arXiv:1811.03179, 2018

# HOW WELL GENERATIVE ADVERSARIAL NETWORKS LEARN DISTRIBUTIONS[1]

Tengyuan Liang
University of Chicago, Booth School of Business

We study in this paper the rates of convergence for learning distributions implicitly with the adversarial framework and Generative Adversarial Networks (GANs), which subsume Wasserstein, Sobolev and MMD GANs as special cases. We study a wide range of parametric and nonparametric target distributions, under a host of objective evaluation metrics. On the nonparametric end, we investigate the minimax optimal rates of the distribution estimation under the adversarial framework. On the parametric end, we establish a theory for general neural network classes (including deep leaky ReLU networks as a special case), that characterizes the interplay on the choice of generator and discriminator pair. We investigate how to obtain a good statistical guarantee for GANs through the lens of regularization. We discover and isolate a new notion of regularization, called the generator-discriminator-pair regularization, that sheds light on the advantage of GANs compared to classical parametric and nonparametric approaches for density estimation. We develop novel oracle inequalities as the main tools for analyzing GANs, which is of independent theoretical interest.

Keywords: Generative Adversarial Networks, Implicit Density Estimation, Simulated Method of Moments, Oracle Inequality, Neural Network Learning, Convergence Rates, Pair Regularization.

## 1. Introduction

Generative models such as Generative Adversarial Networks (GANs) [Arjovsky et al., 2017, Dziugaite et al., 2015, Goodfellow et al., 2014, Li et al., 2015] have recently stood out as an important unsupervised method for learning and efficient sampling from a complex target data distribution. Despite the celebrated empirical success, many questions on the theory [Liang, 2017, Liu and Chaudhuri, 2018, Liu et al., 2017, Singh et al., 2018] and mechanism of GANs [Arora and Zhang, 2017, Arora et al., 2017, Daskalakis et al., 2017, Mescheder et al., 2017] remain to be elucidated.

At the population level, one general formulation of the adversarial framework [Arjovsky et al., 2017, Dziugaite et al., 2015, Li et al., 2015, Liu et al., 2017, Mroueh et al., 2017] considers the following minimax problem,

$$\min_{\mu\in\mathcal{D}_G}\max_{f\in\mathcal{F}_D}\mathop{\mathbb{E}}_{Y\sim\mu} f(Y)-\mathop{\mathbb{E}}_{X\sim\nu} f(X).$$

In plain language, given a target probability distribution $\nu$, one seeks for a simulated probability distribution $\mu$ from a *generator class* $\mathcal{D}_G$, such that it minimizes the loss incurred by the best test function inside a *discriminator class* $\mathcal{F}_D$. In practice, both *the generator and the discriminator classes* are represented by deep neural networks. To be concrete, $\mathcal{D}_G$ quantifies the transformed implicit distributions realized by a neural network pushing-forward random input units (for example, multi-dimensional uniform or Gaussian distributions), and $\mathcal{F}_D$ represents functions realizable by a certain neural network architecture. In practice, one only has access to finite samples of the target distribution $\nu$. Let us denote $\widehat{\nu}^n$ as the empirical distribution based on $n$ i.i.d. samples from $\nu$. Given finite data samples, the adversarial framework solves the following empirical problem

$$\widehat{\mu}\in\mathop{\arg\min}_{\mu\in\mathcal{D}_G}\max_{f\in\mathcal{F}_D}\mathop{\mathbb{E}}_{Y\sim\mu} f(Y)-\mathop{\mathbb{E}}_{X\sim\widehat{\nu}^n} f(X). \tag{1.1}$$

The first expectation over $Y\sim\mu$ can be calculated efficiently using simulations with arbitrary acuracy, since one can simulate samples from $\mu$ directly by pushing-forward random inputs.

tengyuan.liang@chicagobooth.edu

[1]Liang gratefully acknowledges support from the George C. Tiao Fellowship. Liang wishes to thank Maxim Raginsky for pointing out relevant literature on simulated method of moments. This paper was previously posted as "How well can generative adversarial networks learn densities: A nonparametric view" available on arXiv:1712.08244, 2017. The previous version is no longer intended for publication.

