Journal of Machine Learning Research 23 (2022) 1-43 Submitted 7/21; Revised 2/22; Published 4/22

An Error Analysis of Generative Adversarial Networks for Learning Distributions

Jian Huang JIAN-HUANG@UIOWA.EDU
Department of Statistics and Actuarial Science
University of Iowa, Iowa City, Iowa, USA

Yuling Jiao YULINGJIAOMATH@WHU.EDU.CN
School of Mathematics and Statistics
and Hubei Key Laboratory of Computational Science
Wuhan University, Wuhan, China

Zhen Li LISHEN03@GMAIL.COM

Shiao Liu SHIAO-LIU@UIOWA.EDU
Department of Statistics and Actuarial Science
University of Iowa, Iowa City, Iowa, USA

Yang Wang YANGWANG@UST.HK
Department of Mathematics
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong, China

Yunfei Yang * YYANGDC@CONNECT.UST.HK
Department of Mathematics
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, Hong Kong, China

Editor: Mehryar Mohri

Abstract

This paper studies how well generative adversarial networks (GANs) learn probability distributions from finite samples. Our main results establish the convergence rates of GANs under a collection of integral probability metrics defined through Hölder classes, including the Wasserstein distance as a special case. We also show that GANs are able to adaptively learn data distributions with low-dimensional structures or have Hölder densities, when the network architectures are chosen properly. In particular, for distributions concentrated around a low-dimensional set, we show that the learning rates of GANs do not depend on the high ambient dimension, but on the lower intrinsic dimension. Our analysis is based on a new oracle inequality decomposing the estimation error into the generator and discriminator approximation error and the statistical error, which may be of independent interest.

Keywords: Generative adversarial networks, deep neural networks, convergence rate, error decomposition, risk bound

*. Corresponding author.

1. Introduction

Generative adversarial networks (GANs, Goodfellow et al. (2014); Li et al. (2015); Dziugaite et al. (2015); Arjovsky et al. (2017)) have attracted much attention in machine learning and artificial intelligence communities in the past few years. As a powerful unsupervised method for learning and sampling from complex data distributions, GANs have achieved remarkable successes in many machine learning tasks such as image synthesis, medical imaging and natural language generation (Radford et al., 2016; Reed et al., 2016; Zhu et al., 2017; Karras et al., 2018; Yi et al., 2019; Bowman et al., 2016). However, theoretical explanations for their empirical success are not well established. Many problems on the theory and training dynamics of GANs are largely unsolved (Arora et al., 2017; Liang, 2021; Singh et al., 2018).

Different from classical density estimation methods, GANs implicitly learn the data distribution by training a generator and a discriminator against each other. More specifically, to estimate a target distribution μ, one chooses an easy-to-sample source distribution ν (for example, uniform or Gaussian distribution) and find the generator by solving the following minimax optimization problem, at the population level,

$$\min_{g\in\mathcal{G}}\max_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{z\sim\nu}[f(g(z))],$$

where both the generator class $\mathcal{G}$ and the discriminator class $\mathcal{F}$ are often parameterized by neural networks in general. The inner maximization problem can be viewed as that of calculating the Integral Probability Metric (IPM, see Müller (1997)) between the target μ and the generated distribution $g_{\#}\nu$ with respect to the discriminator class $\mathcal{F}$:

$$d_{\mathcal{F}}(\mu, g_{\#}\nu) := \sup_{f\in\mathcal{F}} \mathbb{E}_{\mu}[f] - \mathbb{E}_{g_{\#}\nu}[f] = \sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{z\sim\nu}[f(g(z))],$$

where $g_{\#}\nu$ is the push-forward distribution under g. When only a set of random samples $\{X_i\}_{i=1}^n$ that are independent and identically distributed (i.i.d.) as μ are available in practice, we estimate the expectations by the empirical averages and solve the empirical optimization problem

$$g_n^* = \operatorname*{argmin}_{g\in\mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_{\#}\nu) = \operatorname*{argmin}_{g\in\mathcal{G}} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}_{z\sim\nu}[f(g(z))], \tag{1}$$

where $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution.

One of the fundamental questions in GANs is their generalization capacity: how well can GANs learn a target distribution from finite samples? Recently, much effort has been devoted to answering this question in different aspects. For example, Arora et al. (2017) showed that GANs do not generalize in standard metrics with any polynomial number of examples and provided generalization bounds for neural net distance. Zhang et al. (2018) gave a detailed analysis of neural net distance and extended the results of Arora et al. (2017). Liang (2021) and Singh et al. (2018) analyzed the adversarial framework from a nonparametric density estimation point of view. Chen et al. (2020) studied the convergence properties of GANs when both the target densities and the evaluation class are Hölder classes.

While impressive progress has been made on the theoretical understanding of GANs, there are still some shortcomings in the existing results. For instance, the source and the target distributions are often assumed to have the same ambient dimension in the current theory, while, in practice, GANs are usually trained using a source distribution with ambient dimension much smaller than that of the target distribution. Indeed, an important strength of GANs is their ability to model latent structures of complex high-dimensional distributions using a low-dimensional source distribution. Another issue needs to be addressed is that the generalization bounds often suffer from the curse of dimensionality. In practical applications, the data distributions are of high dimensionality, which makes the convergence rates in theory extremely slow. However, high-dimensional data, such as images, texts and natural languages, often have latent low-dimensional structures, which reduces the complexity of the problem. It is desirable to take into account such structures in the analysis.

1.1 Contributions

In this paper, we provide an error analysis of GANs and establish their convergence rates in various settings. We show that, if the generator and discriminator network architectures are properly chosen, GANs are able to learn any distributions with bounded support. To be concrete, let μ be a probability distribution on $[0,1]^d$ and g_n^* be a solution of the optimization problem (1), then $(g_n^*)_\#\nu$ is an estimate of μ. Informally, our main result shows that the GAN estimator has the convergence rate

$$\mathbb{E}[d_{\mathcal{H}^\beta}(\mu,(g_n^*)_\#\nu)] = \mathcal{O}(n^{-\beta/d} \vee n^{-1/2}\log n),$$

where the expectation is with respect to the random samples. The performance of the estimator is evaluated by the IPM $d_{\mathcal{H}^\beta}$ with respect to some Hölder class $\mathcal{H}^\beta$ of smoothness index $\beta > 0$. These metrics cover a wide range of popular metrics used in the literature, including the Wasserstein distance. In our theory, the ambient dimension of the source distribution ν is allowed to be different from the ambient dimension of the target distribution μ. In particular, it can be much smaller than that of the target distribution, which is the case in practice. Moreover, the convergence rates we derived match the minimax optimal rates of nonparametric density estimation under adversarial losses (Liang, 2021; Singh et al., 2018).

We also adapt our error analysis to three cases: (1) the target distribution concentrates around a low-dimensional set, (2) the target distribution has a density function, and (3) the target distribution has an unbounded support. In particular, we prove that if the target μ is supported on a set with dimension d^*, then the GAN estimator g_n^* has a faster convergence rate:

$$\mathbb{E}[d_{\mathcal{H}^\beta}(\mu,(g_n^*)_\#\nu)] = \mathcal{O}((n^{-\beta/d^*} \vee n^{-1/2})\log n).$$

This implies that the convergence rates of GANs do not depend on nominal high dimensionality of data, but on the lower intrinsic dimension. Our results show that GANs can automatically adapt to the support of the data and overcome the curse of dimensionality.

Our work also makes significant technical contributions to the error analysis of GANs and neural network approximation theory, which may be of independent interest. For example, we develop a new oracle inequality for GAN estimators, which decomposes the estimation error into generator and discriminator approximation error and statistical error. To bound the discriminator approximation error, we establish explicit error bounds on approximating Hölder functions by neural networks, with an explicit upper bound on the Lipschitz constant of the constructed neural network functions. To the best of our knowledge, this is the first approximation result that also controls the regularity of the neural network functions.

1.2 Preliminaries and Notation

Let us first introduce several definitions and notations. The set of positive integers is denoted by $\mathbb{N} = \{1, 2, \dots\}$. We also denote $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ for convenience. Let A and B be two quantities. The maximum and minimum of A and B are denoted by $A \vee B$ and $A \wedge B$ respectively. We use the asymptotic notation $A \precsim B$ and $B \succsim A$ to denote the statement that $A \le CB$ for some constant $C > 0$. We denote $A \asymp B$ when $A \precsim B$ and $A \succsim B$. Let ν be a measure on $\mathbb{R}^k$ and $g : \mathbb{R}^k \to \mathbb{R}^d$ be a measurable mapping. The push-forward measure $g_{\#}\nu$ of a measurable set A is defined as $g_{\#}\nu(A) := \nu(g^{-1}(A))$.

The ReLU function is denoted by $\sigma(x) := x \vee 0$. A neural network function $\phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ is a function that can be parameterized by a ReLU neural network in the following form

$$\phi(x) = T_L(\sigma(T_{L-1}(\cdots \sigma(T_0(x)) \cdots))),$$

where the activation function σ is applied component-wisely and $T_l(x) := A_l x + b_l$ is an affine transformation with $A_l \in \mathbb{R}^{N_{l+1} \times N_l}$ and $b_l \in \mathbb{R}^{N_{l+1}}$ for $l = 0, \dots, L$. The numbers $W = \max\{N_1, \dots, N_L\}$ and L are called the *width* and the *depth* of neural network, respectively. When the input and output dimensions are clear from contexts, we denote by $\mathcal{NN}(W, L)$ the set of functions that can be represented by neural networks with width at most W and depth at most L.

To measure the complexity of neural networks from a learning theory perspective, we use the following notion of combinatorial dimension for a real-valued function class.

Definition 1 (Pseudo-dimension) Let $\mathcal{H}$ be a class of real-valued functions defined on Ω. The *pseudo-dimension* of $\mathcal{H}$, denoted by $\operatorname{Pdim}(\mathcal{H})$, is the largest integer N for which there exist points $x_1, \dots, x_N \in \Omega$ and constants $c_1, \dots, c_N \in \mathbb{R}$ such that

$$|\{\operatorname{sgn}(h(x_1) - c_1), \dots, \operatorname{sgn}(h(x_N) - c_N) : h \in \mathcal{H}\}| = 2^N.$$

Next, let us introduce the notion of regularity for a function. For a multi-index $\alpha = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$, the monomial on $x = (x_1, \dots, x_d)$ is denoted by $x^\alpha := x_1^{\alpha_1} \cdots x_d^{\alpha_d}$, the α-derivative of a function h is denoted by $\partial^\alpha h := \partial^{\|\alpha\|_1} h / \partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}$ with $\|\alpha\|_1 = \sum_{i=1}^d \alpha_i$ as the usual 1-norm for vectors. We use the convention that $\partial^\alpha h := h$ if $\|\alpha\|_1 = 0$.

Definition 2 (Lipschitz functions) Let $\mathcal{X} \subseteq \mathbb{R}^d$ and $h : \mathcal{X} \to \mathbb{R}$, the *Lipschitz constant* of h is denoted by

$$\operatorname{Lip} h := \sup_{x, y \in \mathcal{X}, x \neq y} \frac{|h(x) - h(y)|}{\|x - y\|_2}.$$

We denote $\operatorname{Lip}(\mathcal{X}, K)$ as the set of all functions $h : \mathcal{X} \to \mathbb{R}$ with $\operatorname{Lip} h \le K$. For any $B > 0$, we denote $\operatorname{Lip}(\mathcal{X}, K, B) := \{h \in \operatorname{Lip}(\mathcal{X}, K) : \|h\|_{L^\infty(\mathcal{X})} \le B\}$.

Definition 3 (Hölder classes) For $\beta > 0$ with $\beta = s + r$, where $s \in \mathbb{N}_0$ and $r \in (0, 1]$, and $d \in \mathbb{N}$, we denote the Hölder class $\mathcal{H}^\beta(\mathbb{R}^d)$ as

$$\mathcal{H}^\beta(\mathbb{R}^d) := \left\{ h : \mathbb{R}^d \to \mathbb{R}, \max_{\|\alpha\|_1 \le s} \|\partial^\alpha h\|_\infty \le 1, \max_{\|\alpha\|_1 = s} \sup_{x \ne y} \frac{|\partial^\alpha h(x) - \partial^\alpha h(x)|}{\|x - y\|_2^r} \le 1 \right\}.$$

For any subset $\mathcal{X} \subseteq \mathbb{R}^d$, we denote $\mathcal{H}^\beta(\mathcal{X}) := \{h : \mathcal{X} \to \mathbb{R}, h \in \mathcal{H}^\beta(\mathbb{R}^d)\}$.

It should be noticed that for $\beta = s + 1$, we do *not* assume that $h \in C^{s+1}$. Instead, we only require that $h \in C^s$ and its derivatives of order s are Lipschitz continuous with respect to the metric $\|\cdot\|_2$. Note that, if $\beta \le 1$, then $|h(x) - h(y)| \le \|x - y\|_2^\beta$; if $\beta > 1$, then $|h(x) - h(y)| \le \sqrt{d}\|x - y\|_2$. In particular, with the above definitions, $\mathcal{H}^1([0, 1]^d) = \mathrm{Lip}\,([0, 1]^d, 1, 1)$. We will use the covering number to measure the complexity of a Hölder class.

Definition 4 (Covering number) Let ρ be a pseudo-metric on $\mathcal{M}$ and $S \subseteq \mathcal{M}$. For any $\epsilon > 0$, a set $A \subseteq \mathcal{M}$ is called an ϵ*-covering* of S if for any $x \in S$ there exists $y \in A$ such that $\rho(x, y) \le \epsilon$. The ϵ*-covering number* of S, denoted by $\mathcal{N}(\epsilon, S, \rho)$, is the minimum cardinality of any ϵ-covering of S.

Finally, the composition of two functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^k \to \mathbb{R}^d$ is denoted by $f \circ g(x) := f(g(x))$. We use $\mathcal{F} \circ \mathcal{G} := \{f \circ g : f \in \mathcal{F}, g \in \mathcal{G}\}$ to denote the composition of two function classes. A function class $\mathcal{F}$ is called *symmetric* if $f \in \mathcal{F}$ implies $-f \in \mathcal{F}$.

1.3 Outline

The rest of the paper is organized as follows. Section 2 presents our main result on the error analysis of GANs, where we assume that the target distribution has a compact support. In section 3, we extend the result to three different cases: (1) the target distribution is low-dimensional; (2) the target has a Hölder density; (3) the target has unbounded support. Section 4 discusses related theoretical results of deep neural networks and GANs. Finally, Section 5 gives the proofs of technical lemmas, including error decomposition and bounds on the approximation error and statistical error.

2. Error Analysis of GANs

Let μ be an unknown target probability distribution on $\mathbb{R}^d$, and let ν be a known and easy-to-sample distribution on $\mathbb{R}^k$ such as uniform or Gaussian distribution. Suppose we have n i.i.d. samples $\{X_i\}_{i=1}^n$ from μ and m i.i.d. samples $\{Z_i\}_{i=1}^m$ from ν. Denote the corresponding empirical distributions by $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\widehat{\nu}_m = \frac{1}{m}\sum_{i=1}^m \delta_{Z_i}$, respectively. We consider the following two optimization problems

$$\operatorname*{argmin}_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \nu) = \operatorname*{argmin}_{g \in \mathcal{G}} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}_\nu[f \circ g] \right\}, \tag{2}$$

$$\operatorname*{argmin}_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \widehat{\nu}_m) = \operatorname*{argmin}_{g \in \mathcal{G}} \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^n f(X_i) - \frac{1}{m} \sum_{j=1}^m f(g(Z_i)) \right\}, \tag{3}$$

where the generator class $\mathcal{G}$ is parameterized by a ReLU neural network $\mathcal{NN}(W_1, L_1)$ with width at most W_1 and depth at most L_1, and the discriminator class $\mathcal{F}$ is parameterized by another ReLU neural network $\mathcal{NN}(W_2, L_2)$.

2.1 Convergence Rates of GAN Estimators

We study the convergence rates of the GAN estimators g_n^* and $g_{n,m}^*$ that solve the optimization problems (2) and (3) with optimization error $\epsilon_{opt} \geq 0$. In other words,

$$g_n^* \in \left\{ g \in \mathcal{G} : d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \nu) \leq \inf_{\phi \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, \phi_\# \nu) + \epsilon_{opt} \right\}, \tag{4}$$

$$g_{n,m}^* \in \left\{ g \in \mathcal{G} : d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \widehat{\nu}_m) \leq \inf_{\phi \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, \phi_\# \widehat{\nu}_m) + \epsilon_{opt} \right\}. \tag{5}$$

The performance is evaluated by the IPM between the target μ and the learned distribution $\gamma = (g_n^*)_\# \nu$ or $\gamma = (g_{n,m}^*)_\# \nu$ with respect to some function class $\mathcal{H}$:

$$d_{\mathcal{H}}(\mu, \gamma) := \sup_{h \in \mathcal{H}} \mathbb{E}_{x \sim \mu}[h(x)] - \mathbb{E}_{y \sim \gamma}[h(y)].$$

By specifying $\mathcal{H}$ differently, one can obtain a list of commonly-used metrics:

- when $\mathcal{H} = \mathrm{Lip}\,(\mathbb{R}^d, 1)$ is the 1-Lipschitz function class, then $d_{\mathcal{H}} = \mathcal{W}_1$ is the Wasserstein distance, which is used in the Wasserstein GAN (Arjovsky et al., 2017);
- when $\mathcal{H} = \mathrm{Lip}\,(\mathbb{R}^d, B, B)$ is the bounded Lipschitz function class, then $d_{\mathcal{H}}$ is the Dudley metric, which metricizes weak convergence (Dudley, 2018);
- when $\mathcal{H}$ is the set of continuous function, then $d_{\mathcal{H}}$ is the total variation distance;
- when $\mathcal{H}$ is a Sobolev function class with certain regularity, $d_{\mathcal{H}}$ is used in Sobolev GAN (Mroueh et al., 2018);
- when $\mathcal{H}$ is the unit ball of some reproducing kernel Hilbert space, then $d_{\mathcal{H}}$ is the maximum mean discrepancy (Gretton et al., 2012; Dziugaite et al., 2015; Li et al., 2015).

Here, we consider the case when $\mathcal{H}$ is a Hölder class $\mathcal{H}^\beta(\mathbb{R}^d)$, which covers a wide range of applications. For simplicity, we first consider the case when μ is supported on the compact set $[0,1]^d$ and extend it to different situations in the next section. The main result is summarized in the following theorem.

Theorem 5 *Suppose the target μ is supported on $[0,1]^d$, the source distribution ν is absolutely continuous on $\mathbb{R}$ and the evaluation class is $\mathcal{H} = \mathcal{H}^\beta(\mathbb{R}^d)$. Then, there exist a generator $\mathcal{G} = \{g \in \mathcal{NN}(W_1, L_1) : g(\mathbb{R}) \subseteq [0,1]^d\}$ with*

$$W_1^2 L_1 \precsim n,$$

and a discriminator $\mathcal{F} = \mathcal{NN}(W_2, L_2) \cap \mathrm{Lip}\,(\mathbb{R}^d, K, 1)$ with

$$W_2 L_2 \precsim n^{1/2} \log^2 n, \quad K \precsim (\widetilde{W}_2 \widetilde{L}_2)^{2+\sigma(4\beta-4)/d} \widetilde{L}_2 2^{\widetilde{L}_2^2},$$

where $\widetilde{W}_2 = W_2/\log_2 W_2$ *and* $\widetilde{L}_2 = L_2/\log_2 L_2$*, such that the GAN estimator (4) satisfies*

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu)] - \epsilon_{opt} \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n,$$

where $c(\beta, d) = 1$ *if* $2\beta = d$*, and* $c(\beta, d) = 0$ *otherwise.*

If furthermore $m \succsim n^{2+2\beta/d} \log^6 n$*, then the GAN estimator (5) satisfies*

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_{n,m}^*)_{\#}\nu)] - \epsilon_{opt} \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n.$$

Before proceeding, we make several remarks on the theorem.

