Journal of Machine Learning Research 26 (2025) 1-57 Submitted 1/24; Revised 5/25; Published 5/25

On the Statistical Properties of Generative Adversarial Models for Low Intrinsic Data Dimension

Saptarshi Chakraborty SAPTARSHIC@BERKELEY.EDU
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Peter L. Bartlett PETER@BERKELEY.EDU
Google DeepMind;
Department of Statistics,
Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley
Berkeley, CA 94720-1776, USA

Editor: Joan Bruna

Abstract

Despite the remarkable empirical successes of Generative Adversarial Networks (GANs), the theoretical guarantees for their statistical accuracy remain rather pessimistic. In particular, the data distributions on which GANs are applied, such as natural images, are often hypothesized to have an intrinsic low-dimensional structure in a typically high-dimensional feature space, but this is often not reflected in the derived rates in the state-of-the-art analyses. In this paper, we attempt to bridge the gap between the theory and practice of GANs and their bidirectional variant, Bi-directional GANs (BiGANs), by deriving statistical guarantees on the estimated densities in terms of the intrinsic dimension of the data and the latent space. We analytically show that if one has access to n samples from the unknown target distribution and the network architectures are properly chosen, the expected Wasserstein-1 distance of the estimates from the target scales as $\mathcal{O}\left(n^{-1/d_\mu}\right)$ for GANs and $\tilde{\mathcal{O}}\left(n^{-1/(d_\mu+\ell)}\right)$ for BiGANs, where d_μ and ℓ are the upper Wasserstein-1 dimension of the data-distribution and latent-space dimension, respectively. The theoretical analyses not only suggest that these methods successfully avoid the curse of dimensionality, in the sense that the exponent of n in the error rates does not depend on the data dimension but also serve to bridge the gap between the theoretical analyses of GANs and the known sharp rates from optimal transport literature. Additionally, we demonstrate that GANs can effectively achieve the minimax optimal rate even for non-smooth underlying distributions, with the use of interpolating generator networks.

Keywords: Generative Adversarial Network, Intrinsic Dimension, Wasserstein Dimension, Convergence Rates, ReLU network approximation

1. Introduction

The problem of generating new data from an unknown distribution by observing independent and identically distributed samples from the same has been of great importance to researchers and finds fruitful applications in the fields of statistics, computer vision, biomedical imaging, astronomy, and so on (Alqahtani et al., 2021). Recent developments in deep learning have led to the discovery of Generative Adversarial Networks (GANs) (Goodfellow et al., 2014), which formulates the problem as a (often zero-sum) game between two adversaries, called the generator and the discriminator. GANs have been incredibly successful, especially in the field of computer vision in generating realistic and high-resolution samples resembling natural images. The generator takes input from a relatively low-dimensional white noise, e.g., normal or uniform distribution, and tries to output realistic examples from the target distribution, while the discriminator tries to differentiate between real and fake samples. Both the generator and discriminator are typically realized by some class of deep networks.

Different variants of GANs have been proposed in the past few years, often by varying the underlying divergence measure used and involving additional architectures, to learn different aspects of the data distribution. For example, f-GANs (Nowozin et al., 2016) generalize vanilla GANs by incorporating a f-divergence as a dissimilarity measure. Wasserstein GANs (WGANs)(Arjovsky et al., 2017) incorporate the first-order Wasserstein distance (also known as Kantorovich distance) to deduce a better-behaved GAN-objective. Some popular variants of GAN include MMD-GAN (Dziugaite et al., 2015), LSGAN (Mao et al., 2017), Cycle-GAN (Zhu et al., 2017), DualGAN (Yi et al., 2017), DiscoGAN (Kim et al., 2017) etc. Recent progress in GANs has allowed us to learn not only the map from the latent space to the data space (the generator) but the reverse map as well. This map called the encoder, is useful in finding a proper low-dimensional representation of the samples. Bi-directional GANs (BiGANs) (Donahue et al., 2017) implement an adversarial learning scheme with the discriminator trying to distinguish between the fake data-latent code pair and its real counterpart (see Section 2 for a detailed exposition). Ideally, during the training process, the encoder learns a useful feature representation of the data in the latent space for auxiliary machine-learning tasks and data visualization.

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. (2020) and Belomestny et al. (2021) improved upon the works of Biau et al. (2020) to understand the behavior of GANs for Hölder class of density functions. Arora et al. (2017) showed that generalization 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. In contrast to GANs, the theoretical understanding of BiGANs remains rather limited. While Liu et al. (2021) attempted to establish generalization bounds for the BiGAN problem, it also suffers from the curse of dimensionality as the rates depend on the nominal high-dimensionality of the entire data space.

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 generalization 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 hypothesized that natural images lie on a low-dimensional structure, despite 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 and Li, 2007; Kim et al., 2019), such results are largely unexplored in the context of GANs. Recently, Huang et al. (2022) expressed the generalization 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 Minkowski dimension, although useful, does not fully capture the low-dimensional nature of the underlying distribution (see Section 3), while 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 in (Dahal et al., 2022) is a very strong assumption that might not hold in practice. Furthermore, none of the aforementioned approaches tackle the problem in its full generality and match the sharp convergence rates of the empirical distributions in the optimal transport literature (Weed and Bach, 2019). Additionally, it remains unknown whether the GAN estimates of the target distribution are optimal in the minimax sense.

Contributions In an attempt to overcome the aforementioned drawbacks in the current literature, the major findings of the paper are highlighted below.

- In order to bridge the gap between the theory and practice of such generative models, in this paper, we develop a framework to establish the statistical convergence rates of GANs and BiGANs in terms of the upper Wasserstein dimension (Weed and Bach, 2019) of the underlying target measure.

- Informally, our results guarantee that if one has access to n independent and identically distributed samples (i.i.d.) from μ and if the network architectures are properly chosen, the expected β-Hölder Integral Probability Metric (IPM) between the estimated and target distribution for GANs scales at most at a rate of $\mathbb{E}\|\mu - \hat{G}_\sharp \nu\|_{\mathcal{H}^\beta} \lesssim n^{-\beta/d^*_\beta(\mu)}$. Here, $d^*_\beta(\mu)$ is the β-upper Wasserstein dimension of the target distribution μ (see Definition 4). The recent statistical guarantees of GANs (Huang et al., 2022; Dahal et al., 2022) follow as a direct corollary of our main result. Similarly, for BiGANs, we can guarantee that the expected Wasserstein-1 distance between the estimated joint distributions scales roughly at a rate (ignoring the log-factors) of $\mathbb{E}\mathcal{W}_1\left((\mu, \hat{E}_\sharp \mu), (\hat{D}_\sharp \nu, \nu)\right) \precsim n^{-1/(d^*_1(\mu)+\ell)}$. Here $\hat{E}$ and $\hat{D}$ are the optimal sample encoders and decoders and ℓ is the dimension of the latent space. Notably, when the support encompasses the entire dataspace, this outcome aligns with the findings of Liu et al. (2021).

- We introduce the entropic dimension, characterizing a distribution's intrinsic dimensionality, and demonstrate its relevance in deep learning theory, notably in approximation and generalization bounds. Our investigation into the approximation capabilities of ReLU networks reveals that achieving ϵ-approximation in the $\mathbb{L}_p(\mu)$-norm for α-Hölder functions demands roughly $\mathcal{O}(\epsilon^{-\bar{d}_{\alpha p}(\mu)/\alpha})$ weight terms (refer to Theorem 21). This contrasts with prior estimates of $\mathcal{O}(\epsilon^{-d/\alpha})$ (Yarotsky, 2017) or $\mathcal{O}(\epsilon^{-\overline{\dim}_M(\mu)/\alpha})$ (Nakada and Imaizumi, 2020). Here $\bar{d}_{p\alpha}(\mu) \le \overline{\dim}_M(\mu) \le d$ denote the $p\alpha$-entropic dimension, upper Minkowski dimension of μ and the dimension of the data space, respectively (see Section 3). The result implies sustained approximation capabilities even with smaller ReLU networks. Furthermore, we demonstrate that the $\mathbb{L}_p(\mu)$-metric entropy of β-Hölder functions scales at most at a rate of $\mathcal{O}(\epsilon^{-\bar{d}_{p\beta}(\mu)})$, enhancing the foundational results of Kolmogorov and Tikhomirov (1961).
- Finally, we derive minimax optimal rates for estimating under the α-Hölder IPM, establishing that the GAN estimates can approximately achieve this minimax optimal rate.

Organization The remaining sections are structured as follows: Section 2 revisits necessary notations, and definitions, and outlines the problem statement. In Section 3, we revisit the concept of intrinsic dimension and introduce a novel dimension termed the entropic dimension of a measure, comparing it with commonly used metrics. This entropic dimension proves pivotal in characterizing both the $\mathbb{L}_p$-covering number of Hölder functions (refer to Theorem 13). We show that the Wasserstein dimension determines the convergence rate of the empirical measure to the population in the Hölder Integral IPM in Theorem 14. The subsequent focus shifts to theoretical analyses of GANs and BiGANs in Section 4. We begin by presenting the assumptions in Section 4.1, followed by stating the main result in Section 4.2 and providing a proof sketch in Section 4.3, with detailed proofs available in the Appendix. Section 5 demonstrates that GANs can achieve the minimax optimal rates for estimating distributions, followed by concluding remarks in Section 6.

2. Background

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

Notations We use the notation $x \vee y := \max\{x, y\}$ and $x \wedge y := \min\{x, y\}$. $T_\sharp\mu$ denotes the push-forward of the measure μ by the map T. $B_\varrho(x, r)$ 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, $\text{supp}(\gamma) = \{x : \gamma(B_\varrho(x, r)) > 0, \text{ for all } r > 0\}$. For any function $f : \mathcal{S} \to \mathbb{R}$, and any measure γ on $\mathcal{S}$, let $\|f\|_{\mathbb{L}_p(\gamma)} := \left(\int_{\mathcal{S}} |f(x)|^p d\gamma(x)\right)^{1/p}$, if $0 < p < \infty$. Also let, $\|f\|_{\mathbb{L}_\infty(\gamma)} := \operatorname{ess\,sup}_{x \in \text{supp}(\gamma)} |f(x)|$. For any function class $\mathcal{F}$, and distributions P and Q, $\|P - Q\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\int f dP - \int f dQ|$. For function classes $\mathcal{F}_1$ and $\mathcal{F}_2$, $\mathcal{F}_1 \circ \mathcal{F}_2 = \{f_1 \circ f_2 : f_1 \in \mathcal{F}_1,\, f_2 \in \mathcal{F}_2\}$. We say $A_n \lesssim B_n$ (also written as $A_n = \mathcal{O}(B_n)$) if there exists $C > 0$, independent of n, such that $A_n \le C B_n$. Similarly, the notation, "$\precsim$" (also written as $A_n = \tilde{\mathcal{O}}(B_n)$) ignores poly-log factors in n. We say $A_n \asymp B_n$, if $A_n \lesssim B_n$ and $B_n \lesssim A_n$. For any $k \in \mathbb{N}$, we let $[k] = \{1, \dots, k\}$. For two random variables X and Y, we say that

$X \stackrel{d}{=} Y$, if the random variables have the same distribution. We use bold lowercase letters to denote members of $\mathbb{N}^k$ for $k \in \mathbb{N}$. $\Pi_{\mathcal{A}}$ denotes the set of all probability measures on the set $\mathcal{A}$. $\mathcal{W}_p(\cdot, \cdot)$ denotes the Wasserstein p-distance between probability distributions. $\mu \otimes \nu$ denotes the product distribution of μ and ν.

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

Definition 2 (Neural networks)**.** Let $L \in \mathbb{N}$ and $\{N_i\}_{i \in [L]} \in \mathbb{N}$. Then a L-layer neural network $f : \mathbb{R}^d \to \mathbb{R}^{N_L}$ is defined as,

$$f(x) = A_L \circ \sigma_{L-1} \circ A_{L-1} \circ \dots \circ \sigma_1 \circ A_1(x) \tag{1}$$

Here, $A_i(y) = W_i y + b_i$, with $W_i \in \mathbb{R}^{N_i \times N_{i-1}}$ and $b_i \in \mathbb{R}^{N_{i-1}}$, with $N_0 = d$. Note that σ_j is applied component-wise. Here, $\{W_i\}_{1 \leq i \leq L}$ are known as weights, and $\{b_i\}_{1 \leq i \leq L}$ are known as biases. $\{\sigma_i\}_{1 \leq i \leq L-1}$ are known as the activation functions. Without loss of generality, one can take $\sigma_\ell(0) = 0, \forall\, \ell \in [L-1]$. We define the following quantities: (Depth) $\mathcal{L}(f) := L$ is known as the depth of the network; (Number of weights) the number of weights of the network f is denoted as $\mathcal{W}(f) = \sum_{i=1}^L N_i N_{i-1}$; (maximum weight) $\mathcal{B}(f) = \max_{1 \leq j \leq L}(\|b_j\|_\infty) \vee \|W_j\|_\infty$ to denote the maximum absolute value of the weights and biases.

$$\mathcal{NN}_{\{\sigma_i\}_{1 \leq i \leq L-1}}(L, W, B, R) = \{f \text{ of the form (1)} : \mathcal{L}(f) \leq L,\, \mathcal{W}(f) \leq W, \mathcal{B}(f) \leq B, \\ \sup_{x \in \mathbb{R}^d} \|f(x)\|_\infty \leq R\}.$$

If $\sigma_j(x) = x \vee 0$, for all $j = 1, \dots, L-1$, we use the notation $\mathcal{RN}(L, W, B, R)$ to denote $\mathcal{NN}_{\{\sigma_i\}_{1 \leq i \leq L-1}}(L, W, B, R)$.

Definition 3 (Hölder functions)**.** Let $f : \mathcal{S} \to \mathbb{R}$ be a function, where $\mathcal{S} \subseteq \mathbb{R}^d$. For a multi-index $\boldsymbol{s} = (s_1, \dots, s_d)$, let, $\partial^{\boldsymbol{s}} f = \frac{\partial^{|\boldsymbol{s}|} f}{\partial x_1^{s_1} \dots \partial x_d^{s_d}}$, where, $|\boldsymbol{s}| = \sum_{\ell=1}^d s_\ell$. We say that a function $f : \mathcal{S} \to \mathbb{R}$ is β-Hölder (for $\beta > 0$) if

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

If $f : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$, then we define $\|f\|_{\mathcal{H}^\beta} = \sum_{j=1}^{d_2} \|f_j\|_{\mathcal{H}^\beta}$. For notational simplicity, let, $\mathcal{H}^\beta(\mathcal{S}_1, \mathcal{S}_2, C) = \{f : \mathcal{S}_1 \to \mathcal{S}_2 : \|f\|_{\mathcal{H}^\beta} \leq C\}$. Here, both $\mathcal{S}_1$ and $\mathcal{S}_2$ are subsets of real vector spaces.

2.1 Generative Adversarial Networks (GANs)

Generative adversarial networks or GANs provide a simple yet effective way to generate samples from an unknown data distribution, given i.i.d. samples from the same. Suppose that $\mathcal{X}$ is the data space and μ is a probability distribution on $\mathcal{X}$. For simplicity, we will assume that $\mathcal{X} = [0, 1]^d$. Let $\mathcal{Z}$ be the latent space, from which it is easy to generate samples.

We will take $\mathcal{Z} = [0,1]^\ell$. GANs view the problem by modeling a two-player (often zero-sum) game between the discriminator and the generator. The generator G maps elements of $\mathcal{Z}$ to $\mathcal{X}$. The generator generates fake data points by first simulating a point in $\mathcal{Z}$ from some distribution ν and passing it through the generator G. The discriminator's job is to distinguish between the original and fake samples. Wasserstein GAN's objective is to solve the min-max problem,

$$\inf_G \|\mu - G_\sharp \nu\|_\Phi,$$

where Φ is a class of discriminators.

One often does not have access to μ but, only samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mu$. Moreover one also assumes that one can generate samples, $Z_1, \dots, Z_m$ independently from ν. The empirical distributions of μ and ν are given by, $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\hat{\nu}_m = \frac{1}{n}\sum_{j=1}^m \delta_{Z_j}$, respectively. In practice, one typically solves the following empirical objectives.

$$\inf_G \sup_\phi \frac{1}{n}\sum_{i=1}^n \phi(X_i) - \mathbb{E}_{Z\sim\nu}\phi(G(Z)) \quad \text{or} \quad \inf_G \sup_\phi \frac{1}{n}\sum_{i=1}^n \phi(X_i) - \frac{1}{m}\sum_{j=1}^m \phi(G(Z_j)). \tag{2}$$

One typically realizes $\mathcal{G}$ through a class of ReLU networks. The empirical estimates of the generator are defined as:

$$\hat{G}_n = \operatorname*{argmin}_{G\in\mathcal{G}} \|\hat{\mu}_n - G_\sharp \nu\|_\Phi \quad \text{or} \quad \hat{G}_{n,m} = \operatorname*{argmin}_{G\in\mathcal{G}} \|\hat{\mu}_n - G_\sharp \hat{\nu}_m\|_\Phi. \tag{3}$$

Of course, in practice, one finds an approximation of $\hat{G}$ only up to an optimization error, which we ignore for the simplicity of exposition; however one can incorporate this error into the theoretical results that we derive.

To measure the efficacy of the above generators, one computes the closeness of the target distribution μ and the estimated distribution $G_\sharp\nu$. As a measure of comparison, one takes an IPM for some function class $\mathcal{F}$ and defines the excess risk of the estimator $\hat{G}$ as:

$$\mathfrak{R}_{\mathcal{F}}^{\text{GAN}}(\hat{G}) = \|\mu - \hat{G}_\sharp \nu\|_{\mathcal{F}}.$$

In what follows, we take $\mathcal{F} = \Phi = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$. It should be noted that while the discriminator is trained to differentiate real and fake samples, its role is limited to binary classification. As such, the discriminator network can be replaced by any efficient classifier, in principle. Neural networks are employed to approximate these Lipschitz or Hölder IPMs and enable efficient optimization for practical purposes.

2.2 Bidirectional GANs

For notational simplicity, in the BiGAN context, we refer to the generator as the decoder and denote it by $D : \mathcal{Z} \to \mathcal{X}$. In BiGANs, one also constructs an encoder $E : \mathcal{X} \to \mathcal{Z}$. The discriminator is modified to take values from both the latent space and the data space, i.e. $\psi : \mathcal{X} \times \mathcal{Z} \to [0,1]$. For notational simplicity, $(f_\sharp P, P)$ and $(P, f_\sharp P)$ denote the distributions of $(f(X), X)$ and $(X, f(X))$, respectively, where $X \sim P$. The discriminator in the BiGAN objective also aims to distinguish between the real and fake samples as much as possible, whereas the encoder-decoder pair tries to generate realistic fake data-latent samples to confuse the discriminator. The BiGAN problem is formulated as follows:

$$\min_{D,\, E} \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi, \tag{4}$$

where Ψ denotes the set of all discriminators. In practice, one typically has access only to samples $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \mu$ and $Z_1, \dots, Z_m \overset{\text{i.i.d.}}{\sim} \nu$, and these are used to solve one of the following empirical objectives.

$$\inf_{D,\, E} \sup_{\psi} \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, E(X_i)) - \mathbb{E}_{Z \sim \nu} \psi(D(Z), Z),$$

$$\inf_{D,\, E} \sup_{\psi} \frac{1}{n} \sum_{i=1}^{n} \psi(X_i, E(X_i)) - \frac{1}{m} \sum_{j=1}^{m} \psi(D(Z_j), Z_j).$$

The functions D, E, and ψ are typically realized by deep networks though it is not uncommon to model ψ through functions belonging to certain Hilbert spaces (Li et al., 2017).

We let $\mathcal{D}$ and $\mathcal{E}$ be the set of all decoders and encoders of interest, which we will take to be a class of ReLU networks. The empirical decoder and encoder pairs are defined as:

$$(\hat{D}_n, \hat{E}_n) = \operatorname*{argmin}_{D \in \mathcal{D}, E \in \mathcal{E}} \|(\hat{\mu}_n, E_\sharp \hat{\mu}_n) - (D_\sharp \nu, \nu)\|_\Psi \tag{5}$$

$$(\hat{D}_{n,m}, \hat{E}_{n,m}) = \operatorname*{argmin}_{D \in \mathcal{D}, E \in \mathcal{E}} \|(\hat{\mu}_n, E_\sharp \hat{\mu}_n) - (G_\sharp \hat{\nu}_m, \hat{\nu}_m)\|_\Psi. \tag{6}$$

To measure the efficacy of the above estimates, one computes the closeness of the joint distribution $(\mu, \hat{E}_\sharp \mu)$ to $(\hat{D}_\sharp \nu, \nu)$ as,

$$\mathfrak{R}_{\mathcal{F}}^{\text{BiGAN}}(\hat{D}, \hat{E}) = \|(\mu, \hat{E}_\sharp \mu) - (\hat{D}_\sharp \nu, \nu)\|_{\mathcal{F}}.$$

As before, we take $\mathcal{F} = \Psi = \mathcal{H}^\beta(\mathbb{R}^{d+\ell}, \mathbb{R}, 1)$.

3. Intrinsic Dimension of Data Distribution

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 $(\mathcal{S}, \varrho)$ be a 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. Before, we proceed we recall the (ϵ, τ)-cover of a measure (Posner et al., 1967) as: $\mathcal{N}_\epsilon(\mu, \tau) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : \mu(S) \ge 1 - \tau\}$, i.e. $\mathcal{N}_\epsilon(\mu, \tau)$ counts the minimum number of ϵ-balls required to cover a set S of probability at least $1 - \tau$.

The most simplistic notion of the dimension of a measure is the upper Minkowski dimension of its support, which is defined as:

$$\overline{\dim}_M(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}(\epsilon; \operatorname{supp}(\mu), \ell_\infty)}{\log(1/\epsilon)}.$$

Since this notion of dimensionality depends only on the covering number of the support 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 as fractals. The statistical convergence guarantees of different estimators in terms of the upper Minkowski dimension are well-studied in the literature. Kolmogorov and Tikhomirov (1961) provided a comprehensive study on the dependence of the covering number of different function classes on the underlying upper Minkowski dimension of the support. Nakada and 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. However, one key drawback of using the upper Minkowski dimension is that if the measure spans over the entire sample space, though being concentrated on only some regions, can result in a high value of the dimension. We refer the reader to Examples 10 and 11 for such instances.

To overcome the aforementioned difficulty, as a notion of the intrinsic dimension of a measure μ, Dudley (1969) defined the entropic dimension of a measure as:

$$\kappa_\mu = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon)}{\log(1/\epsilon)}.$$

Inspired by this definition, we define the α-entropic dimension of μ as follows.

Definition 4 (Entropic Dimension)**.** For any $\alpha > 0$, we define the α-entropic dimension of μ as:

$$\bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{\log(1/\epsilon)}.$$

Clearly, $\bar{d}_\alpha(\mu) \geq 0$. Note that Dudley's entropic dimension is $\bar{d}_1(\mu)$. Dudley (1969) showed that when the data is i.i.d. $\mathbb{E}\|\hat{\mu}_n - \mu\|_{\mathcal{H}^1}$ roughly scales as $\mathcal{O}\left(n^{-1/\bar{d}_1(\mu)}\right)$, while $\mathbb{E}\,\mathrm{Dist}_{LP}(\hat{\mu}_n, \mu)$ roughly scales as $\mathcal{O}\left(n^{-1/(2+\bar{d}_1(\mu))}\right)$, where $\mathrm{Dist}_{LP}(\cdot,\cdot)$ denotes the Lévy-Prokhorov metric (Prokhorov, 1956). As shown in Theorems 13 and 21, the entropic dimension characterizes the metric entropy of β-Hölder functions and the approximation capabilities of neural networks with ReLU activation. Weed and Bach (2019) developed upon Dudley's entropic dimension to characterize the expected convergence rate of the Wasserstein-p distance between a measure and its empirical counterpart. The upper and lower Wasserstein dimension of μ is defined as follows:

Definition 5 (Upper and Lower Wasserstein Dimensions (Weed and Bach, 2019))**.** For any $\alpha > 0$, the α-upper dimension of μ is defined by

$$d^*_\alpha(\mu) = \inf\left\{ s \in (2\alpha, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \leq s \right\}.$$

The lower Wasserstein dimension is defined as: $d_*(\mu) = \lim_{\tau \downarrow 0} \liminf_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \tau)}{\log(1/\epsilon)}$.

Weed and Bach (2019) showed that, roughly, $n^{-1/d_*(\mu)} \lesssim \mathbb{E}\mathcal{W}_p(\hat{\mu}_n, \mu) \lesssim n^{-1/d_p^*(\mu)}$. In Proposition 8, we give some elementary properties of these dimensions with proof in Appendix A.1. We also recall the definition of regularity dimensions and packing dimension of a measure (Fraser and Howroyd, 2017).

Definition 6 (Regularity dimensions)**.** The upper and lower regularity dimensions of a measure are defined as:

$$\overline{\dim}_{\text{reg}}(\mu) = \inf\left\{s : \exists C > 0 \text{ such that, for all } 0 < r < R \text{ and } x \in \text{supp}(\mu), \frac{\mu(B(x,R))}{\mu(B(x,r))} \leq C\left(\frac{R}{r}\right)^s\right\},$$

$$\underline{\dim}_{\text{reg}}(\mu) = \sup\left\{s : \exists C > 0 \text{ such that, for all } 0 < r < R \text{ and } x \in \text{supp}(\mu), \frac{\mu(B(x,R))}{\mu(B(x,r))} \geq C\left(\frac{R}{r}\right)^s\right\}.$$

Definition 7 (Upper packing dimension)**.** The upper packing dimension of a measure μ is defined as: $\overline{\dim}_P(\mu) = \operatorname{ess\,sup}\left\{\limsup_{r\to 0} \frac{\log \mu(B(x,r))}{\log r} : x \in \text{supp}(\mu)\right\}$.

Proposition 8. *For any μ and for any $\alpha \geq 0$, the following hold:*

(a) $d_*(\mu) \leq \bar{d}_\alpha(\mu) \leq d_\alpha^*(\mu)$,

(b) *if* $\alpha_1 \leq \alpha_2$, *then,* $\bar{d}_{\alpha_1}(\mu) \leq \bar{d}_{\alpha_2}(\mu)$,

(c) $\bar{d}_\alpha(\mu) \leq \overline{\dim}_M(\mu)$,

(d) *if* $\alpha \in \left(0, \overline{\dim}_P(\mu)/2\right)$, $d_\alpha^*(\mu) \leq \overline{\dim}_P(\mu) \leq \overline{\dim}_{reg}(\mu)$,

(e) $\bar{d}_\alpha(\mu) \leq \overline{\dim}_P(\mu) \leq \overline{\dim}_{reg}(\mu)$,

(f) $\underline{\dim}_{reg}(\mu) \leq d_*(\mu)$.

It was noted by Weed and Bach (2019) that the upper Wasserstein dimension is usually smaller than the upper Minkowski dimension as noted in the following lemma.

