\|\theta_0-\theta_1\|_\infty\geq\mathrm{diam}(\mathbb{M})/2$. Let $P_0(\theta_0)=P_0(\theta_1)=1/2$ and $P_1(\theta_0)=1-P_1(\theta_1)=1/2-\delta$ with $\delta\in(0,1/4)$. Clearly, $\mathrm{TV}(P_0,P_1)=\delta$. Thus, by Lemma 8, we observe that

$$\|P_1-P_0\|_{\mathbb{H}^\beta}\gtrsim(\mathrm{diam}(\mathbb{M})/2)^\beta\delta\gtrsim\delta.$$

Similarly, $\mathbb{W}_p(P_1,P_0)\geq(\mathrm{diam}(\mathbb{M})/2)\delta^{1/p}\gtrsim\delta^{1/p}$. Again,

$$\mathrm{KL}(P_1^{\otimes_n}\|P_0^{\otimes_n})=n\,\mathrm{KL}(P_1\|P_0)\leq n\,\chi^2(P_1\|P_0)=4n\delta^2.$$

By Pinsker's inequality (Tsybakov, 2009, Lemma 2.5), we note that,

$$\mathrm{TV}(P_1^{\otimes_n},P_0^{\otimes_n})\leq\sqrt{\frac{1}{2}\,\mathrm{KL}(P_1^{\otimes_n}\|P_0^{\otimes_n})}=2\delta\sqrt{n}=1/4,$$

if $\delta=\frac{1}{8\sqrt{n}}$. Thus from equation 15.14 of Wainwright (2019), we observe that,

$$\inf_{\hat{\mu}}\sup_{\mu\in\Pi_{\mathbb{M}}}\mathbb{E}_\mu\|\hat{\mu}-\mu\|_{\mathbb{H}^\beta}\gtrsim\delta\asymp 1/\sqrt{n}. \tag{8}$$

$$\text{Similarly,}\quad\inf_{\hat{\mu}}\sup_{\mu\in\Pi_{\mathbb{M}}}\mathbb{E}_\mu\mathbb{W}_p(\hat{\mu},\mu)\gtrsim\delta^{1/p}\asymp n^{-\frac{1}{2p}}. \tag{9}$$

The result now follows from combining (6) and (8). The minimax rates for the Wasserstein distance follows from combining (7) and (9).

4 Conclusion

In this paper, we aimed to study the fundamental question of whether GANs are optimal in providing accurate estimates of target distributions, especially when the data exhibit a low-dimensional structure. We charecterise this notion of low-dimensionality through the so called Minkowski dimension. We have demonstrated that, in scenarios where n independent and identically distributed samples are available from a target distribution on a set $\mathbb{M}$, the convergence rate for any estimator is bounded by $\Omega\left(n^{-\frac{\beta}{d_{\mathbb{M}}-\delta}} \vee n^{-1/2}\right)$ in the β-Hölder IPM and $\Omega\left(n^{-\frac{1}{d_{\mathbb{M}}-\delta}} \vee n^{-1/(2p)}\right)$ in the Wasserstein p-metric. When the support is regular in the Minkowski sense, the convergence rates for GANs closely resemble this lower bound, when the the network are properly chosen. Some future results in this direction might render fruitful avenues understanding similar lower bounds especially with a different notion of the dimensionality of the data distribution, such as the Wasserstein dimension (Weed & Bach, 2019).

References

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*, pp. 224–232. PMLR, 2017.

Sanjeev Arora, Andrej Risteski, and Yi Zhang. Do GANs learn the distribution? some theory and empirics. In *International Conference on Learning Representations*, 2018. URL `https://openreview.net/forum?id=BJehNfW0-`.

Hayk Asatryan, Hanno Gottschalk, Marieke Lippert, and Matthias Rottmann. A convenient infinite dimensional framework for generative adversarial learning. *Electronic Journal of Statistics*, 17(1):391 – 428, 2023. doi: 10.1214/23-EJS2104. URL `https://doi.org/10.1214/23-EJS2104`.

Denis Belomestny, Eric Moulines, Alexey Naumov, Nikita Puchkin, and Sergey Samsonov. Rates of convergence for density estimation with gans. *arXiv preprint arXiv:2102.00199*, 2021.

P K Bhattacharya and Y P Mack. Weak convergence of k-nn density and regression estimators with varying k and applications. *Annals of Statistics*, 15(3):976–994, 1987.

Gérard Biau, Benoît Cadre, Maxime Sangnier, and Ugo Tanielian. Some theoretical properties of gans. *The Annals of Statistics*, 48(3):1539–1566, 2020.

Gérard Biau, Maxime Sangnier, and Ugo Tanielian. Some theoretical insights into wasserstein gans. *Journal of Machine Learning Research*, 22(1):5287–5331, 2021.