Remark 6 If $\beta = 1$, then $\mathcal{H}^1([0,1]^d) = \mathrm{Lip}\,([0,1]^d, 1, 1)$ and $d_{\mathcal{H}^1}$ is the Dudley distance (the Wasserstein distance $\mathcal{W}_1$ on $[0,1]^d$ is IPM with the class $\mathrm{Lip}\,([0,1]^d, 1)$ or $\mathrm{Lip}\,([0,1]^d, 1, \sqrt{d})$, and it satisfies $\mathcal{W}_1(\mu, \gamma) \le \sqrt{d} d_{\mathcal{H}^1}(\mu, \gamma)$). In this case, the required Lipschitz constant of the discriminator network is reduced to $K \precsim \widetilde{W}_2^2 \widetilde{L}_2^3 2^{\widetilde{L}_2^2}$. If we choose the depth L_2 to be a constant, then the Lipschitz constant can be chosen to have the order of $K \precsim \widetilde{W}_2^2 \precsim n \log^2 n$.

Remark 7 For simplicity, we assume that the source distribution ν is on $\mathbb{R}$. This is not a restriction, because any absolutely continuous distribution on $\mathbb{R}^k$ can be projected to an absolutely continuous distribution on $\mathbb{R}$ by linear mapping. Hence, the same result holds for any absolutely continuous source distribution on $\mathbb{R}^k$. The requirement on the generator that $g(\mathbb{R}) \subseteq [0,1]^d$ is easy to satisfy by adding an additional clipping layer to the output and using the fact that

$$\min\{\max\{x, -1\}, 1\} = \sigma(x+1) - \sigma(x-1) - 1, \quad x \in \mathbb{R}.$$

Remark 8 The Lipschitz condition on the discriminator might be difficult to satisfy in practice. It is done by weight clipping in the original Wasserstein GAN (Arjovsky et al., 2017). In the follow-up works (Gulrajani et al., 2017; Kodali et al., 2017; Petzka et al., 2018; Wei et al., 2018; Thanh-Tung et al., 2019), several regularization methods have been applied to Wasserstein GANs. It would be interesting to develop similar error analysis for regularized GAN estimators, and we leave this as future work.

2.2 Error Decomposition

Our proof of Theorem 5 is based on a new error decomposition and estimation of approximation error and statistical error sketched below. The proofs of technical lemmas are deferred to Section 5.

We first introduce a new oracle inequality, which decomposes the estimation error into the generator approximation error, the discriminator approximation error and the statistical error.

Lemma 9 *Assume* $\mathcal{F}$ *is symmetric,* μ *and* $g_{\#}\nu$ *are supported on* $\Omega \subseteq \mathbb{R}^d$ *for all* $g \in \mathcal{G}$*. Let* g_n^* *and* $g_{n,m}^*$ *be the GAN estimators (4) and (5) respectively. Then, for any function class* $\mathcal{H}$ *defined on* Ω*,*

$$d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu) \le \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_{\#}\nu) + d_{\mathcal{F}}(\mu, \widehat{\mu}_n) \wedge d_{\mathcal{H}}(\mu, \widehat{\mu}_n),$$

$$d_{\mathcal{H}}(\mu, (g_{n,m}^*)_{\#}\nu) \le \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_{\#}\nu) + d_{\mathcal{F}}(\mu, \widehat{\mu}_n) \wedge d_{\mathcal{H}}(\mu, \widehat{\mu}_n) + 2 d_{\mathcal{F} \circ \mathcal{G}}(\nu, \widehat{\nu}_m),$$

where $\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega)$ is the approximation error of $\mathcal{H}$ from $\mathcal{F}$ on Ω:

$$\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) := \sup_{h \in \mathcal{H}} \inf_{f \in \mathcal{F}} \|h - f\|_{L^\infty(\Omega)}.$$

Next, we bound each error term separately. We will show that the generator approximation error $\inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \nu) = 0$ as long as the size of the generator network $\mathcal{G}$ is sufficiently large. The discriminator approximation error $\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega)$ can be bounded by constructing neural networks to approximate functions in $\mathcal{H}$. The remaining statistical error terms can be controlled using the empirical process theory.

2.2.1 Bounding Generator Approximation Error

Observe that the empirical distribution $\widehat{\mu}_n$ is supported on at most n points. To bound the generator approximation error $\inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \nu)$, we need to estimate the distance between the generated distribution $\{g_\# \nu : g \in \mathcal{G}\}$ and the set of all discrete distribution supported on at most n points:

$$\mathcal{P}(n) := \left\{ \gamma = \sum_{i=1}^{n} p_i \delta_{x_i} : \sum_{i=1}^{n} p_i = 1, p_i \geq 0, x_i \in \mathbb{R}^d \right\}.$$

Yang et al. (2022) showed that their Wasserstein distance vanishes when the generator class is sufficiently large.

Lemma 10 *Suppose that $W \geq 7d + 1$, $L \geq 2$ and $\mathcal{G} = \mathcal{NN}(W, L)$. Let ν be an absolutely continuous probability distribution on $\mathbb{R}$. If $n \leq \frac{W-d-1}{2} \lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor + 2$, then for any $\gamma \in \mathcal{P}(n)$ and any $\epsilon > 0$, there exists $g \in \mathcal{G}$ such that*

$$\mathcal{W}_1(\gamma, g_\# \nu) < \epsilon.$$

If the support of γ is contained in some convex set $\mathcal{C}$, then g can be chosen to satisfy $g(\mathbb{R}) \subseteq \mathcal{C}$.

Since $\widehat{\mu}_n$ is supported on $[0,1]^d$, if we choose the generator $\mathcal{G} = \{g \in \mathcal{NN}(W_1, L_1) : g(\mathbb{R}) \subseteq [0,1]^d\}$ that satisfies the condition in the Lemma 10, which means we can choose $W_1^2 L_1 \precsim n$, then for any $\mathcal{F} \subseteq \mathrm{Lip}\,([0,1]^d, K)$, we have

$$\inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \nu) \leq K \inf_{g \in \mathcal{G}} \mathcal{W}_1(\widehat{\mu}_n, g_\# \nu) = 0.$$

This shows that the generator approximation error vanishes.

2.2.2 Bounding Discriminator Approximation Error

To bound $\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d)$, we construct a neural network to approximate any given function in $\mathcal{H}^\beta([0,1]^d)$. Our construction is based on the idea in Daubechies et al. (2021); Shen et al. (2020) and Lu et al. (2021). More importantly, we give an upper bound on the Lipschitz constant of the neural network function that achieves small approximation error.

Lemma 11 *Assume $h \in \mathcal{H}^\beta([0,1]^d)$ with $\beta = s + r$, $s \in \mathbb{N}_0$ and $r \in (0,1]$. For any $W \geq 6$, $L \geq 2$, there exists $\phi \in \mathcal{NN}(49(s+1)^2 3^d d^{s+1} W\lceil \log_2 W\rceil, 15(s+1)^2 L\lceil \log_2 L\rceil + 2d)$ such that $\|\phi\|_\infty \leq 1$, $\operatorname{Lip}\phi \leq (s+1)d^{s+1/2}L(WL)^{\sigma(4\beta-4)/d}(1260W^2L^2 2^{L^2} + 19s7^s)$ and*

$$\|\phi - h\|_{L^\infty([0,1]^d)} \leq 6(s+1)^2 d^{(s+\beta/2)\vee 1}\lfloor (WL)^{2/d}\rfloor^{-\beta}.$$

This lemma implies that, for any $h \in \mathcal{H}^\beta([0,1]^d)$, there exists a neural network ϕ with width $\precsim W\log_2 W$ and depth $\precsim L\log_2 L$ such that $\phi \in \operatorname{Lip}(\mathbb{R}^d, K, 1)$ with Lipschitz constant $K \precsim (WL)^{2+\sigma(4\beta-4)/d} L 2^{L^2}$ and $\|\phi - h\|_{L^\infty([0,1]^d)} \precsim (WL)^{-2\beta/d}$. Hence, if we choose $W_2 \asymp W\log_2 W$ and $L_2 \asymp L\log_2 L$, then

$$W \asymp W_2/\log_2 W_2 = \widetilde{W}_2, \quad L \asymp L_2/\log_2 L_2 = \widetilde{L}_2,$$

and $\phi \in \mathcal{NN}(W_2, L_2) \cap \operatorname{Lip}(\mathbb{R}^d, K, 1)$ with

$$K \precsim (WL)^{2+\sigma(4\beta-4)/d} L 2^{L^2} \precsim (\widetilde{W}_2\widetilde{L}_2)^{2+\sigma(4\beta-4)/d}\widetilde{L}_2 2^{\widetilde{L}_2^2}.$$

This shows that, for the discriminator $\mathcal{F} = \mathcal{NN}(W_2, L_2) \cap \operatorname{Lip}(\mathbb{R}^d, K, 1)$,

$$\mathcal{E}(\mathcal{H}^\beta, \mathcal{F}, [0,1]^d) \precsim (W_2L_2/(\log_2 W_2 \log_2 L_2))^{-2\beta/d}.$$

2.2.3 Bounding Statistical Error

For any function class $\mathcal{F}$, the statistical error $\mathbb{E}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)]$ can be bounded by the Rademacher complexity of $\mathcal{F}$, by using the standard symmetrization technique. We can further bound the Rademacher complexity by the covering number of $\mathcal{F}$. The result is summarized in the following lemma.

Lemma 12 *Assume $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq B$, then we have the following entropy integral bound*

$$\mathbb{E}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)] \leq 8\mathbb{E}_{X_{1:n}} \inf_{0<\delta<B/2}\left(\delta + \frac{3}{\sqrt{n}}\int_\delta^{B/2}\sqrt{\log\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty)}d\epsilon\right),$$

where we denote $\mathcal{F}_{|X_{1:n}} = \{(f(X_1), \dots, f(X_n)) : f \in \mathcal{F}\}$ for any i.i.d. samples $X_{1:n} = \{X_i\}_{i=1}^n$ from μ and $\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty)$ is the ϵ-covering number of $\mathcal{F}_{|X_{1:n}} \subseteq \mathbb{R}^n$ with respect to the $\|\cdot\|_\infty$ distance.

For the Hölder class $\mathcal{H} = \mathcal{H}^\beta(\mathbb{R}^d)$, for any i.i.d. samples $X_{1:n} = \{X_i\}_{i=1}^n$ from μ, which is supported on $[0,1]^d$, we have

$$\log\mathcal{N}(\epsilon, \mathcal{H}_{|X_{1:n}}, \|\cdot\|_\infty) \leq \log\mathcal{N}(\epsilon, \mathcal{H}^\beta([0,1]^d), \|\cdot\|_\infty) \precsim \epsilon^{-d/\beta},$$

where the last inequality is from the entropy bound in Kolmogorov and Tikhomirov (1961) (see also Lemma 17). Thus, if we denote $\eta = d/(2\beta)$, then

$$\mathbb{E}[d_{\mathcal{H}}(\mu, \widehat{\mu}_n)] \precsim \inf_{0<\delta<1/2}\left(\delta + n^{-1/2}\int_\delta^{1/2}\epsilon^{-\eta}d\epsilon\right).$$

When $\eta < 1$, one has

$$\mathbb{E}[d_{\mathcal{H}}(\mu, \widehat{\mu}_n)] \precsim \inf_{0<\delta<1/2} \Big(\delta + (1-\eta)^{-1} n^{-1/2} (2^{\eta-1} - \delta^{1-\eta})\Big) \precsim n^{-1/2}.$$

When $\eta = 1$, one has

$$\mathbb{E}[d_{\mathcal{H}}(\mu, \widehat{\mu}_n)] \precsim \inf_{0<\delta<1/2} \Big(\delta + n^{-1/2}(-\log 2 - \log \delta)\Big) \precsim n^{-1/2} \log n,$$

where we take $\delta = n^{-1/2}$ in the last step. When $\eta > 1$, one has

$$\mathbb{E}[d_{\mathcal{H}}(\mu, \widehat{\mu}_n)] \precsim \inf_{0<\delta<1/2} \Big(\delta + (\eta-1)^{-1} n^{-1/2} (\delta^{1-\eta} - 2^{\eta-1})\Big) \precsim n^{-1/(2\eta)} = n^{-\beta/d},$$

where we take $\delta = n^{-1/(2\eta)}$. Combining these cases together, we have

$$\mathbb{E}[d_{\mathcal{H}}(\mu, \widehat{\mu}_n)] \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n, \tag{6}$$

where $c(\beta, d) = 1$ if $2\beta = d$, and $c(\beta, d) = 0$ otherwise.

2.3 Proof of Theorem 5

For the GAN estimator g_n^*, by Lemma 9, we have the error decomposition

$$d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu) \leq \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d) + \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_{\#}\nu) + d_{\mathcal{H}}(\mu, \widehat{\mu}_n). \tag{7}$$

We choose the generator class $\mathcal{G}$ with $W_1^2 L_1 \precsim n$ that satisfies the condition in Lemma 10. Then

$$\inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_{\#}\nu) = 0,$$

since $\mathcal{F} \subseteq \mathrm{Lip}\,([0,1]^d, K)$. By Lemma 11, for our choice of the discriminator class $\mathcal{F}$,

$$\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d) \precsim (W_2 L_2 / (\log_2 W_2 \log_2 L_2))^{-2\beta/d} \precsim n^{-\beta/d},$$

where we can choose $W_2 L_2 \asymp n^{1/2} \log^2 n$ so that the last inequality holds. By Lemma 12,

$$\mathbb{E}[d_{\mathcal{H}}(\mu, \widehat{\mu}_n)] \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n.$$

In summary, by (7), we have

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu)] - \epsilon_{opt} \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n.$$

For the estimator $g_{n,m}^*$, we only need to estimate the extra term $\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)]$ by Lemma 9. We can bound this statistical error by the entropy integral in Lemma 12, and further bound it by the pseudo-dimension $\mathrm{Pdim}\,(\mathcal{F}\circ\mathcal{G})$ of the network $\mathcal{F}\circ\mathcal{G}$ (see corollary 35):

$$\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)] \precsim \sqrt{\frac{\mathrm{Pdim}\,(\mathcal{F}\circ\mathcal{G}) \log m}{m}}.$$

It was shown in Bartlett et al. (2019) that the pseudo-dimension of a ReLU neural network satisfies the bound $\mathrm{Pdim}\left(\mathcal{NN}(W,L)\right) \precsim UL\log U$, where $U \asymp W^2 L$ is the number of parameters. Hence,

$$\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu,\widehat{\nu}_m)] \precsim \sqrt{\frac{(W_1^2 L_1 + W_2^2 L_2)(L_1+L_2)\log(W_1^2 L_1 + W_2^2 L_2)\log m}{m}}.$$

Since we have chosen $W_2 L_2 \precsim n^{1/2}\log^2 n$ and $W_1^2 L_1 \precsim n$, we have

$$\begin{aligned}\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu,\widehat{\nu}_m)] &\precsim \sqrt{\frac{(n+n\log^4 n)(n+n^{1/2}\log^2 n)\log n\log m}{m}}\\ &\precsim \sqrt{\frac{n^2\log^5 n\log m}{m}}.\end{aligned}$$

Hence, if $m \gtrsim n^{2+2\beta/d}\log^6 n$, then $\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu,\widehat{\nu}_m)] \precsim n^{-\beta/d}$ and, by Lemma 9,

$$\mathbb{E}[d_{\mathcal{H}}(\mu,(g^*_{n,m})_\#\nu)] - \epsilon_{opt} \precsim n^{-\beta/d} \vee n^{-1/2}\log^{c(\beta,d)} n,$$

which completes the proof.

We make three remarks on the proof and the technical lemmas.

Remark 13 Our error decomposition for GANs in Lemma 9 is different from the classical bias-variance decomposition for regression in the sense that the statistical error $d_{\mathcal{F}}(\mu,\widehat{\mu}_n)\wedge d_{\mathcal{H}}(\mu,\widehat{\mu}_n) \le d_{\mathcal{H}}(\mu,\widehat{\mu}_n)$ depends on the evaluation class $\mathcal{H}$. The proof of Theorem 5 essentially shows that we can choose the generator class and the discriminator class sufficiently large to reduce the approximation error so that the learning rate of GAN estimator is not slower than that of the empirical distribution.

Remark 14 We give explicit estimate of the Lipschitz constant of the discriminator in Lemma 11, because it is essential in bounding the generator approximation error in our analysis. Alternatively, one can also bound the parameters in the discriminator network and then estimate the Lipschitz constant. For example, by using the construction in Yarotsky (2017), one can bound the weights as $\mathcal{O}(\epsilon^{-\alpha})$ for some $\alpha>0$, where ϵ is the approximation error. Then convergence rates can be obtained for the discriminator network with bounded weights (the bound depends on the sample size n).

Remark 15 The bound on the expectation $\mathbb{E}[d_{\mathcal{H}}(\mu,(g^*_n)_\#\nu)]$ can be turned into a high probability bound by using concentration inequalities (Boucheron et al., 2013; Shalev-Shwartz and Ben-David, 2014; Mohri et al., 2018). For example, by McDiarmid's inequality, one can shows that, for all $t>0$,

$$\mathbb{P}\left(d_{\mathcal{H}}(\mu,\widehat{\mu}_n) \ge \mathbb{E}[d_{\mathcal{H}}(\mu,\widehat{\mu}_n)] + t\right) \le \exp(-nt^2/2), \tag{8}$$

because for any $\{X_i\}_{i=1}^n$ and $\{X_i'\}_{i=1}^n$ that satisfies $X_i'=X_i$ except for $i=j$, we have

$$\left|\sup_{h\in\mathcal{H}}\left(\mathbb{E}_\mu[h]-\frac{1}{n}\sum_{i=1}^n h(X_i)\right)-\sup_{h\in\mathcal{H}}\left(\mathbb{E}_\mu[h]-\frac{1}{n}\sum_{i=1}^n h(X_i')\right)\right| \le \sup_{h\in\mathcal{H}}\frac{1}{n}\left|h(X_j)-h(X_j')\right| \le \frac{2}{n}.$$

Since other error terms in inequality (7) can be bounded independent of the random samples, it holds with probability at least $1-\delta$ that

$$d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu) - \epsilon_{opt} - \sqrt{\frac{2\log(1/\delta)}{n}} \precsim n^{-\beta/d} \vee n^{-1/2}\log^{c(\beta,d)} n,$$

where we choose $\exp(-nt^2/2) = \delta$ in inequality (8).

3. Extensions of the Main Theorem

In this section, we extend the main theorem to the following cases: (1) the target distribution concentrates around a low-dimensional set, (2) the target distribution has a density function and, (3) the target distribution has an unbounded support.

3.1 Learning Low-dimensional Distributions

The convergence rates in Theorem 5 suffer from the curse of dimensionality. In practice, the ambient dimension is usually large, which makes the convergence very slow. However, in many applications, high-dimensional complex data such as images, texts and natural languages, tend to be supported on approximate lower-dimensional manifolds. To take into account this fact, we assume that the target distribution μ has a low-dimensional structure. We introduce the Minkowski dimension (or box-counting dimension) to determine the dimensionality of a set.

Definition 16 (Minkowski dimension) The upper and the lower Minkowski dimensions of a set $A \subseteq \mathbb{R}^d$ are defined respectively as

$$\overline{\dim}_M(A) := \limsup_{\epsilon\to 0} \frac{\log \mathcal{N}(\epsilon, A, \|\cdot\|_2)}{-\log\epsilon},$$
$$\underline{\dim}_M(A) := \liminf_{\epsilon\to 0} \frac{\log \mathcal{N}(\epsilon, A, \|\cdot\|_2)}{-\log\epsilon}.$$

If $\overline{\dim}_M(A) = \underline{\dim}_M(A) = \dim_M(A)$, then $\dim_M(A)$ is called the *Minkowski dimension* of the set A.