Proposition 9. *(Proposition 2 of Weed and Bach (2019)) If $\overline{\dim}_M \geq 2\alpha$, then, $d_\alpha^*(\mu) \leq \overline{\dim}_M(\mu)$.*

Propositions 8 (c) and 9 imply that for high-dimensional data, both the entropic dimension and the upper Wasserstein dimension are no larger than the upper Minkowski dimension of the support. In many cases, strict inequality holds as seen in the following examples.

Example 10. Let the measure μ on $\mathbb{N}$, be such that $\mu(n) = 2^{-n}$ for all $n \in \mathbb{N}$. Clearly, the support of μ is $\mathbb{N}$, which has an upper Minkowski dimension of ∞. To find $\bar{d}_\alpha(\mu)$, we first fix $\epsilon \in (0,1)$. For any $n \in \mathbb{N}$, let $A_n = [n]$. We observe that, $\mu\left(A_n^{\complement}\right) = \frac{1}{2^n}$. We take

$K = \lceil \alpha \log_2(1/\epsilon) \rceil$. Clearly, $\mu(A_K) \geq 1 - \epsilon^\alpha$. We can cover A_K by at most K many intervals of length ϵ. Thus, $\mathcal{N}_\epsilon(\mu, \epsilon^\alpha) \leq K = \lceil \log_2(1/\epsilon) \rceil$. Hence,

$$0 \leq \bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{-\log \epsilon} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \lceil \alpha \log_2(1/\epsilon) \rceil}{-\log \epsilon} = 0.$$

Thus, $\bar{d}_\alpha(\mu) = 0$, for all $\alpha > 0$. Similarly, let $s > 2\alpha$ and take, $K = \lceil \frac{s\alpha}{s-2\alpha} \log_2(1/\epsilon) \rceil$. According to the argument above,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}})}{-\log \epsilon} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \lceil \frac{s\alpha}{s-2\alpha} \log_2(1/\epsilon) \rceil}{-\log \epsilon} = 0 < s.$$

Thus, $d^*_\alpha(\mu) = 2\alpha$. Thus, for this example, we get, $\bar{d}_\alpha(\mu) < d^*_\alpha(\mu) < \overline{\dim}_M(\mu)$.

Even when $\overline{\dim}_M(\mu) < \infty$, it can be the case that $\bar{d}_\alpha(\mu) < d^*_\alpha(\mu) < \overline{\dim}_M(\mu)$ as seen in the following example.

Example 11. Let μ be such that $\mu\left(\left(\frac{1}{n_1}, \dots, \frac{1}{n_d}\right)\right) = \frac{1}{2^n}$ for all $n_1, \dots, n_d \in \mathbb{N}$. We know that $\overline{\dim}_M(\mu) = \frac{d}{2}$, by Example 1.14 of Bishop and Peres (2017). A calculation, similar to Example 10 shows that $\bar{d}_\alpha(\mu) = 0$ and $d^*_\alpha(\mu) = 2\alpha$, for all $\alpha > 0$. Thus, if $\alpha < \frac{d}{4}$, $\bar{d}_\alpha(\mu) < d^*_\alpha(\mu) < \overline{\dim}_M(\mu)$.

There are cases where both $\bar{d}_\alpha(\mu)$ and $d^*_\alpha(\mu)$ are both positive and a strict inequality might hold. We give an example as follows.

Example 12. Suppose that μ is a probability measure on $\mathbb{N}$, such that, $\mu(n) = n^{-2/3} - (n+1)^{-2/3}$. Then for any $\epsilon \in (0, 1)$, let $K = \lceil \epsilon^{-3\alpha/2} \rceil$. Then, $\mu(A_K) \geq 1 - \epsilon^\alpha$. Again, A_K requires at most K intervals of length ϵ to be covered, which implies that $\mathcal{N}_\epsilon(\mu, \epsilon^\alpha) \leq K = \lceil \epsilon^{-3\alpha/2} \rceil$. Hence,

$$\bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{-\log \epsilon} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \lceil \epsilon^{-3\alpha/2} \rceil}{-\log \epsilon} = 3\alpha/2.$$

Further, by definition, $d^*_\alpha(\mu) \geq 2\alpha$. This implies that $\bar{d}_\alpha < d^*_\alpha(\mu) < \overline{\dim}_M(\mu)$.

The entropic dimension can be used to characterize the metric entropy of the set of all Hölder functions in the $\mathbb{L}_p$-norm w.r.t the corresponding measure. We focus on the following theorem that strengthens the seminal result by Kolmogorov and Tikhomirov (1961) (see Lemma 52). It is important to note that when the measure has an intrinsically lower dimension compared to the high dimensionality of the entire feature space, the metric entropy only depends on the $p\beta$-entropic dimension of the underlying measure μ, not the dimension of the feature space, i.e. d.

Theorem 13. *Let $\mathcal{F} = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$. Then, if $s > \bar{d}_{p\beta}(\mu)$, then, there exists an $\epsilon' > 0$, such that, if $0 < \epsilon \leq \epsilon'$, then, $\log \mathcal{N}(\epsilon; \mathcal{F}, \mathbb{L}_p(\mu)) \lesssim \epsilon^{-s/\beta}$.*

We note that if $\overline{\dim}_M(\mu) < \infty$, then, by Proposition 8 (c), $\bar{d}_{p\beta}(\mu) \leq \overline{\dim}_M(\mu)$. Thus as an immediate consequence of Theorem 13, we observe that, if $s > \overline{\dim}_M(\mu)$, then, for ϵ small enough, $\log \mathcal{N}(\epsilon; \mathcal{F}, \mathbb{L}_p(\mu)) \lesssim \epsilon^{-s/\beta}$, recovering a similar result as observed by Kolmogorov and Tikhomirov (1961) (Lemma 52).

Next, we characterize the rate of convergence of $\hat{\mu}_n$ to μ under the β-Hölder IPM. The proof technique stems from Dudley's seminal work (Dudley, 1969). The idea is to construct a dyadic-like partition of the data space and approximate the functions through their Taylor approximations on each of these small pieces. The reader is referred to Appendix A.3 for a proof of this result.

Theorem 14. *Let $\mathcal{F} = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$. Then, for any $d^\star > d^*_\beta(\mu)$, we can find $n_0 \in \mathbb{N}$, such that if $n \geq n_0$,*

$$\mathbb{E}\|\hat{\mu}_n - \mu\|_{\mathcal{F}} \lesssim n^{-\beta/d^\star}.$$

Here n_0 might depend on μ and $d^\star$.

Note that if $\beta, C = 1$, then, $\mathcal{F} = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$ is the set of all 1-bounded and 1-Lipschitz functions (denoted as BL) on $\mathbb{R}^d$. Thus, for any $d^\star > d^*_1(\mu)$, we observe that $\mathbb{E}\|\hat{\mu}_n - \mu\|_{\mathrm{BL}} \lesssim n^{-1/d^\star}$, recovering the bound derived by Dudley (1969). Furthermore, if μ has a bounded support, i.e. there exist $M > 0$, such that, $\mu([-M, M]^d) = 1$, then it is easy to see that $\mathbb{E}\mathcal{W}_1(\hat{\mu}_n, \mu) \leq \mathbb{E}\|\hat{\mu}_n - \mu\|_{\mathcal{H}^\beta([0,1]^d, \mathbb{R}, M\sqrt{d})} \lesssim n^{-1/d^\star}$, which recovers the rates derived by Weed and Bach (2019), for Wasserstein-1 distance.

4. Theoretical Analyses

4.1 Assumptions

To lay the foundation for our analysis of GANs and BiGANs, we introduce a set of assumptions that form the basis of our theoretical investigations. These assumptions encompass the underlying data distribution, the existence of "true" generator and encoder-decoder pairs, and certain smoothness properties. For the purpose of the theoretical analysis, we assume that the data are independent and identically distributed from some unknown target distribution μ on $[0,1]^d$. This is a standard assumption in the literature (Liang, 2021; Huang et al., 2022) and is stated formally as follows:

A1. *We assume that $X_1, \dots, X_n$ are independent and identically distributed according to the probability distribution μ, such that $\mu\left([0,1]^d\right) = 1$.*

Furthermore, to facilitate the analysis of GANs, we introduce the concept of a "true" generator, denoted as $\tilde{G}$. We assume that $\tilde{G}$ belongs to a function space $\mathcal{H}^{\alpha_g}\left(\mathbb{R}^\ell, \mathbb{R}^d, C_g\right)$ and that the target distribution μ can be represented as $\mu = \tilde{G}_\sharp \nu$.

A2. *There exists a $\tilde{G} \in \mathcal{H}^{\alpha_g}\left(\mathbb{R}^\ell, \mathbb{R}^d, C_g\right)$, such that $\mu = \tilde{G}_\sharp \nu$.*

Similarly, for the BiGAN problem, we consider the existence of a smooth encoder-decoder pair, denoted as $\tilde{D}$ and $\tilde{E}$ respectively. These functions belong to the spaces $\mathcal{H}^{\alpha_d}\left(\mathbb{R}^\ell, \mathbb{R}^d, C_d\right)$ and $\mathcal{H}^{\alpha_e}\left(\mathbb{R}^d, \mathbb{R}^\ell, C_e\right)$, respectively, and satisfy the property that the joint distribution of $(X, \tilde{E}(X))$ is equal in distribution to $(\tilde{D}(Z), Z)$, where $X \sim \mu$ and $Z \sim \nu$.

B1. *There exists $\tilde{D} \in \mathcal{H}^{\alpha_d}\left(\mathbb{R}^\ell, \mathbb{R}^d, C_d\right)$ and $\tilde{E} \in \mathcal{H}^{\alpha_e}\left(\mathbb{R}^d, \mathbb{R}^\ell, C_e\right)$, such that, $(X, \tilde{E}(X)) \overset{d}{=} (\tilde{D}(Z), Z)$. Here, $X \sim \mu$ and $Z \sim \nu$.*

A direct consequence of Assumption B1 is that the composition of the true encoder and decoder functions yields the identity map almost surely, as described in Lemma 15. The notation, a.e. $[\mu]$ denotes almost everywhere under the probability measure μ.

Lemma 15. *Under assumption B1, $\tilde{D} \circ \tilde{E}(\cdot) = id(\cdot)$, a.e. $[\mu]$ and $\tilde{E} \circ \tilde{D}(\cdot) = id(\cdot)$, a.e. $[\nu]$.*

4.2 Main Results

Under assumptions A1, A2 and B1, one can bound the expected excess risk of the GAN and BiGAN estimates of the target distribution in terms of the sample size with the exponent only depending on the upper Wasserstein dimension of the target distribution and the exponent of the Hölder IPM considered. The main results of this paper are summarized in the following two theorems. We recall that $\Phi = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$ and $\Psi = \mathcal{H}^\beta(\mathbb{R}^{d+\ell}, \mathbb{R}, 1)$ denote the discriminator classes and $\hat{G}_n$, $\hat{G}_{n,m}$ denote the sample minimizers for the GAN-problem, defined in (3). Similarly, we recall the notations $(\hat{D}_n, \hat{E}_n)$ and $(\hat{D}_{n,m}, \hat{E}_{n,m})$ as the sample minimizers for the BiGAN-problem, defined in (5) and (6), respectively.

Theorem 16 (Error rate for GANs)**.** *Suppose assumptions A1 and A2 hold and let $s > d^*_\beta(\mu)$. There exist constants N, c that might depend on d, ℓ, α_g, β and $\tilde{G}$, such that, if $n \geq N$, we can choose $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ with the network parameters as $L_g \leq c \log n$, $W_g \leq c n^{\frac{\beta \ell}{\alpha_g s (\beta \wedge 1)}} \log n$. Then,*

$$\mathbb{E}\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \lesssim n^{-\beta/s}. \tag{7}$$

Furthermore, if $m \geq \upsilon_n$, where, $\upsilon_n = \inf\left\{m \in \mathbb{N} : \frac{(\log m)^2}{m^{\left(\max\left\{2+\frac{\ell}{\alpha_g(\beta\wedge 1)}, \frac{d}{\beta}\right\}\right)^{-1}}} \leq n^{-\beta/s}\right\}$, and the network parameters are chosen as, $L_g \leq c \log m$, $W_g \leq c m^{\frac{\ell}{2\alpha_g(\beta\wedge 1)+\ell}} \log m$, then

$$\mathbb{E}\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Phi \lesssim n^{-\beta/s}. \tag{8}$$

Theorem 17 (Error rate for BiGANs)**.** *Suppose assumptions A1 and B1 hold and let $s_1 > \bar{d}_{\alpha_e}(\mu)$ and $s_2 > d^*_\beta(\mu)$. There exists constants N, c that might depend on $d, \ell, \alpha_d, \alpha_e, \beta, \tilde{D}$ and $\tilde{E}$, such that, if $n \geq N$, we can choose the networks $\mathcal{E} = \mathcal{RN}(L_e, W_e, B_e, 2C_e)$ and $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ with the network parameters chosen as*

$$L_e \leq c \log n,\ W_e \leq c n^{\frac{s_1}{2\alpha_e(\beta\wedge 1)+s_1}} \log n,\ L_d \leq c \log n,\ W_d \leq c n^{\frac{\beta\ell}{\alpha_d(\beta\wedge 1)(s_2+\ell)}} \log n. \tag{9}$$

Then,

$$\mathbb{E}\|(\mu, (\hat{E}_n)_\sharp \mu) - ((\hat{D}_n)_\sharp \nu, \nu)\|_\Psi \lesssim n^{-\frac{1}{\max\left\{2+\frac{s_1}{\alpha_e(\beta\wedge 1)}, \frac{s_2+\ell}{\beta}\right\}}} (\log n)^2. \tag{10}$$

Furthermore, if the network parameters are chosen as

$$L_e \leq c \log n,\ W_e \leq c n^{\frac{s_1}{2\alpha_e(\beta\wedge 1)+s_1}} \log n,\ L_d \leq c \log m,\ W_d \leq c m^{\frac{\ell}{2\alpha_d(\beta\wedge 1)+\ell}} \log m, \tag{11}$$

then,

$$\begin{aligned}&\mathbb{E}\|(\mu, (\hat{E}_{n,m})_\sharp \mu) - ((\hat{D}_{n,m})_\sharp \nu, \nu)\|_\Psi \\ \lesssim\ & n^{-\frac{1}{\max\left\{2+\frac{s_1}{\alpha_e(\beta\wedge 1)}, \frac{s_2+\ell}{\beta}\right\}}} (\log n)^2 + m^{-\frac{1}{\max\left\{2+\frac{d}{\alpha_g(\beta\wedge 1)}, \frac{\ell+d}{\beta}\right\}}} (\log m)^2.\end{aligned} \tag{12}$$

It should be noted that the dependence on d in N and c may be influenced by a multitude of factors including the constants in Theorems 18 and 21 and thus, can potentially be exponential. Additionally, although the BiGAN problem might seem symmetric on the surface, there are subtle asymmetries that make the bounds on W_e and W_d asymmetric in (9). For example, for the estimation problem (5), one has full access to ν but only n samples from μ. The bounds for W_e and W_d in (11) are, however, more symmetric since one has access to m samples from ν instead of the entire distribution in the estimation problem (6), resulting in a bound that is more symmetric in m and n as shown in (11). Since one has the luxury to generate as many samples as one wants from ν, increasing m matches the rate of (12) to (10) as shown in the following corollary.

Corollary 18. *Let* $\vartheta = \frac{\max\left\{2+\frac{\ell}{\alpha_d(\beta\wedge 1)}, \frac{d+\ell}{\beta}\right\}}{\max\left\{2+\frac{s_1}{\alpha_e(\beta\wedge 1)}, \frac{s_2+\ell}{\beta}\right\}} \vee 1$. *Under the assumptions of Theorem 17, if* $m \geq n^{\vartheta}$, *and the network parameters are chosen as*

$$L_e \leq c\log n,\; W_e \leq cn^{\frac{s_1}{2\alpha_e(\beta\wedge 1)+s_1}}\log n,\; L_d \leq c\log n,\; W_d \leq cm^{\frac{\ell}{2\alpha_d(\beta\wedge 1)+\ell}}\log n, \tag{13}$$

then,

$$\mathbb{E}\|(\mu, (\hat{E}_{n,m})_\sharp\mu) - ((\hat{D}_{n,m})_\sharp\nu, \nu)\|_\Psi \lesssim n^{-\frac{1}{\max\left\{2+\frac{s_1}{\alpha_e(\beta\wedge 1)}, \frac{s_2+\ell}{\beta}\right\}}}(\log n)^2. \tag{14}$$

Implications for BiGANs We note that when the true model for the BiGAN problem is assumed to be Lipschitz, i.e. if $\alpha_d = \alpha_e = 1$ in B1 and we take the $\|\cdot\|_{\mathrm{BL}}$-metric i.e. the IPM w.r.t. $\Psi = \mathcal{H}^1(\mathbb{R}^{d+\ell}, \mathbb{R}, 1)$, then, the excess risk for the BiGAN problem roughly scales as $\tilde{\mathcal{O}}\left(n^{-1/\left(\max\left\{2+\bar{d}_1(\mu), d_1^*(\mu)+\ell\right\}\right)}\right)$, barring the log-factors. Thus, the expected excess risk scales at a rate of $\tilde{\mathcal{O}}\left(n^{-1/\left(\bar{d}_1(\mu)\vee d_1^*(\mu)+\ell\vee 2\right)}\right)$. In practice, one usually takes $\ell \geq 2$, and by the fact that $\bar{d}_1(\mu) \leq d_1^*(\mu)$ (Proposition 8 (a)), we observe that this rate is at most $\tilde{\mathcal{O}}\left(n^{-1/(d_1^*(\mu)+\ell)}\right)$, which is akin to the minimax estimation rate for estimating the joint distribution of $\mu \otimes \nu$ on $\mathbb{R}^{d+\ell}$ (Weed and Berthet, 2022; Singh and Póczos, 2018).

Further, we note that Assumption B1 implies that the supports of μ and ν are homeomorphic, which is rather restrictive. The proof of this result can be found in Appendix B.2.

Proposition 19. *Under Assumption B1, the supports of* μ *and* ν *are homeomorphic.*

It would be interesting to see if this assumption could be lifted to derive an analogous result akin to Theorem 28 for the BiGAN problem as well. The main hindrance is that although for $\hat{\mu}_n$, one can construct D such that $D_\sharp\nu \approx \hat{\mu}_n$, the same cannot be said the other way around, i.e. there might not exist a E such that $E_\sharp\hat{\mu}_n \approx \nu$. Further, both these approximations have to hold jointly, which is even more difficult to show. We believe that this would be an interesting direction for future research.

Comparison of Rates with Recent Literature We also note that for the GAN problem, the expected error for estimating the target density through $\hat{G}_\sharp\mu$ roughly scales as $\mathbb{E}\|\mu - \hat{G}_\sharp\nu\|_{\mathcal{H}^\beta} = \mathcal{O}\left(n^{-\beta/d_\beta^*(\mu)}\right)$. For high-dimensional data, it can be expected that $\overline{\dim}_M(\mu) \gg 2\beta$, which (by Lemma 9) would imply that, $d_\beta^*(\mu) \leq \overline{\dim}_M(\mu)$, with strict inequality holding in many cases (see Examples 10 and 11). We thus observe that the derived rate, derived rate is faster than the ones derived by Huang et al. (2022) and Chen et al. (2020), who showed that the error rates scale as $\tilde{\mathcal{O}}\left(n^{-\beta/\overline{\dim}_M(\mu)} \vee n^{-1/2}\right)$ and $\tilde{\mathcal{O}}\left(n^{-\beta/(2\beta+d)}\right)$, respectively. We note that the removal of the excess log-factor is an artifact of the requirement that $s > d^*_\beta(\mu)$ and one can similarly do away with the log-factors in the rates derived by Huang et al. (2022).

Inference for Data supported on a Manifold Recall that we call a set $\mathcal{A}$ is $\tilde{d}$-regular w.r.t. the $\tilde{d}$-dimensional Hausdorff measure (see Definition 48) $\mathbb{H}^{\tilde{d}}$, if $\mathbb{H}^{\tilde{d}}(B_\varrho(x,r)) \asymp r^{\tilde{d}}$, for all $x \in \mathcal{A}$ (see Definition 6 of Weed and Bach (2019)). It is known (Weed and Bach, 2019, Proposition 8) that if $\text{supp}(\mu)$ is d regular and $\mu \ll \mathbb{H}^{\tilde{d}}$, for $\beta \in [1, \tilde{d}/2]$, $d_*(\mu) = d^*_\beta(\mu) = \tilde{d}$. Thus, by Theorem 16, the error rates for GANs scale at a rate of $\tilde{\mathcal{O}}(n^{-\beta/\tilde{d}})$. Since, compact $\tilde{d}$-dimensional differentiable manifolds are $\tilde{d}$-regular (Weed and Bach, 2019, Proposition 9), this implies that when the support of μ is a compact $\tilde{d}$-dimensional differentiable manifold, the error rates scale as $\tilde{\mathcal{O}}(n^{-\beta/\tilde{d}})$, which recovers a similar result as that by Dahal et al. (2022) as a special case. A similar result holds when $\text{supp}(\mu)$ 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 $\tilde{d}$.

Network Sizes We note that Theorems 16 and 17 suggest that the depths of the encoders and decoders can be chosen as the order of $\log n$ and the number of weights can be chosen as some exponent of n and m. The degrees of these exponents only depend on $\bar{d}_\beta(\mu)$, $d^*_\beta(\mu)$ and ℓ i.e. the dimension of the problem and the smoothness of the true model. The bounds on the number of weights in Theorems 16 and 17 also show that one requires a smaller network if one lets α_e, α_g and α_d grow, i.e. when the true encoder, generator/decoder are very smooth and well-behaved. This is because the optimal choices for W_e and W_g (or W_d) in terms of n in Theorems 16 and 17, decreases as α_e, α_g and α_d increase. This is quite expected that one requires less complicated architectures when the target densities are smooth enough.

Scaling of m We note that in Theorem 16, for the estimator $\hat{G}_{n,m}$, one roughly requires $m \gtrsim n^{\frac{\beta}{d^*_\beta(\mu)}\max\{2+\ell/(\alpha_g(\beta\wedge 1)), d/\beta\}}$. If the true generator is Lipschitz, $\alpha_g = 1$ and one considers the difference in the BL-metric, i.e. $\beta = 1$, then, $m \gtrsim n^{\max\{d,(2+\ell)\}/d^*_\beta(\mu)} = n^{d/d^*_\beta(\mu)}$, for all practical purposes as $d \gg \ell$. Thus, the dimension affects the sample scaling; higher d values necessitate more generated samples to achieve equivalent accuracy. This dimension-dependent scaling also appears in recent works by Huang et al. (2022). It would be interesting to explore whether this dependency can be reduced in future research. Nonetheless, since practitioners can generate as many samples as needed from ν, this dependency may be less limiting in practical settings.

The remainder of Section 4 discusses the proofs of Theorems 16 and 17 by first decomposing the error into misspecification and generalization terms and individually bounding them.

4.3 Proof of the Main Results

4.3.1 Error Decomposition

As a first step towards deriving bounds on the expected risks for both the GAN and BiGAN problems, we now derive the following oracle inequality that bounds this excess risk in terms of the approximation error and a generalization gap with proof in Appendix C.1. A similar result was derived by Huang et al. (2022) for analyzing the GAN problem.

Lemma 20 (Oracle inequalities for GANs and BiGANs)**.** *Suppose that $\Phi = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$ and $\Psi = \mathcal{H}^\beta(\mathbb{R}^{d+\ell}, \mathbb{R}, 1)$, then*

(a) if $\hat{G}_n$ and $\hat{G}_{n,m}$ are the generator estimates defined in (3)*, the following hold:*

$$\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \le \inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi, \tag{15}$$

$$\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Psi \le \inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi + 2\|\nu - \hat{\nu}_m\|_{\Phi \circ \mathcal{G}}. \tag{16}$$

(b) if $(\hat{D}_n, \hat{E}_n)$ and $(\hat{D}_{n,m}, \hat{E}_{n,m})$ are the decoder-encoder estimates defined in (5) *and* (6)*, respectively, the following hold:*

$$\|(\mu, (\hat{E}_n)_\sharp \mu) - (\hat{D}_n)_\sharp \nu, \nu)\|_\Psi \le \inf_{D \in \mathcal{D},\, E \in \mathcal{E}} \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi + 2\|\hat{\mu}_n - \mu\|_{\mathcal{F}_1}, \tag{17}$$

$$\begin{aligned}\|(\mu, (\hat{E}_{n,m})_\sharp \mu) - (\hat{D}_{n,m})_\sharp \nu, \nu)\|_\Psi \le & \inf_{D \in \mathcal{D},\, E \in \mathcal{E}} \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi + 2\|\hat{\mu}_n - \mu\|_{\mathcal{F}_1} \\ & + 2\|\hat{\nu}_m - \nu\|_{\mathcal{F}_2},\end{aligned} \tag{18}$$

where, $\mathcal{F}_1 = \{\psi(\cdot, E(\cdot)) : \psi \in \Psi,\, E \in \mathcal{E}\}$ and $\mathcal{F}_2 = \{\psi(D(\cdot), \cdot) : \psi \in \Psi,\, D \in \mathcal{D}\}$.

Using Lemma 20, we aim to bound each of the terms individually in the following sections. The misspecification errors are controlled by deriving a new approximation result involving ReLU networks, whereas, the generalization gap is tackled through empirical process theory. It is important to note that in Lemma 20, the term $\|\mu - \hat{\mu}_n\|_\Phi$, which scales at a rate of roughly $\mathcal{O}\left(n^{-\beta/d^*_\beta(\mu)}\right)$ (by Theorem 14), dominates the other error terms as can be observed in the proof, leading to the error rate in Theorem 16.

4.3.2 Bounding the Misspecification Error

The approximation capabilities of neural networks have received a lot of attention in the past decade. The seminal works of Cybenko (1989) and Hornik (1991) delve into the universal approximation of networks with sigmoid-like activations to show that wide one-hidden-layer neural networks can approximate any continuous function on a compact set. With the recent advancements in deep learning, there has been a surge of research investigating the approximation capabilities of deep neural networks (Yarotsky, 2017; Petersen and Voigtlaender, 2018; Shen et al., 2022; Nakada and Imaizumi, 2020; Schmidt-Hieber, 2020).