Peter J Bickel and Bo Li. Local polynomial regression on unknown manifolds. *Lecture Notes-Monograph Series*, pp. 177–186, 2007.

Saptarshi Chakraborty, Debolina Paul, and Swagatam Das. Automated clustering of high-dimensional data with a feature weighted mean shift algorithm. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 6930–6938, 2021.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for the cluster tree. *Advances in neural information processing systems*, 23, 2010.

Probal Chaudhuri, Anil K Ghosh, and Hannu Oja. Classification based on hybridization of parametric and nonparametric classifiers. *IEEE transactions on pattern analysis and machine intelligence*, 31(7): 1153–1164, 2008.

Biraj Dahal, Alexander Havrilla, Minshuo Chen, Tuo Zhao, and Wenjing Liao. On deep generative models for approximation and estimation of distributions on manifolds. *Advances in Neural Information Processing Systems*, 35:10615–10628, 2022.

Luc P Devroye and Terry J Wagner. The strong uniform consistency of nearest neighbor density estimates. *The Annals of Statistics*, pp. 536–540, 1977.

David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Density estimation by wavelet thresholding. *The Annals of statistics*, pp. 508–539, 1996.

Arnaud Doucet, Nando De Freitas, and Neil Gordon. An introduction to sequential monte carlo methods. *Sequential Monte Carlo methods in practice*, pp. 3–14, 2001.

Kenneth Falconer. *Fractal geometry: mathematical foundations and applications*. John Wiley & Sons, 2004.

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*, volume 27. Curran Associates, Inc., 2014.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. *Journal of Machine Learning Research*, 23(116):1–43, 2022. URL `http://jmlr.org/papers/v23/21-0732.html`.

Jisu Kim, Jaehyeok Shin, Alessandro Rinaldo, and Larry Wasserman. Uniform convergence rate of the kernel density estimator adaptive to intrinsic volume dimension. In *International Conference on Machine Learning*, pp. 3398–3407. PMLR, 2019.

Benoît R Kloeckner. Empirical measures: regularity is a counter-curse to dimensionality. *ESAIM: Probability and Statistics*, 24:408–434, 2020.

Andrey N Kolmogorov and Vladimir Mikhailovich Tikhomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *Translations of the American Mathematical Society*, 17:277–364, 1961.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Society, 2017.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(1):10366–10406, 2021.

Giulia Luise, Massimiliano Pontil, and Carlo Ciliberto. Generalization properties of optimal transport gans with latent distribution learning. *arXiv preprint arXiv:2007.14641*, 2020.

Pertti Mattila. *Geometry of sets and measures in Euclidean spaces: fractals and rectifiability*. Number 44. Cambridge University Press, 1999.

Daniel McDonald. Minimax density estimation for growing dimension. In *Artificial Intelligence and Statistics*, pp. 194–203. PMLR, 2017.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *Journal of Machine Learning Research*, 21(174):1–38, 2020. URL `http://jmlr.org/papers/v21/20-002.html`.

Jonathan Niles-Weed and Quentin Berthet. Minimax estimation of smooth densities in Wasserstein distance. *The Annals of Statistics*, 50(3):1519 – 1540, 2022. doi: 10.1214/21-AOS2161. URL `https://doi.org/10.1214/21-AOS2161`.

Emanuel Parzen. On estimation of a probability density function and mode. *The annals of mathematical statistics*, 33(3):1065–1076, 1962.

Phil Pope, Chen Zhu, Ahmed Abdelkader, Micah Goldblum, and Tom Goldstein. The intrinsic dimension of images and its impact on learning. In *International Conference on Learning Representations*, 2020.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8(7), 2007.

Alessandro Rinaldo and Larry Wasserman. Generalized density clustering. *Annals of Statistics*, 38(5): 2678–2722, 2010.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Algorithmic Learning Theory*, pp. 1051–1071. PMLR, 2021.

Shashank Singh and Barnabás Póczos. Minimax distribution estimation in wasserstein distance. *arXiv preprint arXiv:1802.08855*, 2018.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *The Annals of Statistics*, 51(3):1282–1308, 2023.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, Springer New York, NY, 1 edition, 2009. ISBN 978-0-387-79051-0. doi: 10.1007/b13794. URL `https://doi.org/10.1007/b13794`. Published: 26 November 2008.

Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric density estimation & convergence rates for gans under besov ipm losses. *Advances in neural information processing systems*, 32, 2019.

Martin J Wainwright. *High-dimensional statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Jonathan Weed and Francis Bach. Sharp Asymptotic and Finite-Sample Rates of Convergence of Empirical Measures in Wasserstein Distance. *Bernoulli*, 25(4A):2620–2648, 2019. doi: 10.3150/18-BEJ1065. URL `https://doi.org/10.3150/18-BEJ1065`.

Puning Zhao and Lifeng Lai. Analysis of knn density estimation. *IEEE Transactions on Information Theory*, 68(12):7971–7995, 2022.