The Minkowski dimension measures how the covering number of A decays when the radius of covering balls converges to zero. When A is a manifold, its Minkowski dimension is the same as the dimension of the manifold. Since the Minkowski dimension only depends on the metric, it can also be used to measure the dimensionality of highly non-regular set, such as fractals (Falconer, 2004). For function classes defined on a set with a small Minkowski dimension, it is intuitive to expect that the covering number only depends on the intrinsic Minkowski dimension, rather than the ambient dimension. Kolmogorov and Tikhomirov (1961) gave a comprehensive study on such problems. We will need the following useful lemma in our analysis.

Lemma 17 (Kolmogorov and Tikhomirov (1961)) *If $\mathcal{X} \subseteq \mathbb{R}^d$ is a compact set with $\dim_M(\mathcal{X}) = d^*$, then*

$$\log \mathcal{N}(\epsilon, \mathcal{H}^\beta(\mathcal{X}), \|\cdot\|_\infty) \precsim \epsilon^{-d^*/\beta}\log(1/\epsilon).$$

If, in addition, $\mathcal{X}$ is connected, then

$$\log \mathcal{N}(\epsilon, \mathcal{H}^\beta(\mathcal{X}), \|\cdot\|_\infty) \precsim \epsilon^{-d^*/\beta}.$$

For regression, Nakada and Imaizumi (2020) showed that deep neural networks can adapt to the low-dimensional structure of data, and the convergence rates do not depend on the nominal high dimensionality of data, but on its lower intrinsic dimension. We will show that similar results hold for GANs by analyzing the learning rates of a target distribution that concentrates on a low-dimensional set.

Assumption 18 The target $X \sim \mu$ has the form $X = \widetilde{X} + \xi$, where $\widetilde{X}$ and ξ are independent, $\widetilde{X} \sim \widetilde{\mu}$ is supported on some compact set $\mathcal{X} \subseteq [0,1]^d$ with $\dim_M(\mathcal{X}) = d^*$, and ξ has zero mean $\mathbb{E}[\xi] = 0$ and bounded variance $V = \mathbb{E}[\|\xi\|_2^2] < \infty$.

The next theorem shows that the convergence rates of the GAN estimators only depend on the intrinsic dimension d^*, when the network architectures are properly chosen.

Theorem 19 *Suppose the target μ satisfies assumption 18, the source distribution ν is absolutely continuous on $\mathbb{R}$ and the evaluation class is $\mathcal{H} = \mathcal{H}^\beta(\mathbb{R}^d)$. Then, there exist a generator $\mathcal{G} = \{g \in \mathcal{NN}(W_1, L_1) : g(\mathbb{R}) \subseteq [0,1]^d\}$ with*

$$W_1^2 L_1 \precsim n,$$

and a discriminator $\mathcal{F} = \mathcal{NN}(W_2, L_2) \cap \mathrm{Lip}\,(\mathbb{R}^d, K, 1)$ with

$$W_2 L_2 \precsim n^{d/(2d^*)} \log^2 n, \quad K \precsim (\widetilde{W}_2 \widetilde{L}_2)^{2+\sigma(4\beta-4)/d} \widetilde{L}_2 2^{\widetilde{L}_2^2},$$

where $\widetilde{W}_2 = W_2/\log_2 W_2$ and $\widetilde{L}_2 = L_2/\log_2 L_2$, such that the GAN estimator (4) satisfies

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_\# \nu)] - \epsilon_{opt} - 2\sqrt{d} V^{(\beta \wedge 1)/2} \precsim (n^{-\beta/d^*} \vee n^{-1/2}) \log n.$$

If furthermore

$$m \succsim \begin{cases} n^{(3d+4\beta)/(2d^*)} \log^6 n & d^* \le d/2, \\ n^{1+(d+2\beta)/d^*} \log^4 n & d^* > d/2, \end{cases}$$

then the GAN estimator (5) satisfies

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_{n,m}^*)_\# \nu)] - \epsilon_{opt} - 2\sqrt{d} V^{(\beta \wedge 1)/2} \precsim (n^{-\beta/d^*} \vee n^{-1/2}) \log n.$$

Proof For any i.i.d. observations $X_{1:n} = \{X_i\}_{i=1}^n$ from μ, where $X_i = \widetilde{X}_i + \xi_i$ with $\widetilde{X}_i \sim \widetilde{\mu}$, we denote $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ and $\widehat{\widetilde{\mu}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\widetilde{X}_i}$. As in the proof of Theorem 5, by Lemma 25, we have

$$\begin{aligned} \mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_\# \nu)] &\le d_{\mathcal{H}}(\mu, \widetilde{\mu}) + \mathbb{E}[d_{\mathcal{H}}(\widetilde{\mu}, (g_n^*)_\# \nu)] \\ &\le d_{\mathcal{H}}(\mu, \widetilde{\mu}) + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d) + \mathbb{E}[d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\widetilde{\mu}}_n)] + \epsilon_{opt}, \end{aligned}$$

and there exists a discriminator $\mathcal{F}$ with $W_2 L_2 \asymp n^{d/(2d^*)} \log^2 n$ such that

$$\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d) \precsim (W_2 L_2/(\log_2 W_2 \log_2 L_2))^{-2\beta/d} \precsim n^{-\beta/d^*}.$$

For the term $d_{\mathcal{H}}(\mu, \widetilde{\mu})$, we can bound it as

$$d_{\mathcal{H}}(\mu, \widetilde{\mu}) = \sup_{h \in \mathcal{H}} \mathbb{E}_{\xi}[\mathbb{E}_{\widetilde{X}}[h(\widetilde{X} + \xi) - h(\widetilde{X})]] \leq \sqrt{d}\mathbb{E}_{\xi}[\|\xi\|_2^{\beta \wedge 1}] \leq \sqrt{d} V^{(\beta \wedge 1)/2}, \tag{9}$$

where we use the Lipschitz inequality $|h(\widetilde{X} + \xi) - h(\widetilde{X})| \leq \sqrt{d}\|\xi\|_2^{\beta \wedge 1}$ for the second inequality, and Jensen's inequality for the last inequality.

For the statistical error, we have

$$\mathbb{E}_{X_{1:n}}[d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\mu}_n)] \leq \mathbb{E}_{\widetilde{X}_{1:n}} d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\widetilde{\mu}}_n) + \mathbb{E}_{\xi_{1:n}} \mathbb{E}_{\widetilde{X}_{1:n}} d_{\mathcal{H}}(\widehat{\widetilde{\mu}}_n, \widehat{\mu}_n).$$

Using Lipschitz continuity of h, we have

$$\begin{aligned}
\mathbb{E}_{\xi_{1:n}} \mathbb{E}_{\widetilde{X}_{1:n}} d_{\mathcal{H}}(\widehat{\widetilde{\mu}}_n, \widehat{\mu}_n) &= \mathbb{E}_{\xi_{1:n}} \mathbb{E}_{\widetilde{X}_{1:n}} \sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} h(\widetilde{X}_i + \xi_i) - h(\widetilde{X}_i) \\
&\leq \sqrt{d}\mathbb{E}_{\xi_{1:n}} \frac{1}{n} \sum_{i=1}^{n} \|\xi_i\|_2^{\beta \wedge 1} \\
&\leq \sqrt{d} V^{(\beta \wedge 1)/2}.
\end{aligned} \tag{10}$$

To estimate $\mathbb{E}_{\widetilde{X}_{1:n}} d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\widetilde{\mu}}_n)$, recall that we have denoted $\mathcal{H}_{|\widetilde{X}_{1:n}} := \{(h(\widetilde{X}_1), \ldots, h(\widetilde{X}_n)) : h \in \mathcal{H}\} \subseteq \mathbb{R}^n$. Since $\widetilde{\mu}$ is supported on $\mathcal{X}$ with $\dim_M(\mathcal{X}) = d^*$ by Assumption 18, the covering number of $\mathcal{H}_{|\widetilde{X}_{1:n}}$ with respect to the distance $\|\cdot\|_\infty$ on $\mathbb{R}^n$ can be bounded by the covering number of $\mathcal{H}$ with respect to the $L^\infty(\mathcal{X})$ distance. Hence,

$$\log \mathcal{N}(\epsilon, \mathcal{H}_{|\widetilde{X}_{1:n}}, \|\cdot\|_\infty) \leq \log \mathcal{N}(\epsilon, \mathcal{H}^\beta(\mathcal{X}), \|\cdot\|_\infty) \precsim \epsilon^{-d^*/\beta} \log(1/\epsilon),$$

by Lemma 17. Therefore, by Lemma 12,

$$\begin{aligned}
\mathbb{E}_{\widetilde{X}_{1:n}} d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\widetilde{\mu}}_n) &\leq 8\mathbb{E}_{\widetilde{X}_{1:n}} \inf_{0<\delta<1/2} \left(\delta + \frac{3}{\sqrt{n}} \int_\delta^{1/2} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{H}_{|\widetilde{X}_{1:n}}, \|\cdot\|_\infty)} d\epsilon \right) \\
&\precsim \inf_{0<\delta<1/2} \left(\delta + n^{-1/2} \int_\delta^{1/2} \epsilon^{-d^*/(2\beta)} \log(1/\epsilon) d\epsilon \right) \\
&\precsim \inf_{0<\delta<1/2} \left(\delta + n^{-1/2} \log(1/\delta) \int_\delta^{1/2} \epsilon^{-d^*/(2\beta)} d\epsilon \right).
\end{aligned}$$

A calculation similar to the inequality (6) gives

$$\mathbb{E}_{\widetilde{X}_{1:n}} d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\widetilde{\mu}}_n) \precsim (n^{-\beta/d^*} \vee n^{-1/2}) \log n.$$

Therefore,

$$\mathbb{E}_{X_{1:n}}[d_{\mathcal{H}}(\widetilde{\mu}, \widehat{\mu}_n)] - \sqrt{d} V^{(\beta \wedge 1)/2} \precsim (n^{-\beta/d^*} \vee n^{-1/2}) \log n.$$

In summary, we obtain the desired bound

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu)] - \epsilon_{opt} - 2\sqrt{d} V^{(\beta \wedge 1)/2} \precsim (n^{-\beta/d^*} \vee n^{-1/2}) \log n.$$

For the estimator $g^*_{n,m}$, we use the pseudo-dimension to bound $\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)]$. Since we have chosen $W_2 L_2 \precsim n^{d/(2d^*)} \log^2 n$ and $W_1^2 L_1 \precsim n$,

$$\begin{aligned}
\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)] &\precsim \sqrt{\frac{(W_1^2 L_1 + W_2^2 L_2)(L_1 + L_2)\log(W_1^2 L_1 + W_2^2 L_2)\log m}{m}} \\
&\precsim \sqrt{\frac{(n + n^{d/d^*}\log^4 n)(n + n^{d/(2d^*)}\log^2 n)\log n \log m}{m}} \\
&\precsim \sqrt{\frac{n^{d/d^*}(n + n^{d/(2d^*)}\log^2 n)\log^5 n \log m}{m}}.
\end{aligned}$$

By our choice of m, we always have $\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)] \precsim n^{-\beta/d^*} \log n$. The result then follows from Lemma 9. ■

Remark 20 In the proof, we actually show that the same convergence rate holds for $\widetilde{\mu}$: $\mathbb{E}[d_{\mathcal{H}}(\widetilde{\mu}, (g^*_n)_{\#}\nu)] - \epsilon_{opt} - \sqrt{d} V^{(\beta\wedge 1)/2} \precsim (n^{-\beta/d^*} \vee n^{-1/2}) \log n$. Note that the constant $\sqrt{d}$ is due to the Lipschitz constant of the evaluation class $\mathcal{H}^\beta$. When $\beta = 1$, we have a better Lipschitz inequality $|h(\widetilde{X}+\xi) - h(\widetilde{X})| \le \|\xi\|_2$ in inequalities (9) and (10). As a consequence, one can check that, for the Dudley metric,

$$\mathbb{E}[d_{\mathcal{H}^1}(\mu, (g^*_n)_{\#}\nu)] - \epsilon_{opt} - 2V^{1/2} \precsim (n^{-1/d^*} \vee n^{-1/2}) \log n.$$

This bound is useful only when the variance term $V^{1/2}$ is negligible, i.e. the data distribution is really low-dimensional. One can regard the variance as a "measure" of how well the low-dimension assumption is fulfilled. It is numerically confirmed that several well-known real data have small intrinsic dimensions, while their nominal dimensions are very large (Nakada and Imaizumi, 2020).

3.2 Learning Distributions with Densities

When the target distribution μ has a density function $p_\mu \in \mathcal{H}^\alpha([0,1]^d)$, it was proved in Liang (2021); Singh et al. (2018) that the minimax convergence rates of nonparametric density estimation satisfy

$$\inf_{\widetilde{\mu}_n} \sup_{p_\mu \in \mathcal{H}^\alpha([0,1]^d)} \mathbb{E} d_{\mathcal{H}^\beta([0,1]^d)}(\mu, \widetilde{\mu}_n) \asymp n^{-(\alpha+\beta)/(2\alpha+d)} \vee n^{-1/2},$$

where the infimum is taken over all estimator $\widetilde{\mu}_n$ with density $p_{\widetilde{\mu}_n} \in \mathcal{H}^\alpha([0,1]^d)$ based on n i.i.d. samples $\{X_i\}_{i=1}^n$ of μ. Ignoring the logarithmic factor, Theorem 5 gives the same convergence rate with $\alpha = 0$, which reveals the optimality of the result (since we do not assume the target has density in Theorem 5).

Under a priori that $p_\mu \in \mathcal{H}^\alpha$ for some $\alpha > 0$, it is not possible for the GAN estimators (4) and (5) to learn the regularity of the target, because the empirical distribution $\widehat{\mu}_n$ do not inherit the regularity. However, we can use certain regularized empirical distribution

$\widetilde{\mu}_n$ as the plug-in for GANs and consider the estimators

$$\widetilde{g}_n^* \in \left\{ g \in \mathcal{G} : d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) \leq \inf_{\phi \in \mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, \phi_{\#}\nu) + \epsilon_{opt} \right\}, \tag{11}$$

$$\widetilde{g}_{n,m}^* \in \left\{ g \in \mathcal{G} : d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\widehat{\nu}_m) \leq \inf_{\phi \in \mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, \phi_{\#}\widehat{\nu}_m) + \epsilon_{opt} \right\}. \tag{12}$$

By choosing the regularized distribution $\widetilde{\mu}_n$, the generator $\mathcal{G}$ and the discriminator $\mathcal{F}$ properly, we show that $\widetilde{g}_n^*$ and $\widetilde{g}_{n,m}^*$ can achieve faster convergence rates than the GAN estimators (4) and (5), which use the empirical distribution $\widehat{\mu}_n$ as the plug-in. The result can be seen as a complement to the nonparametric results in (Liang, 2021, Theorem 3).

Theorem 21 *Suppose the target μ has a density function $p_\mu \in \mathcal{H}^\alpha([0,1]^d)$ for some $\alpha > 0$, the source distribution ν is absolutely continuous on $\mathbb{R}$ and the evaluation class is $\mathcal{H} = \mathcal{H}^\beta(\mathbb{R}^d)$. Then, there exist a regularized empirical distribution $\widetilde{\mu}_n$ with density $p_{\widetilde{\mu}_n} \in \mathcal{H}^\alpha([0,1]^d)$, a generator $\mathcal{G} = \{g \in \mathcal{NN}(W_1, L_1) : g(\mathbb{R}) \subseteq [0,1]^d\}$ with*

$$W_1^2 L_1 \precsim n^{\frac{\alpha+\beta}{2\alpha+d} \frac{d+\beta+\sigma(2\beta-2)}{\beta} d},$$

and a discriminator $\mathcal{F} = \mathcal{NN}(W_2, L_2) \cap \mathrm{Lip}\,(\mathbb{R}^d, K, 1)$ with

$$W_2 / \log_2 W_2 \precsim n^{\frac{\alpha+\beta}{2\alpha+d} \frac{d}{2\beta}}, \quad L_2 \asymp 1, \quad K \precsim (W_2/\log_2 W_2)^{2+\sigma(4\beta-4)/d} \precsim n^{\frac{\alpha+\beta}{2\alpha+d} \frac{d+\sigma(2\beta-2)}{\beta}},$$

such that the GAN estimator (11) satisfies

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (\widetilde{g}_n^*)_{\#}\nu)] - \epsilon_{opt} \precsim n^{-(\alpha+\beta)/(2\alpha+d)} \vee n^{-1/2}.$$

If furthermore $m \succsim n^{\frac{2\alpha+2\beta}{2\alpha+d}(\frac{d+\beta+\sigma(2\beta-2)}{\beta} d+1)} \log^2 n$, then the GAN estimator (12) satisfies

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (\widetilde{g}_{n,m}^*)_{\#}\nu)] - \epsilon_{opt} \precsim n^{-(\alpha+\beta)/(2\alpha+d)} \vee n^{-1/2}.$$

Proof Liang (2021) and Singh et al. (2018) showed the existence of regularized empirical distribution $\widetilde{\mu}_n$ with density $p_{\widetilde{\mu}_n} \in \mathcal{H}^\alpha([0,1]^d)$ that satisfies

$$\mathbb{E} d_{\mathcal{H}}(\mu, \widetilde{\mu}_n) \precsim n^{-(\alpha+\beta)/(2\alpha+d)} \vee n^{-1/2}.$$

Similar to Lemma 9, we can decompose the error as (see Lemma 25)

$$d_{\mathcal{H}}(\mu, (\widetilde{g}_n^*)_{\#}\nu) \leq \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d) + \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + d_{\mathcal{H}}(\mu, \widetilde{\mu}_n).$$

By Lemma 11, we can choose a discriminator $\mathcal{F}$ that satisfies the condition in the theorem such that the discriminator approximation error can be bounded by

$$\mathcal{E}(\mathcal{H}, \mathcal{F}, [0,1]^d) \precsim (W_2 L_2 / (\log W_2 \log L_2))^{-2\beta/d} \precsim n^{-(\alpha+\beta)/(2\alpha+d)}.$$

For the generator approximation error, since $\mathcal{F} \subseteq \mathrm{Lip}\,([0,1]^d, K)$,

$$\inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) \leq K \inf_{g \in \mathcal{G}} \mathcal{W}_1(\widetilde{\mu}_n, g_{\#}\nu).$$

It was shown in Yang et al. (2022) that (see also Corollary 27)

$$\inf_{g\in\mathcal{G}} \mathcal{W}_1(\widetilde{\mu}_n, g_{\#}\nu) \precsim (W_1^2 L_1)^{-1/d}.$$

Hence, there exists a generator $\mathcal{G}$ with $W_1^2 L_1 \asymp n^{\frac{\alpha+\beta}{2\alpha+d}\frac{d+\beta+\sigma(2\beta-2)}{\beta}d}$ such that

$$\inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) \precsim K(W_1^2 L_1)^{-1/d} \precsim n^{-(\alpha+\beta)/(2\alpha+d)}.$$

In summary, we have

$$\mathbb{E} d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu) - \epsilon_{opt} \precsim n^{-(\alpha+\beta)/(2\alpha+d)} \vee n^{-1/2}.$$

For the estimator $\widetilde{g}_{n,m}^*$, we only need to further bound $d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)$ due to Lemma 25. By corollary 35, we can bound it using the pseudo-dimension of $\mathcal{F}\circ\mathcal{G}$:

$$\begin{aligned}
\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)] &\precsim \sqrt{\frac{(W_1^2 L_1 + W_2^2 L_2)(L_1+L_2)\log(W_1^2 L_1 + W_2^2 L_2)\log m}{m}} \\
&\precsim \sqrt{\frac{(n^{\frac{\alpha+\beta}{2\alpha+d}\frac{d+\beta+\sigma(2\beta-2)}{\beta}d} + n^{\frac{\alpha+\beta}{2\alpha+d}\frac{d}{\beta}}\log n) n^{\frac{\alpha+\beta}{2\alpha+d}\frac{d+\beta+\sigma(2\beta-2)}{\beta}d} \log n \log m}{m}} \\
&\precsim n^{\frac{\alpha+\beta}{2\alpha+d}\frac{d+\beta+\sigma(2\beta-2)}{\beta}d} \sqrt{\frac{\log n \log m}{m}}.
\end{aligned}$$

Since $m \succsim n^{\frac{2\alpha+2\beta}{2\alpha+d}(\frac{d+\beta+\sigma(2\beta-2)}{\beta}d+1)} \log^2 n$, we have $\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)] \precsim n^{-(\alpha+\beta)/(2\alpha+d)}$, which completes the proof. ■

As we noted in Remark 13, the proof essentially shows that the convergence rates of $\widetilde{g}_n^*$ and $\widetilde{g}_{n,m}^*$ are not worse than the convergence rate of $\mathbb{E} d_{\mathcal{H}}(\mu, \widetilde{\mu}_n)$ if we choose the network architectures properly.