In this section, we show how a ReLU network with a large enough depth and width can approximate any function lying on a low-dimensional structure. We suppose that $f \in \mathcal{H}^\alpha\left(\mathbb{R}^d, \mathbb{R}, C\right)$ and γ is a measure on $\mathbb{R}^d$. We show that for any $\epsilon > 0$ and $s > \bar{d}_{\alpha p}(\gamma)$, we can find a ReLU network $\hat{f}$ of depth at most $\mathcal{O}(\log(1/\epsilon))$ and number of weights at most $\mathcal{O}(\epsilon^{-s/\alpha}\log(1/\epsilon))$ such that $\|f-\hat{f}\|_{\mathbb{L}_p(\gamma)} \le \epsilon$. Note that when $\text{supp}(\mu)$ has a finite Minkowski dimension, by Proposition 8 (c), we observe that $\bar{d}_{\alpha p} \le \overline{\dim}_M(\mu)$. Thus, the number of weights required for an ϵ-approximation, in the $\mathbb{L}_p$ sense, requires at most $\mathcal{O}(\epsilon^{-\overline{\dim}_M(\mu)/\alpha}\log(1/\epsilon))$, recovering a similar result as derived by Nakada and Imaizumi (2020) as a special case. Note that the required number of weights for low-dimensional data, i.e. when $\bar{d}_{\alpha p}(\gamma) \ll d$, is much smaller than $\mathcal{O}(\epsilon^{-d/\alpha}\log(1/\epsilon))$, that holds when approximating on the entire space w.r.t. ℓ_∞-norm (Yarotsky, 2017; Chen et al., 2019). The idea is to approximate the Taylor series expansion of the corresponding Hölder functions. The general proof technique was developed by Yarotsky (2017). Our result is formally stated in the following theorem.

Theorem 21. *Suppose that $f \in \mathcal{H}^\alpha(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$ and let $s > \bar{d}_{\alpha p}(\mu)$. Then, we can find constants ϵ_0 and a, that might depend on α, d and C, such that, for any $\epsilon \in (0, \epsilon_0]$, there exists a ReLU network, $\hat{f}$ with $\mathcal{L}(\hat{f}) \le a\log(1/\epsilon)$, $\mathcal{W}(\hat{f}) \le a\log(1/\epsilon)\epsilon^{-s/\alpha}$, $\mathcal{B}(\hat{f}) \le a\epsilon^{-1/\alpha}$ and $\mathcal{R}(\hat{f}) \le 2C$, that satisfies, $\|f-\hat{f}\|_{\mathbb{L}_p(\gamma)} \le \epsilon$.*

Applying the above theorem, one can control the model-misspecification error for GANs and BiGANs as follows. It is important to note that none of the approximations require the number of weights of the approximating network to increase exponentially with d, i.e. the dimension of the entire data space.

Lemma 22. *Suppose assumption A2 holds. There exists an ϵ_0, such that, for any $0 < \epsilon \le \epsilon_0$, we can take $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ such that $L_g \lesssim \log(1/\epsilon)$, $W_g \lesssim \epsilon^{-\ell/\alpha_g}\log(1/\epsilon)$, $B_g \lesssim \epsilon^{-1/\alpha_g}$ and $\inf_{G\in\mathcal{G}} \|\mu - G_\sharp\nu\|_\Phi \lesssim \epsilon^{\beta\wedge 1}$.*

Lemma 23. *Suppose assumption B1 holds and let $s > \bar{d}_{\alpha_e}(\mu)$. There exists $\epsilon_0 > 0$, such that, for any $0 < \epsilon_1, \epsilon_2 \le \epsilon_0$, we can take $\mathcal{E} = \mathcal{RN}(L_e, W_e, B_e, 2C_e)$ and $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ such that*

$$L_e \lesssim \log(1/\epsilon_e),\ W_e \lesssim \epsilon_e^{-s/\alpha_e}\log(1/\epsilon_e),\ \text{ and } B_e \lesssim \epsilon^{-1/\alpha_e};$$
$$L_d \lesssim \log(1/\epsilon_d),\ W_d \lesssim \epsilon_d^{-\ell/\alpha_d}\log(1/\epsilon_d),\ \text{ and } B_d \lesssim \epsilon_d^{-1/\alpha_d}.$$

Then, $\inf_{G\in\mathcal{G},\, E\in\mathcal{E}} \|(\mu, E_\sharp\mu) - (G_\sharp\nu, \nu)\|_\Psi \lesssim \epsilon_e^{\beta\wedge 1} + \epsilon_d^{\beta\wedge 1}$.

4.3.3 Generalization Gap

The third step to bounding the excess risk is to bound the generalization gaps w.r.t. the function classes discussed in Lemma 20. To do so, we first derive a bound on the metric entropy of networks from $\mathbb{R}^d \to \mathbb{R}^{d'}$ that have piece-wise polynomial activations, extending the results of Bartlett et al. (2019) to vector-valued networks. Recall the notation $\mathcal{F}_{|_{X_{1:n}}} = \{(f(X_1), \dots, f(X_n))^\top \in \mathbb{R}^{n\times d'} : f \in \mathcal{F}\}$.

Lemma 24. *Suppose that $n \ge 6$ and $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^{d'}$ be a class of bounded neural networks with depth at most L and the number of weights at most W. Furthermore, the activation functions are piece-wise polynomial activation with the number of pieces and degree at most*

$k \in \mathbb{N}$. *Then, there exists positive constants* θ *and* ϵ_0*, such that, if* $n \geq \theta(W + 6d' + 2d'L)(L+3)\left(\log(W + 6d' + 2d'L) + L + 3\right)$ *and* $\epsilon \in (0, \epsilon_0]$*,*

$$\log \mathcal{N}(\epsilon; \mathcal{F}_{|X_{1:n}}, \ell_\infty) \lesssim (W + 6d' + 2d'L)(L+3)\left(\log(W + 6d' + 2d'L) + L + 3\right) \log\left(\frac{nd'}{\epsilon}\right),$$

where d' *is the output dimension of the networks in* $\mathcal{F}$*.*

As a corollary of the above result, we can bound the metric entropies of the function classes in Lemma 20 as a function of the number of samples used and the size of the networks classes $\mathcal{G}$ and $\mathcal{E}$. Apart from appealing to Lemma 24, one also uses the bound on the metric entropy of β-Hölder functions as in Theorem 13. We recall that, $\mathcal{F}_1 = \{\psi(\cdot, E(\cdot)) : \psi \in \Psi,\, E \in \mathcal{E}\}$ and $\mathcal{F}_2 = \{\psi(G(\cdot), \cdot) : \psi \in \Psi,\, G \in \mathcal{G}\}$.

Corollary 25. *Suppose that* $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ *and* $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ *with* $L_d, L_g \geq 3$*,* $W_g \geq 6d + 2dL_g$ *and* $W_d \geq 6d + 2dL_d$*. Then, there is a constant* c*, such that,* $m \geq c\,(W_g L_g(\log W_g + L_g)) \vee (W_d L_d(\log W_d + L_d))$*, then,*

$$\log \mathcal{N}\left(\epsilon; (\Psi \circ G)_{|Z_{1:m}}, \ell_\infty\right) \lesssim \epsilon^{-\frac{d}{\beta}} + W_g L_g \left(\log W_g + L_g\right) \log\left(\frac{md}{\epsilon}\right),$$
$$\log \mathcal{N}\left(\epsilon; (\mathcal{F}_2)_{|Z_{1:m}}, \ell_\infty\right) \lesssim \epsilon^{-\frac{d+\ell}{\beta}} + W_d L_d \left(\log W_d + L_d\right) \log\left(\frac{md}{\epsilon}\right).$$

We note that though the above metric entropies depend exponentially on the data dimension d, this is not a problem in finding the generalization error as this exponential dependence is only in terms of m, the number of generated fake samples, which the practitioner can increase to tackle this curse of dimensionality. Using Corollary 25, we can bound stochastic errors of these function classes in Lemma 26.

Lemma 26. *Suppose that* $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ *and* $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ *with* $L_d, L_g \geq 3$*,* $W_g \geq 6d + 2dL_g$ *and* $W_d \geq 6d + 2dL_d$*. Then, there is a constant* c*, such that,* $m \geq c\,(W_g L_g(\log W_g + L_g)) \vee (W_d L_d(\log W_d + L_d))$*, then,*

$$\mathbb{E}\|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}} \lesssim m^{-\frac{\beta}{d}} \vee m^{-1/2} \log m + \sqrt{\frac{W_g L_g(\log W_g + L_g)\log(md)}{m}},$$
$$\mathbb{E}\|\hat{\nu}_m - \nu\|_{\mathcal{F}_2} \lesssim m^{-\frac{\beta}{d+\ell}} \vee (m^{-1/2} \log m) + \sqrt{\frac{W_d L_d(\log W_d + L_d)\log(md)}{m}}.$$

Here c *is the same as in Corollary 25.*

Next, we focus on deriving a uniform concentration bound w.r.t. the function class $\mathcal{F}_1$. This is a little trickier, compared to the ones derived in Lemma 26, in the sense that one does not have direct control over the ℓ_∞-metric entropy of the function class Ψ (and in turn, $\mathcal{F}_1$) in terms of the Wasserstein dimension of the data. We resolve the issue by performing a one-step discretization and appealing to the sub-Gaussian property of the data for a fixed $E \in \mathcal{E}$. The proof is detailed in Appendix E.4.

Lemma 27. *Suppose that* $\mathcal{E} = \mathcal{RN}(L_e, W_e, B_e, 2C_e)$*. Then, for any* $s > d^*_\beta(\mu)$*, we can find an* $n' \in \mathbb{N}$*, such that if* $n \geq n'$*,*

$$\mathbb{E}\|\hat{\mu}_n - \mu\|_{\mathcal{F}_1} \lesssim n^{-1/2}\left(W_e \log\left(2L_e B_e^{L_e}(W_e+1)^{L_e} n^{\frac{1}{2(\beta\wedge 1)}}\right)\right)^{1/2} + n^{-\frac{\beta}{s+\ell}}.$$

To complete the proof, the idea is to combine the pieces discussed throughout Sections 4.3.2 and 4.3.3. The goal is to choose the sizes of the neural network classes, i.e. the class of generators, encoders and decoders, such that both the misspecification errors and generalization errors are small. A large network size implies the misspecification error will be small enough but will make the bounds in Lemma 26 loose and vice versa. Since the network sizes are expressed in terms of ϵ_e, ϵ_g and ϵ_d in Lemmas 22 and 23, expressing them in terms of n in an optimal way that serves the purpose of finding a trade-off between the two errors to minimize the bound on the excess risk in Lemma 20. The detailed proofs are given in Appendix F.

5. Optimal Bounds for GANs with Interpolating Generators

In this section, we show that GAN estimates can (almost) achieve the minimax optimal rate for estimating distributions with a low intrinsic dimensional structure. To generalize the scenario further, we drop assumption A2 and only work with A1. We replace assumption A2 and assume that ν is an absolutely continuous distribution on $[0,1]^\ell$. This is satisfied by the uniform and normal distributions on the latent space, which are the commonly used choices in practice. It is important to note that lifting assumption A2, requires that the generator network have many more parameters than compared to the case when the target distribution is smooth (i.e. under A2), which is expected. Following a similar analysis as done in Section 4, we can arrive at the following theorem, which states that the generator network can be selected in such a way as to obtain a rate of convergence of roughly, $\mathcal{O}\left(n^{-\beta/d^*_\beta(\mu)}\right)$, with proof in Appendix G.1. The networks class in Theorem 28 is constructed so that some members of the class can (almost) linearly interpolate the data, as opposed to the Taylor series approximation used in the proof of Theorem 21. Compared to Theorem 16, the networks required in Theorem 28 are shallower, with a constant depth as opposed to a $\mathcal{O}(\log n)$ depth. However, one does not have any control over the maximum value of the weight of these nearly interpolating networks.

Theorem 28. *Suppose Assumption A1 holds and let* ν *be absolutely continuous on* $[0,1]^\ell$*. Then, if* $s > d^*_\beta(\mu)$*, one can choose* $\mathcal{G} = \mathcal{RN}(L_g, W_g, \infty, 1)$ *with the network parameters as* $L_g \geq 2$ *as a constant and* $W_g \asymp n$*, such that* $\mathbb{E}\|\mu - (\hat{G}_n)_\sharp\nu\|_\Phi \lesssim n^{-\beta/s}$*. Furthermore, if* $m \geq n^{d/s+1}$*, then* $\mathbb{E}\|\mu - (\hat{G}_{n,m})_\sharp\nu\|_\Phi \lesssim n^{-\beta/s}$*.*

To understand whether the bounds in Theorem 28 are optimal, we derive a minimax lower bound for the expected excess risk for the GAN problem for estimating distributions whose upper Wasserstein dimension is upper bounded by some constant. Assuming $d \geq 2\beta$, we fix a constant $d^\star \in [2\beta, d]$ and consider the family of probability measures,

$$\mathbb{M}^{d^\star,\beta} = \{\mu \in \Pi_{[0,1]^d} : d^*_\beta(\mu) \leq d^\star\},$$

the set of all distributions on $[0,1]^d$, whose upper Wasserstein dimension is at most $d^\star$. We note that this collection of distributions contains distributions on specific manifolds of dimension $d^\star$ or less. Here, we use the notation $\Pi_{\mathcal{A}}$ to denote the set of all probability measures on $\mathcal{A}$. The minimax expected risk for this problem is given by,

$$\mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \mathbb{M}^{d^\star,\beta}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)},$$

where the infimum is taken over all measurable estimates of μ, i.e. on $\{\hat{\mu} : X_{1:n} \to \Pi_{[0,1]^d} : \hat{\mu} \text{ is measurable}\}$. Here, we write $\mathbb{E}_\mu$ to denote that the expectation is taken with respect to the joint distribution of $X_1, \dots, X_n$, which are i.i.d. μ. $X_{1:n}$ denotes the data $X_1, \dots, X_n$. Theorem 29 characterizes this minimax rate, which states that the rate cannot be made any faster than roughly, $\mathcal{O}\left(n^{-\beta/d^\star}\right)$.

Theorem 29. *Suppose that $d \ge 2\beta$ and let $d^\star \in [2\beta, d]$. Then for any $s < d^\star$,*

$$\mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \mathbb{M}^{d^\star,\beta}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \gtrsim n^{-\beta/s},$$

where the infimum is taken over all measurable estimates of μ, based on the data.

By Theorem, 28, we observe that for any $\mu \in \mathbb{M}^{d^\star,\beta}$ and any $\overline{s} > d^\star$, $\mathbb{E}_\mu \|\hat{G}_\sharp \nu - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)} \lesssim n^{-\beta/\overline{s}}$. Thus, in $\mathbb{M}^{d^\star}$, the GAN estimator, $\hat{G}_\sharp \nu$, roughly achieves the optimal rate.

Suppose that $\mathbb{M}_{d_\star} = \{\mu : d_*(\mu) \le d_\star\}$, then it is clear that $\mathbb{M}^{d_\star,\beta} \subseteq \mathbb{M}_{d_\star}$ by Proposition 8. Hence, if $2\beta \le d_\star \le d$ and $\underline{s} < d_\star$ then,

$$\inf_{\hat{\mu}} \sup_{\mu \in \mathbb{M}_{d_\star}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \ge \inf_{\hat{\mu}} \sup_{\mu \in \mathbb{M}^{d_\star,\beta}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \gtrsim n^{-\beta/\underline{s}}$$

Following a similar proof as that of Theorem 29, one can do away with the condition of $d_\star \ge 2\beta$ and arrive at the following theorem.

Theorem 30. *Suppose $d_\star \in [0, d]$. Then for any $s < d_\star$,*

$$\mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \mathbb{M}_{d_\star}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \gtrsim n^{-\beta/s} \vee n^{-1/2},$$

where the infimum is taken over all measurable estimates of μ, based on the data.

It is worth observing that for the special case of $\beta = 1$, i.e. under the Wasserstein-1 distance, this rate matches with the ones derived in literature (Weed and Berthet, 2022; Singh and Póczos, 2018), i.e. $\mathcal{O}(n^{-1/d})$.

6. Discussions and Conclusion

This paper delves into the theoretical properties of GANs and Bi-directional GANs when the data exhibits an intrinsically low-dimensional structure within a high-dimensional feature space. We propose to characterize the low-dimensional nature of the data distribution through its upper Wasserstein dimension and the so-called α-entropic dimension, which we develop by extending Dudley's notion of entropic dimension (Dudley, 1969). Specifically, we not only show that the classical result by Kolmogorov and Tikhomirov (1961) can be strengthened by incorporating this entropic dimension in the metric entropy but also show that the convergence rates of the empirical distribution to the target population in the β-Hölder IPM scales as $\mathcal{O}\left(n^{-\beta/d^*_\beta(\mu)}\right)$, extending the results by Weed and Bach (2019). Furthermore, we improve upon the existing results on the approximation capabilities of ReLU networks. By balancing the generalization gap and approximation errors, we establish that under the assumption that the true generator and encoders are Hölder continuous, the excess risk in terms of an Hölder IPM can be bounded in terms of the ambient upper Wasserstein dimension of the target measure and the latent space. This bypasses the curse of dimensionality of the full data space, strengthens the known results in GANs, and leads to novel bounds for BiGANs. The derived results also match the sharp convergence rates for the empirical distribution available in the optimal transport literature. We also show that the GAN estimate of the target distribution can roughly achieve the minimax optimal rates for estimating intrinsically low-dimensional distributions.

While our results provide insights into the theoretical properties of GANs and Bidirectional GANs, it is important to acknowledge that estimating the full error of these generative models used in practice involves considering an optimization error term. Unfortunately, accurately estimating this term remains a challenging task due to the non-convex and complex nature of the minimax optimization. However, it is worth pointing out that our error analyses are independent of optimization and can be seamlessly combined with optimization analyses. An additional area that we have not addressed in this paper is that the β-Hölder discriminators are typically also realized by ReLU networks in practice. While this is an important consideration, it requires imposing restrictions on the networks to closely resemble smooth functions, which is typically achieved through various regularization techniques. However, ensuring that these regularization techniques provably lead to well-behaved discriminators remains a challenging task from a theoretical perspective. We believe that this is a promising avenue for future research. One can also attempt to extend our approximation results to more general functions representable on some other basis than non-smooth functions such as Besov spaces or Fourier series approximations, for instance using ideas of Suzuki (2019) or Bresler and Nagaraj (2020).

Acknowledgments

We gratefully acknowledge the support of the NSF and the Simons Foundation for the Collaboration on the Theoretical Foundations of Deep Learning through awards DMS-2031883 and #814639 and the NSF's support of FODSI through grant DMS-2023505.

Appendices

Appendix A. Proofs from Section 3

This section provides the proofs of the results from Section 3. For notational simplicity, we write $\text{fat}(S, \epsilon) = \{y : \inf_{x \in S} \varrho(x, y) \le \epsilon\}$ for the ϵ-fattening of the set S, w.r.t. the metric ϱ.

A.1 Proof of Proposition 8

Before we prove Proposition 8, we first show the following key lemma.

Lemma 31. *For any* $0 < \tau_1 \le \tau_2 < 1$, $\mathcal{N}_\epsilon(\mu, \tau_1) \ge \mathcal{N}_\epsilon(\mu, \tau_2)$.

Proof It is easy to see that $\{S : \mu(S) \geq 1 - \tau_1\} \subseteq \{S : \mu(S) \geq 1 - \tau_2\}$. Hence, $\mathcal{N}_\epsilon(\mu, \tau_1) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : \mu(S) \geq 1 - \tau_1\} \geq \inf\{\mathcal{N}(\epsilon; S, \varrho) : \mu(S) \geq 1 - \tau_2\} = \mathcal{N}_\epsilon(\mu, \tau_2)$. ■

We are now ready to prove Proposition 8 as shown below.

Proposition 8. *For any μ and for any $\alpha \geq 0$, the following hold:*

(a) $d_*(\mu) \leq \bar{d}_\alpha(\mu) \leq d^*_\alpha(\mu)$,

(b) *if* $\alpha_1 \leq \alpha_2$, *then,* $\bar{d}_{\alpha_1}(\mu) \leq \bar{d}_{\alpha_2}(\mu)$,

(c) $\bar{d}_\alpha(\mu) \leq \overline{\dim}_M(\mu)$,

(d) *if* $\alpha \in \big(0, \overline{\dim}_P(\mu)/2\big)$, $d^*_\alpha(\mu) \leq \overline{\dim}_P(\mu) \leq \overline{\dim}_{reg}(\mu)$,

(e) $\bar{d}_\alpha(\mu) \leq \overline{\dim}_P(\mu) \leq \overline{\dim}_{reg}(\mu)$,

(f) $\underline{\dim}_{reg}(\mu) \leq d_*(\mu)$.

Proof **Proof of part (a)**: Suppose that

$$\mathcal{A} = \left\{ s \in (2\alpha, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \leq s \right\}.$$

Fix $s \in \mathcal{A}$. By definition, $s > 2\alpha$. Since for $\epsilon < 1$, $\epsilon^\alpha \geq \epsilon^{\frac{s\alpha}{s-2\alpha}}$. Thus, by Lemma 31, $\mathcal{N}_\epsilon(\mu, \epsilon^\alpha) \leq \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)$, when $\epsilon < 1$. Fix $\tau \in (0, 1)$. Hence,

$$\bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{-\log \epsilon} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}})}{-\log \epsilon} \leq s.$$

Taking infimum over $s \in \mathcal{A}$ gives us, $\bar{d}_\alpha(\mu) \leq d^*_\alpha(\mu)$.

To observe the other implication, we first note that for any fixed $\tau \in (0, 1)$, $\epsilon^\alpha \leq \tau$, if $\epsilon \leq \tau^{1/\alpha}$. Thus, by Lemma 31, $\mathcal{N}_\epsilon(\mu, \epsilon^\alpha) \geq \mathcal{N}_\epsilon(\mu, \tau)$, if $\epsilon \leq \tau^{1/\alpha}$. Hence,

$$\bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{-\log \epsilon} \geq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \tau)}{-\log \epsilon} \geq \liminf_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \tau)}{-\log \epsilon}.$$

Taking limit on both sides as $\tau \downarrow 0$, gives us, $\bar{d}_\alpha(\mu) \geq d_*(\mu)$.

Proof of part (b): If $\alpha_1 \leq \alpha_2$, then, we observe that $\mathcal{N}_\epsilon(\mu, \epsilon^{\alpha_1}) \leq \mathcal{N}_\epsilon(\mu, \epsilon^{\alpha_2})$, for $\epsilon < 1$. Thus, $\bar{d}_{\alpha_1}(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^{\alpha_1})}{-\log \epsilon} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^{\alpha_2})}{-\log \epsilon} = \bar{d}_{\alpha_2}(\mu)$.

Proof of part (c): For any $\alpha > 0$, $\mathcal{N}_\epsilon(\mu, \epsilon^\alpha) \leq \mathcal{N}_\epsilon(\mu, 0) = \mathcal{N}(\epsilon; \mathrm{supp}(\mu), \varrho)$. Thus taking $\limsup$ as $\epsilon \downarrow 0$ gives us the result.

Proof of part (d) Suppose that $\mathcal{A} = \left\{ s \in (2\alpha, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \leq s \right\}$. Let $0 < \epsilon < 1$, $s > \overline{\dim}_P(\mu)$ and $\tau = \epsilon^{\frac{s\alpha}{s-2\alpha}}$. S be such that $\mu(S) \geq 1 - \tau$ and $\mathcal{N}(\epsilon; S, \varrho) = \mathcal{N}_\epsilon(\mu, \tau)$. We let $R = \mathrm{diam}(S) \vee 1$. Let $\{x_1, \dots, x_M\}$ be an optimal 2ϵ-packing of $S \cap \mathrm{supp}(\mu)$.

By the definition of the upper packing dimension, for any $s > \overline{\dim}_P(\mu)$ we can find $r_0 < 1$, such that,

$$\frac{\log \mu(B(x,r))}{\log r} \le s, \, \forall r \le r_0 \text{ and } x \in \text{supp}(\mu)$$
$$\implies \mu(B(x,r)) \ge r^s, \, \forall r \le r_0 \text{ and } x \in \text{supp}(\mu).$$

Thus, if $\epsilon \le r_0$, $1 \ge \mu\left(\cup_{i=1}^M B(x_i,\epsilon)\right) = \sum_{i=1}^M \mu\left(B(x_i,\epsilon)\right) \ge M\epsilon^s \implies M \le \epsilon^{-s}$. By Lemma 51, we know that $\mathcal{N}_\epsilon(\mu,\tau) = \mathcal{N}(\epsilon; S, \varrho) \le M \le \epsilon^{-s}$. Thus,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{-\log \epsilon} \le s \implies s \in \mathcal{A} \implies d^*_\alpha(\mu) \le s.$$

Since $d^*_\alpha(\mu) \le s$, for all $s > \overline{\dim}_P(\mu)$, we get, $d^*_\alpha(\mu) \le \overline{\dim}_P(\mu)$. The inequality $\overline{\dim}_P(\mu) \le \overline{\dim}_{\text{reg}}(\mu)$ follows from Fraser and Howroyd (2017, Theorem 2.1).

Proof of part (e):

Let $0 < \epsilon < 1$, $s > \overline{\dim}_P(\mu)$ and $\tau = \epsilon^\alpha$. S be such that $\mu(S) \ge 1 - \tau$ and $\mathcal{N}(\epsilon; S, \varrho) = \mathcal{N}_\epsilon(\mu,\tau)$. We let $R = \text{diam}(S) \vee 1$. Let $\{x_1, \dots, x_M\}$ be an optimal 2ϵ-packing of $S \cap \text{supp}(\mu)$. By the definition of the upper packing dimension, for any $s > \overline{\dim}_P(\mu)$ we can find $r_0 < 1$, such that, $\frac{\log \mu(B(x,r))}{\log r} \le s$, $\forall r \le r_0$ and $x \in \text{supp}(\mu) \implies \mu(B(x,r)) \ge r^s$, $\forall r \le r_0$ and $x \in \text{supp}(\mu)$. Thus, if $\epsilon \le r_0$, $1 \ge \mu\left(\cup_{i=1}^M B(x_i,\epsilon)\right) = \sum_{i=1}^M \mu\left(B(x_i,\epsilon)\right) \ge M\epsilon^s \implies M \le \epsilon^{-s}$. By Lemma 51, we know that $\mathcal{N}_\epsilon(\mu,\tau) = \mathcal{N}(\epsilon; S, \varrho) \le M \le \epsilon^{-s}$. Thus,

$$\bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^\alpha\right)}{-\log \epsilon} \le s.$$

Since $\bar{d}_\alpha(\mu) \le s$, for all $s > \overline{\dim}_P(\mu)$, we get, $\bar{d}_\alpha(\mu) \le \overline{\dim}_P(\mu)$. The inequality $\overline{\dim}_P(\mu) \le \overline{\dim}_{\text{reg}}(\mu)$ follows from Theorem 2.1 of Fraser and Howroyd (2017).

Proof of part (f): Let $0 < \epsilon < 1$ and S be such that $\mu(S) \ge 1 - \tau$ and $\mathcal{N}(\epsilon; S, \varrho) = \mathcal{N}_\epsilon(\mu,\tau)$. We let $R = \text{diam}(S) \vee 1$. Let $\{x_1, \dots, x_N\}$ be an optimal ϵ-net of $S \cap \text{supp}(\mu)$. By the definition of the upper regularity dimension, for any $s < \overline{\dim}_{\text{reg}}(\mu)$ we observe that, $\mu(B(x_i,\epsilon)) \le \frac{\mu(B(x_i,R))}{CR^s}\epsilon^s \le \frac{(1-\tau)}{CR^s}\epsilon^s$. We observe that, $\mu(\cup_{i=1}^N B(x_i,\epsilon)) \ge \mu(S) \ge 1 - \tau \implies \sum_{i=1}^N \mu(B(x_i,\epsilon)) \ge 1 - \tau$. This implies that,

$$N\epsilon^s \frac{1-\tau}{CR^s} \ge \sum_{i=1}^N \mu(B(x_i,\epsilon)) \ge 1 - \tau \implies N \ge CR^s\epsilon^{-s}.$$

Thus, $\mathcal{N}(\epsilon; S, \varrho) \ge CR^s(2\epsilon)^{-s} \implies \liminf_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu,\tau)}{-\log \epsilon} \ge s$, which implies that $d_*(\mu) = \lim_{\tau \downarrow 0} \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu,\tau)}{-\log \epsilon} \ge s$. Hence, $d_*(\mu) \ge s$, for any $s < \overline{\dim}_{\text{reg}}(\mu)$, which gives us the desired result. ■

A.2 Proof of Theorem 13

Next, we give a proof of Theorem 13. Before proceeding, we consider the following lemma.

Lemma 32. *For any $s > \bar{d}_\alpha(\mu)$, we can find an $\epsilon' \in (0,1)$, such that if $\epsilon \in (0, \epsilon']$ and a set S, such that $\mu(S) \geq 1 - \epsilon^\alpha$ and $\mathcal{N}(\epsilon; S, \varrho) \leq \epsilon^{-s}$.*

Proof By definition of $\bar{d}_\alpha(\mu) = \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{\log(1/\epsilon)}$, if $s > \bar{d}_\alpha(\mu)$, we can find an $\epsilon' \in (0,1)$, such that if $\epsilon \in (0, \epsilon']$, $\frac{\log \mathcal{N}_\epsilon(\mu, \epsilon^\alpha)}{\log(1/\epsilon)} \leq s$, which implies that $\mathcal{N}_\epsilon(\mu, \epsilon^\alpha) \leq \epsilon^{-s}$. By definition of $\mathcal{N}_\epsilon(\mu, \cdot)$, we can find S that satisfies the conditions of the Lemma. ■

By construction, for $s > \bar{d}_{p\beta}(\mu)$, we can find $\epsilon' > 0$, such that if $\epsilon \in (0, \epsilon']$, we can find a bounded $S \subset \mathbb{R}^d$, such that $\mathcal{N}(\epsilon; S, \ell_\infty) \leq \epsilon^{-s}$ and $\mu(S) \geq 1 - \epsilon^{p\beta}$. We fix such an $\epsilon \in (0, \epsilon']$.

Let $M = \sup_{x \in S} \|x\|_\infty$ and $K = \lceil \frac{M}{\epsilon} \rceil$. For any $\boldsymbol{i} \in [K]^d$, let $\theta^{\boldsymbol{i}} = (-M + i_1 \epsilon, \dots, -M + i_d \epsilon)$. We also let, $\mathcal{P}_\epsilon = \{B_{\ell_\infty}(\theta^{\boldsymbol{i}}, \epsilon) : \boldsymbol{i} \in [K]^d\}$. By construction, the sets in $\mathcal{P}_\epsilon$ are disjoint. We first claim the following:

Lemma 33. $|\{A \in \mathcal{P}_\epsilon : A \cap S \neq \emptyset\}| \leq 2^d \epsilon^{-s}$.

Proof Let, $r = \mathcal{N}(\epsilon; S, \ell_\infty)$ and suppose that $\{a_1, \dots, a_r\}$ be an ϵ-net of S and $\mathcal{P}_\epsilon^* = \{B_{\ell_\infty}(a_i, \epsilon) : i \in [r]\}$ be an optimal ϵ-cover of S. Note that each box in $\mathcal{P}_\epsilon^*$ can intersect at most 2^d boxes in $\mathcal{P}_\epsilon$. This implies that, $|\mathcal{P}_\epsilon \cap S| \leq |\mathcal{P}_\epsilon \cap (\cup_{i=1}^r B_{\ell_\infty}(a_i, \epsilon))| = |\cup_{i=1}^r (\mathcal{P}_\epsilon \cap B_{\ell_\infty}(a_i, \epsilon))| \leq 2^d r$, which concludes the proof. ■

Theorem 13. *Let $\mathcal{F} = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$. Then, if $s > \bar{d}_{p\beta}(\mu)$, then, there exists an $\epsilon' > 0$, such that, if $0 < \epsilon \leq \epsilon'$, then, $\log \mathcal{N}(\epsilon; \mathcal{F}, \mathbb{L}_p(\mu)) \lesssim \epsilon^{-s/\beta}$.*

Proof Let $\mathcal{I} = \{\boldsymbol{i} \in [k]^d : B_{\ell_\infty}(\theta^{\boldsymbol{i}}, \epsilon) \in \mathcal{P}_\epsilon \text{ and } B_{\ell_\infty}(\theta^{\boldsymbol{i}}, \epsilon) \cap S \neq \emptyset\}$. For any $x \in \mathbb{R}^d$, let, $h_{x,r}(y) = \mathbb{1}\{y \in B_{\ell_\infty}(x, r)\}$. Fix any $f \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$. We define, $P_\theta^f(x) = \sum_{|\boldsymbol{s}| \leq \lfloor \beta \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta)}{\boldsymbol{s}!} (x - \theta)^{\boldsymbol{s}}$. Since, $|\partial^{\boldsymbol{s}} f(\theta)| \leq C$, we can define a δ-net, U_δ, of $[-C, C]$ of size at most C/δ. We define the function class

$$\mathscr{F} = \left\{ \sum_{\boldsymbol{i} \in \mathcal{I}} \sum_{|\boldsymbol{s}| \leq \lfloor \beta \rfloor} \frac{b_{\boldsymbol{s}, \theta^{\boldsymbol{i}}}}{\boldsymbol{s}!} (x - \theta^{\boldsymbol{i}})^{\boldsymbol{s}} h_{\theta^{\boldsymbol{i}}, \epsilon}(x) : b_{\boldsymbol{s}, \theta^{\boldsymbol{i}}} \in U_\delta \right\}.$$

Clearly, for any $f \in \mathscr{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, we can find, $\hat{f} \in \mathscr{F}$, such that,

$$\hat{f} = \sum_{\boldsymbol{i} \in \mathcal{I}} \sum_{|\boldsymbol{s}| \leq \lfloor \beta \rfloor} \frac{b_{\boldsymbol{s}, \theta^{\boldsymbol{i}}}}{\boldsymbol{s}!} (x - \theta^{\boldsymbol{i}})^{\boldsymbol{s}} h_{\theta^{\boldsymbol{i}}, \epsilon}(x),$$

with $|b_{\boldsymbol{s}, \theta^{\boldsymbol{i}}} - \partial^{\boldsymbol{s}} f(\theta^{\boldsymbol{i}})| \leq \delta$. Thus, for any $x \in \text{fat}(S, \epsilon)$, let θ be such that $\|\theta - x\| \leq \epsilon$ and $\theta \in \{\theta^{\boldsymbol{i}} : \boldsymbol{i} \in \mathcal{I}\}$.

$$|f(x) - \hat{f}(x)|$$

$$
\begin{aligned}
&\leq \left|\sum_{\boldsymbol{i}\in\mathcal{I}}\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{b_{\boldsymbol{s},\theta^{\boldsymbol{i}}}}{\boldsymbol{s}!}(x-\theta^{\boldsymbol{i}})^{\boldsymbol{s}}h_{\theta^{\boldsymbol{i}},\epsilon}(x)-f(x)\right|\\
&=\left|\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{b_{\boldsymbol{s},\theta}}{\boldsymbol{s}!}(x-\theta^{\boldsymbol{i}})^{\boldsymbol{s}}-f(x)\right|\\
&\leq\left|\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{\partial^{\boldsymbol{s}}f(\theta)}{\boldsymbol{s}!}(x-\theta)^{\boldsymbol{s}}-f(x)\right|+\left|\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{b_{\boldsymbol{s},\theta}}{\boldsymbol{s}!}(x-\theta)^{\boldsymbol{s}}-\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{\partial^{\boldsymbol{s}}f(\theta)}{\boldsymbol{s}!}(x-\theta)^{\boldsymbol{s}}\right|\\
&\leq\left|\sum_{|\boldsymbol{s}|=\lfloor\beta\rfloor}\frac{(x-\theta)^{\boldsymbol{s}}}{\boldsymbol{s}!}(\partial^{\boldsymbol{s}}f(x)-\partial^{\boldsymbol{s}}f(\theta'))\right|+\left|\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{|b_{\boldsymbol{s},\theta}-\partial^{\boldsymbol{s}}f(\theta)|}{\boldsymbol{s}!}(x-\theta)^{\boldsymbol{s}}\right|\\
&\leq\sum_{|\boldsymbol{s}|=\lfloor\beta\rfloor}\frac{\|x-\theta\|_\infty^{\boldsymbol{s}}}{\boldsymbol{s}!}\|x-\theta\|_\infty^{\beta-\lfloor\beta\rfloor}+\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{\delta}{\boldsymbol{s}!}\\
&\leq\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{\epsilon^\beta}{\boldsymbol{s}!}+\sum_{|\boldsymbol{s}|\leq\lfloor\beta\rfloor}\frac{\delta}{\boldsymbol{s}!}\\
&\lesssim\epsilon^\beta+\delta
\end{aligned}
$$

In the above calculations, θ' lies on the line-segment joining x and θ. We take $\delta=\epsilon^\beta$. It is easy to see that $|\mathcal{F}|\leq\left(\frac{C}{\delta}\right)^{|\mathcal{I}|\lfloor\beta\rfloor^d}$. By definition $\hat{f}(x)=0$ if $x\not\in\text{fat}(S,\epsilon)$. Thus,

$$
\begin{aligned}
\int|f(x)-\hat{f}(x)|^p d\mu(x)&=\int_{\text{fat}(S,\epsilon)}|f(x)-\hat{f}(x)|^p d\mu(x)+\int_{\text{fat}(S,\epsilon)^\complement}|f(x)-\hat{f}(x)|^p d\mu(x)\\
&\leq\sup_{x\in\text{fat}(S,\epsilon)}|f(x)-\hat{f}(x)|^p+\int_{\text{fat}(S,\epsilon)^\complement}|f(x)|^p d\mu(x)\\
&\lesssim\epsilon^{p\beta}+\mu(\text{fat}(S,\epsilon)^\complement)\\
&\leq\epsilon^{p\beta}+\mu(S^\complement)\\
&\leq\epsilon^{p\beta}+\epsilon^{p\beta}\\
&\lesssim\epsilon^{p\beta}.
\end{aligned}
$$

Clearly, $\mathcal{F}$ forms a $c\epsilon^\beta$-cover of $\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)$ in the $\mathbb{L}_p(\mu)$-norm, for some constant $c>0$. Thus,

$$
\log\mathcal{N}\left(c\epsilon^\beta;\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C),\|\cdot\|_{\mathbb{L}_p(\mu)}\right)\leq|\mathcal{I}|\lfloor\beta\rfloor^d\log(C/\delta)\lesssim\epsilon^{-s}\log(1/\epsilon).
$$

Replacing ϵ with $(\epsilon/c)^{1/\beta}$ gives us,

$$
\log\mathcal{N}\left(\epsilon;\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C),\|\cdot\|_{\mathbb{L}_p(\mu)}\right)\lesssim\epsilon^{-s/\beta}\log(1/\epsilon)\leq\epsilon^{-s'/\beta},
$$

for any $s'>s$. Now rewriting s for s' gives us the desired result. ■

A.3 Proof of Theorem 14

We will prove a more general version of Theorem 14. We begin by generalizing the notion of (upper) Wasserstein dimension for a class of distributions. Suppose that Λ is a family of distributions on $[0,1]^d$. For any $\tau \in [0,1]$, we define,

$$\mathcal{N}_\epsilon(\Lambda, \tau) = \inf\{\mathcal{N}(\epsilon; S, \varrho) : \mu(S) \geq 1 - \tau, \, \forall \mu \in \Lambda\}.$$

The interpretation of $\mathcal{N}_\epsilon(\cdot, \tau)$ is the same as that for the case of a single measure, defined in Section 3. We define the upper Wasserstein dimension of this family as:

$$d_\alpha^*(\Lambda) = \inf\left\{ s \in (2\alpha, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \leq s \right\}.$$

The proof of Theorem 14 requires some supporting lemmas. We sequentially state and prove these lemmas as we proceed. We first prove in Lemma 34 that the set $\mathcal{A}$ (defined below) takes the shape of an interval (left open or closed).

Lemma 34. *Suppose that* $\mathcal{A} = \left\{ s \in (2\alpha, \infty) : \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \leq s \right\}$. *Then,* $\mathcal{A} \supseteq (d_\alpha^*(\mu), \infty)$.

Proof We begin by claiming the following:

$$\textbf{Claim}: \text{ If } s_1 \in \mathcal{A} \text{ then } s_2 \in \mathcal{A}, \text{ for all } s_2 \geq s_1. \tag{19}$$

To observe this, we note that, if $s_2 \geq s_1 > 2\alpha$ and $\epsilon \in (0,1)$,

$$\frac{s_1\alpha}{s_1 - 2\alpha} \geq \frac{s_2\alpha}{s_2 - 2\alpha} \implies \epsilon^{\frac{s_1\alpha}{s_1-2\alpha}} \leq \epsilon^{\frac{s_2\alpha}{s_2-2\alpha}} \implies \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s_1\alpha}{s_1-2\alpha}}\right) \geq \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s_2\alpha}{s_2-2\alpha}}\right).$$

Here the last implication follows from Lemma 31. Thus,

$$\limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s_2\alpha}{s_2-2\alpha}}\right)}{\log(1/\epsilon)} \leq \limsup_{\epsilon \downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s_1\alpha}{s_1-2\alpha}}\right)}{\log(1/\epsilon)} \leq s_1 \leq s_2.$$

Hence, $s_2 \in \mathcal{A}$.

Let $s > d_\alpha^*(\Lambda)$, then by definition of infimum, we note that we can find $s' \in [d_\alpha^*(\Lambda), s)$, such that, $s' \in \mathcal{A}$. Since, $s > s' \in \mathcal{A}$, by Claim (19), $s \in \mathcal{A}$. Thus, for any $s > d_\alpha^*(\Lambda)$, $s \in \mathcal{A}$, which proves the lemma.

■

An immediate corollary of Lemma 34 is as follows.

Corollary 35. *Let* $s > d_\alpha^*(\Lambda)$. *Then, there exists* $\epsilon' \in (0,1]$, *such that if* $0 < \epsilon \leq \epsilon'$, *then, there exists a set* S, *such that* $\mathcal{N}(\epsilon; S, \varrho) \leq \epsilon^{-s}$ *and* $\mu(S) \geq 1 - \epsilon^{\frac{s\alpha}{s-2\alpha}}$, *for all* $\mu \in \Lambda$.

Proof Let $\delta = s - d^*_\alpha(\mu)$. By Lemma 34, we observe that $s' = d^*_\alpha(\mu) + \delta/2 \in \mathcal{A}$. Now by definition of $\limsup$, one can find a $\epsilon' > 0$, such that, $\frac{\log \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s'\alpha}{s'-2\alpha}}\right)}{\log(1/\epsilon)} \le s' + \delta/2 = s$, for all $\epsilon \in (0, \epsilon']$. The result now follows from observing that $\mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right) \le \mathcal{N}_\epsilon\left(\mu, \epsilon^{\frac{s'\alpha}{s'-2\alpha}}\right) \le \epsilon^{-s}$. ■

We suppose that $\delta = d^\star - d^*_\beta(\mu)$ and let $d' = d^*_\beta(\mu) + \delta/2$. By Corollary 35, we can find an $\epsilon' \in (0,1]$, such that if $\epsilon \in (0, \epsilon']$, we can find S_ϵ, such that $\mathcal{N}(\epsilon; S_\epsilon, \ell_\infty) \le \epsilon^{-d'}$ and $\mu(S_\epsilon) \ge 1 - \epsilon^{\frac{d'\beta}{d'-2\beta}}$, for all $\mu \in \Lambda$. As a first step of constructing a dyadic-like partition of $[0,1]^d$, we state and prove the following lemma that helps us create the base of this sequential partitioning of the data space. For notational simplicity, we use $\operatorname{diam}(A) = \sup_{x,y \in A} \varrho(x,y)$ to denote the diameter of a set w.r.t. the metric ϱ.

Lemma 36. *For any $r \ge \lceil \log_3(1/\epsilon') - 2 \rceil$, we can find disjoint sets $S_{r,0}, \dots, S_{r,m_r}$, such that, $\cup_{j=0}^{m_r} S_{r,j} = \mathbb{R}^d$. Furthermore, $m_r \le 3^{d'(r+2)}$, $\operatorname{diam}(S_{r,j}) \le 3^{-(r+1)}$, for all $j = 1, \dots, m_r$ and $\mu(S_{r,0}) \le 3^{-\frac{d'(r+2)\beta}{d'-2\beta}}$ $\forall \mu \in \Lambda$.*

Proof We take $\epsilon = 3^{-(r+2)}$. Clearly, $0 < \epsilon \le \epsilon'$. We take $S_{r,0} = S_\epsilon^\complement$. By definition of covering numbers, we can find a minimal ϵ-net $\{x_1, \dots, x_{m_r}\}$, such that $S \subseteq \cup_{j=1}^{m_r} B_{\ell_\infty}(x_i, \epsilon)$ and $m_r \le \epsilon^{-d'} = 3^{d'(r+2)}$. We construct $S_{r,1}, \dots S_{r,m_r}$ as follows:

- Take $S_{r,1} = B_{\ell_\infty}(x_1, \epsilon) \setminus S_{r,0}$.
- For any $j = 2, \dots, m_r$, we take $S_{r,j} = B_{\ell_\infty}(x_j, \epsilon) \setminus \left(\cup_{j'=0}^{j-1} S_{r,j'}\right)$.

By construction $\{S_{r,j}\}_{j=0}^{m_r}$ are disjoint. Moreover, $\mu(S_{r,0}) = 1 - \mu(S_\epsilon) \le \epsilon^{\frac{d'\beta}{d'-2\beta}} = 3^{-\frac{d'(r+2)\beta}{d'-2\beta}}$, for all $\mu \in \Lambda$. Furthermore since, $S_{r,j} \subseteq B_{\ell_\infty}(x_j, \epsilon)$,

$$\operatorname{diam}(S_{r,j}) \le \operatorname{diam}(B_{\ell_\infty}(x_j, \epsilon)) = 2\epsilon = 2 \times 3^{-(r+2)} \le 3^{-(r+1)}.$$

■

We now construct a sequence of collection of sets $\{\mathcal{Q}^\ell\}_{\ell=1}^r$ as follows:

- Take $\mathcal{Q}^r = \{S_{r,j}\}_{j=1}^{m_r}$.
- Given $\ell + 1$, let, $Q_1^\ell = \bigcup_{\substack{Q \in \mathcal{Q}^{\ell+1}, \\ Q \cap S_{\ell,1} \neq \emptyset}} (Q \setminus S_{\ell,0})$ and if $2 \le j \le m_\ell$, we let,

$$Q_j^\ell = \left(\bigcup_{\substack{Q \in \mathcal{Q}^{\ell+1}, \\ Q \cap S_{r,j} \neq \emptyset}} (Q \setminus S_{\ell,0})\right) \setminus \left(\cup_{j'=1}^{j-1} Q_{j'}^\ell\right).$$

  Take $\mathcal{Q}^\ell = \{Q_j^\ell\}_{j=1}^{m_\ell}$.

Clearly for any $Q \in \mathcal{Q}^\ell$, $\sup_{Q \in \mathcal{Q}^\ell} \operatorname{diam}(Q) \le 2 \sup_{Q' \in \mathcal{Q}^{\ell+1}} \operatorname{diam}(Q') + \sup_{1 \le j \le m_\ell} \operatorname{diam}(S_{\ell,j}) \le 3 \times 3^{-(\ell+1)} = 3^{-\ell}$, by induction.

Also, if $Q \in \mathcal{Q}^{\ell+1}$, we can find a $Q' \in \mathcal{Q}^\ell$, such that $Q \subseteq Q' \cup S_{\ell,0}$. Also if $Q_1, Q_2 \in \mathcal{Q}^\ell$, $Q_1 \cap Q_2 = \emptyset$. Furthermore, $|\mathcal{Q}^\ell| \le m_\ell$, by construction.

Let $\epsilon = n^{-1/d'}$. Recall that, $\delta = d^\star - d^*_\beta(\mu)$ and $d' = d^\star_\beta(\mu) + \delta/2$. We take n_0, large enough such that $\epsilon \le \epsilon'$, if $n \ge n_0$. Let t be the smallest integer such that $3^{-t} \le \epsilon$. Also, let s be the smallest integer such that $3^{-s} \le \epsilon^{\frac{d'-2\beta}{d'}}$. Clearly, $s \le t$. Also, $3^t \le \frac{3}{\epsilon}$ and $3^s \le 3\epsilon^{\frac{2\beta-d'}{d'}}$.

Let $P^f_\theta(x) = \operatorname{Clip}\left(\sum_{|\boldsymbol{s}| \le \lfloor \beta \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta)}{\boldsymbol{s}!}(x-\theta)^{\boldsymbol{s}}, -C, C\right)$, where $\operatorname{Clip}(x,a,b) = (x \wedge b) \vee a$. We choose a point $x_{r,j} \in Q^r_j$, $s \le r \le t$ and $j \in [m_r]$. We also define the bivariate function $q(\cdot,\cdot)$ as $q(\ell,j) = j'$ if $Q^\ell_j \cap Q^{\ell-1}_{j'} \ne \emptyset$. This j' is unique, since we can find a $Q^{\ell-1}_{j'}$ which contains Q^ℓ_j and since sets in $\mathcal{Q}^{\ell-1}$ are disjoint.

We denote, $S_{s:t,0} = \cup_{r=s}^t S_{r,0}$ and let, $M_r = \sum_{j=1}^{m_r} |(\mu_n - \mu)(Q^r_j)|$. The following result (Lemma 37) helps us control $\mathbb{E} M_r$.

Lemma 37. *Suppose that $\{A_1, \dots, A_m\}$ forms a partition of a set T. Then,*

$$\mathbb{E} M_r = \mathbb{E} \sum_{j=1}^m |\mu_n(A_j) - \mu(A_j)| \le \sqrt{\frac{m\mu(T)}{n}}.$$

Proof We begin by noting that $n\mu_n(A_j) \sim \operatorname{Bin}(n, \mu(A_j))$. Thus, $\operatorname{Var}(\mu_n(A_j)) = \frac{\mu(A_j)(1-\mu(A_j))}{n}$. Hence,

$$\begin{aligned}
\mathbb{E} \sum_{j=1}^m |\mu_n(A_j) - \mu(A_j)| \overset{(i)}{\le} & \sqrt{m} \left(\mathbb{E} \sum_{j=1}^m (\mu_n(A_j) - \mu(A_j))^2 \right)^{1/2} \\
= & \sqrt{m} \left(\sum_{j=1}^m \operatorname{Var}(\mu_n(A_j)) \right)^{1/2} \\
= & \sqrt{m} \left(\sum_{j=1}^m \frac{\mu(A_j)(1-\mu(A_j))}{n} \right) \\
\le & \sqrt{m} \sqrt{\sum_{j=1}^m \frac{\mu(A_j)}{n}} \\
= & \sqrt{\frac{m\mu(T)}{n}}.
\end{aligned}$$

In the above calculations, (i) follows from Cauchy-Schwartz inequality. ■

Lemma 38. *Suppose that Poly(k,d,α) denotes the set of all polynomials from $[0,1]^d \to \mathbb{R}$, with degree at most k and absolutely bounded by α. Also let $A_1, \dots, A_m$ are disjoint and suppose that $\mathfrak{P} = \{f = \sum_{j=1}^m f_j \mathbb{1}_{A_j} : f_j \in$ Poly$(k,d,\alpha)\}$. Then,*

$$\gamma(\alpha, m, n) := \mathbb{E} \sup_{f \in \mathfrak{P}} \left(\int f d\mu_n - \int f d\mu \right) \lesssim \alpha \sqrt{\frac{m(k+1)^d \log n}{n}}.$$

Proof It is easy to note that, $\mathrm{Pdim}(\mathfrak{P}) \le m(k+1)^d$. Here $\mathrm{Pdim}(\mathcal{F})$ denotes the pseudo-dimension of the real-valued function class $\mathcal{F}$ (see Anthony and Bartlett (2009, Definition 11.2)). Applying Dudley's chaining, we recall that,

$$\begin{aligned}
\mathbb{E}_\sigma \sup_{f \in \mathfrak{P}} \sum_{i=1}^n \sigma_i f(x_i) \lesssim & \int_0^\alpha \sqrt{\frac{\log \mathcal{N}(\epsilon; \mathfrak{P}, \|\cdot\|_{\mathbb{L}_2(\mu_n)})}{n}} d\epsilon \\
\le & \int_0^\alpha \sqrt{\frac{\log \mathcal{N}(\epsilon; \mathfrak{P}, \|\cdot\|_{\mathbb{L}_\infty(\mu_n)})}{n}} d\epsilon \\
\le & \int_0^\alpha \sqrt{\frac{\mathrm{Pdim}(\mathfrak{P})}{n} \log(2\alpha e n/\epsilon)} d\epsilon \qquad (20) \\
\lesssim & \alpha \sqrt{\frac{\mathrm{Pdim}(\mathfrak{P}) \log n}{n}} \\
\le & \alpha \sqrt{\frac{m(k+1)^d \log n}{n}}.
\end{aligned}$$

Here inequality (20) follows from Lemma 54. Thus, by symmetrization,

$$\gamma(\alpha, m, n) = \mathbb{E} \sup_{f \in \mathfrak{P}} \left(\int f d\mu_n - \int f d\mu \right) \le 2\mathbb{E} \sup_{f \in \mathfrak{P}} \sum_{i=1}^n \sigma_i f(x_i) \lesssim \alpha \sqrt{\frac{m(k+1)^d \log n}{n}}.$$

■

We now state and prove a more general version of Theorem 14. Theorem 14 follows as a corollary of Theorem 39 for the special case when $\Lambda = \{\mu\}$.