3.3 Learning Distributions with Unbounded Supports

So far, we have assumed that the target distribution has a compact support. In this section, we show how to generalize the results to target distributions with unbounded supports. For simplicity, we only consider the case when the target μ is sub-exponential in the sense that

$$\mu(\{x \in \mathbb{R}^d : \|x\|_\infty > \log t\}) \precsim t^{-a/d}, \tag{13}$$

for some $a > 0$. The basic idea is to truncate the target distribution and apply the error analysis to the truncated distribution.

Theorem 22 *Suppose the target μ satisfies condition (13), the source distribution ν is absolutely continuous on $\mathbb{R}$ and the evaluation class is $\mathcal{H} = \mathcal{H}^\beta(\mathbb{R}^d)$. Then, there exist a generator $\mathcal{G} = \{g \in \mathcal{NN}(W_1, L_1) : g(\mathbb{R}) \subseteq [-\beta a^{-1}\log n, \beta a^{-1}\log n]^d\}$ with*

$$W_1^2 L_1 \precsim n$$

and a discriminator $\mathcal{F} = \mathcal{NN}(W_2, L_2) \cap \mathrm{Lip}\,(\mathbb{R}^d, K, 1)$ *with*

$$W_2 L_2 \lesssim n^{1/2} \log^{2+d/2} n, \quad K \lesssim (\widetilde{W}_2 \widetilde{L}_2)^{2+\sigma(4\beta-4)/d} \widetilde{L}_2 2^{\widetilde{L}_2^2} (2\beta a^{-1} \log n)^{\beta-1},$$

where $\widetilde{W}_2 = W_2 / \log_2 W_2$ *and* $\widetilde{L}_2 = L_2 / \log_2 L_2$*, such that the GAN estimator (4) satisfies*

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu)] - \epsilon_{opt} \lesssim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n,$$

where $c(\beta, d) = 1$ *if* $2\beta = d$*, and* $c(\beta, d) = 0$ *otherwise.*

If furthermore $m \gtrsim n^{2+2\beta/d} \log^{6+d} n$*, then the GAN estimator (5) satisfies*

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_{n,m}^*)_{\#}\nu)] - \epsilon_{opt} \lesssim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n.$$

Proof Without loss of generality, we assume $a = 1$ in (13). Denote $A_n = [-\beta \log n, \beta \log n]^d$, then $1 - \mu(A_n) \lesssim n^{-\beta/d}$ by (13). We define an operator $\mathcal{T}_n : \mathcal{P}(\mathbb{R}^d) \to \mathcal{P}(A_n)$ on the set $\mathcal{P}(\mathbb{R}^d)$ of all probability distributions on $\mathbb{R}^d$ by

$$\mathcal{T}_n \gamma = \gamma|_{A_n} + (1 - \gamma(A_n))\delta_0, \quad \gamma \in \mathcal{P}(\mathbb{R}^d),$$

where $\mu|_{A_n}$ is the restriction to A_n and δ_0 is the point measure on the zero vector. Since any function $h \in \mathcal{H}$ is bounded $\|h\|_\infty \le 1$, we have

$$\begin{aligned} d_{\mathcal{H}}(\mu, \mathcal{T}_n \mu) &= \sup_{h \in \mathcal{H}} \int_{\mathbb{R}^d} h(x) d\mu(x) - \int_{\mathbb{R}^d} h(x) d\mathcal{T}_n \mu(x) \\ &= \sup_{h \in \mathcal{H}} \int_{\mathbb{R}^d \setminus A_n} h(x) d\mu(x) - (1 - \mu(A_n)) h(0) \\ &\le 2(1 - \mu(A_n)) \lesssim n^{-\beta/d}. \end{aligned}$$

As a consequence, by the triangle inequality,

$$d_{\mathcal{H}}(\mu, (g_n^*)_{\#}\nu) - d_{\mathcal{H}}(\mathcal{T}_n \mu, (g_n^*)_{\#}\nu) \le d_{\mathcal{H}}(\mu, \mathcal{T}_n \mu) \lesssim n^{-\beta/d}.$$

Since $\mathcal{T}_n \mu$ and $g_{\#}\nu$ are supported on A_n for all $g \in \mathcal{G}$, by Lemma 25,

$$d_{\mathcal{H}}(\mathcal{T}_n \mu, (g_n^*)_{\#}\nu) \le \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, A_n) + \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_{\#}\nu) + d_{\mathcal{H}}(\mathcal{T}_n \mu, \widehat{\mu}_n).$$

For the discriminator approximation error, we need to approximate any function $h \in \mathcal{H}^\beta(A_n)$. We can consider the function $\widetilde{h} \in \mathcal{H}^\beta([0,1]^d)$ defined by

$$\widetilde{h}(x) = \frac{1}{(2\beta \log n)^\beta} h(\beta \log n (2x - 1)).$$

By Lemma 11, there exists $\widetilde{\phi} \in \mathcal{NN}(W_2, L_2 - 1) \cap \mathrm{Lip}\,(\mathbb{R}^d, K/(2\beta \log n)^{\beta-1}, 1)$ such that $\|\widetilde{h} - \widetilde{\phi}\|_{L^\infty([0,1]^d)} \lesssim (W_2 L_2 / (\log W_2 \log L_2))^{-2\beta/d}$. Define

$$\begin{aligned} \phi_0(x) &:= (2\beta \log n)^\beta \widetilde{\phi}\left(\tfrac{x}{2\beta \log n} + \tfrac{1}{2}\right), \\ \phi(x) &:= \min\{\max\{\phi_0(x), -1\}, 1\} = \sigma(\phi_0(x) + 1) - \sigma(\phi_0(x) - 1) - 1, \end{aligned}$$

then $\phi \in \mathcal{NN}(W_2, L_2) \cap \mathrm{Lip}\,(\mathbb{R}^d, K, 1)$ and

$$\|h - \phi\|_{L^\infty(A_n)} \precsim (W_2 L_2/(\log W_2 \log L_2))^{-2\beta/d} \log^\beta n.$$

This shows that, if we choose $W_2 L_2 \asymp n^{1/2} \log^{2+d/2} n$,

$$\mathcal{E}(\mathcal{H}, \mathcal{F}, A_n) \precsim (W_2 L_2/(\log W_2 \log L_2))^{-2\beta/d} \log^\beta n \precsim n^{-\beta/d}.$$

For the generator approximation error,

$$\inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widehat{\mu}_n, g_\# \nu) \le d_{\mathcal{F}}(\widehat{\mu}_n, \mathcal{T}_n \widehat{\mu}_n) + \inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\mathcal{T}_n \widehat{\mu}_n, g_\# \nu).$$

By Lemma 10, we can choose a generator $\mathcal{G}$ with $W_1^2 L_1 \precsim n$ such that the last term vanishes. Since $\|f\|_\infty \le 1$ for any $f \in \mathcal{F}$, we have

$$\mathbb{E} d_{\mathcal{F}}(\widehat{\mu}_n, \mathcal{T}_n \widehat{\mu}_n) \le \mathbb{E}[2\widehat{\mu}_n(\mathbb{R}^d \setminus A_n)] = 2\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n 1_{\{X_i \notin A_n\}}\right] = 2\mu(\mathbb{R}^d \setminus A_n) \precsim n^{-\beta/d}.$$

For the statistical error, by Lemma 12,

$$\mathbb{E} d_{\mathcal{H}}(\mathcal{T}_n \mu, \widehat{\mu}_n) \le d_{\mathcal{H}}(\mathcal{T}_n \mu, \mu) + \mathbb{E} d_{\mathcal{H}}(\mu, \widehat{\mu}_n) \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n.$$

In summary, we have shown that

$$\mathbb{E}[d_{\mathcal{H}}(\mu, (g_n^*)_\# \nu)] - \epsilon_{opt} \precsim n^{-\beta/d} \vee n^{-1/2} \log^{c(\beta,d)} n.$$

The error bound for $g_{n,m}^*$ can be estimated in a similar way. By Lemma 25, we only need to further bound $\mathbb{E}[d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m)]$, which can be done as in the proof of Theorem 5. ■

Remark 23 When $\beta = 1$, $\mathcal{H}^1 = \mathrm{Lip}\,(\mathbb{R}^d, 1, 1)$, the metric $d_{\mathcal{H}^1}$ is the Dudley metric. For the Wasserstein distance $\mathcal{W}_1$, we let $A_n = [2a^{-1}\log n, 2a^{-1}\log n]^d$, then

$$\begin{aligned}
\mathcal{W}_1(\mu, \mathcal{T}_n \mu) &= \sup_{\mathrm{Lip}\, h \le 1} \int_{\mathbb{R}^d \setminus A_n} h(x) - h(0) d\mu(x) \le \int_{\mathbb{R}^d \setminus A_n} \|x\|_2 d\mu(x) \\
&\le \sqrt{d}\,\mathbb{E}[\|X\|_\infty 1_{\{X \notin A_n\}}] = \sqrt{d} \int_0^\infty \mu(\|X\|_\infty 1_{\{X\notin A_n\}} > t) dt \\
&\precsim \int_0^{2a^{-1}\log n} n^{-2/d} dt + \int_{2a^{-1}\log n}^\infty 2^{-at/d} dt \\
&\precsim n^{-2/d} \log n.
\end{aligned}$$

If we choose the generator $\mathcal{G} = \{g \in \mathcal{NN}(W_1, L_1) : g(\mathbb{R}) \subseteq A_n\}$ and the discriminator $\mathcal{F} = \mathcal{NN}(W_2, L_2) \cap \mathrm{Lip}\,(\mathbb{R}^d, K, 2a^{-1}\sqrt{d}\log n)$ satisfying the conditions in Theorem 22 with $\beta = 1$, one can show that

$$\mathbb{E}[\mathcal{W}_1(\mu, (g_n^*)_\# \nu)] - \epsilon_{opt} \precsim n^{-1/d} \vee n^{-1/2} \log^{c(1,d)} n,$$

where the same convergence rate holds for $\mathbb{E}\mathcal{W}_1(\mu, \widehat{\mu}_n)$ by Fournier and Guillin (2015). When m is chosen properly, the same rate holds for the estimator $g_{n,m}^*$.

4. Discussion and Related Works

It is well-known that one-hidden-layer neural networks can approximate any continuous function on a compact set (Cybenko, 1989; Hornik, 1991; Pinkus, 1999). Recent breakthroughs of deep learning have motivated many studies on the approximation capacity of deep neural networks (Yarotsky, 2017, 2018; Yarotsky and Zhevnerchuk, 2020; Shen et al., 2019, 2020; Lu et al., 2021; Petersen and Voigtlaender, 2018). These works quantify the approximation error of deep ReLU networks in terms of the number of parameters or neurons. Our result on bounding discriminator approximation error uses ideas similar to those in these papers. An important feature of Lemma 11 is that it gives an explicit bound on the Lipschitz constant required for approximating Hölder functions, which is new in the literature.

In contrast to the vast amount of studies on function approximation by neural networks, there are only a few papers estimating the generator approximation error (Lee et al., 2017; Bailey and Telgarsky, 2018; Perekrestenko et al., 2020; Lu and Lu, 2020; Chen et al., 2020; Yang et al., 2022). The existing studies often assume that the source distribution and the target distribution have the same ambient dimension (Lu and Lu, 2020; Chen et al., 2020) or the distributions have some special form (Lee et al., 2017; Bailey and Telgarsky, 2018; Perekrestenko et al., 2020). However, these assumptions are not satisfied in practical applications. Our analysis of generator approximation is based on Yang et al. (2022), which has the minimal requirement on the source and the target distributions.

The generalization errors of GANs have been studied in several recent works. Arora et al. (2017) showed that, in general, GANs do not generalize under the Wasserstein distance and the Jensen-Shannon divergence with any polynomial number of samples. Alternatively, they estimated the generalization bound under the "neural net distance", which is the IPM with respect to the discriminator network. Zhang et al. (2018) improved the generalization bound in Arora et al. (2017) by explicitly quantifying the complexity of the discriminator network. However, these generalization theories make the assumption that the generator can approximate the data distribution well under the neural net distance, while the construction of such generator network is unknown. Also, the neural net distance is too weak that it can be small when two distributions are not very close (Arora et al., 2017, corollary 3.2). In contrast, our results explicitly state the network architectures and provide convergence rates of GANs under the Wasserstein distance.

Similar to our results, Bai et al. (2019) showed that GANs are able to learn distributions in Wasserstein distance, if the discriminator class has strong distinguishing power against the generator class. But their theory requires each layer of the neural network generator to be invertible, and hence the width of the generator has to be the same with the input dimension, which is not the usual practice in applications. In contrast, we do not make any invertibility assumptions, and allow the discriminator and the generator networks to be wide. The work of Chen et al. (2020) is the most related to ours. They studied statistical properties of GANs and established convergence rate $\mathcal{O}(n^{-\beta/(2\beta+d)}\log^2 n)$ for distributions with Hölder densities, when the evaluation class is another Hölder class $\mathcal{H}^\beta$. Their estimation on generator approximation is based on the optimal transport theory, which requires that the input and the output dimensions of the generator to be the same. In this paper, we study the same problem as Chen et al. (2020) and improve the convergence rate to

$\mathcal{O}(n^{-\beta/d} \vee n^{-1/2} \log n)$ for general probability distributions without any restrictions on the input and the output dimensions of the generator. Furthermore, our results circumvent the curse of dimensionality if the data distribution has a low-dimensional structure, and establish the convergence rate $\mathcal{O}((n^{-\beta/d^*} \vee n^{-1/2}) \log n)$ when the distribution concentrates around a set with Minkowski dimension d^*. The recent work of Schreuder et al. (2021) also consider learning low-dimensional distributions by GANs. However, in their setting, the data distribution is generated from some smooth function and their GAN estimators are defined by directly minimizing Hölder IPMs, rather than using a discriminator network. Hence, our results are more general and practical.

There is another line of work (Liang, 2021; Singh et al., 2018; Uppal et al., 2019) concerning the non-parametric density estimation under IPMs. For example, Liang (2021) and Singh et al. (2018) established the minimax optimal rate $\mathcal{O}(n^{-(\alpha+\beta)/(2\alpha+d)} \vee n^{-1/2})$ for learning a Sobolev class with smoothness index $\alpha > 0$, when the evaluation class is another Sobolev class with smoothness β. Uppal et al. (2019) generalized the minimax rate to Besov IPMs, where both the target density and the evaluation classes are Besov classes. Our main result matches this optimal rate with $\alpha = 0$ without any assumption on the regularity of the data distribution. Theorem 21 shows that GAN is able to achieve the optimal rate by using a suitable regularized empirical distribution.

As we noted in Remark 8, the Lipschitz constraint on the discriminator network may be difficult to satisfy in practical applications. Several regularization techniques (Gulrajani et al., 2017; Kodali et al., 2017; Petzka et al., 2018; Wei et al., 2018; Thanh-Tung et al., 2019) have been applied to GANs and shown to have good empirical performance. It is interesting to see how these regularization techniques affect the convergence rates of GANs. We leave this problem for the future studies.

Finally, we note that there is an optimization error term in our results of convergence rates. So, in order to estimate the full error of GANs used in practice, one also need to estimate the optimization error, which is still a very difficult problem at present. Fortunately, our error analysis is independent of the optimization, so it is possible to combine it with other analysis of optimization. In our main theorems, we give bounds on the network size so that GANs can achieve the optimal convergence rates of learning distributions. In practice, as the network size and sample size get larger, the training becomes more difficult and hence the optimization error may become larger. So there is a trade-off between the optimization error and the bounds derived in this paper. This trade-off can provide some guide on the choice of network size in practice.

5. Proofs of Technical Lemmas

This section provides the proofs of technical lemmas used in the error analysis of GANs. We will first give a general error decomposition of the estimation error in Subsection 5.1, and then bound the generator approximation error in Subsection 5.2, the discriminator approximation error in Subsection 5.3 and the statistical error in Subsection 5.4.

5.1 Error Decomposition

In this subsection, we prove the error decomposition Lemma 9. Before the proof, we introduce the following useful lemma, which states that for any two probability distributions, the difference in IPMs with respect to two distinct evaluation classes will not exceed two times the approximation error between the two evaluation classes. Recall that, for any $\Omega \subseteq \mathbb{R}^d$ and function classes $\mathcal{F}$ and $\mathcal{H}$ defined on Ω, we denote

$$\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) := \sup_{h\in\mathcal{H}} \inf_{f\in\mathcal{F}} \|h - f\|_{L^\infty(\Omega)}.$$

Lemma 24 *For any probability distributions μ and γ supported on $\Omega \subseteq \mathbb{R}^d$,*

$$d_{\mathcal{H}}(\mu, \gamma) \le d_{\mathcal{F}}(\mu, \gamma) + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega).$$

Proof For any $\epsilon > 0$, there exists $h_\epsilon \in \mathcal{H}$ such that

$$d_{\mathcal{H}}(\mu, \gamma) = \sup_{h\in\mathcal{H}} \{\mathbb{E}_\mu[h] - \mathbb{E}_\gamma[h]\} \le \mathbb{E}_\mu[h_\epsilon] - \mathbb{E}_\gamma[h_\epsilon] + \epsilon.$$

Choose $f_\epsilon \in \mathcal{F}$ such that $\|h_\epsilon - f_\epsilon\|_{L^\infty(\Omega)} \le \inf_{f\in\mathcal{F}} \|h_\epsilon - f\|_{L^\infty(\Omega)} + \epsilon$, then

$$\begin{aligned} d_{\mathcal{H}}(\mu, \gamma) \le& \mathbb{E}_\mu[h_\epsilon - f_\epsilon] - \mathbb{E}_\gamma[h_\epsilon - f_\epsilon] + \mathbb{E}_\mu[f_\epsilon] - \mathbb{E}_\gamma[f_\epsilon] + \epsilon \\ \le& 2\|h_\epsilon - f_\epsilon\|_{L^\infty(\Omega)} + \mathbb{E}_\mu[f_\epsilon] - \mathbb{E}_\gamma[f_\epsilon] + \epsilon \\ \le& 2 \inf_{f\in\mathcal{F}} \|h_\epsilon - f\|_{L^\infty(\Omega)} + 2\epsilon + d_{\mathcal{F}}(\mu, \gamma) + \epsilon \\ \le& 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + d_{\mathcal{F}}(\mu, \gamma) + 3\epsilon, \end{aligned}$$

where we use the assumption that μ and γ are supported on Ω in the second inequality, and use the definition of IPM $d_{\mathcal{F}}$ in the third inequality. Letting $\epsilon \to 0$, we get the desired result. ■

The next lemma gives an error decomposition of GAN estimators associated with an estimator $\widetilde{\mu}_n$ of the target distribution μ. Lemma 9 is a special case of this lemma with $\widetilde{\mu}_n = \widehat{\mu}_n$ being the empirical distribution. In the proof, we use two properties of IPM: the triangle inequality $d_{\mathcal{F}}(\mu, \gamma) \le d_{\mathcal{F}}(\mu, \tau) + d_{\mathcal{F}}(\tau, \gamma)$ and, if $\mathcal{F}$ is symmetric, then $d_{\mathcal{F}}(\mu, \gamma) = d_{\mathcal{F}}(\gamma, \mu)$. These properties can be proved easily using the definition.

Lemma 25 *Assume $\mathcal{F}$ is symmetric, μ and $g_\# \nu$ are supported on $\Omega \subseteq \mathbb{R}^d$ for all $g \in \mathcal{G}$. For any probability distribution $\widetilde{\mu}_n$ supported on Ω, let $\widetilde{g}_n^*$ and $\widetilde{g}_{n,m}^*$ be the associated GAN estimators defined by*

$$\begin{aligned} \widetilde{g}_n^* &\in \left\{ g \in \mathcal{G} : d_{\mathcal{F}}(\widetilde{\mu}_n, g_\# \nu) \le \inf_{\phi\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, \phi_\# \nu) + \epsilon_{opt} \right\}, \\ \widetilde{g}_{n,m}^* &\in \left\{ g \in \mathcal{G} : d_{\mathcal{F}}(\widetilde{\mu}_n, g_\# \widehat{\nu}_m) \le \inf_{\phi\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, \phi_\# \widehat{\nu}_m) + \epsilon_{opt} \right\}. \end{aligned}$$

Then, for any function class $\mathcal{H}$ defined on Ω,

$$\begin{aligned} d_{\mathcal{H}}(\mu, (\widetilde{g}_n^*)_\# \nu) &\le \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + \inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_\# \nu) + d_{\mathcal{F}}(\mu, \widetilde{\mu}_n) \wedge d_{\mathcal{H}}(\mu, \widetilde{\mu}_n), \\ d_{\mathcal{H}}(\mu, (\widetilde{g}_{n,m}^*)_\# \nu) &\le \epsilon_{opt} + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + \inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_\# \nu) + d_{\mathcal{F}}(\mu, \widetilde{\mu}_n) \wedge d_{\mathcal{H}}(\mu, \widetilde{\mu}_n) \\ &\quad + 2 d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m). \end{aligned}$$

Proof By lemma 24 and the triangle inequality, for any $g \in \mathcal{G}$,

$$\begin{aligned} d_{\mathcal{H}}(\mu, g_{\#}\nu) &\leq 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + d_{\mathcal{F}}(\mu, g_{\#}\nu) \\ &\leq 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + d_{\mathcal{F}}(\mu, \widetilde{\mu}_n) + d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu). \end{aligned}$$

Alternatively, we can apply the triangle inequality first and then use lemma 24:

$$\begin{aligned} d_{\mathcal{H}}(\mu, g_{\#}\nu) &\leq d_{\mathcal{H}}(\mu, \widetilde{\mu}_n) + d_{\mathcal{H}}(\widetilde{\mu}_n, g_{\#}\nu) \\ &\leq d_{\mathcal{H}}(\mu, \widetilde{\mu}_n) + d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega). \end{aligned}$$

Combining these two bounds, we have

$$d_{\mathcal{H}}(\mu, g_{\#}\nu) \leq 2\mathcal{E}(\mathcal{H}, \mathcal{F}, \Omega) + d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + d_{\mathcal{F}}(\mu, \widetilde{\mu}_n) \wedge d_{\mathcal{H}}(\mu, \widetilde{\mu}_n). \tag{14}$$

Letting $g = \widetilde{g}_n^*$ and observing that $d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_n^*)_{\#}\nu) \leq \inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + \epsilon_{opt}$, we get the bound for $d_{\mathcal{H}}(\mu, (\widetilde{g}_n^*)_{\#}\nu)$.

For $\widetilde{g}_{n,m}^*$, we only need to bound $d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_{n,m}^*)_{\#}\nu)$. By the triangle inequality,

$$d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_{n,m}^*)_{\#}\nu) \leq d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_{n,m}^*)_{\#}\widehat{\nu}_m) + d_{\mathcal{F}}((\widetilde{g}_{n,m}^*)_{\#}\widehat{\nu}_m, (\widetilde{g}_{n,m}^*)_{\#}\nu).$$

By the definition of IPM, the last term can be bounded as

$$d_{\mathcal{F}}((\widetilde{g}_{n,m}^*)_{\#}\widehat{\nu}_m, (\widetilde{g}_{n,m}^*)_{\#}\nu) \leq d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\nu}_m, \nu).$$

By the definition of $\widetilde{g}_{n,m}^*$ and the triangle inequality, we have, for any $g \in \mathcal{G}$,

$$\begin{aligned} d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_{n,m}^*)_{\#}\widehat{\nu}_m) - \epsilon_{opt} \leq d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\widehat{\nu}_m) &\leq d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + d_{\mathcal{F}}(g_{\#}\nu, g_{\#}\widehat{\nu}_m) \\ &\leq d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m). \end{aligned}$$

Taking infimum over all $g \in \mathcal{G}$, we have

$$d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_{n,m}^*)_{\#}\widehat{\nu}_m) \leq \epsilon_{opt} + \inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m).$$

Therefore,

$$d_{\mathcal{F}}(\widetilde{\mu}_n, (\widetilde{g}_{n,m}^*)_{\#}\nu) \leq \epsilon_{opt} + \inf_{g\in\mathcal{G}} d_{\mathcal{F}}(\widetilde{\mu}_n, g_{\#}\nu) + 2d_{\mathcal{F}\circ\mathcal{G}}(\nu, \widehat{\nu}_m).$$

Combining this with the inequality (14), we get the bound for $d_{\mathcal{H}}(\mu, (\widetilde{g}_{n,m}^*)_{\#}\nu)$. ■

5.2 Bounding Generator Approximation Error

For completeness, we sketch the proof of Lemma 10, whose detailed proof can be found in Yang et al. (2022). The proof is essentially based on the fact that ReLU neural networks can express any piece-wise linear functions. The following lemma is a quantified description of this fact.

Lemma 26 (Yang et al. (2022), Lemma 3.1) *Suppose that* $W \geq 7d+1$, $L \geq 2$ *and* $N \leq (W-d-1)\lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor$. *For any* $z_0 < z_1 < \cdots < z_N < z_{N+1}$, *let* $\mathcal{S}^d(z_0, \dots, z_{N+1})$ *be the set of all continuous piece-wise linear functions* $g : \mathbb{R} \to \mathbb{R}^d$ *which have breakpoints only at* z_i, $0 \leq i \leq N+1$, *and are constant on* $(-\infty, z_0)$ *and* (z_{N+1}, ∞). *Then* $\mathcal{S}^d(z_0, \dots, z_{N+1}) \subseteq \mathcal{NN}(W, L)$.

This lemma essentially says that $N \precsim W^2L/d$ is sufficient for $\mathcal{S}^d(z_0, \dots, z_{N+1}) \subseteq \mathcal{NN}(W, L)$. One can also show that it is also a necessary condition. To see this, we denote the number of parameters in $\mathcal{NN}(W, L)$ by $n(W, L) \precsim W^2L$, and consider the function $F : \mathbb{R}^{n(W,L)} \to \mathbb{R}^{d(N+2)}$ defined by $F(\theta) := (f_\theta(z_0), \dots, f_\theta(z_{N+1}))$, where $f_\theta \in \mathcal{NN}(W, L)$ denote the neural network function parameterized by $\theta \in \mathbb{R}^{n(W,L)}$. Since F is a piece-wise multivariate polynomial of θ, it is Lipschitz continuous on any compact sets, hence it does not increase the Hausdorff dimension (Evans and Garzepy, 2018, Theorem 2.8). If $\mathcal{S}^d(z_0, \dots, z_{N+1}) \subseteq \mathcal{NN}(W, L)$, then F is surjective, which implies $d(N+2) \leq n(W, L)$. Thus, $N \precsim n(W, L)/d \precsim W^2L/d$ is necessary for $\mathcal{S}^d(z_0, \dots, z_{N+1}) \subseteq \mathcal{NN}(W, L)$.

Now, we sketch the proof of Lemma 10. For any $\gamma \in \mathcal{P}(n)$, we can assume $\gamma = \sum_{i=1}^n p_i \delta_{x_i}$ with $\sum_{i=1}^n p_i = 1$, $p_i > 0$ and $x_i \in \mathbb{R}^d$. For any absolutely continuous probability measure ν on $\mathbb{R}$, we can choose $2n-2$ points

$$z_{3/2} < z_2 < z_{5/2} < \cdots < z_{n-1/2} < z_n$$

such that $\nu((z_i, z_{i+1/2})) \approx p_i$ for $1 \leq i \leq n$, where we set $z_1 = -\infty$ and $z_{n+1/2} = \infty$ for convenient. Then, we can construct a continuous piece-wise linear function g such that $g(z) = x_i$ for $z \in (z_i, z_{i+1/2})$ and g is linear on $(z_{i+1/2}, z_{i+1})$. For such a function g, $g_\# \nu$ is supported on a union of line segments that pass through all x_i, and $g_\# \nu(\{x_i\}) \approx p_i$ for all $1 \leq i \leq n$. Since $g \in \mathcal{S}^d(z_{3/2}, z_2, \dots, z_n)$ with $2n-2 \leq (W-d-1)\lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor$ breakpoints, Lemma 26 tells us that $g \in \mathcal{NN}(W, L)$. Using this construction, one can show that for any given $\epsilon > 0$, there exists $g \in \mathcal{NN}(W, L)$ such that

$$\mathcal{W}_1(\gamma, g_\# \nu) < \epsilon.$$

Furthermore, in our construction, $g(\mathbb{R}) = \cup_{i=1}^{n-1} g([z_i, z_{i+1}])$ is a union of line segments with endpoints x_i and x_{i+1}. Hence, $g(\mathbb{R})$ must be contained in the convex hull of $\{x_i : 1 \leq i \leq n\}$. Thus, if the support of γ is in a convex set $\mathcal{C}$, g can be chosen to satisfy $g(\mathbb{R}) \subseteq \mathcal{C}$.

Using Lemma 10, we can also bound the generator approximation error of a distribution with bounded support.

Corollary 27 *Let* ν *be an absolutely continuous probability distribution on* $\mathbb{R}$. *Assume that* μ *is a probability distribution on* $[0,1]^d$. *Then, for any* $W \geq 7d+1$ *and* $L \geq 2$, *for the generator* $\mathcal{G} = \{g \in \mathcal{NN}(W, L) : g(\mathbb{R}) \subseteq [0,1]^d\}$, *one has*

$$\inf_{g \in \mathcal{G}} \mathcal{W}_1(\mu, g_\# \nu) \leq C_d (W^2L)^{-1/d},$$

where C_d *is a constant depending only on* d.

Proof Given any $k \in \mathbb{N}$, we denote $A_k := \{(i_1, \dots, i_d)/k : i_j \in \mathbb{N}, i_j \leq k, j = 1, 2, \dots, d\}$, whose cardinality is k^d. It is easy to see that there exists a partition $[0,1]^d = \cup_{x_i \in A_k} Q_{x_i}$

such that for any $x_i \in A_k$ and $x \in Q_{x_i}$, $\|x - x_i\|_2 \leq \sqrt{d}/k$. We consider the discrete distribution

$$\gamma_k := \sum_{x_i \in A_k} \mu(Q_{x_i})\delta_{x_i}.$$

Then,

$$\begin{aligned}
\mathcal{W}_1(\mu, \gamma_k) &= \sup_{\operatorname{Lip} h \leq 1} \int_{[0,1]^d} h(x)d\mu(x) - \sum_{x_i \in A_k} \mu(Q_{x_i})h(x_i) \\
&= \sup_{\operatorname{Lip} h \leq 1} \sum_{x_i \in A_k} \int_{Q_{x_i}} h(x) - h(x_i)d\mu(x) \\
&\leq \sum_{x_i \in A_k} \mu(Q_{x_i})\frac{\sqrt{d}}{k} = \frac{\sqrt{d}}{k}.
\end{aligned}$$

For any $W \geq 7d+1$ and $L \geq 2$, we choose the largest $k \in \mathbb{N}$ such that $k^d \leq \frac{W-d-1}{2}\lfloor \frac{W-d-1}{6d} \rfloor \lfloor \frac{L}{2} \rfloor +$ 2, then by triangle inequality and Lemma 10,

$$\inf_{g \in \mathcal{G}} \mathcal{W}_1(\mu, g_\# \nu) \leq \mathcal{W}_1(\mu, \gamma_k) + \inf_{g \in \mathcal{G}} \mathcal{W}_1(\gamma_k, g_\# \nu) \leq \frac{\sqrt{d}}{k} \leq C_d(W^2L)^{-1/d},$$

for some constant C_d depending only on d. ∎

5.3 Bounding Discriminator Approximation Error

This subsection considers the discriminator approximation error. Our goal is to construct a neural network to approximate a function $h \in \mathcal{H}^\beta([0,1]^d)$ with $\beta = s + r \geq 1$, $s \in \mathbb{N}_0$ and $r \in (0,1]$. The main idea is to approximate the Taylor expansion of h. By Petersen and Voigtlaender (2018, Lemma A.8), for any $x, x_0 \in [0,1]^d$,

$$\left| h(x) - \sum_{\|\alpha\|_1 \leq s} \frac{\partial^\alpha h(x_0)}{\alpha!}(x - x_0)^\alpha \right| \leq d^s \|x - x_0\|_2^\beta.$$

The approximation of the Taylor expansion can be divided into three parts:

- Partition $[0,1]^d$ into small cubes $\cup_\theta Q_\theta$, and construct a network ψ that approximately maps each $x \in Q_\theta$ to a fixed point $x_\theta \in Q_\theta$. Hence, ψ approximately discretize $[0,1]^d$.

- For any α, construct a network ϕ_α that approximates the Taylor coefficient $x \in Q_\theta \mapsto \partial^\alpha h(x_\theta)$. Once $[0,1]^d$ is discretized, this approximation is reduced to a data fitting problem.

- Construct a network $P_\alpha(x)$ to approximate the monomial x^α. In particular, we can construct a network $\phi_\times$ that approximates the product function.

Then our construction of neural network can be written in the form

$$\phi(x) = \sum_{\|\alpha\|_1 \le s} \phi_\times \left(\frac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x - \psi(x)) \right).$$

We collect the required preliminary results in next two subsections and give a proof of Lemma 11 in Subsection 5.3.3.

5.3.1 Data Fitting

Given any $N+2$ samples $\{(x_i, y_i) \in \mathbb{R}^2 : i = 0, 1, \ldots, N+1\}$ with $x_0 < x_1 < \cdots < x_N < x_{N+1}$, there exists a unique piece-wise linear function ϕ that satisfies the following three condition

1. $\phi(x_i) = y_i$ for $i = 0, 1, \ldots, N+1$.
2. ϕ is linear on each interval $[x_i, x_{i+1}]$, $i = 0, 1, \ldots, N$
3. $\phi(x) = y_0$ for $x \in (-\infty, x_0)$ and $\phi(x) = y_{N+1}$ for $x \in (x_{N+1}, \infty)$.

We say ϕ is the linear interpolation of the given samples. Note that for any $x \in \mathbb{R}$,

$$\min_{0 \le i \le N+1} y_i \le \phi(x) \le \max_{0 \le i \le N+1} y_i, \quad \text{and} \quad \operatorname{Lip} \phi \le \max_{0 \le i \le N} \left| \frac{y_{i+1} - y_i}{x_{i+1} - x_i} \right|.$$

The next lemma estimates the required size of network to interpolate the given samples. Note that this lemma is a special case of Lemma 26, which is from Yang et al. (2022, Lemma 3.1) and Daubechies et al. (2021, lemma 3.4).

Lemma 28 *For any $W \ge 6$, $L \in \mathbb{N}$ and any samples $\{(x_i, y_i) \in \mathbb{R}^2 : i = 0, 1, \ldots, N+1\}$ with $x_0 < x_1 < \cdots < x_N < x_{N+1}$, where $N \le \lfloor W/6 \rfloor WL$, the linear interpolation of these samples $\phi \in \mathcal{NN}(W+2, 2L)$.*

As an application of Lemma 28, we show how to use a ReLU neural network to approximately discretize the input space $[0,1]^d$.

Proposition 29 *For any integers $W \ge 6$, $L \ge 2$, $d \ge 1$ and $0 < \delta \le \frac{1}{3K}$ with $K = \lfloor (WL)^{2/d} \rfloor$, there exists a one-dimensional ReLU network $\phi \in \mathcal{NN}(4W+3, 4L)$ such that $\phi(x) \in [0,1]$ for all $x \in \mathbb{R}$, $\operatorname{Lip} \phi \le \frac{2L}{K^2 \delta^2}$ and*

$$\phi(x) = \tfrac{k}{K}, \quad \textit{if } x \in \left[\tfrac{k}{K}, \tfrac{k+1}{K} - \delta \cdot 1_{\{k < K-1\}}\right], k = 0, 1, \ldots, K-1.$$

Proof The proof is divided into two cases: $d = 1$ and $d \ge 2$.

Case 1: $d = 1$. We have $K = W^2 L^2$ and denote $M = W^2 L$. Then we consider the sample set

$$\{(\tfrac{m}{M}, m) : m = 0, 1, \ldots, M-1\} \cup \{(\tfrac{m+1}{M} - \delta, m) : m = 0, 1, \ldots, M-2\} \cup \{(1, M-1)\}.$$

Its cardinality is $2M = 2W^2L \le \lfloor 4W/6 \rfloor (4W)L + 2$. By Lemma 28, the linear interpolation of these samples $\phi_1 \in \mathcal{NN}(4W+2, 2L)$. In particular, $\phi_1(x) \in [0, M-1]$ for all $x \in \mathbb{R}$, $\operatorname{Lip}\phi_1 = 1/\delta$ and

$$\phi_1(x) = m, \quad \text{if } x \in \left[\tfrac{m}{M}, \tfrac{m+1}{M} - \delta \cdot 1_{\{m<M-1\}}\right], m = 0, 1, \dots, M-1.$$

Next, we consider the sample set

$$\{(\tfrac{l}{ML}, l) : l = 0, 1, \dots, L-1\} \cup \{(\tfrac{l+1}{ML} - \delta, l) : l = 0, 1, \dots, L-2\} \cup \{(\tfrac{1}{M}, L-1)\}.$$

Its cardinality is $2L$. By Lemma 28, the linear interpolation of these samples $\phi_2 \in \mathcal{NN}(8, 2L)$. In particular, $\phi_2(x) \in [0, L-1]$ for all $x \in \mathbb{R}$, $\operatorname{Lip}\phi_2 = 1/\delta$ and for $m = 0, 1, \dots, M-1$, $l = 0, 1, \dots, L-1$, we have

$$\phi_2\left(x - \tfrac{1}{M}\phi_1(x)\right) = \phi_2\left(x - \tfrac{m}{M}\right) = l, \quad \text{if } x \in \left[\tfrac{mL+l}{ML}, \tfrac{mL+l+1}{ML} - \delta \cdot 1_{\{mL+l<ML-1\}}\right].$$

Define $\phi(x) := \frac{1}{M}\phi_1(x) + \frac{1}{ML}\phi_2\left(\sigma(x) - \frac{1}{M}\phi_1(x)\right) \in [0,1]$. Then, it is easy to see that $\phi \in \mathcal{NN}(4W+3, 4L)$. For each $x \in \left[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{\{k<K-1\}}\right]$ with $k \in \{0, 1, \dots, K-1\} = \{0, 1, \dots, ML-1\}$, there exists a unique representation $k = mL + l$ for $m \in \{0, 1, \dots, M-1\}$, $l \in \{0, 1, \dots, L-1\}$, and we have

$$\phi(x) = \tfrac{1}{M}\phi_1(x) + \tfrac{1}{ML}\phi_2\left(\sigma(x) - \tfrac{1}{M}\phi_1(x)\right) = \tfrac{mL+l}{ML} = \tfrac{k}{K}.$$

Observing that the Lipschitz constant of the function $x \mapsto \sigma(x) - \frac{1}{M}\phi_1(x)$ is $\frac{1}{M\delta}$, the Lipschitz constant of ϕ is at most $\frac{1}{M}\frac{1}{\delta} + \frac{1}{ML}\frac{1}{\delta}\frac{1}{M\delta} \le \frac{2L}{K^2\delta^2}$.