Theorem 39. *Let $\mathcal{F} = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$. Then if $d^\star > d^*_\beta(\Lambda)$, we can find an n_0, such that, if $n \ge n_0$,*

$$\sup_{\mu \in \Lambda} \mathbb{E} \|\hat{\mu}_n - \mu\|_{\mathcal{F}} \lesssim n^{-\beta/d^\star}.$$

Proof We fix any $\mu \in \Lambda$. Recall that $P_\theta^f(x) = \sum_{|\boldsymbol{s}| \le \lfloor \beta \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta)}{\boldsymbol{s}!} (x - \theta)^{\boldsymbol{s}}$. We note that since, $\|P_\theta^f\|_\infty \le C$,

$$\begin{aligned}
|P_\theta^f(x) - f(x)| \le & \left| \sum_{|\boldsymbol{s}| \le \lfloor \beta \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta)}{\boldsymbol{s}!} (x - \theta)^{\boldsymbol{s}} - f(x) \right| \\
= & \left| \sum_{\boldsymbol{s} : |\boldsymbol{s}| = \lfloor \beta \rfloor} \frac{(x - \theta)^{\boldsymbol{s}}}{\boldsymbol{s}!} (\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta)) \right| \qquad (21)
\end{aligned}$$

$$
\begin{aligned}
&\le \|x-\theta\|_\infty^{\lfloor\beta\rfloor} \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor\beta\rfloor} \frac{1}{\boldsymbol{s}!}|\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta)| \\
&\le 2C\|x-\theta\|_\infty^{\lfloor\beta\rfloor}\|y-\theta\|_\infty^{\beta-\lfloor\beta\rfloor} \\
&\le 2C\|x-\theta\|_\infty^{\beta} \qquad (22)
\end{aligned}
$$

Equation (21) follows from Taylor's theorem where y lies in the line segment joining x and θ. Furthermore,

$$
\begin{aligned}
|P_\theta^f(x) - P_{\theta'}^f(x)| \le& |P_\theta^f(x) - f(x)| + |f(x) - P_{\theta'}^f(x)| \\
\le& 2C\left(\|x-\theta\|_\infty^\beta + \|x-\theta'\|_\infty^\beta\right) \qquad (23)
\end{aligned}
$$

For notational simplicity, let $\xi_r = \sup_{f \in \mathrm{Poly}(\lfloor\beta\rfloor, d, \alpha_r,)} \sum_{j=1}^{m_r} \int_{Q_j^r} f d(\mu_n - \mu)$, where, $\alpha_r = \max_{1 \le j' \le m_{t-1}} \|P_{x_{t,j}}^f - P_{x_{t-1,j'}}^f\|_{\mathbb{L}_\infty(Q_{j'}^{t-1})}$. We can control $\mathbb{E}\xi_r$ through Lemma 38. We note that, for any $f \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$,

$$
\begin{aligned}
&\left|\int f d(\mu_n - \mu)\right| \\
=& \left|\sum_{j=0}^{m_t} \int_{S_{t,j}} f d(\mu_n - \mu)\right| \\
\le& C(\mu_n + \mu)(S_{s:t,0}) + \left|\sum_{j=1}^{m_t} \int_{Q_j^t \setminus S_{s:t,0}} (f - P_{x_{t,j}}^f + P_{x_{t,j}}^f) d(\mu_n - \mu)\right| \\
\le& C(\mu_n + \mu)(S_{s:t,0}) + 2\max_{1 \le j \le m_t} \|f - P_{x_{t,j}}^f\|_{\mathbb{L}_\infty(Q_j^t)} + \left|\sum_{j=1}^{m_t} \int_{Q_j^t \setminus S_{s:t,0}} P_{x_{t,j}}^f d(\mu_n - \mu)\right| \\
\le& C(\mu_n + \mu)(S_{s:t,0}) + 2C3^{-t\beta} + \left|\sum_{j=1}^{m_t} \int_{Q_j^t \setminus S_{s:t,0}} P_{x_{t,j}}^f d(\mu_n - \mu)\right| \qquad (24) \\
=& C(\mu_n + \mu)(S_{s:t,0}) + 2C3^{-t\beta} \\
&+ \left|\sum_{j'=1}^{m_{t-1}} \sum_{j:q(t,j)=j'} \int_{Q_j^t \setminus S_{s:t,0}} (P_{x_{t,j}}^f - P_{x_{t-1,j'}}^f + P_{x_{t-1,j'}}^f) d(\mu_n - \mu)\right| \\
\le& C(\mu_n + \mu)(S_{s:t,0}) + 2C3^{-t\beta} + \xi_t + \left|\sum_{j'=1}^{m_{t-1}} \int_{Q_{j'}^t \setminus S_{s:t,0}} P_{x_{t-1,j'}}^f d(\mu_n - \mu)\right| \qquad (25) \\
\le& C(\mu_n + \mu)(S_{s:t,0}) + 2C3^{-t\beta} + CM_s + \sum_{r=s}^t \xi_r \qquad (26)
\end{aligned}
$$

Inequality (24) follows from observing that $\|f - P^f_{x_{j,t}}\|_{\mathbb{L}_\infty(Q^t_j)} \le C \sup_{x \in \mathbb{L}_\infty(Q^t_j)} \|x - x_{j,t}\|_\infty \le C\text{diam}(Q^t_j) \le C3^{-t}$. Inequality (25) follows from observing that

$$\left|\sum_{j'=1}^{m_{t-1}} \sum_{j:q(t,j)=j'} \int_{Q^t_j \setminus S_{s:t,0}} (P^f_{x_{t,j}} - P^f_{x_{t-1,j'}}) d(\mu_n - \mu)\right| \le \xi_t.$$

Equation (26) follows easily from an inductive argument and observing that $\|P^f_a\|_\infty \le C$, for any $a \in \mathbb{R}^d$, by construction. To bound the expectation of ξ_r, we note that,

$$\begin{aligned}
\mathbb{E}\xi_r \le & \gamma\left(\max_{1 \le j' \le m_{t-1}} \|P^f_{x_{t,j}} - P^f_{x_{t-1,j'}}\|_{\mathbb{L}_\infty(Q^{t-1}_{j'})}, m_r, n\right) \\
\lesssim & \max_{1 \le j' \le m_{t-1}} \|P^f_{x_{t,j}} - P^f_{x_{t-1,j'}}\|_{\mathbb{L}_\infty(Q^{t-1}_{j'})} \sqrt{\frac{m_r \log n}{n}} \quad \text{(from Lemma 38)} \\
\le & 4C3^{-\beta(t-1)} \sqrt{\frac{m_r \log n}{n}}.
\end{aligned}$$

Thus, taking expectations on both sides of (26), we get,

$$\begin{aligned}
& \mathbb{E}\left|\int f d(\mu_n - \mu)\right| \\
\lesssim & 3^{-t\beta} + 2C \sum_{r=s}^{t} 3^{-\frac{d'(r+2)\beta}{d'-2\beta}} + C\sqrt{m_s/n} + 4C \sum_{r=s}^{t} 3^{-\beta r} \sqrt{\frac{m_r \log n}{n}} \qquad (27) \\
\le & 3^{-t\beta} + 2C \sum_{r=s}^{t} 3^{-\frac{d'(r+2)\beta}{d'-2\beta}} + C\sqrt{\frac{3^{d'(s+2)}}{n}} + 4C \sum_{r=s}^{t} 3^{-\beta r} \sqrt{\frac{3^{d'(r+2)} \log n}{n}} \qquad (28) \\
\le & 3^{-t\beta} + 2C \sum_{r=s}^{\infty} 3^{-\frac{d'(r+2)\beta}{d'-2\beta}} + C\sqrt{\frac{\log n}{n}} \left(3^{\frac{d'(s+2)}{2}} + 4 \sum_{r=s}^{t} 3^{\frac{2d' + r(d'-2\beta)}{2}}\right) \\
= & 3^{-t\beta} + 2C \frac{3^{\frac{-d'(s+2)\beta}{d'-2\beta}}}{1 - 3^{-\frac{d'\beta}{d'-2\beta}}} + C\sqrt{\frac{\log n}{n}} \left(3^{\frac{d'(s+2)}{2}} + 4 \frac{3^{\frac{2d' + s(d'-2\beta)}{2}}((3^{\frac{d'-2\beta}{2}})^{t-s+1} - 1)}{3^{\frac{d'-2\beta}{2}} - 1}\right) \\
\le & 3^{-t\beta} + 2C \frac{3^{\frac{-2d'\beta}{d'-2}}}{1 - 3^{-\frac{d'\beta}{d'-2\beta}}} \times (3^{-s})^{\frac{d'\beta}{d'-2\beta}} \\
& + C\sqrt{\frac{\log n}{n}} \left(3^{\frac{d'(s+2)}{2}} + 4 \frac{3^{\frac{2d' + s(d'-2\beta)}{2}}((3^{\frac{d'-2\beta}{2}})^{t-s+1} - 1)}{3^{\frac{d'-2\beta}{2}} - 1}\right) \\
\le & 3^{-t\beta} + 2C \frac{\epsilon^\beta}{1 - 3^{-\frac{d'\beta}{d'-2\beta}}} + C\sqrt{\frac{\log n}{n}} \left(3^{\frac{d'(s+1)}{2}} + 4 \frac{3^{\frac{2d' + s(d'-2\beta)}{2}}((3^{\frac{d'-2\beta}{2}})^{t-s+1} - 1)}{3^{\frac{d'-2\beta}{2}} - 1}\right) \\
\lesssim & \epsilon^\beta + \sqrt{\frac{\log n}{n}} \left(3^{\frac{d's}{2}} + 3^{\frac{(d'-2\beta)t}{2}}\right) \\
\lesssim & \epsilon^\beta + \sqrt{\frac{\log n}{n}} \epsilon^{\beta - d'/2}
\end{aligned}$$

$$\lesssim n^{-\beta/d'}\sqrt{\log n}$$
$$\lesssim n^{-\beta/d^\star}.$$

Equation (27) follows from Lemma 37, where as (28) follows from Lemma 36. ■

Appendix B. Proofs from Section 4.1

B.1 Proof of Lemma 15

Lemma 15. *Under assumption B1, $\tilde{D}\circ\tilde{E}(\cdot) = id(\cdot)$, a.e. $[\mu]$ and $\tilde{E}\circ\tilde{D}(\cdot) = id(\cdot)$, a.e. $[\nu]$.*

Proof To show the first implication, it is enough to show that $\int \|\tilde{D}(\tilde{E}(x)) - x\|_2^2 d\mu(x) = 0$. To see this, we observe that,

$$\int \|\tilde{D}\circ\tilde{E}(x) - x\|_2^2 d\mu(x) = \int \|\tilde{D}(z) - \tilde{D}(z)\|_2^2 d\nu(z) = 0.$$

Here the first equality follows from the fact that $(X, \tilde{E}(X)) \stackrel{d}{=} (\tilde{D}(Z), Z)$. The other statement follows similarly. ■

B.2 Proof of Proposition 19

Proposition 19. *Under Assumption B1, the supports of μ and ν are homeomorphic.*

Proof We use the so-called good set principle to prove this result. Suppose $A = \{x \in [0,1]^d : \tilde{D}\circ\tilde{E}(x) = x\}$ and let $x \in \text{supp}(\mu)$. Clearly, by Lemma 15, $\mu(A) = 1$. By definition of the support, for any $n \in \mathbb{N}$, $\mu(B(x, 1/n)\cap A) = \mu(B(x, 1/n)) > 0$, which implies that $B(x, 1/n)\cap A$ is non-empty. Let $x_n \in B(x, 1/n)\cap A$. Thus, for any $x \in \text{supp}(\mu)$, there exist a sequence $\{x_n(x)\}_{n\in\mathbb{N}} \subseteq A$, such that $x_n(x) \to x$ as $n \to \infty$. We first show that $\tilde{E}$, restricted to the support of μ is a bijection. To show injectivity, let x, $x' \in \text{supp}(\mu)$ and $\tilde{E}(x) = \tilde{E}(x')$. Thus, $\tilde{D}\circ\tilde{E}(x) = \tilde{D}\circ\tilde{E}(x') \implies \tilde{D}\circ\tilde{E}(\lim_{n\to\infty} x_n(x)) = \tilde{D}\circ\tilde{E}(\lim_{n\to\infty} x_n(x'))$. Owing to the uniform continuity of $\tilde{D}$ and $\tilde{E}$ and the fact that $x_n(x)$ and $x_n(x')$ are in A, this implies that $\lim_{n\to\infty}\tilde{D}\circ\tilde{E}(x_n(x)) = \lim_{n\to\infty}\tilde{D}\circ\tilde{E}(x_n(x')) \implies \lim_{n\to\infty} x_n(x) = \lim_{n\to\infty} x_n(x') \implies x = x'$.

To show surjectivity, let $z \in \text{supp}(\nu)$. Let $A' = \{z \in [0,1]^\ell : \tilde{E}\circ\tilde{D}(z) = z\}$. By a similar argument, there exists a sequence $z_n \to z$, such that $\{z_n\}_{n\in\mathbb{N}} \subseteq A'$. We consider $\tilde{D}(z)$. Clearly, $\tilde{E}\circ\tilde{D}(z) = \tilde{E}\circ\tilde{D}(\lim_{n\to\infty} z_n) = \lim_{n\to\infty}\tilde{E}\circ\tilde{D}(z_n) = \lim_{n\to\infty} z_n = z$. The only thing remaining is to show that $\tilde{D}(z) \in \text{supp}(\mu)$. To show this, let $z' \in B(z, \epsilon)$. Clearly, $\nu(B(z,\epsilon)) > 0$. $\|\tilde{D}(z) - \tilde{D}(z')\| \lesssim \epsilon^{\alpha_d\wedge 1}$, since, $\tilde{D} \in \mathcal{H}^{\alpha_d}\left(\mathbb{R}^\ell, \mathbb{R}^d, C_d\right)$. Thus, there exists a $c > 0$ such that $\|\tilde{D}(z) - \tilde{D}(z')\| \le c\,\epsilon^{\alpha_d\wedge 1}$. Hence, $B(\tilde{D}(z), c\epsilon^{\alpha_d\wedge 1}) \supseteq \{\tilde{D}(z') : z' \in B(z,\epsilon)\} \implies \mu\left(B(\tilde{D}(z), c\epsilon^{\alpha_d\wedge 1})\right) \ge \mu\left(\{D(z') : z' \in B(z,\epsilon)\}\right) = \nu(B(z,\epsilon)) > 0$, for any $\epsilon > 0$. Thus, for any $\delta > 0$, taking $\epsilon = (\delta/c)^{1/(\alpha_d\wedge 1)}$, gives that $\mu\left(B(\tilde{D}(z),\delta)\right) > 0$, for any $\delta > 0$, which implies that $D(z) \in \text{supp}(\mu)$.

Now to show that $\tilde{E}$ is the inverse of $\tilde{D}$ on the support of μ, we let $x \in \text{supp}(\mu)$. By the argument above, there exists a sequence $\{x_n\}_{n \in \mathbb{N}} \subseteq A$, such that $\lim_{n \to \infty} x_n = x$. Thus, $\tilde{D} \circ \tilde{E}(x) = \tilde{D} \circ \tilde{E}(\lim_{n \to \infty} x_n) = \lim_{n \to \infty} \tilde{D} \circ \tilde{E}(x_n) = \lim_{n \to \infty} x_n = x$. Similarly, one can show that $\tilde{E} \circ \tilde{D}(z) = z$, for any $z \in \text{supp}(\nu)$. Clearly, $\tilde{D}$ and $\tilde{E}$ are continuous. Thus, $\text{supp}(\mu)$ and $\text{supp}(\nu)$ are homeomorphic. ■

Appendix C. Proofs from Section 4.3.1

C.1 Proof of Lemma 20

Lemma 20 (Oracle inequalities for GANs and BiGANs)**.** *Suppose that* $\Phi = \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$ *and* $\Psi = \mathcal{H}^\beta(\mathbb{R}^{d+\ell}, \mathbb{R}, 1)$*, then*

(a) if $\hat{G}_n$ *and* $\hat{G}_{n,m}$ *are the generator estimates defined in* (3)*, the following hold:*

$$\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \le \inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi, \tag{15}$$

$$\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Psi \le \inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi + 2\|\nu - \hat{\nu}_m\|_{\Phi \circ \mathcal{G}}. \tag{16}$$

(b) if $(\hat{D}_n, \hat{E}_n)$ *and* $(\hat{D}_{n,m}, \hat{E}_{n,m})$ *are the decoder-encoder estimates defined in* (5) *and* (6)*, respectively, the following hold:*

$$\|(\mu, (\hat{E}_n)_\sharp \mu) - (\hat{D}_n)_\sharp \nu, \nu)\|_\Psi \le \inf_{D \in \mathcal{D},\, E \in \mathcal{E}} \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi + 2\|\hat{\mu}_n - \mu\|_{\mathcal{F}_1}, \tag{17}$$

$$\begin{aligned}\|(\mu, (\hat{E}_{n,m})_\sharp \mu) - (\hat{D}_{n,m})_\sharp \nu, \nu)\|_\Psi \le& \inf_{D \in \mathcal{D},\, E \in \mathcal{E}} \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi + 2\|\hat{\mu}_n - \mu\|_{\mathcal{F}_1} \\ &+ 2\|\hat{\nu}_m - \nu\|_{\mathcal{F}_2},\end{aligned} \tag{18}$$

where, $\mathcal{F}_1 = \{\psi(\cdot, E(\cdot)) : \psi \in \Psi,\, E \in \mathcal{E}\}$ *and* $\mathcal{F}_2 = \{\psi(D(\cdot), \cdot) : \psi \in \Psi,\, D \in \mathcal{D}\}$*.*

Proof **Proof of part (a)**: For any $G \in \mathcal{G}$, we observe that,

$$\begin{aligned}\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \le& \|\hat{\mu}_n - (\hat{G}_n)_\sharp \nu\|_\Phi + \|\mu - \hat{\mu}_n\|_\Phi \\ \le& \|\hat{\mu}_n - G_\sharp \nu\|_\Phi + \|\mu - \hat{\mu}_n\|_\Phi \\ \le& \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi\end{aligned}$$

Taking infimum on both sides w.r.t. $G \in \mathcal{G}$ gives us the desired result.

Similarly, for the estimator $\hat{G}_{n,m}$, we note that,

$$\begin{aligned}\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Phi \le& \|\hat{\mu}_n - (\hat{G}_{n,m})_\sharp \hat{\nu}_m\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + \|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}} \\ \le& \|\hat{\mu}_n - G_\sharp \hat{\nu}_m\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + \|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}} \\ \le& \|\mu - G_\sharp \nu\|_\Phi + 2\|\hat{\mu}_n - \mu\|_\Phi + 2\|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}}\end{aligned}$$

Again taking infimum on both sides w.r.t. $G \in \mathcal{G}$ gives us the desired result.

Proof of part (b): For any $E \in \mathcal{E}$ and $D \in \mathcal{D}$,

$$\|(\mu, (\hat{E}_n)_\sharp \mu) - ((\hat{D}_n)_\sharp \nu, \nu)\|_\Psi$$

$$\begin{aligned}
\leq & \|(\hat{\mu}_n, (\hat{E}_n)_\sharp \hat{\mu}_n) - ((\hat{D}_n)_\sharp \nu, \nu)\|_\Psi + \|(\hat{\mu}_n, (\hat{E}_n)_\sharp \hat{\mu}_n) - (\mu, (\hat{E}_n)_\sharp \mu)\|_\Psi \\
\leq & \|(\hat{\mu}_n, E_\sharp \hat{\mu}_n) - (D_\sharp \nu, \nu)\|_\Psi + \|\hat{\mu}_n - \hat{\mu}\|_{\mathcal{F}_1} \\
\leq & \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi + 2\|\hat{\mu}_n - \hat{\mu}\|_{\mathcal{F}_1}.
\end{aligned}$$

Taking infimum over E and D gives us the desired result.

Similarly,

$$\begin{aligned}
& \|(\mu, (\hat{E}_{n,m})_\sharp \mu) - ((\hat{D}_{n,m})_\sharp \nu, \nu)\|_\Psi \\
\leq & \|(\hat{\mu}_n, (\hat{E}_{n,m})_\sharp \hat{\mu}_n) - ((\hat{D}_{n,m})_\sharp \hat{\nu}_m, \hat{\nu}_m)\|_\Psi + \|\hat{\mu}_n - \hat{\mu}\|_{\mathcal{F}_1} + \|\hat{\nu}_m - \hat{\nu}\|_{\mathcal{F}_2} \\
\leq & \|(\hat{\mu}_n, E_\sharp \hat{\mu}_n) - (D_\sharp \hat{\nu}_m, \hat{\nu}_m)\|_\Psi + \|\hat{\mu}_n - \hat{\mu}\|_{\mathcal{F}_1} \\
\leq & \|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi + 2\|\hat{\mu}_n - \hat{\mu}\|_{\mathcal{F}_1} + 2\|\hat{\nu}_m - \hat{\nu}\|_{\mathcal{F}_2}.
\end{aligned}$$

Again taking infimum over E and D gives us the desired result. ■

Appendix D. Proofs from Section 4.3.2

D.1 Proof of Theorem 21

We begin by proving Theorem 21, which discusses the approximation capabilities of ReLU networks when the underlying measure has a low entropic dimension. Some supporting lemmas required for the proof are stated in Appendix D.2.

For $s > \bar{d}_{p\beta}(\mu)$, by Lemma 32, we can find an $\epsilon' \in (0,1)$, such that if $\epsilon \in (0, \epsilon']$ and a set S, such that $\mu(S) \geq 1 - \epsilon^{p\beta}$ and $\mathcal{N}(\epsilon; S, \ell_\infty) \leq \epsilon^{-s}$. In what follows, we take $\epsilon \in (0, \epsilon']$.

Let $K = \lceil \frac{1}{2\epsilon} \rceil$. For any $\boldsymbol{i} \in [K]^d$, let $\theta^{\boldsymbol{i}} = (\epsilon + 2(i_1 - 1)\epsilon, \dots, \epsilon + 2(i_d - 1)\epsilon)$. We also let, $\mathcal{P}_\epsilon = \{B_{\ell_\infty}(\theta^{\boldsymbol{i}}, \epsilon) : \boldsymbol{i} \in [K]^d\}$. By construction, $\theta^{\boldsymbol{i}}$'s are at least 2ϵ apart, making the sets in $\mathcal{P}_\epsilon$ disjoint. Following the proof of Lemma 33, we make the following claim.

Lemma 40. $|\{A \in \mathcal{P}_\epsilon : A \cap S \neq \emptyset\}| \leq 2^d \epsilon^{-s}$.

Theorem 21. *Suppose that $f \in \mathcal{H}^\alpha(\mathbb{R}^d, \mathbb{R}, C)$, for some $C > 0$ and let $s > \bar{d}_{\alpha p}(\mu)$. Then, we can find constants ϵ_0 and a, that might depend on α, d and C, such that, for any $\epsilon \in (0, \epsilon_0]$, there exists a ReLU network, $\hat{f}$ with $\mathcal{L}(\hat{f}) \leq a \log(1/\epsilon)$, $\mathcal{W}(\hat{f}) \leq a \log(1/\epsilon) \epsilon^{-s/\alpha}$, $\mathcal{B}(\hat{f}) \leq a\epsilon^{-1/\alpha}$ and $\mathcal{R}(\hat{f}) \leq 2C$, that satisfies, $\|f - \hat{f}\|_{\mathbb{L}_p(\gamma)} \leq \epsilon$.*

Proof We also let $\mathcal{I} = \left\{\boldsymbol{i} \in [K]^d : B_{\ell_\infty}(\theta^{\boldsymbol{i}}, \epsilon) \cap S \neq \emptyset\right\}$. We also let $\mathcal{I}^\dagger = \{\boldsymbol{j} \in [K]^d : \min_{\boldsymbol{i} \in \mathcal{I}} \|\boldsymbol{i} - \boldsymbol{j}\|_1 \leq 1\}$. We know that $|\mathcal{I}^\dagger| \leq 3^d |\mathcal{I}| \leq 6^d \mathcal{N}(\epsilon; S, \ell_\infty)$. For $0 < b \leq a$, let,

$$\xi_{a,b}(x) = \mathrm{ReLU}\left(\frac{x+a}{a-b}\right) - \mathrm{ReLU}\left(\frac{x+b}{a-b}\right) - \mathrm{ReLU}\left(\frac{x-b}{a-b}\right) + \mathrm{ReLU}\left(\frac{x-a}{a-b}\right).$$

A pictorial view of this function is given in Figure 1 and can be implemented by a ReLU network of depth two and width four. Thus, $\mathcal{L}(\xi_{a,b}) = 2$ and $\mathcal{W}(\xi_{a,b}) = 12$. Suppose that $\delta = \epsilon/3$ and let, $\zeta(x) = \prod_{\ell=1}^d \xi_{\epsilon+\delta,\delta}(x_\ell)$. Clearly, $\mathcal{B}(\xi_{\epsilon+\delta,\delta}) \leq \frac{1}{\delta}$. It is easy to observe that $\{\zeta(\cdot - \theta^{\boldsymbol{i}}) : \boldsymbol{i} \in \mathcal{I}^\dagger\}$ forms a partition of unity on S, i.e. $\sum_{\boldsymbol{i} \in \mathcal{I}^\dagger} \zeta(x - \theta^{\boldsymbol{i}}) = 1, \forall x \in S$.

We consider the Taylor approximation of f around θ as,

$$P_\theta(x) = \sum_{|\boldsymbol{s}| \le \lfloor \alpha \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta)}{\boldsymbol{s}!} (x - \theta)^{\boldsymbol{s}} .$$

Note that for any $x \in [0,1]^d$, $f(x) - P_\theta(x) = \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor \alpha \rfloor} \frac{(x-\theta)^{\boldsymbol{s}}}{\boldsymbol{s}!}(\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta))$, for some y, which is a convex combination of x and θ. Thus,

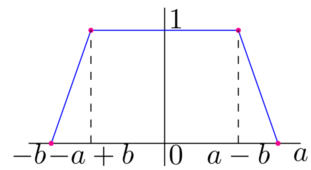

Figure 1: Plot of $\xi_{a,b}(\cdot)$.

$$\begin{aligned} f(x) - P_\theta(x) = & \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor \alpha \rfloor} \frac{(x-\theta)^{\boldsymbol{s}}}{\boldsymbol{s}!}(\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta)) \\ \le & \|x-\theta\|_\infty^{\lfloor \alpha \rfloor} \sum_{\boldsymbol{s}:|\boldsymbol{s}|=\lfloor \alpha \rfloor} \frac{1}{\boldsymbol{s}!} |\partial^{\boldsymbol{s}} f(y) - \partial^{\boldsymbol{s}} f(\theta)| \\ \le & 2C \|x-\theta\|_\infty^{\lfloor \alpha \rfloor} \|y-\theta\|_\infty^{\alpha - \lfloor \alpha \rfloor} \\ \le & 2C \|x-\theta\|_\infty^{\alpha}. \end{aligned} \tag{29}$$

Next we define $\tilde{f}(x) = \sum_{\boldsymbol{i} \in \mathcal{I}^\dagger} \zeta(x - \theta^{\boldsymbol{i}}) P_{\theta^{\boldsymbol{i}}}(x)$. Thus, if $x \in S$,

$$\begin{aligned} |f(x) - \tilde{f}(x)| = \left| \sum_{\boldsymbol{i} \in \mathcal{I}^\dagger} \zeta(x-\theta^{\boldsymbol{i}})(f(x) - P_{\theta^{\boldsymbol{i}}}(x)) \right| \le & \sum_{\boldsymbol{i} \in \mathcal{I}^\dagger : \|x - \theta^{\boldsymbol{i}}\|_\infty \le 2\epsilon} |f(x) - P_{\theta^{\boldsymbol{i}}}(x)| \\ \le & C 2^{d+1} (2\epsilon)^\alpha \\ = & C 2^{d+\alpha+1} \epsilon^\alpha. \end{aligned} \tag{30}$$

We note that, $\tilde{f}(x) = \sum_{\boldsymbol{i} \in \mathcal{I}^\dagger} \zeta(x - \theta^{\boldsymbol{i}}) P_{\theta^{\boldsymbol{i}}}(x) = \sum_{\boldsymbol{i} \in \mathcal{I}^\dagger} \sum_{|\boldsymbol{s}| \le \lfloor \alpha \rfloor} \frac{\partial^{\boldsymbol{s}} f(\theta^{\boldsymbol{i}})}{\boldsymbol{s}!} \zeta(x - \theta^{\boldsymbol{i}}) \left(x - \theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}$. Let $a_{\boldsymbol{i},\boldsymbol{s}} = \frac{\partial^{\boldsymbol{s}} f(\theta^{\boldsymbol{i}})}{\boldsymbol{s}!}$ and

$$\begin{aligned} \hat{f}_{\boldsymbol{i},\boldsymbol{s}}(x) = & \mathrm{prod}_m^{(d+|\boldsymbol{s}|)}(\xi_{\epsilon_1,\delta_1}(x_1 - \theta^{\boldsymbol{i}}_1), \dots, \xi_{\epsilon_d,\delta_d}(x_d - \theta^{\boldsymbol{i}}_d), \underbrace{(x_1 - \theta^{\boldsymbol{i}}_1), \dots, (x_1 - \theta^{\boldsymbol{i}}_1)}_{s_1 \text{ times}}, \\ & \dots, \underbrace{(x_1 - \theta^{\boldsymbol{i}}_d), \dots, (x_d - \theta^{\boldsymbol{i}}_d)}_{s_d \text{ times}}), \end{aligned}$$

where $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}(\cdot)$ is defined in Lemma 43 and approximates the product function for $d + |\boldsymbol{s}|$ real numbers. We note that $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}$ has at most $d + |\boldsymbol{s}| \le d + \lfloor \alpha \rfloor$ many inputs. By Lemma 43, $\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}$ can be implemented by a ReLU network with $\mathcal{L}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)})$, $\mathcal{W}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}) \le c_3 m$ and $\mathcal{B}(\mathrm{prod}_m^{(d+|\boldsymbol{s}|)}) \le 4 \vee (d-1)^2$. Thus, $\mathcal{L}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le c_3 m + 2$ and $\mathcal{W}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le c_3 m + 8d + 4|s| \le c_3 m + 8d + 4\lfloor \alpha \rfloor$. Furthermore, $\mathcal{B}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le 4 \vee \frac{1}{\delta} \le 1/\delta$, when δ is small enough. With this $\hat{f}_{\boldsymbol{i},\boldsymbol{s}}$, we observe from Lemma 43 that,

$$\left| \hat{f}_{\boldsymbol{i},\boldsymbol{s}}(x) - \zeta(x - \theta^{\boldsymbol{i}}) \left(x - \theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}} \right| \le \frac{d + \lfloor \alpha \rfloor}{2^{2m-1}}, \forall x \in S. \tag{31}$$

Finally, let, $\hat{f}(x) = \sum_{\boldsymbol{i}\in\mathcal{I}^\dagger}\sum_{|\boldsymbol{s}|\le\lfloor\alpha\rfloor} a_{\boldsymbol{i},\boldsymbol{s}}\hat{f}_{\boldsymbol{i},\boldsymbol{s}}(x)$. Clearly, $\mathcal{L}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le c_3 m + 3$ and $\mathcal{W}(\hat{f}_{\boldsymbol{i},\boldsymbol{s}}) \le \lfloor\alpha\rfloor^d\,(c_3 m + 8d + 4\lfloor\alpha\rfloor)$. This implies that,

$$\begin{aligned}|\hat{f}(x) - \tilde{f}(x)| \le & \sum_{\boldsymbol{i}\in\mathcal{I}^\dagger:\|x-\theta^{\boldsymbol{i}}\|_\infty\le 2\epsilon}\ \sum_{|\boldsymbol{s}|\le\lfloor\alpha\rfloor} |a_{\boldsymbol{i},\boldsymbol{s}}|\zeta(x-\theta^{\boldsymbol{i}})|\hat{f}_{\boldsymbol{i}\boldsymbol{s}}(x) - \left(x-\theta^{\boldsymbol{i}}\right)^{\boldsymbol{s}}| \\ \le & 2^d \sum_{|\boldsymbol{s}|\le\lfloor\alpha\rfloor} |a_{\theta,\boldsymbol{s}}| \left|\hat{f}_{\theta^{\boldsymbol{i}(x)},\boldsymbol{s}}(x) - \zeta_{\boldsymbol{\epsilon},\boldsymbol{\delta}}(x-\theta^{(\boldsymbol{i}(x))})\left(x-\theta^{\boldsymbol{i}(x)}\right)^{\boldsymbol{s}}\right| \\ \le & \frac{(d+\lfloor\alpha\rfloor)C}{2^{2m-d-1}}.\end{aligned}$$

We thus get that if $x \in S$,

$$|f(x) - \hat{f}(x)| \le |f(x) - \tilde{f}(x)| + |\hat{f}(x) - \tilde{f}(x)| \le C2^{d+\alpha+1}\epsilon^\alpha + \frac{(d+\lfloor\alpha\rfloor)C}{2^{2m-d-1}}. \tag{32}$$

We also note that by construction $\|\hat{f}\|_\infty \le C'$, for some constant C' that might depend on C, α and d. Hence,

$$\begin{aligned}\|f-\hat{f}\|^p_{\mathbb{L}_p(\mu)} = & \int_S |f(x)-\hat{f}(x)|^p d\mu(x) + \int_{S^\complement} |f(x)-\hat{f}(x)|^p d\mu(x) \\ \le & \left(C2^{d+\alpha+1}\epsilon^\alpha + \frac{(d+\lfloor\alpha\rfloor)^3 C}{2^{2m-d-1}}\right)^p \mu(S) + (C+C')\mu(S^\complement) \\ \lesssim & \left(C2^{d+\alpha+1}\epsilon^\alpha + \frac{(d+\lfloor\alpha\rfloor)^3 C}{2^{2m-d-1}}\right)^p + \epsilon^{p\alpha} \\ \implies \|f-\hat{f}\|_{\mathbb{L}_p(\mu)} \lesssim & C2^{d+\alpha+1}\epsilon^\alpha + \frac{(d+\lfloor\alpha\rfloor)C}{2^{2m-d-1}} + \epsilon^\alpha \lesssim \epsilon^\alpha + 4^{-m}.\end{aligned}$$

Taking $\epsilon \asymp \eta^{1/\alpha}$ and $m \asymp \log(1/\eta)$ ensures that $\|f-\hat{f}\|_{\mathbb{L}_p(\mu)} \le \eta$. We note that $\hat{f}$ has $|\mathcal{I}^\dagger| \le 6^d\mathcal{N}_\epsilon(S) \lesssim 6^d\epsilon^{-s}$ many networks with depth $c_3 m + 3$ and number of weights $\lfloor\alpha\rfloor^d\,(c_3 m + 8d + 4\lfloor\alpha\rfloor)$. Thus, $\mathcal{L}(\hat{f}) \le c_3 m + 4$ and $\mathcal{W}(\hat{f}) \le \epsilon^{-s}(6\lfloor\alpha\rfloor)^d\,(c_3 m + 8d + 4\lfloor\alpha\rfloor)$. we thus get, $\mathcal{L}(\hat{f}) \le c_3 m + 4 \le c_4 \log\left(\frac{1}{\eta}\right)$, where c_4 is a function of δ, $\lfloor\alpha\rfloor$ and d. Similarly,

$$\mathcal{W}(\hat{f}) \le \epsilon^{-s}(6\lfloor\alpha\rfloor)^d\,(c_3 m + 8d + 4\lfloor\alpha\rfloor) \le c_6 \log(1/\eta)\eta^{-s/\alpha}.$$

Taking $a = c_4 \vee c_6$ gives the result. Furthermore, by construction, we note that, $\mathcal{B}(\hat{f}) \lesssim 1/\delta = 3/\epsilon \lesssim \eta^{-1/\alpha}$. ■

D.2 Additional Results

Lemma 41 (Proposition 2 of Yarotsky (2017))**.** *The function $f(x) = x^2$ on the segment $[0,1]$ can be approximated with any error by a ReLU network, $sq_m(\cdot)$, such that,*

1. *$\mathcal{L}(sq_m)$, $\mathcal{W}(sq_m) \le c_1 m$.*
2. *$sq_m\left(\frac{k}{2^m}\right) = \left(\frac{k}{2^m}\right)^2$, for all $k = 0, 1, \ldots, 2^m$.*

3. $\|sq_m - x^2\|_{\mathcal{L}_\infty([0,1])} \le \frac{1}{2^{2m+2}}$.

4. $\mathcal{B}(sq_m) \le 4$.

Lemma 42. *Let $M \ge 1$, then we can find a ReLU network $prod_m^{(2)}$, such that,*

1. $\mathcal{L}(prod_m^{(2)}), \mathcal{W}(prod_m^{(2)}) \le c_2 m$*, for some absolute constant* c_2.

2. $\|prod_m^{(2)} - xy\|_{\mathcal{L}_\infty([-M,M]\times[-M,M])} \le \frac{M^2}{2^{2m+1}}$.

3. $\mathcal{B}(prod_m^{(2)}) \le 4 \vee M^2$.

Proof Let $\operatorname{prod}_m^{(2)}(x,y) = M^2\left(\operatorname{sq}_m\left(\frac{|x+y|}{2M}\right) - \operatorname{sq}_m\left(\frac{|x-y|}{2M}\right)\right)$. Clearly, $\operatorname{prod}_m^{(2)}(x,y) = 0$, if $xy = 0$. We note that, $\mathcal{L}(\operatorname{prod}_m^{(2)}) \le c_1 m + 1 \le c_2 m$ and $\mathcal{W}(\operatorname{prod}_m^{(2)}) \le 2c_1 m + 2 \le c_2 m$, for some absolute constant c_2. Clearly,

$$\|\operatorname{prod}_m^{(2)} - xy\|_{\mathbb{L}_\infty([-M,M]\times[-M,M])} \le 2M^2\|\operatorname{sq}_m - x^2\|_{\mathbb{L}_\infty([0,1])} \le \frac{M^2}{2^{2m+1}}.$$

Part 3 of the Lemma follows easily by applying part 4 of Lemma 41. ■

Lemma 43. *For any $m \ge \frac{1}{2}(\log_2(4d)-1)$, we can construct a ReLU network $prod_m^{(d)} : \mathbb{R}^d \to \mathbb{R}$, such that for any $x_1, \dots, x_d \in [-1,1]$, $\|prod_m^{(d)}(x_1,\dots,x_d) - x_1 \dots x_d\|_{\mathcal{L}_\infty([-1,1]^d)} \le \frac{1}{2^m}$. Furthermore,*

1. $\mathcal{L}(prod_m^{(d)}) \le c_3 m$, $\mathcal{W}(prod_m^{(d)}) \le c_3 m$.

2. $\mathcal{B}(prod_m^{(d)}) \le 4$.

Proof Let $M = 1$ and $d \ge 2$. We define

$$\operatorname{prod}_m^{(k)}(x_1,\dots,x_k) = \operatorname{prod}_m^{(2)}(\operatorname{prod}_m^{(k-1)}(x_1,\dots,x_{k-1}), x_d),\ k \ge 3.$$

Clearly $\mathcal{W}(\operatorname{prod}_m^{(d)}), \mathcal{L}(\operatorname{prod}_m^{(d)}) \le c_3 dm$, for some absolute constant c_3. Suppose that $m \ge \frac{1}{2}(\log_2(4d)-1)$. We show that $|\operatorname{prod}_m^{(k)}(x_1,\dots,x_k)| \le 2$, for all $2 \le k \le d$. Clearly, the statement holds for $k = 2$. Suppose the statement holds for some $2 \le k < d$. Then,

$$\begin{aligned}
|\operatorname{prod}_m^{(k+1)}(x_1,\dots,x_{k+1})| \le & \frac{2^2}{2^{2m+1}} + |x_{k+1}||\operatorname{prod}_m^{(k)}(x_1,\dots,x_k)| \\
\le & \frac{1}{d} + |\operatorname{prod}_m^{(k)}(x_1,\dots,x_k)| \\
\le & \frac{k-2}{d} + |\operatorname{prod}_m^{(2)}(x_1,x_2)| \\
\le & \frac{k-1}{d} + 1 \\
\le & 2.
\end{aligned}$$

Thus, by induction, it is easy to see that, $|\text{prod}_m^{(d)}| \le 2$. Taking $M = 2$, we get that,

$$\begin{aligned}
&\|\text{prod}_m^{(d)}(x_1, \dots, x_d) - x_1 \dots x_d\|_{\mathcal{L}_\infty([-1,1]^d)} \\
=&\|\text{prod}_m^{(2)}(\text{prod}_m^{(d-1)}(x_1, \dots, x_{d-1}), x_d) - x_1 \dots x_d\|_{\mathcal{L}_\infty([-1,1]^d)} \\
\le&\|\text{prod}_m^{(d-1)}(x_1, \dots, x_{d-1}) - x_1 \dots x_{d-1}\|_{\mathcal{L}_\infty([-1,1]^d)} + \frac{M^2}{2^{2m+1}} \\
\le&\frac{dM^2}{2^{2m+1}} \\
=&\frac{d}{2^{2m-1}}.
\end{aligned}$$

■

D.3 Proofs of Lemmata 22 and 23

D.3.1 Proof of Lemma 22

Lemma 22. *Suppose assumption A2 holds. There exists an ϵ_0, such that, for any $0 < \epsilon \le \epsilon_0$, we can take $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ such that $L_g \lesssim \log(1/\epsilon)$, $W_g \lesssim \epsilon^{-\ell/\alpha_g} \log(1/\epsilon)$, $B_g \lesssim \epsilon^{-1/\alpha_g}$ and $\inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi \lesssim \epsilon^{\beta \wedge 1}$.*

Proof We begin by noting that,

$$\begin{aligned}
\|\mu - G_\sharp \nu\|_\Phi = \|\tilde{G}_\sharp \nu - G_\sharp \nu\|_\Phi =& \sup_{\phi \in \Phi} \left(\int \phi(\tilde{G}(x)) d\nu(x) - \int \phi(G(x)) d\nu(x) \right) \\
=& \sup_{\phi \in \Phi} \left(\int (\phi(\tilde{G}(x)) - \phi(G(x))) d\nu(x) \right) \\
\le& \int |\tilde{G}(x) - G(x)|^{\beta \wedge 1} d\nu(x) &(33)\\
\le& \|\tilde{G} - G\|_{\mathbb{L}_1(\nu)}^{\beta \wedge 1}. &(34)
\end{aligned}$$

Here, inequality (33) follows from observing that $\phi \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, 1)$ and (34) follows from Jensen's inequality. The Lemma immediately follows from observing that $\tilde{G} \in \mathcal{H}^{\alpha_g}(\mathbb{R}^\ell, \mathbb{R}, C_g)$ (by assumption A2) and applying Theorem 21 to approximate the i-th coordinate of $\tilde{G}$ ($i \in [d]$) and stacking the networks parallelly. ■

D.3.2 Proof of Lemma 23

Lemma 23. *Suppose assumption B1 holds and let $s > \bar{d}_{\alpha_e}(\mu)$. There exists $\epsilon_0 > 0$, such that, for any $0 < \epsilon_1, \epsilon_2 \le \epsilon_0$, we can take $\mathcal{E} = \mathcal{RN}(L_e, W_e, B_e, 2C_e)$ and $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ such that*

$$L_e \lesssim \log(1/\epsilon_e),\ W_e \lesssim \epsilon_e^{-s/\alpha_e} \log(1/\epsilon_e),\ \text{ and } B_e \lesssim \epsilon^{-1/\alpha_e};$$
$$L_d \lesssim \log(1/\epsilon_d),\ W_d \lesssim \epsilon_d^{-\ell/\alpha_d} \log(1/\epsilon_d),\ \text{ and } B_d \lesssim \epsilon_d^{-1/\alpha_d}.$$

Then, $\inf_{G \in \mathcal{G},\, E \in \mathcal{E}} \|(\mu, E_\sharp \mu) - (G_\sharp \nu, \nu)\|_\Psi \lesssim \epsilon_e^{\beta \wedge 1} + \epsilon_d^{\beta \wedge 1}$.

Proof We begin by noting that,

$$\begin{aligned}
&\|(\mu, E_\sharp \mu) - (D_\sharp \nu, \nu)\|_\Psi \\
=&\|(\mu, E_\sharp \mu) - (\mu, \tilde{E}_\sharp \mu)\|_\Psi + \|(\tilde{D}_\sharp \nu, \nu) - (D_\sharp \nu, \nu)\|_\Psi + \|(\mu, \tilde{E}_\sharp \mu) - (\tilde{D}_\sharp \nu, \nu)\|_\Psi \\
=&\|(\mu, E_\sharp \mu) - (\mu, \tilde{E}_\sharp \mu)\|_\Psi + \|(\tilde{D}_\sharp \nu, \nu) - (D_\sharp \nu, \nu)\|_\Psi \\
=&\sup_{\psi \in \Psi} \int \Big(\psi(x, E(x)) - \psi(x, \tilde{E}(x))\Big) d\mu(x) + \sup_{\psi \in \Psi} \int \Big(\psi(D(z), z) - \psi(\tilde{D}(z), z)\Big) d\nu(z) \\
\le& \int |E(x) - \tilde{E}(x)|^{\beta \wedge 1} d\mu(x) + \int |D(z) - \tilde{D}(z)|^{\beta \wedge 1} d\nu(z) \\
\le&\|E - \tilde{E}\|_{\mathbb{L}_1(\mu)}^{\beta \wedge 1} + \|D - \tilde{D}\|_{\mathbb{L}_1(\nu)}^{\beta \wedge 1}.
\end{aligned}$$

The Lemma follows from observing that $\tilde{E} \in \mathcal{H}^{\alpha_e}(\mathbb{R}^d, \mathbb{R}, C_e)$ and $\tilde{D} \in \mathcal{H}^{\alpha_d}(\mathbb{R}^\ell, \mathbb{R}, C_d)$ (by assumption B1) and applying Theorem 21 to approximate the coordinate-wise and stacking the networks parallelly. ■

Appendix E. Proofs from Section 4.3.3

E.1 Proof of Lemma 24

Lemma 24. *Suppose that* $n \ge 6$ *and* $\mathcal{F} : \mathbb{R}^d \to \mathbb{R}^{d'}$ *be a class of bounded neural networks with depth at most* L *and the number of weights at most* W*. Furthermore, the activation functions are piece-wise polynomial activation with the number of pieces and degree at most* $k \in \mathbb{N}$*. Then, there exists positive constants* θ *and* ϵ_0*, such that, if* $n \ge \theta(W + 6d' + 2d'L)(L + 3)\left(\log(W + 6d' + 2d'L) + L + 3\right)$ *and* $\epsilon \in (0, \epsilon_0]$*,*

$$\log \mathcal{N}(\epsilon; \mathcal{F}_{|X_{1:n}}, \ell_\infty) \lesssim (W + 6d' + 2d'L)(L + 3)\left(\log(W + 6d' + 2d'L) + L + 3\right) \log\left(\frac{nd'}{\epsilon}\right),$$

where d' *is the output dimension of the networks in* $\mathcal{F}$*.*

Proof We let, $h(x, y) = y^\top f(x)$ and let $\mathcal{H} = \{h(x, y) = y^\top f(x) : f \in \mathcal{F}\}$. Also, let, $\mathcal{T} = \{(h(X_i, e_\ell)|_{i \in [n], \ell \in [d']}) \in \mathbb{R}^{nd'} : h \in \mathcal{H}\}$. By construction of $\mathcal{T}$, it is clear that, $\mathcal{N}(\epsilon; \mathcal{F}_{|X_{1:n}}, \ell_\infty) = \mathcal{N}(\epsilon; \mathcal{T}, \ell_\infty)$. We observe that,

$$h(x, y) = \frac{1}{4}(\|y + f(x)\|_2^2 - \|y - f(x)\|_2^2)$$

Clearly, h can be implemented by a network with $\mathcal{L}(h) = \mathcal{L}(f) + 3$ and $\mathcal{W}(h) = \mathcal{W}(f) + 6d' + 2d'\mathcal{L}(f)$ (see Figure 2 for such a construction). Let $B = \max_{1 \le i \le d} \sup_{f \in \mathcal{F}} \|f_j\|_\infty$, where f_j denotes the j-th component of f. Thus, from Theorem 12.9 of Anthony and Bartlett (2009) (see Lemma 54), we note that, if $n \ge \text{Pdim}(\mathcal{H})$, $\mathcal{N}(\epsilon; \mathcal{T}, \ell_\infty) \le \left(\frac{2eBnd'}{\epsilon \text{Pdim}(\mathcal{H})}\right)^{\text{Pdim}(\mathcal{H})}$. Furthermore, by Lemma 55, $\text{Pdim}(\mathcal{H}) \lesssim (W + 6d' + 2d'L)(L + 3)\left(\log(W + 6d' + 2d'L) + L + 3\right)$. This implies that,

$$\log \mathcal{N}(\epsilon; \mathcal{H}, \ell_\infty) \lesssim \text{Pdim}(\mathcal{H}) \log\left(\frac{2eBnd'}{\epsilon \text{Pdim}(\mathcal{H})}\right)$$

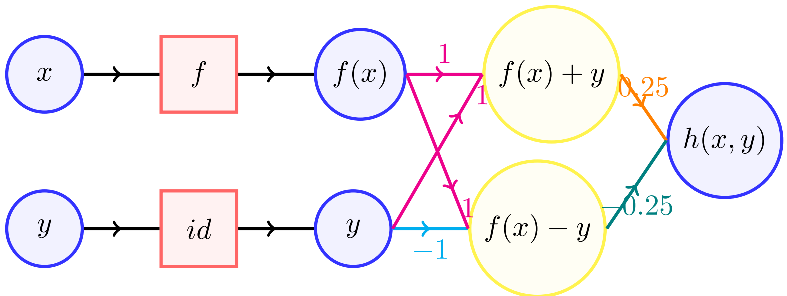

Figure 2: A representation of the network $h(\cdot,\cdot)$. The magenta lines represent d' weights of value 1. Similarly, cyan lines represent d' weights of value -1. Finally, the orange and teal lines represent d' weights (each) with values $+0.25$ and -0.25, respectively. The identity map takes $2d'\mathcal{L}(f)$ many weights (Nakada and Imaizumi, 2020, Remark 15 (iv)). The magenta, cyan, orange and teal connections take $6d'$ many weights. All activations are taken to be ReLU, except the output of the yellow nodes, whose activation is $\sigma(x) = x^2$.

$$
\begin{aligned}
&\lesssim \mathrm{Pdim}(\mathcal{H}) \log\left(\frac{nd'}{\epsilon}\right) \\
&\lesssim (W + 6d' + 2d'L)(L+3)\left(\log(W + 6d' + 2d'L) + L + 3\right)\log\left(\frac{nd'}{\epsilon}\right).
\end{aligned}
$$

■

E.2 Proof of Corollary 25

Corollary 25. *Suppose that $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ and $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ with $L_d, L_g \geq 3$, $W_g \geq 6d + 2dL_g$ and $W_d \geq 6d + 2dL_d$. Then, there is a constant c, such that, $m \geq c\left(W_g L_g(\log W_g + L_g)\right) \vee \left(W_d L_d(\log W_d + L_d)\right)$, then,*

$$
\log \mathcal{N}\left(\epsilon; (\Psi \circ G)_{|Z_{1:m}}, \ell_\infty\right) \lesssim \epsilon^{-\frac{d}{\beta}} + W_g L_g\left(\log W_g + L_g\right)\log\left(\frac{md}{\epsilon}\right),
$$

$$
\log \mathcal{N}\left(\epsilon; (\mathcal{F}_2)_{|Z_{1:m}}, \ell_\infty\right) \lesssim \epsilon^{-\frac{d+\ell}{\beta}} + W_d L_d\left(\log W_d + L_d\right)\log\left(\frac{md}{\epsilon}\right).
$$

Proof Let $\mathcal{V} = \{v_1, \dots, v_r\}$ be an optimal ℓ_∞ ϵ-cover of $\mathcal{G}_{|Z_{1:m}}$. Clearly, $\log r \lesssim W_g L_g(\log W_g + L_g)\log\left(\frac{md}{\epsilon}\right)$, by Lemma 24. Suppose that $\mathcal{F}_\delta = \{f_1, \dots, f_k\}$ be an optimal δ-cover of Φ w.r.t. the ℓ_∞-norm. By Lemma 52, we know that, $\log k \leq \delta^{-\frac{d}{\beta}}$. We note that, for any $f \in \mathcal{F}$, we can find $\tilde{f} \in \mathcal{F}_\delta$, such that, $\|f - \tilde{f}\|_\infty \leq \delta$. For any $Z = Z_{1:m}$, let $\|Z - v^Z\|_\infty \leq \epsilon$, with $v^Z \in \mathcal{V}$.

$$
\begin{aligned}
|f(Z_j) - \tilde{f}(v_j^Z)| \leq& |f(Z_j) - f(v_j^Z)| + |f(v_j^Z) - \tilde{f}(v_j^Z)| \\
\lesssim& \|Z_j - v_j^Z\|_\infty^{\beta \wedge 1} + \|f - \tilde{f}\|_\infty \lesssim \epsilon^{\beta\wedge 1} + \delta.
\end{aligned}
$$

Taking $\delta \asymp \eta/2$ and $\epsilon \asymp \left(\frac{\eta}{2}\right)^{1/(\beta\wedge 1)}$, we conclude that $\max_{1\le j\le m}|f(Z_j)-\tilde{f}(v_j^Z)|\le \eta$. Thus, $\{(\tilde{f}(v_1),\dots,\tilde{f}(v_m)) : \tilde{f}\in\mathcal{F}_\delta,\, v\in\mathcal{V}\}$ constitutes an η-net of $(\Phi\circ\mathcal{G})_{|Z_{1:m}}$. Hence,

$$\begin{aligned}\log\mathcal{N}(\eta;(\Phi\circ\mathcal{G})_{|Z_{1:m}},\ell_\infty) &\le \log\left(\mathcal{N}(\delta;\Phi,\ell_\infty)\times\mathcal{N}(\epsilon;\mathcal{G}_{|Z_{1:m}},\ell_\infty)\right)\\ &\lesssim \eta^{-\frac{d}{\beta}} + W_g L_g(\log W_g + L_g)\log\left(\frac{md}{\eta^{1/(\beta\wedge 1)}}\right)\\ &\lesssim \eta^{-\frac{d}{\beta}} + W_g L_g(\log W_g + L_g)\log\left(\frac{md}{\eta}\right)\end{aligned}$$

The latter part of the lemma follows from a similar calculation. ■

E.3 Proof of Lemma 26

Before we discuss the proof of Lemma 26, we state and prove the following result.