Case 2: $d \ge 2$. We consider the sample set

$$\{(\tfrac{k}{K}, \tfrac{k}{K}) : k = 0, 1, \dots, K-1\} \cup \{(\tfrac{k+1}{K} - \delta, \tfrac{k}{K}) : k = 0, 1, \dots, K-1\} \cup \{(1, \tfrac{K-1}{K})\}.$$

Its cardinality is $2K \le 2W^{2/d}L^{2/d} \le \lfloor 4W/6 \rfloor (4W)L + 2$. By Lemma 28, the linear interpolation of these samples $\phi \in \mathcal{NN}(4W+2, 2L)$. In particular, $\phi(x) \in [0,1]$ for all $x \in \mathbb{R}$,

$$\phi(x) = \tfrac{k}{K}, \quad \text{if } x \in \left[\tfrac{k}{K}, \tfrac{k+1}{K} - \delta \cdot 1_{\{k<K-1\}}\right], k = 0, 1, \dots, K-1,$$

and the Lipschitz constant of ϕ is $\frac{1}{K\delta} \le \frac{2L}{K^2\delta^2}$. ■

Lemma 28 shows that a network $\mathcal{NN}(W, L)$ can exactly fit $N \asymp W^2L$ samples. We are going to show that it can approximately fit $N \asymp (W/\log_2 W)^2(L/\log_2 L)^2$ samples. The construction is based on the bit extraction technique (Bartlett et al., 1998, 2019). The following lemma shows how to extract a specific bit using ReLU neural networks. For convenient, we denote the binary representation as

$$\operatorname{Bin} 0.x_1x_2\dots x_L := \sum_{j=1}^{L} x_j 2^{-j} \in [0,1],$$

where $x_j \in \{0,1\}$ for all $j = 1, 2, \dots, L$.

Lemma 30 *For any $L \in \mathbb{N}$, there exists $\phi \in \mathcal{NN}(8, 2L)$ such that $\phi(x, l) = x_l$ for $x = \mathrm{Bin}\, 0.x_1 x_2 \dots x_L$ with $x_j \in \{0, 1\}$ and $l = 1, 2, \dots, L$. Furthermore, $|\phi(x, l) - \phi(x', l')| \le 2 \cdot 2^{L^2} |x - x'| + L|l - l'|$ for any $x, x', l, l' \in \mathbb{R}$.*

Proof For any $x = \mathrm{Bin}\, 0.x_1 x_2 \dots x_L$, we define $\xi_j := \mathrm{Bin}\, 0.x_j x_{j+1} \dots x_L$ for $j = 1, 2, \dots, L$. Then $\xi_1 = x$ and $\xi_{j+1} = 2\xi_j - x_j = \sigma(2\sigma(\xi_j) - x_j)$ for $j = 1, 2, \dots, L-1$. Let

$$T(x) := \sigma(2^L x - 2^{L-1} + 1) - \sigma(2^L x - 2^{L-1}) = \begin{cases} 0 & x \le 1/2 - 2^{-L}, \\ \text{linear} & 1/2 - 2^{-L} < x < 1/2, \\ 1 & x \ge 1/2. \end{cases}$$

It is easy to check that $x_j = T(\xi_j)$.

Denote $\delta_j = 1$ if $j = 0$ and $\delta_j = 0$ if $j \neq 0$ is an integer. Observing that

$$\delta_j = \sigma(j+1) + \sigma(j-1) - 2\sigma(j),$$

and $t_1 t_2 = \sigma(t_1 + t_2 - 1)$ for any $t_1, t_2 \in \{0, 1\}$, we have

$$x_l = \sum_{j=1}^{L} \delta_{l-j} x_j = \sum_{j=1}^{L} \sigma\left(\sigma(l-j+1) + \sigma(l-j-1) - 2\sigma(l-j) + x_j - 1\right). \tag{15}$$

If we denote the partial sum $s_{l,j} = \sum_{i=1}^{j} \sigma(\sigma(l-i+1) + \sigma(l-i-1) - 2\sigma(l-i) + x_i - 1)$, then $x_l = s_{l,L}$.

For any $t_1, t_2, t_3 \in \mathbb{R}$, we define a function $\psi(t_1, t_2, t_3) = (y_1, y_2, y_3) \in \mathbb{R}^3$ by

$$\begin{aligned} y_1 &:= \sigma(2\sigma(t_1) - T(t_1)), \\ y_2 &:= \sigma(t_2) + \sigma(\sigma(t_3) + \sigma(t_3 - 2) - 2\sigma(t_3 - 1) + T(t_1) - 1), \\ y_3 &:= \max\{t_3 - 1, -L\} = \sigma(t_3 - 1 + L) - L. \end{aligned}$$

Then, it is easy to check that $\psi \in \mathcal{NN}(8, 2)$. Using the expressions (15) we have derived for x_l, one has

$$\psi(\xi_j, s_{l,j-1}, l - j + 1) = (\xi_{j+1}, s_{l,j}, l - j), \quad l, j = 1, \dots, L,$$

where $s_{l,0} := 0$ and $\xi_{L+1} := 0$. Hence, by composing ψ L times, we can construct a network $\phi = \psi \circ \cdots \circ \psi \in \mathcal{NN}(8, 2L)$ such that $\phi(x, l) = \psi \circ \cdots \circ \psi(x, 0, l) = s_{l,L} = x_l$ for $l = 1, 2, \dots, L$, where we drop the first and the third outputs of ψ in the last layer.

It remains to estimate the Lipschitz constant. For any $t_1, t_2, t_3, t_1', t_2', t_3' \in \mathbb{R}$, suppose $(y_1, y_2, y_3) = \psi(t_1, t_2, t_3)$ and $(y_1', y_2', y_3') = \psi(t_1', t_2', t_3')$. Then $|y_1 - y_1'| \le 2^L |t_1 - t_1'|$, $|y_3 - y_3'| \le |t_3 - t_3'|$ and $|y_2 - y_2'| \le |t_2 - t_2'| + 2^L |t_1 - t_1'| + |t_3 - t_3'|$. Therefore, by induction,

$$\begin{aligned} |\phi(x, l) - \phi(x', l')| &\le (2^L + 2^{2L} + 2^{3L} + \cdots + 2^{L^2}) |x - x'| + L|l - l'| \\ &\le 2 \cdot 2^{L^2} |x - x'| + L|l - l'|, \end{aligned}$$

for any $x, x', l, l' \in \mathbb{R}$. ■

Using the bit extraction technique, the next lemma shows a network $\mathcal{NN}(W, L)$ can exactly fit $N \asymp W^2 L^2$ binary samples.

Lemma 31 *Given any $W \geq 6$, $L \geq 2$ and any $\theta_i \in \{0,1\}$ for $i = 0,1,\dots,W^2L^2-1$, there exists $\phi \in \mathcal{NN}(8W+4,4L)$ such that $\phi(i) = \theta_i$ for $i = 0,1,\dots,W^2L^2-1$ and $\operatorname{Lip}\phi \leq 2\cdot 2^{L^2} + L^2$.*

Proof Denote $M = W^2L$, then, for each $i = 0,1,\dots,W^2L^2-1$, there exists a unique representation $i = mL + l$ with $m = 0,1,\dots,M-1$ and $l = 0,1,\dots,L-1$. So we define $b_{m,l} := \theta_i$, where $i = mL+l$. We further set $y_m := \operatorname{Bin} 0.b_{m,0}b_{m,1}\dots b_{m,L-1} \in [0,1]$ and $y_M = 1$. By Lemma 30, there exists $\psi \in \mathcal{NN}(8,2L)$ such that $\psi(y_m, l+1) = b_{m,l}$ for any $m = 0,1,\dots,M-1$, and $l = 0,1,\dots,L-1$.

We consider the sample set

$$\{(mL, y_m) : m = 0,1,\dots,M\} \cup \{(mL-1, y_{m-1}) : m = 1,\dots,M\}.$$

Its cardinality is $2M+1 = 2W^2L+1 \leq \lfloor 4W/6 \rfloor (4W)L + 2$. By Lemma 28, the linear interpolation of these samples $\phi_1 \in \mathcal{NN}(4W+2,2L)$. In particular, $\operatorname{Lip}\phi_1 \leq 1$ and $\phi_1(i) = y_m$, when $i = mL+l$, for $m = 0,1,\dots,M-1$, and $l = 0,1,\dots,L-1$.

Similarly, for the sample set

$$\{(mL,0) : m = 0,1,\dots,M\} \cup \{(mL-1, L-1) : m = 1,\dots,M\},$$

the linear interpolation of these samples $\phi_2 \in \mathcal{NN}(4W+2,2L)$. In particular, $\operatorname{Lip}\phi_2 = L-1$ and $\phi_2(i) = l$, when $i = mL+l$, for $m = 0,1,\dots,M-1$, and $l = 0,1,\dots,L-1$.

We define $\phi(x) := \psi(\phi_1(x), \phi_2(x)+1)$, then $\phi \in \mathcal{NN}(8W+4,4L)$ and

$$\phi(i) = \psi(\phi_1(i), \phi_2(i)+1) = \psi(y_m, l+1) = b_{m,l} = \theta_i$$

for $i = mL+l$ with $m = 0,1,\dots,M-1$, and $l = 0,1,\dots,L-1$. By Lemma 30, we have

$$|\phi(x) - \phi(x')| \leq 2\cdot 2^{L^2}|\phi_1(x) - \phi_1(x')| + L|\phi_2(x) - \phi_2(x')| \leq (2\cdot 2^{L^2} + L^2)|x - x'|$$

for any $x, x' \in \mathbb{R}$. ■

As an application of Lemma 31, we show that a network $\mathcal{NN}(W,L)$ can approximately fit $N \asymp (W/\log_2 W)^2 (L/\log_2 L)^2$ samples.

Proposition 32 *For any $W \geq 6$, $L \geq 2$, $s \in \mathbb{N}$ and any $\xi_i \in [0,1]$ for $i = 0,1,\dots,W^2L^2-1$, there exists $\phi \in \mathcal{NN}(8s(2W+1)\lceil \log_2(2W) \rceil + 2, 4L\lceil \log_2(2L)\rceil + 1)$ such that $\operatorname{Lip}\phi \leq 4\cdot 2^{L^2} + 2L^2$, $|\phi(i) - \xi_i| \leq (WL)^{-2s}$ for $i = 0,1,\dots,W^2L^2-1$ and $\phi(t) \in [0,1]$ for all $t \in \mathbb{R}$.*

Proof Denote $J = \lceil 2s\log_2(WL) \rceil$. For each $\xi_i \in [0,1]$, there exist $b_{i,1}, b_{i,2}, \dots, b_{i,J} \in \{0,1\}$ such that

$$|\xi_i - \operatorname{Bin} 0.b_{i,1}b_{i,2}\dots b_{i,J}| \leq 2^{-J}.$$

By Lemma 31, there exist $\phi_1, \phi_2, \dots, \phi_J \in \mathcal{NN}(8W+4,4L)$ such that $\operatorname{Lip}\phi_j \leq 2\cdot 2^{L^2} + L^2$ and $\phi_j(i) = b_{i,j}$ for $i = 0,1,\dots,W^2L^2-1$ and $j = 1,2,\dots,J$. We define

$$\tilde{\phi}(t) := \sum_{j=1}^{J} 2^{-j}\phi_j(t), \quad t \in \mathbb{R}.$$

Then, for $i = 0, 1, \dots, W^2L^2 - 1$,

$$|\tilde{\phi}(i) - \xi_i| = \left| \sum_{j=1}^{J} 2^{-j} b_{i,j} - \xi_i \right| = |\operatorname{Bin} 0.b_{i,1}b_{i,2}\dots b_{i,J} - \xi_i| \le 2^{-J} \le (WL)^{-2s}.$$

Since $J \le 1 + 2s\log_2(WL) \le 2(1 + s\log_2 W)(1 + \log_2 L) \le 2s\log_2(2W)\log_2(2L)$, $\tilde{\phi}$ can be implemented to be a network with width $8s(2W+1)\lceil\log_2(2W)\rceil+2$ and depth $4L\lceil\log_2(2L)\rceil$, where we use two neurons in each hidden layer to remember the input and intermediate summation. Furthermore, for any $t, t' \in \mathbb{R}$,

$$|\tilde{\phi}(t) - \tilde{\phi}(t')| \le \sum_{j=1}^{J} 2^{-j} \operatorname{Lip} \phi_j |t - t'| \le (4 \cdot 2^{L^2} + 2L^2)|t - t'|.$$

Finally, we define

$$\phi(t) := \min\{\max\{\tilde{\phi}(t), 0\}, 1\} = \sigma(\tilde{\phi}(t)) - \sigma(\tilde{\phi}(t) - 1) \in [0, 1].$$

Then $\phi \in \mathcal{NN}(8s(2W+1)\lceil\log(2W)\rceil, 4L\lceil\log(2L)\rceil + 1)$, $\operatorname{Lip}\phi \le \operatorname{Lip}\tilde{\phi}$ and $\phi(i) = \tilde{\phi}(i)$ for $i = 0, 1, \dots, W^2L^2 - 1$. ■

5.3.2 Approximation of Polynomials

The approximation of polynomials by ReLU neural networks is well-known (Yarotsky, 2017; Lu et al., 2021). The next lemma gives an estimate of the approximation error of the product function.

Lemma 33 *For any $W, L \in \mathbb{N}$, there exists $\phi \in \mathcal{NN}(9W + 1, L)$ such that for any $x, x', y, y' \in [-1, 1]$,*

$$|xy - \phi(x, y)| \le 6W^{-L},$$
$$|\phi(x, y) - \phi(x', y')| \le 7|x - x'| + 7|y - y'|.$$

Proof We follow the construction in Lu et al. (2021). We first construct a neural network ψ that approximates the function $f(x) = x^2$ on $[0, 1]$. Denote

$$T_1(x) := \begin{cases} 2x, & x \in [0, 1/2], \\ 2(1 - x), & x \in (1/2, 1], \end{cases}$$

and $T_i(x) := T_{i-1}(T_1(x))$ for $x \in [0, 1]$ and $i = 2, 3, \cdots$. We note that T_i can be implemented by a one-hidden-layer ReLU network with width 2^i. Let $f_k : [0, 1] \to [0, 1]$ be the piece-wise linear function such that $f_k(\frac{j}{2^k}) = \left(\frac{j}{2^k}\right)^2$ for $j = 0, 1, \dots, 2^k$, and f_k is linear on $[\frac{j-1}{2^k}, \frac{j}{2^k}]$ for $j = 1, 2, \dots, 2^k$. Then, using the fact $\frac{(x-h)^2 + (x+h)^2}{2} - x^2 = h^2$, we have

$$|x^2 - f_k(x)| \le 2^{-2(k+1)}, \quad x \in [0, 1], k \in \mathbb{N}.$$

Furthermore, $f_{k-1}(x) - f_k(x) = \frac{T_k(x)}{2^{2k}}$ and $x - f_1(x) = \frac{T_1(x)}{4}$. Hence,

$$f_k(x) = x - (x - f_1(x)) - \sum_{i=2}^{k}(f_{i-1}(x) - f_i(x)) = x - \sum_{i=1}^{k} \frac{T_i(x)}{2^{2i}}, \quad x \in [0,1], k \in \mathbb{N}.$$

Given $W \in \mathbb{N}$, there exists a unique $n \in \mathbb{N}$ such that $(n-1)2^{n-1} + 1 \le W \le n2^n$. For any $L \in \mathbb{N}$, it was showed in Lu et al. (2021, Lemma 5.1) that f_{nL} can be implemented by a network ψ with width $3W$ and depth L. Hence,

$$|x^2 - \psi(x)| \le |x^2 - f_{nL}(x)| \le 2^{-2(nL+1)} = 2^{-2nL}/4 \le W^{-L}/4, \quad x \in [0,1],$$

where we use $W \le n2^n \le 2^{2n}$ in the last inequality.

Using the fact that

$$xy = 2\left(\left(\tfrac{x+y}{2}\right)^2 - \left(\tfrac{x}{2}\right)^2 - \left(\tfrac{y}{2}\right)^2\right), \quad x, y \in \mathbb{R},$$

we can approximate the function $f(x,y) = xy$ by

$$\phi_0(x,y) := 2\left(\psi\left(\tfrac{x+y}{2}\right) - \psi\left(\tfrac{x}{2}\right) - \psi\left(\tfrac{y}{2}\right)\right).$$

Then, $\phi_0 \in \mathcal{NN}(9W, L)$ and for $x, y \in [0,1]$,

$$|xy - \phi_0(x,y)| \le 2\left|\left(\tfrac{x+y}{2}\right)^2 - \psi\left(\tfrac{x+y}{2}\right)\right| + 2\left|\left(\tfrac{x}{2}\right)^2 - \psi\left(\tfrac{x}{2}\right)\right| + 2\left|\left(\tfrac{y}{2}\right)^2 - \psi\left(\tfrac{y}{2}\right)\right| \le \tfrac{3}{2}W^{-L}$$

Furthermore, for any $x, x', y, y' \in [0,1]$,

$$\begin{aligned}
&|\phi_0(x,y) - \phi_0(x',y')| \\
\le& 2\left|f_{nL}\left(\tfrac{x+y}{2}\right) - f_{nL}\left(\tfrac{x'+y'}{2}\right)\right| + 2\left|f_{nL}\left(\tfrac{x}{2}\right) - f_{nL}\left(\tfrac{x'}{2}\right)\right| + 2\left|f_{nL}\left(\tfrac{y}{2}\right) - f_{nL}\left(\tfrac{y'}{2}\right)\right| \\
\le& 4\left|\tfrac{x+y}{2} - \tfrac{x'+y'}{2}\right| + 2\left|\tfrac{x}{2} - \tfrac{x'}{2}\right| + 2\left|\tfrac{y}{2} - \tfrac{y'}{2}\right| \\
\le& 3|x - x'| + 3|y - y'|,
\end{aligned}$$

where we use $|f_{nL}(t) - f_{nL}(t')| \le 2|t - t'|$ for $t, t' \in [0,1]$ and $|f_{nL}(t) - f_{nL}(t')| \le |t - t'|$ for $t, t' \in [0, 1/2]$.

For any $x, y \in [-1,1]$, set $x_0 = (x+1)/2 \in [0,1]$ and $y_0 = (y+1)/2 \in [0,1]$, then $xy = 4x_0y_0 - x - y - 1$. Using this fact, we define the target function by

$$\phi(x,y) = 4\phi_0(\tfrac{x+1}{2}, \tfrac{y+1}{2}) - \sigma(x+y+2) + 1.$$

Then, $\phi \in \mathcal{NN}(9W+1, L)$ and for $x, y \in [-1,1]$,

$$|xy - \phi(x,y)| \le 4|\tfrac{x+1}{2}\tfrac{y+1}{2} - \phi_0(\tfrac{x+1}{2}, \tfrac{y+1}{2})| \le 6W^{-L}.$$

Furthermore, for any $x, x', y, y' \in [-1,1]$,

$$\begin{aligned}
|\phi(x,y) - \phi(x',y')| \le& 4|\phi_0(\tfrac{x+1}{2}, \tfrac{y+1}{2}) - \phi_0(\tfrac{x'+1}{2}, \tfrac{y'+1}{2})| + |x + y - x' - y'| \\
\le& 7|x - x'| + 7|y - y'|,
\end{aligned}$$

which completes the proof. ■

By applying the approximation of the product function, we can approximate any monomials by neural networks.