Lemma 44. $\inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\int_\delta^B \epsilon^{-\tau}d\epsilon\right)\lesssim\begin{cases} n^{-1/2}, & \text{if } \tau<1\\ n^{-1/2}\log n, & \text{if } \tau=1\\ n^{-\frac{1}{2\tau}}, & \text{if } \tau>1.\end{cases}$

Proof If $\tau<1$,

$$\begin{aligned}\inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\int_\delta^B\epsilon^{-\tau}d\epsilon\right) &= \inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\frac{\epsilon^{-\tau+1}}{-\tau+1}\Big|_\delta^B\right)\\ &= \inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\frac{B^{1-\tau}-\delta^{1-\tau}}{1-\tau}\right)\\ &\lesssim \frac{1}{\sqrt{n}}.\end{aligned}$$

If $\tau=1$, we observe that

$$\begin{aligned}\inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\int_\delta^B\epsilon^{-\tau}d\epsilon\right) &= \inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\log\epsilon\Big|_\delta^B\right)\\ &= \inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}(\log B-\log\delta)\right)\\ &\lesssim \frac{\log n}{\sqrt{n}},\end{aligned}$$

when $\delta\asymp n^{-1/2}$. Furthermore, if $\tau>1$,

$$\inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\int_\delta^B\epsilon^{-\tau}d\epsilon\right) = \inf_{0<\delta\le B}\left(\delta+\frac{1}{\sqrt{n}}\frac{\delta^{1-\tau}-B^{1-\tau}}{\tau-1}\right)\lesssim\ n^{-\frac{1}{2\tau}}, \text{when, } \delta\asymp n^{-\frac{1}{2\tau}}.$$

■

Lemma 26. *Suppose that* $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ *and* $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ *with* $L_d, L_g \geq 3$, $W_g \geq 6d + 2dL_g$ *and* $W_d \geq 6d + 2dL_d$. *Then, there is a constant* c*, such that,* $m \geq c\left(W_g L_g(\log W_g + L_g)\right) \vee \left(W_d L_d(\log W_d + L_d)\right)$*, then,*

$$\mathbb{E}\|\hat{\nu}_m - \nu\|_{\Phi\circ\mathcal{G}} \lesssim m^{-\frac{\beta}{d}} \vee m^{-1/2}\log m + \sqrt{\frac{W_g L_g(\log W_g + L_g)\log(md)}{m}},$$
$$\mathbb{E}\|\hat{\nu}_m - \nu\|_{\mathcal{F}_2} \lesssim m^{-\frac{\beta}{d+\ell}} \vee (m^{-1/2}\log m) + \sqrt{\frac{W_d L_d(\log W_d + L_d)\log(md)}{m}}.$$

Here c *is the same as in Corollary 25.*

Proof From Dudley's chaining argument, we note that,

$$\begin{aligned}
&\mathbb{E}\|\hat{\nu}_m - \nu\|_{\Phi\circ\mathcal{G}} \\
\lesssim &\inf_{0<\delta\leq 1/2}\left(\delta + \frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\log\mathcal{N}(\epsilon; \Phi\circ\mathcal{G}, \mathbb{L}_2(\hat{\nu}_m))}d\epsilon\right) \\
\leq &\inf_{0<\delta\leq 1/2}\left(\delta + \frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\log\mathcal{N}(\epsilon; (\Phi\circ\mathcal{G})_{|Z_{1:m}}, \ell_\infty)}d\epsilon\right) \\
\lesssim &\inf_{0<\delta\leq 1/2}\left(\delta + \frac{1}{\sqrt{m}}\int_\delta^{1/2}\sqrt{\epsilon^{-\frac{d}{\beta}} + W_g L_g(\log W_g + L_g)\log\left(\frac{md}{\epsilon}\right)}d\epsilon\right) \qquad (35) \\
\leq &\inf_{0<\delta\leq 1/2}\left(\delta + \frac{1}{\sqrt{m}}\int_\delta^{1/2}\left(\sqrt{\epsilon^{-\frac{d}{\beta}}} + \sqrt{W_g L_g(\log W_g + L_g)\log\left(\frac{md}{\epsilon}\right)}\right)d\epsilon\right) \\
\leq &\inf_{0<\delta\leq 1/2}\left(\delta + \frac{1}{\sqrt{m}}\int_\delta^{1/2}\epsilon^{-\frac{d}{2\beta}}d\epsilon + \frac{1}{\sqrt{m}}\int_0^{1/2}\sqrt{W_g L_g(\log W_g + L_g)\log\left(\frac{md}{\epsilon}\right)}d\epsilon\right) \\
\lesssim &\inf_{0<\delta\leq 1/2}\left(\delta + \frac{1}{\sqrt{m}}\int_\delta^{1/2}\epsilon^{-\frac{d}{2\beta}}d\epsilon + \sqrt{\frac{W_g L_g(\log W_g + L_g)\log(md)}{m}}\right) \\
\lesssim &m^{-\beta/d} \vee (m^{-1/2}\log m) + \sqrt{\frac{W_g L_g(\log W_g + L_g)\log(md)}{m}} \qquad (36)
\end{aligned}$$

In the above calculations, (35) follows from Corollary 25 and (36) follows from Lemma 44. The latter part of the lemma follows from a similar calculation as above. ■

E.4 Proof of Lemma 27

We first let $\Lambda = \{(\mu, E_\sharp\mu) : E \in \mathcal{E}\}$. We first argue that $d^*_\alpha(\Lambda) \leq d^*_\alpha(\mu)$. To see this, let $s > d^*_\alpha(\mu) + \ell$. We note that for any $\epsilon > 0$, we can find $S \subseteq [0,1]^d$, such that $\mu(S) \geq 1 - \epsilon^{\frac{s\alpha}{s-2\alpha}}$ and $\mathcal{N}(\epsilon; S, \ell_\infty) = \mathcal{N}_\epsilon(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}})$. We let $S' = S \times [0,1]^\ell$. Clearly, $[0,1]^\ell$ can be covered with $\epsilon^{-\ell}$-many ℓ_∞-balls, which implies that S' can be covered with at most $\mathcal{N}_\epsilon(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}) \times \epsilon^{-\ell}$-many ℓ_∞-balls. Thus, $\mathcal{N}_\epsilon(\Lambda, \epsilon^{\frac{s\alpha}{s-2\alpha}}) \leq \mathcal{N}_\epsilon(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}}) \times \epsilon^{-\ell}$. Thus,

$$\limsup_{\epsilon\downarrow 0}\frac{\log\mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \leq \limsup_{\epsilon\downarrow 0}\frac{\log\mathcal{N}_\epsilon(\mu, \epsilon^{\frac{s\alpha}{s-2\alpha}})}{\log(1/\epsilon)} + \ell \leq s + \ell.$$

Here, the last inequality follows from the fact that $s > d^*_\alpha(\mu)$ and applying Lemma 34. We now note that, $\frac{(s+\ell)\alpha}{s+\ell-2\alpha} \le \frac{s\alpha}{s-2\alpha} \implies \epsilon^{\frac{(s+\ell)\alpha}{s+\ell-2\alpha}} \ge \epsilon^{\frac{s\alpha}{s-2\alpha}} \implies \mathcal{N}_\epsilon(\Lambda, \epsilon^{\frac{(s+\ell)\alpha}{s+\ell-2\alpha}}) \le \mathcal{N}_\epsilon(\Lambda, \epsilon^{\frac{s\alpha}{s-2\alpha}})$. Thus,

$$\limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{(s+\ell)\alpha}{s+\ell-2\alpha}}\right)}{\log(1/\epsilon)} \le \limsup_{\epsilon\downarrow 0} \frac{\log \mathcal{N}_\epsilon\left(\Lambda, \epsilon^{\frac{s\alpha}{s-2\alpha}}\right)}{\log(1/\epsilon)} \le s+\ell \implies d^*_\alpha(\Lambda) \le s+\ell.$$

Thus, for any $s > d^*_\alpha(\mu)$, $d^*_\alpha(\Lambda) \le s+\ell$.

Lemma 27. *Suppose that $\mathcal{E} = \mathcal{RN}(L_e, W_e, B_e, 2C_e)$. Then, for any $s > d^*_\beta(\mu)$, we can find an $n' \in \mathbb{N}$, such that if $n \ge n'$,*

$$\mathbb{E}\|\hat{\mu}_n - \mu\|_{\mathcal{F}_1} \lesssim n^{-1/2}\left(W_e \log\left(2L_e B_e^{L_e}(W_e+1)^{L_e} n^{\frac{1}{2(\beta\wedge 1)}}\right)\right)^{1/2} + n^{-\frac{\beta}{s+\ell}}.$$

Proof Let, $Y_E = \sup_{\psi\in\Psi}\left(\frac{1}{n}\sum_{i=1}^n \psi(X_i, E(X_i)) - \int \psi(x, E(x))d\mu(x)\right)$. Clearly, if one replaces X_i with X_i', the change in Y_E is at most $\frac{2C_e}{n}$. Thus, by the bounded difference inequality,

$$\mathbb{P}\left(|Y_E - \mathbb{E}(Y_E)| \ge t\right) \le 2\exp\left(-\frac{nt^2}{2C_e^2}\right). \tag{37}$$

Inequality (37) implies that:

$$\mathbb{E}\exp\left(\lambda(Y_E - \mathbb{E}Y_E)\right) \le \exp\left(\frac{c'\lambda^2}{n}\right), \tag{38}$$

for some $c' > 0$, by Proposition 2.5.2 of Vershynin (2018). We thus note that,

$$\begin{aligned}
&\mathbb{E}\sup_{\psi\in\Psi, E\in\mathcal{E}} \frac{1}{n}\sum_{i=1}^n \psi(X_i, E(X_i)) - \int \psi(x, E(x))d\mu(x)\\
=&\mathbb{E}\sup_{E\in\mathcal{E}} Y_E\\
=&\mathbb{E}\sup_{E\in\mathcal{E}} \left(Y_E - \mathbb{E}(Y_E) + \mathbb{E}(Y_E)\right)\\
\le&\mathbb{E}\sup_{E\in\mathcal{E}} \left(Y_E - \mathbb{E}(Y_E)\right) + \sup_{E\in\mathcal{E}}\mathbb{E}Y_E\\
=&\mathbb{E}\sup_{\tilde{E}\in\mathcal{C}(\epsilon;\mathcal{E},\ell_\infty)}\ \sup_{E\in B_{\ell_\infty}(\tilde{E},\epsilon)} \left(Y_E - \mathbb{E}(Y_E)\right) + \sup_{E\in\mathcal{E}}\mathbb{E}Y_E\\
\lesssim&\mathbb{E}\sup_{\tilde{E}\in\mathcal{C}(\epsilon;\mathcal{E},\ell_\infty)} \left(Y_{\tilde{E}} - \mathbb{E}(Y_{\tilde{E}})\right) + \epsilon^{\beta\wedge 1} + \sup_{E\in\mathcal{E}}\mathbb{E}Y_E\\
\lesssim&\sqrt{\frac{\mathcal{N}(\epsilon;\mathcal{E},\ell_\infty)}{n}} + 4C\epsilon^{\beta\wedge 1} + \sup_{E\in\mathcal{E}}\mathbb{E}Y_E && (39)\\
\lesssim&\left(\frac{W_e}{n}\log\left(\frac{2L_e B_e^{L_e}(W_e+1)^{L_e}}{\epsilon}\right)\right)^{1/2} + 4C\epsilon^{\beta\wedge 1} + \sup_{E\in\mathcal{E}}\mathbb{E}Y_E && (40)
\end{aligned}$$

Here, (39) follows bounding the expectation of the maximum of subgaussian random variables (Boucheron et al., 2013, Theorem 2.5). Inequality (40) follows from Lemma 53. For any $E \in \mathcal{E}$, by applying Theorem 39, we note that,

$$\mathbb{E}Y_E = \mathbb{E}\sup_{\psi \in \Psi} \int \psi(x, E(x)) d\mu_n(x) - \int \psi(x, E(x)) d\mu(x) \lesssim n^{-\frac{\beta}{s+\ell}}.$$

Plugging this into (40) and taking $\epsilon = n^{-\frac{1}{2(\beta \wedge 1)}}$, we get that,

$$\begin{aligned}
&\mathbb{E} \sup_{\psi \in \Psi, E \in \Psi} \frac{1}{n}\sum_{i=1}^{n} \psi(X_i, E(X_i)) - \int \psi(x, E(x)) d\mu(x) \\
\lesssim &\sqrt{\frac{W_e}{n} \log\left(2L_e B_e^{L_e} (W_e+1)^{L_e} n^{\frac{1}{2(\beta \wedge 1)}}\right)} + \frac{1}{\sqrt{n}} + n^{-\frac{\beta}{s+\ell}} \\
\lesssim &\sqrt{\frac{W_e}{n} \log\left(2L_e B_e^{L_e} (W_e+1)^{L_e} n^{\frac{1}{2(\beta \wedge 1)}}\right)} + n^{-\frac{\beta}{s+\ell}}.
\end{aligned}$$

The last inequality follows since $s > d^*_\beta(\mu) \geq 2\beta$. ∎

Appendix F. Proofs of the Main Results (Section 4.2)

F.1 Proof of Theorem 16

Theorem 16 (Error rate for GANs)**.** *Suppose assumptions A1 and A2 hold and let $s > d^*_\beta(\mu)$. There exist constants N, c that might depend on d, ℓ, α_g, β and $\tilde{G}$, such that, if $n \geq N$, we can choose $\mathcal{G} = \mathcal{RN}(L_g, W_g, B_g, 2C_g)$ with the network parameters as $L_g \leq c \log n$, $W_g \leq c n^{\frac{\beta \ell}{\alpha_g s(\beta \wedge 1)}} \log n$. Then,*

$$\mathbb{E}\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \lesssim n^{-\beta/s}. \tag{7}$$

Furthermore, if $m \geq \upsilon_n$, where, $\upsilon_n = \inf\left\{m \in \mathbb{N} : \frac{(\log m)^2}{m^{\left(\max\left\{2+\frac{\ell}{\alpha_g(\beta\wedge 1)}, \frac{d}{\beta}\right\}\right)^{-1}}} \leq n^{-\beta/s}\right\}$, and the network parameters are chosen as, $L_g \leq c \log m$, $W_g \leq c m^{\frac{\ell}{2\alpha_g(\beta \wedge 1)+\ell}} \log m$, then

$$\mathbb{E}\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Phi \lesssim n^{-\beta/s}. \tag{8}$$

Proof We note from the oracle inequality (Lemma 20, inequality (15))

$$\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \leq \inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi$$

If $\mathcal{G}$ is chosen as in Lemma 22, then, for large n,

$$\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \lesssim \epsilon^{\beta \wedge 1} + n^{-\beta/s}.$$

We take $\epsilon = n^{-\frac{\beta}{(\beta \wedge 1)s}}$. This makes, $L_g \lesssim \log n$, $W_g \lesssim n^{\frac{\ell\beta}{s\alpha_g(\beta\wedge 1)}} \log n$ and gives the desired bound (7).

To derive (8), we note that, for the estimate $\hat{G}_{n,m}$, we note that, for the choice of $\mathcal{G}$ as in Lemma 22,

$$\begin{aligned}
&\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Psi \\
\le & \inf_{G \in \mathcal{G}} \|\mu - G_\sharp \nu\|_\Phi + 2\|\mu - \hat{\mu}_n\|_\Phi + 2\|\nu - \hat{\nu}_m\|_{\Phi \circ \mathcal{G}} \\
\lesssim & \epsilon^{\beta \wedge 1} + n^{-\beta/s} + m^{-\beta/d} \vee (m^{-1/2} \log m) + \sqrt{\frac{W_g L_g (\log W_g + L_g) \log(md)}{m}} \\
\lesssim & \epsilon^{\beta \wedge 1} + n^{-\beta/s} + m^{-\beta/d} \vee (m^{-1/2} \log m) + \sqrt{\frac{\epsilon^{-\ell/\alpha_g} (\log(1/\epsilon))^3 \log(md)}{m}}
\end{aligned}$$

We choose, $\epsilon \asymp m^{-\frac{1}{2(\beta\wedge 1) + \frac{\ell}{\alpha_g}}} = m^{-\frac{\alpha_g}{2\alpha_g(\beta\wedge 1)+\ell}}$. This makes,

$$\begin{aligned}
\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Psi \lesssim & n^{-\beta/s} + m^{-\beta/d} \vee (m^{-1/2} \log m) + (\log m)^2 m^{-\frac{\alpha_g(\beta\wedge 1)}{2\alpha_g(\beta\wedge 1)+\ell}} \\
\lesssim & n^{-\beta/s} + (\log m)^2 m^{-\frac{1}{\max\left\{2 + \frac{\ell}{\alpha_g(\beta\wedge 1)}, \frac{d}{\beta}\right\}}}.
\end{aligned}$$

We take $\upsilon_n = \inf \left\{ m \in \mathbb{N} : (\log m)^2 m^{-\frac{1}{\max\left\{2 + \frac{\ell}{\alpha_g(\beta\wedge 1)}, \frac{d}{\beta}\right\}}} \le n^{-\beta/s} \right\}$. Clearly, this if $m \ge \upsilon_n$, the choice of ϵ gives us the desired bound, i.e. (8) and also satisfies the bounds on L_g and W_g. ■

F.2 Proof of Theorem 17

Theorem 17 (Error rate for BiGANs). *Suppose assumptions A1 and B1 hold and let $s_1 > \bar{d}_{\alpha_e}(\mu)$ and $s_2 > d^*_\beta(\mu)$. There exists constants N, c that might depend on $d, \ell, \alpha_d, \alpha_e, \beta, \tilde{D}$ and $\tilde{E}$, such that, if $n \ge N$, we can choose the networks $\mathcal{E} = \mathcal{RN}(L_e, W_e, B_e, 2C_e)$ and $\mathcal{D} = \mathcal{RN}(L_d, W_d, B_d, 2C_d)$ with the network parameters chosen as*

$$L_e \le c \log n,\ W_e \le c n^{\frac{s_1}{2\alpha_e(\beta\wedge 1)+s_1}} \log n,\ L_d \le c \log n,\ W_d \le c n^{\frac{\beta\ell}{\alpha_d(\beta\wedge 1)(s_2+\ell)}} \log n. \tag{9}$$

Then,

$$\mathbb{E}\|(\mu, (\hat{E}_n)_\sharp \mu) - ((\hat{D}_n)_\sharp \nu, \nu)\|_\Psi \lesssim n^{-\frac{1}{\max\left\{2 + \frac{s_1}{\alpha_e(\beta\wedge 1)}, \frac{s_2+\ell}{\beta}\right\}}} (\log n)^2. \tag{10}$$

Furthermore, if the network parameters are chosen as

$$L_e \le c \log n,\ W_e \le c n^{\frac{s_1}{2\alpha_e(\beta\wedge 1)+s_1}} \log n,\ L_d \le c \log m,\ W_d \le c m^{\frac{\ell}{2\alpha_d(\beta\wedge 1)+\ell}} \log m, \tag{11}$$

then,

$$\begin{aligned}
&\mathbb{E}\|(\mu, (\hat{E}_{n,m})_\sharp \mu) - ((\hat{D}_{n,m})_\sharp \nu, \nu)\|_\Psi \\
\lesssim & n^{-\frac{1}{\max\left\{2 + \frac{s_1}{\alpha_e(\beta\wedge 1)}, \frac{s_2+\ell}{\beta}\right\}}} (\log n)^2 + m^{-\frac{1}{\max\left\{2 + \frac{d}{\alpha_g(\beta\wedge 1)}, \frac{\ell+d}{\beta}\right\}}} (\log m)^2.
\end{aligned} \tag{12}$$

Proof From Lemma 20, (17), we note that,

$$\begin{aligned}\|(\mu,(\hat{E}_n)_\sharp\mu)-(\hat{D}_n)_\sharp\nu,\nu)\|_\Psi \leq & \inf_{D\in\mathcal{D},\, E\in\mathcal{E}} \|(\mu,E_\sharp\mu)-(D_\sharp\nu,\nu)\|_\Psi + 2\|\hat{\mu}_n-\mu\|_{\mathcal{F}_1} \\ \lesssim & \epsilon_e^{\beta\wedge 1}+\epsilon_d^{\beta\wedge 1}+2\|\hat{\mu}_n-\mu\|_{\mathcal{F}_1},\end{aligned}$$

if $\mathcal{E}=\mathcal{R}(L_e,W_e,B_e,2C_e)$ and $\mathcal{D}=\mathcal{RN}(L_d,W_d,B_d,2C_d)$, with

$$\begin{aligned}&L_e\lesssim \log(1/\epsilon_e),\ W_e\lesssim \epsilon_e^{-s_1/\alpha_e}\log(1/\epsilon_e),\ \text{and } B_e\lesssim \epsilon^{-1/\alpha_e};\\ &L_d\lesssim \log(1/\epsilon_d),\ W_d\lesssim \epsilon_d^{-\ell/\alpha_d}\log(1/\epsilon_d),\ \text{and } B_d\lesssim \epsilon_d^{-1/\alpha_d},\end{aligned}$$

by Lemma 23. Applying Lemma 27, we get,

$$\begin{aligned}&\|(\mu,(\hat{E}_n)_\sharp\mu)-(\hat{D}_n)_\sharp\nu,\nu)\|_\Psi\\ \leq & \epsilon_e^{\beta\wedge 1}+\epsilon_d^{\beta\wedge 1}+\sqrt{\frac{W_e}{n}\log\left(2L_eB_e^{L_e}(W_e+1)^{L_e}n^{\frac{1}{2(\beta\wedge 1)}}\right)}+n^{-\frac{\beta}{s_2+\ell}}.\end{aligned}$$

We choose $\epsilon_e\asymp n^{-\frac{1}{2(\beta\wedge 1)+\frac{s_1}{\alpha_e}}}$ and $\epsilon_d\asymp n^{-\frac{\beta}{(\beta\wedge 1)(s+\ell)}}$. This makes,

$$\begin{aligned}\|(\mu,(\hat{E}_n)_\sharp\mu)-(\hat{D}_n)_\sharp\nu,\nu)\|_\Psi \lesssim & (\log n)^2 n^{-\frac{1}{2+\frac{s_1}{\alpha_e(\beta\wedge 1)}}}+n^{-\frac{\beta}{s_2+\ell}}\\ \lesssim & n^{-\frac{1}{\max\left\{2+\frac{s_1}{\alpha_e(\beta\wedge 1)},\frac{s_2+\ell}{\beta}\right\}}}(\log n)^2.\end{aligned}$$

For the latter part of the theorem, we note that,

$$\begin{aligned}&\|(\mu,(\hat{E}_{n,m})_\sharp\mu)-(\hat{D}_{n,m})_\sharp\nu,\nu)\|_\Psi\\ \leq & \inf_{D\in\mathcal{D},\, E\in\mathcal{E}}\|(\mu,E_\sharp\mu)-(D_\sharp\nu,\nu)\|_\Psi+2\|\hat{\mu}_n-\mu\|_{\mathcal{F}_1}+2\|\hat{\nu}_m-\nu\|_{\mathcal{F}_2}\\ \lesssim & \epsilon_e^{\beta\wedge 1}+\epsilon_d^{\beta\wedge 1}+\sqrt{\frac{W_e}{n}\log\left(2L_eB_e^{L_e}(W_e+1)^{L_e}n^{\frac{1}{2(\beta\wedge 1)}}\right)}+n^{-\frac{\beta}{s_2+\ell}}\\ &+m^{-\frac{\beta}{d+\ell}}\vee(m^{-1/2}\log m)+\sqrt{\frac{W_dL_d(\log W_d+L_d)\log(md)}{m}}\end{aligned}$$

We choose $\epsilon_e\asymp n^{-\frac{1}{2(\beta\wedge 1)+\frac{s_1}{\alpha_e}}}$ and $\epsilon_d\asymp m^{-\frac{1}{2(\beta\wedge 1)+\frac{\ell}{\alpha_d}}}$. This makes,

$$\begin{aligned}&\|(\mu,(\hat{E}_{n,m})_\sharp\mu)-(\hat{D}_{n,m})_\sharp\nu,\nu)\|_\Psi\\ \lesssim & (\log n)^2 n^{-\frac{1}{2+\frac{s_1}{\alpha_e(\beta\wedge 1)}}}+n^{-\frac{\beta}{s_2+\ell}}+m^{-\frac{\beta}{d+\ell}}\vee(m^{-1/2}\log m)+(\log m)^2 m^{-\frac{1}{2+\frac{\ell}{\alpha_d(\beta\wedge 1)}}}.\end{aligned}$$

■

Appendix G. Proofs from Section 5

G.1 Proof of Theorem 28

Lemma 45. *Suppose that γ is a measure supported on k points on $[0,1]^d$ and ν be an absolutely continuous distribution on $[0,1]^\ell$. Then, we can choose W and L, such that, $W \gtrsim k$ and $\inf_{f \in \mathcal{RN}(L,W,\infty,1)} \mathcal{W}_1(\gamma, f_\sharp \nu) = 0$.*

Proof We take the first layer of f to be mapping to $\mathbb{R}$ with the output being the first coordinate variable. We then pass this output through a network of width w and depth L, such that $k \le \frac{w-d-1}{2} \lfloor (w-d-1)/(6d) \rfloor \lfloor L/2 \rfloor + 2$. Since the output of the first layer is still absolutely continuous, the network can make the Wasserstein distance between μ and $f_\sharp \nu$, arbitrarily small by Lemma 3.1 of Yang et al. (2022). By the choice of the network width and depth, we note that, $k \lesssim w^2 L \asymp W$. Since, by construction, the networks can be taken as linear interpolators taking values in $[0,1]^d$, the network output remains bounded by 1 in the ℓ_∞-norm. ■

Lemma 46. *For the GAN estimators $\hat{G}_n$ and $\hat{G}_{n,m}$, we observe that,*

$$\begin{aligned}
\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \le & \|\hat{\mu}_n - G_\sharp \nu\|_\Phi + \|\mu - \hat{\mu}_n\|_\Phi \\
\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Phi \le & \|\mu_n - G_\sharp \nu\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + 2\|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}}.
\end{aligned}$$

Proof For any $G \in \mathcal{G}$, we observe that,

$$\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \le \|\hat{\mu}_n - (\hat{G}_n)_\sharp \nu\|_\Phi + \|\mu - \hat{\mu}_n\|_\Phi \le \|\hat{\mu}_n - G_\sharp \nu\|_\Phi + \|\mu - \hat{\mu}_n\|_\Phi$$

Taking infimum on both sides w.r.t. $G \in \mathcal{G}$ gives us the desired result. For the estimator $\hat{G}_{n,m}$, we note that,

$$\begin{aligned}
\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Phi \le & \inf_{G \in \mathcal{G}} \|\hat{\mu}_n - (\hat{G}_{n,m})_\sharp \hat{\nu}_m\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + \|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}} \\
\le & \inf_{G \in \mathcal{G}} \|\hat{\mu}_n - G_\sharp \hat{\nu}_m\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + \|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}} \\
\le & \|\mu_n - G_\sharp \nu\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + 2\|\hat{\nu}_m - \nu\|_{\Phi \circ \mathcal{G}}
\end{aligned}$$

Again taking infimum on both sides w.r.t. $G \in \mathcal{G}$ gives us the desired result. ■

Theorem 28. *Suppose Assumption A1 holds and let ν be absolutely continuous on $[0,1]^\ell$. Then, if $s > d^*_\beta(\mu)$, one can choose $\mathcal{G} = \mathcal{RN}(L_g, W_g, \infty, 1)$ with the network parameters as $L_g \ge 2$ as a constant and $W_g \asymp n$, such that $\mathbb{E}\|\mu - (\hat{G}_n)_\sharp \nu\|_\Phi \lesssim n^{-\beta/s}$. Furthermore, if $m \ge n^{d/s+1}$, then $\mathbb{E}\|\mu - (\hat{G}_{n,m})_\sharp \nu\|_\Phi \lesssim n^{-\beta/s}$.*

Proof We first note that for any two random variables, V_1 and V_2, that follow the distributions γ_1 and γ_2, respectively,

$$\|\gamma_1 - \gamma_2\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} = \sup_{f \in \mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \mathbb{E}(f(V_1) - f(V_2)) \le \mathbb{E}\|V_1 - V_2\|^{\beta \wedge 1} \overset{(i)}{\le} (\mathbb{E}\|V_1 - V_2\|)^{\beta \wedge 1}$$

Here (i) follows from applying Jensen's inequality. Taking an infimum with respect to the joint distribution of V_1 and V_2 yields, $\|\gamma_1 - \gamma_2\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \le \mathcal{W}_1(V_1, V_2)^{\beta\wedge 1}$.

From Lemmas 45 and 46, we note that for the GAN estimator, $\hat{G}_n$, we note that, if $\mathcal{G} = \mathcal{RN}(L_g, W_g, \infty, R_g)$, with $W_g \ge n$

$$\begin{aligned}\mathbb{E}\|\mu - (\hat{G}_n)_\sharp\nu\|_\Phi \le & \mathbb{E} \inf_{G\in\mathcal{G}} \|\hat{\mu}_n - G_\sharp\nu\|_\Phi + \mathbb{E}\|\mu - \hat{\mu}_n\|_\Phi \\ \le & \mathbb{E} \inf_{G\in\mathcal{G}} \mathcal{W}_1(\hat{\mu}_n, G_\sharp\nu)^{\beta\wedge 1} + \mathbb{E}\|\mu - \hat{\mu}_n\|_\Phi \\ = & \mathbb{E}\|\mu - \hat{\mu}_n\|_\Phi \lesssim n^{-\frac{\beta}{s}}\end{aligned}$$

Similarly, for the GAN estimator, $\hat{G}_{n,m}$, if $\mathcal{G} = \mathcal{RN}(L_g, W_g, \infty, 1)$, with $L_g \ge 2$ as constant and $W_g \asymp n$, we observe that,

$$\begin{aligned}\mathbb{E}\|\mu - (\hat{G}_{n,m})_\sharp\nu\|_\Phi \le & \mathbb{E} \inf_{G\in\mathcal{G}} \|\mu_n - G_\sharp\nu\|_\Phi + \|\hat{\mu}_n - \mu\|_\Phi + 2\|\hat{\nu}_m - \nu\|_{\Phi\circ\mathcal{G}} \\ \lesssim & n^{-\beta/s} + m^{-\beta/d} \vee m^{-1/2}\log m + \sqrt{\frac{W_g L_g(\log W_g + L_g)\log(md)}{m}} \\ \lesssim & n^{-\beta/s} + m^{-\beta/d} \vee m^{-1/2}\log m + \sqrt{\frac{n\log n\log(md)}{m}} \\ \lesssim & n^{-\beta/s},\end{aligned}$$

taking $m \ge n^{d/s+1}$. ■

G.2 Proof of Theorems 29 and 30

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, $\text{sep}(\Xi) = \inf_{\xi,\xi'\in\Xi:\xi\ne\xi'} \|\xi - \xi'\|_\infty$.

Lemma 47. *Let Ξ be a finite subset of $\mathbb{R}^p$ and let, $P, Q \in \Pi_\Xi$. Then,*

$$\|P - Q\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)} \gtrsim (sep(\Xi))^\beta \|P - Q\|_{TV}.$$

Proof Let $b(x) = \exp\left(\frac{1}{x^2-1}\right)\mathbb{1}\{|x| \le 1\}$ 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 $ab(x) \in \mathcal{H}^\beta(\mathbb{R}, \mathbb{R}, C)$. It is easy to observe that $h_\delta \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$. Let P and Q be two distributions on $\Xi = \{\xi_1, \dots, \xi_k\}$. Let $\delta = \frac{1}{3}\min_{i\ne j} \|\xi_i - \xi_j\|_\infty$ 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 $\mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$ and have disjoint supports, $h^\star \in \mathcal{H}^\beta(\mathbb{R}^d, \mathbb{R}, C)$. We take $\alpha_i = 2\mathbb{1}\{P(\xi_i) \ge Q(\xi_i)\} - 1$. Thus,

$$\|P - Q\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)} \ge \int h^* dP - \int h^* dQ = \sum_{i=1}^k a\delta^\beta \alpha_i (P(\xi_i) - Q(\xi_i))$$

$$
\begin{aligned}
&= a\delta^\beta \sum_{i=1}^k |P(\xi_i) - Q(\xi_i)| \\
&= 2a\delta^\beta \|P - Q\|_{\text{TV}} \\
&\gtrsim (\text{sep}(\Xi))^\beta \|P - Q\|_{\text{TV}}.
\end{aligned}
$$

■

To derive our main result, we also need to show that there indeed exists a distribution $\mu_\star \in \mathbb{M}_{d_\star}$, such that, $d_*(\mu_\star) = d_\star$. Proposition 49 gives a construction that ensures the existence of such a distribution. Before proceeding, we recall the definition of the Hausdorff dimension of a measure. We adopt the definition stated by Weed and Bach (2019).

Definition 48 (Hausdorff dimension (Weed and Bach, 2019))**.** *The d-Hausdroff measure of a set S is defined as,*

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

The Hausdorff dimension of the set S is defined as, $\dim_H(S) := \inf\{d : \mathbb{H}^d(S) = 0\}$. *For a measure μ, then the Hausdorff dimension of μ is defined as:*

$$
\dim_H(S) := \inf \left\{ \dim_H(S) : \mu(S) = 1 \right\}.
$$

Proposition 49. *For any $d_\star \in [0, d]$, one can find a distribution $\mu_\star$ on $[0,1]^d$, such that, $d_*(\mu_\star) = \overline{dim}_M(\mu_\star) = d_\star$.*

Proof If $d_\star \in \mathbb{N}$, then, the distribution which has is i.i.d. uniform on the first $d_\star$ coordinates and is identically 0, otherwise satisfies $d_*(\mu_\star) = d_\star$. Now suppose that $d_\star \notin \mathbb{N}$. Let $\alpha = 1 - \exp\left(\left(1 - \frac{1}{d_\star - \lfloor d_\star \rfloor}\right) \log 2\right)$. It is well known that the Cantor set C_α, which is constructed by removing the middle α (instead of 1/3-rd in regular Cantor set) has a Hausdorff and upper Minkowski dimension of $\frac{\log 2}{\log 2 - \log(1-\alpha)} = d_\star - \lfloor d_\star \rfloor$. A proof of this result can be found in (Bishop and Peres, 2017, Example 1.1.3) and (Ziemer and Torres, 2017, Chapter 4.8). Let $\text{Cf}_\alpha : [0,1] \to [0,1]$ be the corresponding Cantor function. It is easy to check that Cf_α satisfies all the properties of a cumulative distribution function (CDF). We let U_1 be a random variable, whose CDF is Cf_α. It is known (Presnell, 2022) that the support of U_1 is C_α. To construct $\mu_\star$, we define $U_2, \dots, U_{\lfloor d_\star \rfloor + 1} \overset{i.i.d.}{\sim} \text{Unif}([0,1])$ and let, $U = (U_1, \dots, U_{\lfloor d_\star \rfloor + 1}, 0, \dots, 0)$. The distribution of U is denoted as $\mu_\star$. First note that,

$$
\overline{\dim}_M(\mu) = \overline{\dim}_H(C_\alpha \times [0,1]^{\lfloor d_\star \rfloor}) \le \overline{\dim}_M(C_\alpha) + \lfloor d_\star \rfloor = d_\star.
$$

Similarly, we note that,

$$
\dim_H(\mu) = \dim_H(C_\alpha \times [0,1]^{\lfloor d_\star \rfloor}) \ge \dim_H(C_\alpha) + \lfloor d_\star \rfloor = d_\star.
$$

Applying Proposition 9, we observe that $d_*(\mu_\star) = d_\star$. ■

Corollary 50. *Suppose that $d \geq 2\beta$ and $d^\star \in [2\beta, d]$. Then, one can find a distribution $\mu^\star$ on $[0,1]^d$, such that, $d^*_\beta(\mu^\star) = d_*(\mu^\star) = d^\star$.*

Proof It is easy to see that by Proposition 49, one can find $\mu^\star$, such that $d_*(\mu^\star) = \overline{\dim}_M(\mu^\star) = d^\star$. If $d^\star \in [2\beta, d]$, then, by Proposition 9, we note that, $d^\star = d_*(\mu^\star) \leq d^*_\beta(\mu^\star) \leq \overline{\dim}_M(\mu^\star) = d^\star$. Hence the result. ■

Theorem 29. *Suppose that $d \geq 2\beta$ and let $d^\star \in [2\beta, d]$. Then for any $s < d^\star$,*

$$\mathfrak{M}_n = \inf_{\hat{\mu}} \sup_{\mu \in \mathbb{M}^{d^\star,\beta}} \mathbb{E}_\mu \|\hat{\mu} - \mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \gtrsim n^{-\beta/s},$$

where the infimum is taken over all measurable estimates of μ, based on the data.

Proof Let $\mu^\star \in \mathbb{M}^{d^\star,\beta}$ be such that $d_*(\mu^\star) = d^\star$ (this existence is guaranteed by Corollary 50). Now let, $s < d_*(\mu) = d^\star$. We can find a $\tau^\star$ and $\epsilon^\star$, such that $\mathcal{N}_\epsilon(\mu^\star, \tau) \geq \epsilon^{-s}$, for any $\epsilon \in (0, \epsilon^\star]$ and $\tau \in (0, \tau^\star]$. We can thus find a set $S^\star \subseteq [0,1]^d$, such that $\mathcal{M}(\epsilon; S^\star, \varrho) \geq \mathcal{N}(\epsilon; S^\star, \varrho) \geq \epsilon^{-s}$. We take $n \geq (128(\epsilon^\star)^{-s}) \vee 8192$. Suppose $\epsilon = (n/128)^{-1/s}$. Let $\Theta = \{\theta_1, \dots, \theta_k\}$ be a ϵ-separated set of $S^\star$. 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 $j = 1, \dots, \lfloor k/2 \rfloor$. 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, $d^*_\beta(P_{\boldsymbol{\omega}}) \leq d^\star$. Furthermore, $\|P_{\boldsymbol{\omega}} - P_{\boldsymbol{\omega}'}\|_{\text{TV}} = \frac{\delta_k}{k}\|\boldsymbol{\omega} - \boldsymbol{\omega}'\|_1$. By the Varshamov-Gilbert bound (Lemma 56), 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,

$$\|P_{\boldsymbol{\omega}} - P_{\boldsymbol{\omega}'}\|_{\text{TV}} \geq \frac{\delta_k \lfloor k/2 \rfloor}{8k}. \tag{41}$$

Hence, by Lemma 47, $\|P_{\boldsymbol{\omega}} - P_{\boldsymbol{\omega}'}\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \gtrsim \epsilon^\beta \frac{\delta_k \lfloor k/2 \rfloor}{k}$. Furthermore, we observe that

$$\begin{aligned} \text{KL}(P_{\boldsymbol{\omega}}^{\otimes n} \| P_{\boldsymbol{\omega}'}^{\otimes n}) = n\,\text{KL}(P_{\boldsymbol{\omega}} \| P_{\boldsymbol{\omega}'}) \leq \; n\,\chi^2(P_{\boldsymbol{\omega}} \| P_{\boldsymbol{\omega}'}) =& n \sum_{i=1}^k \frac{(P_{\boldsymbol{\omega}}(\xi_i) - P_{\boldsymbol{\omega}'(\xi_i)})^2}{P_{\boldsymbol{\omega}}(\xi_i)} \\ \leq& 2nk \sum_{i=1}^k (P_{\boldsymbol{\omega}}(\xi_i) - P_{\boldsymbol{\omega}'}(\xi_i))^2 \\ \leq& 2nk \lfloor k/2 \rfloor (2\delta_k/k)^2 \\ =& 8\frac{n \lfloor k/2 \rfloor \delta_k^2}{k}. \end{aligned}$$

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

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

$$\text{Thus, } \frac{I(Z;J)+\log 2}{\log|\Omega|} \le 8\frac{I(Z;J)+\log 2}{\lfloor k/2\rfloor\log 2} \le 64\frac{n\delta_k^2}{k\log 2}+\frac{8}{\lfloor k/2\rfloor} \le 64\frac{n\delta_k^2}{k\log 2}+\frac{1}{4}. \tag{42}$$

The last inequality follows since $k\ge 64$. We take $\delta_k = \frac{1}{16}\sqrt{\frac{k\log 2}{n}}$. Clearly, $\epsilon\le 1/2$ as $n\ge 64k$. This choice of ϵ makes, $\frac{I(Z;J)+\log 2}{\log|\Omega|}\le\frac{1}{2}$. Thus, by Theorem 15.2 of Wainwright (2019),

$$\begin{aligned}\inf_{\hat\mu}\sup_{\mu\in\mathbb{M}}\|\hat\mu-\mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)} \gtrsim \frac{\epsilon^\beta\delta_k\lfloor k/2\rfloor}{k} = \frac{\epsilon^\beta\lfloor k/2\rfloor}{16k}\sqrt{\frac{k\log 2}{n}} &\ge \frac{\epsilon^\beta\lfloor k/2\rfloor}{16k}\sqrt{\frac{\log 2}{128}}\\ &\gtrsim \epsilon^\beta\\ &\asymp n^{-\beta/s}.\end{aligned}$$

■

The proof of Theorem 30 is similar to that of Theorem 29. For the sake of completeness, the proof is provided below.

Theorem 30. *Suppose $d_\star\in[0,d]$. Then for any $s<d_\star$,*

$$\mathfrak{M}_n = \inf_{\hat\mu}\sup_{\mu\in\mathbb{M}_{d_\star}}\mathbb{E}_\mu\|\hat\mu-\mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)} \gtrsim n^{-\beta/s}\vee n^{-1/2},$$

where the infimum is taken over all measurable estimates of μ, based on the data.

Proof Let $\mu_\star\in\mathbb{M}_{d_\star}$ be such that $d_*(\mu_\star)=d_\star$ (this existence is guaranteed by Proposition 49). Now let, $s<d_*(\mu)=d_\star$. We can find a $\tau_\star$ and $\epsilon_\star$, such that $\mathcal{N}_\epsilon(\mu_\star,\tau)\ge\epsilon^{-s}$, for any $\epsilon\in(0,\epsilon_\star]$ and $\tau\in(0,\tau_\star]$. We can thus find a set $S_\star\subseteq[0,1]^d$, such that $\mathcal{M}(\epsilon;S_\star,\varrho)\ge\mathcal{N}(\epsilon;S_\star,\varrho)\ge\epsilon^{-s}$. We take $n\ge(128\epsilon_\star^{-s})\vee 8192$. Suppose $\epsilon=(n/128)^{-1/s}$. Let $\Theta=\{\theta_1,\dots,\theta_k\}$ be a ϵ-separated set of $S_\star$. For the above choices of n and ϵ, we observe that, $k=\epsilon^{-s}=n/128\ge 64$ and $n\ge 64k$,

Let $p_j(x)=\mathbb{1}\{x=\theta_j\}-\mathbb{1}\{x=\theta_{\lfloor k/2\rfloor+j}\}$, for all $j=1,\dots,\lfloor k/2\rfloor$. 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 $\delta_k\in(0,1/2]$. Furthermore, $\|P_{\boldsymbol{\omega}}-P_{\boldsymbol{\omega}'}\|_{\mathrm{TV}}=\frac{\delta_k}{k}\|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_1$. By the Varshamov-Gilbert Bound (Lemma 56), let $\Omega\subseteq\{0,1\}^{\lfloor k/2\rfloor}$ be such that $|\Omega|\ge 2^{\frac{1}{8}\lfloor k/2\rfloor}$ and $\|\boldsymbol{\omega}-\boldsymbol{\omega}'\|_1\ge\frac{1}{8}\lfloor k/2\rfloor$, for all $\boldsymbol{\omega}\ne\boldsymbol{\omega}'$ both in Ω. Thus for any $\boldsymbol{\omega}\ne\boldsymbol{\omega}'$, both in Ω, $\|P_{\boldsymbol{\omega}}-P_{\boldsymbol{\omega}'}\|_{\mathrm{TV}}\ge\frac{\delta_k\lfloor k/2\rfloor}{8k}$. Hence, by Lemma 47, $\|P_{\boldsymbol{\omega}}-P_{\boldsymbol{\omega}'}\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},1)}\gtrsim\epsilon^\beta\frac{\delta_k\lfloor k/2\rfloor}{k}$. Furthermore, we observe that $\mathrm{KL}(P_{\boldsymbol{\omega}}^{\otimes n}\|P_{\boldsymbol{\omega}'}^{\otimes n})\le\frac{n\lfloor k/2\rfloor\delta_k^2}{k}$.

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

$$\frac{I(Z;J)+\log 2}{\log|\Omega|}\le 8\frac{I(Z;J)+\log 2}{\lfloor k/2\rfloor\log 2}\le 64\frac{n\delta_k^2}{k\log 2}+\frac{8}{\lfloor k/2\rfloor}\le 64\frac{n\delta_k^2}{k\log 2}+\frac{1}{4}. \tag{43}$$

The last inequality follows since $k\ge 64$. We take $\delta_k=\frac{1}{16}\sqrt{\frac{k\log 2}{n}}$. Clearly, $\epsilon\le 1/2$ as $n\ge 64k$. This choice of ϵ makes, $\frac{I(Z;J)+\log 2}{\log|\Omega|}\le\frac{1}{2}$. Thus, by Theorem 15.2 of Wainwright (2019),

$$\begin{aligned}\inf_{\hat\mu}\sup_{\mu\in\mathbb{M}_{d_\star}}\|\mu-\hat\mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)}\gtrsim\epsilon^\beta\frac{\delta_k\lfloor k/2\rfloor}{k}&=\frac{\epsilon^\beta\lfloor k/2\rfloor}{16k}\sqrt{\frac{k\log 2}{n}}\\&\ge\frac{\epsilon^\beta\lfloor k/2\rfloor}{16k}\sqrt{\frac{\log 2}{128}}\\&\gtrsim\epsilon^\beta\asymp n^{-\beta/s}.\end{aligned}$$

To show that $\inf_{\hat\mu}\sup_{\mu\in\mathbb{M}_{d_\star}}\|\hat\mu-\mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)}\gtrsim n^{-1/2}$, we use Le Cam's method (Wainwright, 2019, Chapter 15.2). Let $\theta_0,\theta_1\in[0,1]^d$ be such that $\|\theta_0-\theta_1\|_\infty\ge 1/2$. 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, $P_0,P_1\in\mathbb{M}_{d_\star}$. Further, $\mathrm{TV}(P_0,P_1)=\delta$. Thus, by Lemma 47, we observe that

$$\|P_1-P_0\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)}\gtrsim 2^{-\beta}\delta\gtrsim\delta.$$

Again,

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

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

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

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

$$\inf_{\hat\mu}\sup_{\mu\in\mathbb{M}_{d_\star}}\mathbb{E}_\mu\|\hat\mu-\mu\|_{\mathcal{H}^\beta(\mathbb{R}^d,\mathbb{R},C)}\gtrsim\delta\asymp 1/\sqrt{n}. \tag{44}$$

■

Appendix H. Supporting Results

This section states some of the supporting results used in this paper.

Lemma 51 (Lemma 5.5 of Wainwright (2019)). *For any metric space,* (S,ϱ) *and* $\epsilon>0$, $M(2\epsilon;S,\varrho)\le\mathcal{N}(\epsilon;S,\varrho)\le M(\epsilon;S,\varrho)$.

Lemma 52 (Theorem XIV of Kolmogorov and Tikhomirov (1961); also see page 114 of Shiryayev (1992)). *Suppose that $\mathcal{M}$ be a subset of a real vector space, then for any $s > \overline{dim}_M(\mathcal{M})$, there exists $\epsilon_0 \in (0,1)$, such that if $\epsilon \in (0, \epsilon_0]$,*

$$\log \mathcal{N}\left(\epsilon; \mathcal{H}^\beta(\mathcal{M}), \|\cdot\|_\infty\right) \lesssim \epsilon^{-s/\beta}.$$

Furthermore, $\log \mathcal{N}\left(\epsilon; \mathcal{H}^\beta([0,1]^d), \|\cdot\|_\infty\right) \lesssim \epsilon^{-d/\beta}$.

Lemma 53 (Lemma 21 of Nakada and Imaizumi (2020)). *Let $\mathcal{F} = \mathcal{RN}(W, L, B)$ be a space of ReLU networks with the number of weights, the number of layers, and the maximum absolute value of weights bounded by W, L, and B respectively. Then,*

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty) \le W \log\left(2LB^L(W+1)^L \frac{1}{\epsilon}\right).$$

Lemma 54 (Theorem 12.2 of Anthony and Bartlett (2009)). *Assume for all $f \in \mathcal{F}$, $\|f\|_\infty \le M$. Denote the pseudo-dimension of $\mathcal{F}$ as $Pdim(\mathcal{F})$, then for $n \ge Pdim(\mathcal{F})$, we have for any ϵ and any $X_1, \dots, X_n$, $\mathcal{N}(\epsilon; \mathcal{F}_{|X_{1:n}}, \ell_\infty) \le \left(\frac{2eMn}{\epsilon Pdim(\mathcal{F})}\right)^{Pdim(\mathcal{F})}$.*

Lemma 55 (Theorem 6 of Bartlett et al. (2019)). *Consider the function class computed by a feed-forward neural network architecture with W many weight parameters and U many computation units arranged in L layers. Suppose that all non-output units have piecewise-polynomial activation functions with $p+1$ pieces and degrees no more than d, and the output unit has the identity function as its activation function. Then the VC-dimension and pseudo-dimension are upper-bounded as*

$$VCdim(\mathcal{F}), Pdim(\mathcal{F}) \le C \cdot LW \log(pU) + L^2 W \log d.$$

Lemma 56 (Lemma 2.9 of Tsybakov (2009)). *Let $m \ge 8$. Then there exists a subset $\{\omega_0, \dots, \omega_M\}$ of $\{0,1\}^m$ such that $\omega_0 = (0, \dots, 0)$, $\|\omega_j - \omega_k\|_1 \ge \frac{m}{8}$, for all $0 \le j < k \le M$, and $M \ge 2^{m/8}$.*

References

Hamed Alqahtani, Manolya Kavakli-Thorne, and Gulshan Kumar. Applications of generative adversarial networks (gans): An updated review. *Archives of Computational Methods in Engineering*, 28(2):525–552, 2021.

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

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

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. 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. *arXiv preprint arXiv:2011.12087*, 2020.

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

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.

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. *The 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*, pages 177–186, 2007.

Christopher J Bishop and Yuval Peres. *Fractals in probability and analysis*, volume 162. Cambridge University Press, 2017.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

Guy Bresler and Dheeraj Nagaraj. Sharp representation theorems for ReLU networks with precise dependence on depth. *Advances in Neural Information Processing Systems*, 33: 10697–10706, 2020.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Efficient approximation of deep ReLU networks for functions on low dimensional manifolds. *Advances in neural information processing systems*, 32, 2019.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Distribution approximation and statistical estimation guarantees of generative adversarial networks. *arXiv preprint arXiv:2002.03938*, 2020.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

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.

Jeff Donahue, Philipp Krähenbühl, and Trevor Darrell. Adversarial feature learning. In *International Conference on Learning Representations*, 2017. URL `https://openreview.net/forum?id=BJtNZAFgg`.

Richard Mansfield Dudley. The speed of mean glivenko-cantelli convergence. *The Annals of Mathematical Statistics*, 40(1):40–50, 1969.

Gintare Karolina Dziugaite, Daniel M Roy, and Zoubin Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, pages 258–267, 2015.

Jonathan M Fraser and Douglas C Howroyd. On the upper regularity dimensions of measures. *arXiv preprint arXiv:1706.09340*, 2017.

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.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

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*, pages 3398–3407. PMLR, 2019.

Taeksoo Kim, Moonsu Cha, Hyunsoo Kim, Jung Kwon Lee, and Jiwon Kim. Learning to discover cross-domain relations with generative adversarial networks. In *International conference on machine learning*, pages 1857–1865. PMLR, 2017.

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

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabás Póczos. Mmd gan: Towards deeper understanding of moment matching network. *Advances in neural information processing systems*, 30, 2017.

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

Shiao Liu, Yunfei Yang, Jian Huang, Yuling Jiao, and Yang Wang. Non-asymptotic error bounds for bidirectional gans. *Advances in Neural Information Processing Systems*, 34: 12328–12339, 2021.

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

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

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`.

Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. *Advances in neural information processing systems*, 29, 2016.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

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.

Edward C Posner, Eugene R Rodemich, and Howard Rumsey Jr. Epsilon entropy of stochastic processes. *The Annals of Mathematical Statistics*, pages 1000–1020, 1967.

Brett Presnell. A geometric derivation of the cantor distribution. *The American Statistician*, 76(1):73–77, 2022.

Yu. V. Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956. doi: 10.1137/1101016. URL `https://doi.org/10.1137/1101016`.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4):1875 – 1897, 2020. doi: 10.1214/19-AOS1875. URL `https://doi.org/10.1214/19-AOS1875`.

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

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Optimal approximation rate of relu networks in terms of width and depth. *Journal de Mathématiques Pures et Appliquées*, 157: 101–135, 2022.

Albert Nikolaevich Shiryayev. *Selected works of AN Kolmogorov: Volume III Information Theory and the Theory of Algorithms*, volume XXVIII. Springer Science & Business Media, 1992.

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

Taiji Suzuki. Adaptivity of deep ReLU network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=H1ebTsActm.

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.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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.

Jonathan Weed and Quentin Berthet. Minimax estimation of smooth densities in wasserstein distance. *The Annals of Statistics*, 50(3):1519–1540, 2022.

Yunfei Yang, Zhen Li, and Yang Wang. On the capacity of deep generative networks for approximating distributions. *Neural Networks*, 145:144–154, 2022.

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

Zili Yi, Hao Zhang, Ping Tan, and Minglun Gong. Dualgan: Unsupervised dual learning for image-to-image translation. In *Proceedings of the IEEE international conference on computer vision*, pages 2849–2857, 2017.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

William P Ziemer and Monica Torres. *Modern Real Analysis*, volume 278. Springer, 2017.