Corollary 34 *Let $P(x) = x^\alpha = x_1^{\alpha_1} x_2^{\alpha_2} \cdots x_d^{\alpha_d}$ for $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_d) \in \mathbb{N}_0^d$ with $\|\alpha\|_1 = k \geq 2$. For any $W, L \in \mathbb{N}$, there exists $\phi \in \mathcal{NN}(9W + k - 1, (k-1)(L+1))$ such that for any $x, y \in [-1,1]^d$, $\phi(x) \in [-1,1]$ and*

$$
\begin{aligned}
|\phi(x) - P(x)| &\leq 6(k-1)W^{-L}, \\
|\phi(x) - \phi(y)| &\leq 7^{k-1} \|\alpha\|_\infty \|x - y\|_1.
\end{aligned}
$$

Proof For any $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$, let $z = (z_1, z_2, \ldots, z_k) \in \mathbb{R}^k$ be the vector such that $z_i = x_j$ if $\sum_{l=1}^{j-1} \alpha_l < i \leq \sum_{l=1}^{j} \alpha_l$ for $j = 1, 2, \ldots, d$. Then $P(x) = x^\alpha = z_1 z_2 \cdots z_k$ and there exists a linear map $\phi_0 : \mathbb{R}^d \to \mathbb{R}^k$ such that $\phi_0(x) = z$.

Let $\psi_1 \in \mathcal{NN}(9W+1, L)$ be the neural network in Lemma 33. We define

$$
\psi_2(x,y) := \min\{\max\{\psi_1(x,y), -1\}, 1\} = \sigma(\psi_1(x,y) + 1) - \sigma(\psi_1(x,y) - 1) - 1 \in [-1,1],
$$

then $\psi_2 \in \mathcal{NN}(9W+1, L+1)$ and ψ_2 also satisfies the inequalities in Lemma 33. For $i = 3, 4, \ldots, k$, we define $\psi_i : [-1,1]^i \to [-1,1]$ inductively by

$$
\psi_i(z_1, \ldots, z_i) := \psi_2(\psi_{i-1}(z_1, \ldots, z_{i-1}), z_i).
$$

Since $z_i = \sigma(z_i + 1) - 1$ for $z_i \in [-1,1]$, it is easy to see that ψ_i can be implemented by a network with width $9W + i - 1$ and depth $(i-1)(L+1)$ by induction. Furthermore,

$$
\begin{aligned}
&|\psi_i(z_1, \ldots, z_i) - z_1 \cdots z_i| \\
\leq &|\psi_2(\psi_{i-1}(z_1, \ldots, z_{i-1}), z_i) - \psi_{i-1}(z_1, \ldots, z_{i-1}) z_i| + |\psi_{i-1}(z_1, \ldots, z_{i-1}) z_i - z_1 \cdots z_i| \\
\leq &6W^{-L} + |\psi_{i-1}(z_1, \ldots, z_{i-1}) - z_1 \cdots z_{i-1}| \\
\leq &\cdots \leq (i-2)6W^{-L} + |\psi_2(z_1, z_2) - z_1 z_2| \\
\leq &(i-1)6W^{-L}.
\end{aligned}
$$

And for any $z = (z_1, z_2, \ldots, z_k), z' = (z_1', z_2', \ldots, z_k') \in [-1,1]^k$,

$$
\begin{aligned}
|\psi_i(z_1, \ldots, z_i) - \psi_i(z_1', \ldots, z_i')| &\leq 7|\psi_{i-1}(z_1, \ldots, z_{i-1}) - \psi_{i-1}(z_1', \ldots, z_{i-1}')| + 7|z_i - z_i'| \\
&\leq \cdots \leq 7^{i-2} |\psi_2(z_1, z_2) - \psi_2(z_1', z_2')| + \sum_{j=3}^{i} 7^{i-j+1} |z_j - z_j'| \\
&\leq 7^{i-1} \|z - z'\|_1.
\end{aligned}
$$

We define the target function as $\phi(x) := \psi_k(\phi_0(x))$, then $\phi \in \mathcal{NN}(9W + k - 1, (k-1)(L+1))$. And for $x, y \in [-1,1]^d$, denote $z = \phi_0(x)$ and $z' = \phi_0(y)$, we have

$$
\begin{aligned}
|\phi(x) - P(x)| &= |\psi_k(z) - z_1 z_2 \cdots z_k| \leq 6(k-1)W^{-L}, \\
|\phi(x) - \phi(y)| &= |\psi_k(z) - \psi_k(z')| \leq 7^{k-1} \|z - z'\|_1 \leq 7^{k-1} \|\alpha\|_\infty \|x - y\|_1.
\end{aligned}
$$

So we finish the proof. ∎

5.3.3 Proof of Lemma 11

Now, we can bound the discriminator approximation error. We recall Lemma 11 in the following and give a proof.

Lemma 11 *Assume $h \in \mathcal{H}^\beta([0,1]^d)$ with $\beta = s + r$, $s \in \mathbb{N}_0$ and $r \in (0,1]$. For any $W \geq 6$, $L \geq 2$, there exists $\phi \in \mathcal{NN}(49(s+1)^2 3^d d^{s+1} W \lceil \log_2 W \rceil, 15(s+1)^2 L \lceil \log_2 L \rceil + 2d)$ such that $\|\phi\|_\infty \leq 1$, $\operatorname{Lip} \phi \leq (s+1) d^{s+1/2} L (WL)^{\sigma(4\beta-4)/d} (1260 W^2 L^2 2^{L^2} + 19 s 7^s)$ and*

$$\|\phi - h\|_{L^\infty([0,1]^d)} \leq 6(s+1)^2 d^{(s+\beta/2)\vee 1} \lfloor (WL)^{2/d} \rfloor^{-\beta}.$$

Proof We divide the proof into four steps as follows.
Step 1: Discretization.

Let $K = \lfloor (WL)^{2/d} \rfloor$ and $\delta = \frac{1}{3K^{\beta \vee 1}} \leq \frac{1}{3K}$. For each $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \{0, 1, \ldots, K-1\}^d$, we define

$$Q_\theta := \left\{ x = (x_1, x_2, \ldots, x_d) : x_i \in \left[\frac{\theta_i}{K}, \frac{\theta_i + 1}{K} - \delta \cdot 1_{\{\theta_i < K-1\}} \right], i = 1, 2, \ldots, d \right\}.$$

By Proposition 29, there exists $\psi_1 \in \mathcal{NN}(4W+3, 4L)$ such that

$$\psi_1(t) = \frac{k}{K}, \quad \text{if } t \in \left[\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{\{k < K-1\}} \right], k = 0, 1, \ldots, K-1,$$

and $\operatorname{Lip} \psi_1 \leq 2LK^{-2}\delta^{-2}$. We define

$$\psi(x) := (\psi_1(x_1), \ldots, \psi_1(x_d)), \quad x = (x_1, \ldots, x_d) \in \mathbb{R}^d.$$

Then, $\psi \in \mathcal{NN}(d(4W+3), 4L)$ and $\psi(x) = \frac{\theta}{K}$ for $x \in Q_\theta$.
Step 2: Approximation of Taylor coefficients.

Since $\theta \in \{0, 1, \ldots, K-1\}^d$ is one-to-one correspondence to $i_\theta := \sum_{j=1}^d \theta_j K^{j-1} \in \{0, 1, \ldots, K^d - 1\}$, we define

$$\psi_0(x) := (K, K^2, \ldots, K^d) \cdot \psi(x) = \sum_{j=1}^d \psi_1(x_j) K^j \quad x \in \mathbb{R}^d,$$

then $\psi_0 \in \mathcal{NN}(d(4W+3), 4L)$ and

$$\psi_0(x) = \sum_{j=1}^d \theta_j K^{j-1} = i_\theta \quad \text{if } x \in Q_\theta,\ \theta \in \{0, 1, \ldots, K-1\}^d.$$

For any $x, x' \in \mathbb{R}^d$, we have

$$|\psi_0(x) - \psi_0(x')| \leq \sum_{j=1}^d K^j |\psi_1(x_j) - \psi_1(x_j')| \leq \sqrt{d} K^d \operatorname{Lip} \psi_1 \|x - x'\|_2 \leq 2\sqrt{d} L K^{d-2} \delta^{-2} \|x - x'\|_2.$$

For any $\alpha \in \mathbb{N}_0^d$ satisfying $\|\alpha\|_1 \leq s$ and each $i = i_\theta \in \{0, 1, \ldots, K^d - 1\}$, we denote $\xi_{\alpha,i} := (\partial^\alpha h(\theta/K) + 1)/2 \in [0,1]$. Since $K^d \leq W^2 L^2$, by Proposition 32, there exists

$\varphi_\alpha \in \mathcal{NN}(8(s+1)(2W+1)\lceil\log_2(2W)\rceil+2, 4L\lceil\log_2(2L)\rceil+1)$ such that $\operatorname{Lip}\varphi_\alpha \le 4\cdot 2^{L^2} + 2L^2 \le 5\cdot 2^{L^2}$ and $|\varphi_\alpha(i)-\xi_{\alpha,i}| \le (WL)^{-2(s+1)}$ for all $i\in\{0,1,\dots,K^d-1\}$. We define

$$\phi_\alpha(x) := 2\varphi_\alpha(\psi_0(x)) - 1 \in [-1,1], \quad x\in\mathbb{R}^d.$$

Then ϕ_α can be implemented by a network with width $8d(s+1)(2W+1)\lceil\log_2(2W)\rceil+2 \le 40d(s+1)W\lceil\log_2 W\rceil$ and depth $4L+4L\lceil\log_2(2L)\rceil+1 \le 13L\lceil\log_2 L\rceil$. And we have

$$\operatorname{Lip}\phi_\alpha \le 2\operatorname{Lip}\varphi_\alpha \operatorname{Lip}\psi_0 \le 20\sqrt{d}LK^{d-2}\delta^{-2}2^{L^2}, \tag{16}$$

and for any $\theta\in\{0,1,\dots,K-1\}^d$, if $x\in Q_\theta$,

$$|\phi_\alpha(x) - \partial^\alpha h(\theta/K)| = 2|\varphi_\alpha(i_\theta)-\xi_{\alpha,i_\theta}| \le 2(WL)^{-2(s+1)}. \tag{17}$$

Step 3: Approximation of h on $\bigcup_{\theta\in\{0,1,\dots,K-1\}^d} Q_\theta$.

Let $\varphi(t)=\min\{\max\{t,0\},1\}=\sigma(t)-\sigma(t-1)$ for $t\in\mathbb{R}$. We extend its definition to $\mathbb{R}^d$ coordinate-wisely, so $\varphi:\mathbb{R}^d\to[0,1]^d$ and $\varphi(x)=x$ for any $x\in[0,1]^d$.

By Lemma 33, there exists $\phi_\times \in \mathcal{NN}(9W+1, 2(s+1)L)$ such that for any $t_1,t_2,t_3,t_4\in[-1,1]$,

$$|t_1t_2 - \phi_\times(t_1,t_2)| \le 6W^{-2(s+1)L}, \tag{18}$$

$$|\phi_\times(t_1,t_2)-\phi_\times(t_3,t_4)| \le 7|t_1-t_3| + 7|t_2-t_4|. \tag{19}$$

By corollary 34, for any $\alpha\in\mathbb{N}_0^d$ with $2\le\|\alpha\|_1\le s$, there exists $P_\alpha\in\mathcal{NN}(9W+s-1,(s-1)(2(s+1)L+1))$ such that for any $x,y\in[-1,1]^d$, $P_\alpha(x)\in[-1,1]$ and

$$|P_\alpha(x)-x^\alpha| \le 6(s-1)W^{-2(s+1)L}, \tag{20}$$

$$|P_\alpha(x)-P_\alpha(y)| \le 7^{s-1}s\|x-y\|_1. \tag{21}$$

When $\|\alpha\|_1=1$, it is easy to implemented $P_\alpha(x)=x^\alpha$ by a neural network with Lipschitz constant at most one. Hence, the inequalities (20) and (21) hold for $1\le\|\alpha\|_1\le s$.

For any $x\in Q_\theta$, $\theta\in\{0,1,\dots,K-1\}^d$, we can approximate $h(x)$ by a Taylor expansion. Thanks to Petersen and Voigtlaender (2018, Lemma A.8), we have the following error estimation for $x\in Q_\theta$,

$$\left| h(x) - h(\tfrac{\theta}{K}) - \sum_{1\le\|\alpha\|_1\le s} \frac{\partial^\alpha h(\frac{\theta}{K})}{\alpha!}(x-\tfrac{\theta}{K})^\alpha \right| \le d^s\|x-\tfrac{\theta}{K}\|_2^\beta \le d^{s+\beta/2}K^{-\beta}. \tag{22}$$

Motivated by this, we define

$$\widetilde{\phi}_0(x) := \phi_{\mathbf{0}_d}(x) + \sum_{1\le\|\alpha\|_1\le s} \phi_\times\left(\tfrac{\phi_\alpha(x)}{\alpha!}, P_\alpha(\varphi(x)-\psi(x))\right),$$

$$\phi_0(x) := \sigma(\widetilde{\phi}_0(x)+1) - \sigma(\widetilde{\phi}_0(x)-1) - 1 \in [-1,1],$$

where we denote $\mathbf{0}_d=(0,\dots,0)\in\mathbb{N}_0^d$. Observe that the number of terms in the summation can be bounded by

$$\sum_{\alpha\in\mathbb{N}_0^d,\|\alpha\|_1\le s} 1 = \sum_{j=0}^{s}\sum_{\alpha\in\mathbb{N}_0^d,\|\alpha\|_1=j} 1 \le \sum_{j=0}^{s} d^j \le (s+1)d^s.$$

Recall that $\varphi \in \mathcal{NN}(2d,1)$, $\psi \in \mathcal{NN}(d(4W+3),4L)$, $P_\alpha \in \mathcal{NN}(9W+s-1, 2(s^2-1)L+s-1)$, $\phi_\alpha \in \mathcal{NN}(40d(s+1)W\lceil \log_2 W\rceil, 13L\lceil \log_2 L\rceil)$ and $\phi_\times \in \mathcal{NN}(9W+1, 2(s+1)L)$. Hence, by our construction, ϕ_0 can be implemented by a neural network with width $49(s+1)^2 d^{s+1} W\lceil \log_2 W\rceil$ and depth $15(s+1)^2 L\lceil \log_2 L\rceil$.

For any $1 \le \|\alpha\|_1 \le s$ and $x,y \in \mathbb{R}^d$, since $\phi_\alpha(x), \phi_\alpha(y), \varphi(x)-\psi(x), \varphi(y)-\psi(y) \in [-1,1]$, by inequalities (16), (19) and (21), we have

$$
\begin{aligned}
&\left|\phi_\times\left(\tfrac{\phi_\alpha(x)}{\alpha!}, P_\alpha(\varphi(x)-\psi(x))\right) - \phi_\times\left(\tfrac{\phi_\alpha(y)}{\alpha!}, P_\alpha(\varphi(y)-\psi(y))\right)\right| \\
\le& 7|\phi_\alpha(x)-\phi_\alpha(y)| + 7|P_\alpha(\varphi(x)-\psi(x)) - P_\alpha(\varphi(y)-\psi(y))| \\
\le& 7\operatorname{Lip}\phi_\alpha \|x-y\|_2 + s7^s\|\varphi(x)-\varphi(y)\|_1 + s7^s\|\psi(x)-\psi(y)\|_1 \\
\le& 140\sqrt{d}LK^{d-2}\delta^{-2}2^{L^2}\|x-y\|_2 + s7^s\sqrt{d}\|x-y\|_2 + 2s7^s\sqrt{d}LK^{-2}\delta^{-2}\|x-y\|_2 \\
\le& \sqrt{d}LK^{2(\beta\vee 1)-2}(1260K^d 2^{L^2} + 19s7^s)\|x-y\|_2.
\end{aligned}
$$

One can check that the bound also holds for $\|\alpha\|_1 = 0$ and $s=0$. Hence,

$$
\begin{aligned}
\operatorname{Lip}\phi_0 \le \operatorname{Lip}\widetilde{\phi}_0 &\le \sum_{\|\alpha\|_1\le s} \sqrt{d}LK^{2(\beta\vee 1)-2}(1260K^d 2^{L^2} + 19s7^s) \\
&\le (s+1)d^{s+1/2}L(WL)^{\sigma(4\beta-4)/d}(1260W^2L^2 2^{L^2} + 19s7^s).
\end{aligned}
$$

We can estimate the error $|h(x)-\phi_0(x)|$ as follows. For any $x \in Q_\theta$, we have $\varphi(x) = x$ and $\psi(x) = \frac{\theta}{K}$. Hence, by the triangle inequality and inequality (22),

$$
\begin{aligned}
&|h(x)-\phi_0(x)| \le |h(x) - \widetilde{\phi}_0(x)| \\
\le& |h(\tfrac{\theta}{K}) - \phi_{\mathbf{0}_d}(x)| + \sum_{1\le\|\alpha\|_1\le s}\left|\frac{\partial^\alpha h(\frac{\theta}{K})}{\alpha!}(x-\tfrac{\theta}{K})^\alpha - \phi_\times\left(\tfrac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x-\tfrac{\theta}{K})\right)\right| + d^{s+\beta/2}K^{-\beta} \\
=:& \sum_{\|\alpha\|_1\le s}\mathcal{E}_\alpha + d^{s+\beta/2}\lfloor (WL)^{2/d}\rfloor^{-\beta}.
\end{aligned}
$$

Using the inequality $|t_1t_2 - \phi_\times(t_3,t_4)| \le |t_1t_2 - t_3t_2| + |t_3t_2 - t_3t_4| + |t_3t_4 - \phi_\times(t_3,t_4)| \le |t_1-t_3| + |t_2-t_4| + |t_3t_4 - \phi_\times(t_3,t_4)|$ for any $t_1,t_2,t_3,t_4 \in [-1,1]$ and the inequalities (17), (18) and (20), we have for $1\le\|\alpha\|_1\le s$,

$$
\begin{aligned}
\mathcal{E}_\alpha \le& \tfrac{1}{\alpha!}\left|\partial^\alpha h(\tfrac{\theta}{K}) - \phi_\alpha(x)\right| + \left|(x-\tfrac{\theta}{K})^\alpha - P_\alpha(x-\tfrac{\theta}{K})\right| \\
&+ \left|\tfrac{\phi_\alpha(x)}{\alpha!}P_\alpha(x-\tfrac{\theta}{K}) - \phi_\times\left(\tfrac{\phi_\alpha(x)}{\alpha!}, P_\alpha(x-\tfrac{\theta}{K})\right)\right| \\
\le& 2(WL)^{-2(s+1)} + 6(s-1)W^{-2(s+1)L} + 6W^{-2(s+1)L} \\
\le& (6s+2)(WL)^{-2(s+1)}.
\end{aligned}
$$

It is easy to check that the bound is also true for $\|\alpha\|_1 = 0$ and $s=0$. Therefore,

$$
\begin{aligned}
|h(x)-\phi_0(x)| &\le \sum_{\|\alpha\|_1\le s}(6s+2)(WL)^{-2(s+1)} + d^{s+\beta/2}\lfloor (WL)^{2/d}\rfloor^{-\beta} \\
&\le (s+1)d^s(6s+2)(WL)^{-2(s+1)} + d^{s+\beta/2}\lfloor (WL)^{2/d}\rfloor^{-\beta} \\
&\le (6s+3)(s+1)d^{s+\beta/2}\lfloor (WL)^{2/d}\rfloor^{-\beta} \\
&=: \mathcal{E},
\end{aligned}
$$

for any $x \in \bigcup_{\theta\in\{0,1,\dots,K-1\}^d} Q_\theta$.

Step 4: Approximation of h on $[0,1]^d$.

Next, we construct a neural network ϕ that uniformly approximates h on $[0,1]^d$. To present the construction, we denote $\mathrm{mid}\,(t_1,t_2,t_3)$ as the function that returns the middle value of three inputs $t_1,t_2,t_3 \in \mathbb{R}$. It is easy to check that

$$\max\{t_1,t_2\} = \frac{1}{2}(\sigma(t_1+t_2) - \sigma(-t_1-t_2) + \sigma(t_1-t_2) + \sigma(t_2-t_1))$$

Thus, $\max\{t_1,t_2,t_3\} = \max\{\max\{t_1,t_2\}, \sigma(t_3)-\sigma(-t_3)\}$ can be implemented by a network with width 6 and depth 2. Similar construction holds for $\min\{t_1,t_2,t_3\}$. Since

$$\mathrm{mid}\,(t_1,t_2,t_3) = \sigma(t_1+t_2+t_3) - \sigma(-t_1-t_2-t_3) - \max\{t_1,t_2,t_3\} - \min\{t_1,t_2,t_3\},$$

it is easy to see $\mathrm{mid}\,(\cdot,\cdot,\cdot) \in \mathcal{NN}(14,2)$.

Recall that $\phi_0 \in \mathcal{NN}(49(s+1)^2 d^{s+1} W\lceil\log_2 W\rceil, 15(s+1)^2 L\lceil\log_2 L\rceil)$. Let $\{e_i\}_{i=1}^d$ be the standard basis in $\mathbb{R}^d$. We inductively define

$$\phi_i(x) := \mathrm{mid}\,(\phi_{i-1}(x-\delta e_i), \phi_{i-1}(x), \phi_{i-1}(x+\delta e_i)) \in [-1,1], \quad i = 1,2,\dots,d.$$

Then $\phi_d \in \mathcal{NN}(49(s+1)^2 3^d d^{s+1} W\lceil\log_2 W\rceil, 15(s+1)^2 L\lceil\log_2 L\rceil + 2d)$. For any $x,x' \in \mathbb{R}^d$, the functions $\phi_{i-1}(\cdot - \delta e_i)$, $\phi_{i-1}(\cdot)$ and $\phi_{i-1}(\cdot + \delta e_i)$ are piece-wise linear on the segment that connecting x and x'. Hence, the Lipschitz constant of these functions on the segment is the maximum absolute value of the slopes of linear parts. Since the middle function does not increase the maximum absolute value of the slopes, it does not increase the Lipschitz constant, which shows that $\mathrm{Lip}\,\phi_d \le \mathrm{Lip}\,\phi_0$.

Denote $Q(K,\delta) := \bigcup_{k=0}^{K-1} [\frac{k}{K}, \frac{k+1}{K} - \delta \cdot 1_{\{k<K-1\}}]$ and define, for $i = 0,1,\dots,d$,

$$E_i := \{(x_1,x_2,\dots,x_d) \in [0,1]^d : x_j \in Q(K,\delta), j > i\},$$

then $E_0 = \bigcup_{\theta\in\{0,1,\dots,K-1\}^d} Q_\theta$ and $E_d = [0,1]^d$. We assert that

$$|\phi_i(x) - h(x)| \le \mathcal{E} + i\delta^{\beta\wedge 1}, \quad \forall x \in E_i, i = 0,1,\dots,d.$$

We prove the assertion by induction. By construction, it is true for $i=0$. Assume the assertion is true for some i, we will prove that it is also holds for $i+1$. For any $x \in E_{i+1}$, at least two of $x - \delta e_{i+1}$, x and $x + \delta e_{i+1}$ are in E_i. Therefore, by assumption and the inequality $|h(x) - h(x \pm \delta e_{i+1})| \le \delta^{\beta\wedge 1}$, at least two of the following inequalities hold

$$\begin{aligned}
|\phi_i(x-\delta e_{i+1}) - h(x)| &\le |\phi_i(x-\delta e_{i+1}) - h(x-\delta e_{i+1})| + \delta^{\beta\wedge 1} \le \mathcal{E} + (i+1)\delta^{\beta\wedge 1}, \\
|\phi_i(x) - h(x)| &\le \mathcal{E} + i\delta^{\beta\wedge 1}, \\
|\phi_i(x+\delta e_{i+1}) - h(x)| &\le |\phi_i(x+\delta e_{i+1}) - h(x+\delta e_{i+1})| + \delta^{\beta\wedge 1} \le \mathcal{E} + (i+1)\delta^{\beta\wedge 1}.
\end{aligned}$$

In other words, at least two of $\phi_i(x-\delta e_{i+1})$, $\phi_i(x)$ and $\phi_i(x+\delta e_{i+1})$ are in the interval $[h(x) - \mathcal{E} - (i+1)\delta^{\beta\wedge 1}, h(x) + \mathcal{E} + (i+1)\delta^{\beta\wedge 1}]$. Hence, their middle value $\phi_{i+1}(x) = \mathrm{mid}\,(\phi_i(x-\delta e_{i+1}), \phi_i(x), \phi_i(x+\delta e_{i+1}))$ must be in the same interval, which means

$$|\phi_{i+1}(x) - h(x)| \le \mathcal{E} + (i+1)\delta^{\beta\wedge 1}.$$

So the assertion is true for $i+1$.

Recall that

$$\delta^{\beta\wedge 1} = \left(\frac{1}{3K^{\beta\vee 1}}\right)^{\beta\wedge 1} = \begin{cases} \frac{1}{3}K^{-\beta} & \beta \geq 1, \\ (3K)^{-\beta} & \beta < 1, \end{cases}$$

and $K = \lfloor (WL)^{2/d} \rfloor$. Since $E_d = [0,1]^d$, let $\phi := \phi_d$, we have

$$\begin{aligned} \|\phi - h\|_{L^\infty([0,1]^d)} &\leq \mathcal{E} + d\delta^{\beta\wedge 1} \\ &\leq (6s+3)(s+1)d^{s+\beta/2}\lfloor (WL)^{2/d} \rfloor^{-\beta} + d\lfloor (WL)^{2/d} \rfloor^{-\beta} \\ &\leq 6(s+1)^2 d^{(s+\beta/2)\vee 1}\lfloor (WL)^{2/d} \rfloor^{-\beta}, \end{aligned}$$

which completes the proof. ■

5.4 Bounding Statistical Error

The technique for bounding the statistical error is rather standard (Anthony and Bartlett, 2009; Shalev-Shwartz and Ben-David, 2014; Mohri et al., 2018). We first show that the statistical error $\mathbb{E}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)]$ of a function class $\mathcal{F}$ can be bounded by the Rademacher complexity, and then bound the Rademacher complexity by Dudley's entropy integral (Dudley, 1967). We restate Lemma 12 here for convenience.

Lemma 12 *Assume* $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq B$*, then we have the following entropy integral bound*

$$\mathbb{E}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)] \leq 8\mathbb{E}_{X_{1:n}} \inf_{0<\delta<B/2} \left(\delta + \frac{3}{\sqrt{n}} \int_\delta^{B/2} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty)}d\epsilon\right),$$

where we denote $\mathcal{F}_{|X_{1:n}} = \{(f(X_1),\dots,f(X_n)) : f \in \mathcal{F}\}$ *for any i.i.d. samples* $X_{1:n} = \{X_i\}_{i=1}^n$ *from* μ *and* $\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty)$ *is the* ϵ*-covering number of* $\mathcal{F}_{|X_{1:n}} \subseteq \mathbb{R}^n$ *with respect to the* $\|\cdot\|_\infty$ *distance.*

Proof Recall that we have n i.i.d. samples $X_{1:n} := \{X_i\}_{i=1}^n$ from μ and $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$. We introduce a ghost data set $X'_{1:n} = \{X'_i\}_{i=1}^n$ drawn i.i.d. from μ, then

$$\begin{aligned} \mathbb{E}_{X_{1:n}}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)] &= \mathbb{E}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{x\sim\mu}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \\ &= \mathbb{E}_{X_{1:n}}\left[\sup_{f\in\mathcal{F}} \mathbb{E}_{X'_{1:n}} \frac{1}{n}\sum_{i=1}^n f(X'_i) - \frac{1}{n}\sum_{i=1}^n f(X_i)\right] \\ &\leq \mathbb{E}_{X_{1:n},X'_{1:n}}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n (f(X'_i) - f(X_i))\right]. \end{aligned}$$

Let $\xi = \{\xi_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher variables independent of $X_{1:n}$ and $X'_{1:n}$. Then, by symmetrization, we can bound $\mathbb{E}_{X_{1:n}}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)]$ by the Rademacher complexity of $\mathcal{F}$:

$$\begin{aligned}\mathbb{E}_{X_{1:n}}[d_{\mathcal{F}}(\mu,\widehat{\mu}_n)] &\leq \mathbb{E}_{X_{1:n},X'_{1:n}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n (f(X'_i)-f(X_i))\right]\\ &= \mathbb{E}_{X_{1:n},X'_{1:n},\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \xi_i(f(X'_i)-f(X_i))\right]\\ &\leq \mathbb{E}_{X_{1:n},X'_{1:n},\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \xi_i f(X'_i)+\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n -\xi_i f(X_i)\right]\\ &= 2\mathbb{E}_{X_{1:n},\xi}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \xi_i f(X_i)\right],\end{aligned}$$

where the last equality is due to the fact that X_i and X'_i have the same distribution and the fact that ξ_i and $-\xi_i$ have the same distribution.

For any $A\subseteq\mathbb{R}^n$, we denote the Rademacher complexity of A by

$$\mathcal{R}(A) := \mathbb{E}_{\xi}\left[\sup_{(a_1,\dots,a_n)\in A}\frac{1}{n}\sum_{i=1}^n \xi_i a_i\right].$$

Then, if we denote $\mathcal{F}_{|X_{1:n}} = \{(f(X_1),\dots,f(X_N)) : f\in\mathcal{F}\}$ for any fixed $X_{1:n}=\{X_i\}_{i=1}^n$, we have shown

$$\mathbb{E}_{X_{1:n}}[d_{\mathcal{F}}(\mu,\widehat{\mu}_n)] \leq 2\mathbb{E}_{X_{1:n}}[\mathcal{R}(\mathcal{F}_{|X_{1:n}})].$$

We define a distance of two vectors $x,y\in\mathbb{R}^n$ by

$$d_2(x,y) := \left(\frac{1}{n}\sum_{i=1}^n (x_i-y_i)^2\right)^{1/2} = \frac{1}{\sqrt{n}}\|x-y\|_2.$$

The corresponding ϵ-covering number of the set $\mathcal{F}_{|X_{1:n}}\subseteq\mathbb{R}^n$ is denoted by $\mathcal{N}(\epsilon,\mathcal{F}_{|X_{1:n}},d_2)$. By chaining technique (see Shalev-Shwartz and Ben-David (2014, Lemma 27.4)), one can show that for any integer $K\geq 0$,

$$\begin{aligned}\mathcal{R}(\mathcal{F}_{|X_{1:n}}) &\leq 2^{-K}B + \frac{6B}{\sqrt{n}}\sum_{k=1}^K 2^{-k}\sqrt{\log\mathcal{N}(2^{-k}B,\mathcal{F}_{|X_{1:n}},d_2)}\\ &\leq 2^{-K}B + \frac{12}{\sqrt{n}}\int_{2^{-K-1}B}^{B/2}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F}_{|X_{1:n}},d_2)}d\epsilon.\end{aligned}$$

Now, for any $\delta\in(0,B/2)$, there exists an integer K such that $2^{-K-2}B\leq\delta<2^{-K-1}B$. Therefore, we have

$$\begin{aligned}\mathbb{E}_{X_{1:n}}[d_{\mathcal{F}}(\mu,\widehat{\mu}_n)] &\leq 2\mathbb{E}_{X_{1:n}}\mathcal{R}(\mathcal{F}_{|X_{1:n}})\\ &\leq 2\mathbb{E}_{X_{1:n}}\inf_{0<\delta<B/2}\left(4\delta+\frac{12}{\sqrt{n}}\int_{\delta}^{B/2}\sqrt{\log\mathcal{N}(\epsilon,\mathcal{F}_{|X_{1:n}},d_2)}d\epsilon\right).\end{aligned}$$

Since $d_2(x,y) \le \|x-y\|_\infty$, we have $\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, d_2) \le \mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty)$, which completes the proof. ∎

When the function class $\mathcal{F}$ has a finite pseudo-dimension, we can further bound the covering number by the pseudo-dimension of $\mathcal{F}$.

Corollary 35 *Assume* $\sup_{f\in\mathcal{F}} \|f\|_\infty \le B$ *and the pseudo-dimension of* $\mathcal{F}$ *is* $\mathrm{Pdim}\,(\mathcal{F}) < \infty$*, then*

$$\mathbb{E}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)] \le CB\sqrt{\frac{\mathrm{Pdim}\,(\mathcal{F})\log n}{n}}$$

for some universal constant $C > 0$*.*

Proof If $n \ge \mathrm{Pdim}\,(\mathcal{F})$, we have the following bound from Anthony and Bartlett (2009, Theorem 12.2),

$$\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty) \le \left(\frac{2eBn}{\epsilon\,\mathrm{Pdim}\,(\mathcal{F})}\right)^{\mathrm{Pdim}\,(\mathcal{F})}.$$

If $n < \mathrm{Pdim}\,(\mathcal{F})$, since $\mathcal{F}_{|X_{1:n}} \subseteq \{x\in\mathbb{R}^n : \|x\|_\infty \le B\}$ can be covered by at most $\lceil\frac{2B}{\epsilon}\rceil^n$ balls with radius ϵ in $\|\cdot\|_\infty$ distance, we always have $\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty) \le \lceil\frac{2B}{\epsilon}\rceil^n$. In any cases,

$$\log\mathcal{N}(\epsilon, \mathcal{F}_{|X_{1:n}}, \|\cdot\|_\infty) \le \mathrm{Pdim}\,(\mathcal{F})\log\frac{2eBn}{\epsilon}.$$

As a consequence,

$$\begin{aligned}
\mathbb{E}[d_{\mathcal{F}}(\mu, \widehat{\mu}_n)] &\le \inf_{0<\delta<B/2}\left(8\delta + 24\sqrt{\frac{\mathrm{Pdim}\,(\mathcal{F})}{n}}\int_\delta^{B/2}\sqrt{\log(2eBn/\epsilon)}d\epsilon\right) \\
&\le \inf_{0<\delta<B/2}\left(8\delta + 12B\sqrt{\frac{\mathrm{Pdim}\,(\mathcal{F})\log(2eBn/\delta)}{n}}\right) \\
&\le CB\sqrt{\frac{\mathrm{Pdim}\,(\mathcal{F})\log n}{n}}
\end{aligned}$$

for some universal constant $C > 0$. ∎

Acknowledgments

The work of Y. Jiao is supported in part by the National Science Foundation of China under Grant 11871474 and by the research fund of KLATASDSMOE. The research of Y. Wang is supported by the HK RGC grant 16308518, the HK Innovation Technology Fund Grant ITS/044/18FX and the Guangdong-Hong Kong-Macao Joint Laboratory for Data Driven Fluid Dynamics and Engineering Applications (Project 2020B1212030001). We thank the editor and reviewers for their feedback on our manuscript.

References

Martin Anthony and Peter L. Bartlett. *Neural network learning: Theoretical foundations*. Cambridge University Press, 2009.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 224–232, 2017.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in GANs. In *7th International Conference on Learning Representations*, 2019.

Bolton Bailey and Matus Telgarsky. Size-noise tradeoffs in generative networks. In *Advances in Neural Information Processing Systems*, volume 31, pages 6490–6500, 2018.

Peter L. Bartlett, Vitaly Maiorov, and Ron Meir. Almost linear VC-dimension bounds for piecewise polynomial networks. *Neural Computation*, 10(8):2159–2173, 1998.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and Pseudodimension bounds for piecewise linear neural networks. *Journal of Machine Learning Research*, 20(63):1–17, 2019.

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

Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. Generating sentences from a continuous space. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 10–21, 2016.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical guarantees of generative adversarial networks for distribution estimation. *arXiv: 2002.05933*, 2020.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals, and Systems*, 2(4):303–314, 1989.

Ingrid Daubechies, Ronald DeVore, Simon Foucart, Boris Hanin, and Guergana Petrova. Nonlinear approximation and (deep) relu networks. *Constructive Approximation*, pages 1–46, 2021.

Richard M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290–330, 1967.

Richard M. Dudley. *Real analysis and probability*. Cambridge University Press, second edition, 2018.

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.

Lawrence C. Evans and Ronald F. Garzepy. *Measure theory and fine properties of functions*. Routledge, 2018.

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

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3):707–738, 2015.

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, pages 2672–2680, 2014.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25): 723–773, 2012.

Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved training of Wasserstein GANs. In *Advances in Neural Information Processing Systems*, volume 30, pages 5767–5777, 2017.

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

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *6th International Conference on Learning Representations*, 2018.

Naveen Kodali, Jacob Abernethy, James Hays, and Zsolt Kira. On convergence and stability of GANs. *arXiv: 1705.07215*, 2017.

Andrey N. Kolmogorov and Vladimir M. Tikhomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations: Series 2*, 17:277–364, 1961.

Holden Lee, Rong Ge, Tengyu Ma, Andrej Risteski, and Sanjeev Arora. On the ability of neural nets to express distributions. In *Proceedings of the 30th Conference on Learning Theory*, volume 65, pages 1271–1296, 2017.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pages 1718–1727, 2015.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021.

Jianfeng Lu, Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Yulong Lu and Jianfeng Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. In *Advances in Neural Information Processing Systems*, volume 33, pages 3094–3105, 2020.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT Press, second edition, 2018.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. In *6th International Conference on Learning Representations*, 2018.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, pages 429–443, 1997.

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.

Dmytro Perekrestenko, Stephan Müller, and Helmut Bölcskei. Constructive universal high-dimensional distribution generation through deep ReLU networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 7610–7619, 2020.

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

Henning Petzka, Asja Fischer, and Denis Lukovnikov. On the regularization of Wasserstein GANs. In *6th International Conference on Learning Representations*, 2018.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *4th International Conference on Learning Representations*, 2016.

Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1060–1069, 2016.

Nicolas Schreuder, Victor-Emmanuel Brunel, and Arnak Dalalyan. Statistical guarantees for generative models without domination. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, volume 132, pages 1051–1071, 2021.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

Zuowei Shen, Haizhao Yang, and Shijun Zhang. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. In *Advances in Neural Information Processing Systems*, volume 31, pages 10246–10257, 2018.

Hoang Thanh-Tung, Truyen Tran, and Svetha Venkatesh. Improving generalization and stability of generative adversarial networks. In *7th International Conference on Learning Representations*, 2019.

Ananya Uppal, Shashank Singh, and Barnabás Póczos. Nonparametric density estimation & convergence rates for GANs under Besov IPM losses. In *Advances in Neural Information Processing Systems*, volume 32, pages 9086–9097, 2019.

Xiang Wei, Boqing Gong, Zixia Liu, Wei Lu, and Liqiang Wang. Improving the improved training of Wasserstein GANs: A consistency term and its dual effect. In *6th International Conference on Learning Representations*, 2018.

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.

Dmitry Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In *Proceedings of the 31st Conference on Learning Theory*, volume 75, pages 639–649, 2018.

Dmitry Yarotsky and Anton Zhevnerchuk. The phase diagram of approximation rates for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 13005–13015, 2020.

Xin Yi, Ekta Walia, and Paul S. Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 58:101552, 2019.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeoff in GANs. In *6th International Conference on Learning Representations*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2242–2251, 2017.