In machine learning, the adversarial loss is also referred to as the Integral Probability Metric (IPM). Define the IPM for a symmetric function class $\mathcal{F}$ as

$$d_{\mathcal{F}}(\mu,\nu) := \sup_{f\in\mathcal{F}} \mathop{\mathbb{E}}_{Y\sim\mu} f(Y) - \mathop{\mathbb{E}}_{X\sim\nu} f(X) = \sup_{f\in\mathcal{F}} \int_{\Omega} f(d\mu - d\nu).$$

By choosing different $\mathcal{F}$'s, the adversarial framework can express a host of commonly-used metrics. To name a few, (1) Wasserstein GAN [Arjovsky et al., 2017]: $\mathcal{F}$ consists of Lipschitz-1 functions, and the IPM is the Wasserstein-1 metric $d_W(\cdot,\cdot)$. (2) Maximum Mean Discrepancy (MMD) GAN [Arbel et al., 2018, Dziugaite et al., 2015, Li et al., 2015]: let $\mathcal{H}$ be a Reproducing Kernel Hilbert Space (RKHS), and $\mathcal{F}$ consists of functions with bounded RKHS norm $\mathcal{F} = \{f \in \mathcal{H} \mid \|f\|_{\mathcal{H}} \leq 1\}$. (3) Sobolev GAN [Mroueh et al., 2017]: $\mathcal{F}$ is the Sobolev space with certain smoothness. (4) Total Variation metric $d_{TV}(\cdot,\cdot)$: $\mathcal{F}$ represents all functions bounded by 1. We refer the readers to Liu et al. [2017] for other related formulations of GANs. Conceptually, the discriminator function class induces a collection of "moment conditions" accessing the closeness of probability distributions, in the viewpoint of Generalized Method of Moments (GMM) [Hansen, 1982].

In the statistical literature, explicit distribution estimation, or density estimation, has been a fundamental topic in nonparametric statistics [Nemirovski, 2000, Tsybakov, 2009, Wassermann, 2006] and in parametric models [Brown, 1986]. In the parametric case, learning density simply reduces to parameter estimation. In the parametric case, learning density reduces to parameter estimation. In the nonparametric case, the optimal minimax rates have been established for a wide range of density function classes quantified by the smoothness property [Stone, 1982]. However, it is not practical to simulate samples efficiently from these minimax optimal explicit density estimators, especially for multi-dimensional data.

The econometrics literature has explored an alternative implicit distribution estimation approach [Back and Brown, 1993, Imbens et al., 1995] using the Method of (Simulated) Moments (MSM) [McFadden, 1989, Pakes and Pollard, 1989]. Such an MSM approach turns out to be a special case of GANs in formulation (1.1): MSM implicitly estimates the target distribution with a simulated distribution (from a certain parametric class $\mu_{\widehat{\theta}} \in \mathcal{D}_G$), by matching the moment conditions (induced by functions $f \in \mathcal{F}_D$) to the empirical distribution $\widehat{\nu}^n$. In the classic method of moments with finite $K$ moment conditions $\{f_k(x), k \in [K]\}$, $\mathcal{F}_D$ consists of functions satisfying a quadratic constraint $\{f(x) = \sum_{k\in[K]} \omega_k f_k(x) \mid \omega^\top \mathbf{W}^{-1}\omega \leq 1, \omega \in \mathbb{R}^K\}$ with a given symmetric positive-definite weight matrix $\mathbf{W} \in \mathbb{R}^{K\times K}$. In this language, the adversarial framework in (1.1) extends the MSM where the moment conditions are induced by a rich class of functions $\mathcal{F}_D$. More recently, Athey et al. [2019] conducted a systematic empirical study to learn the distribution of real economic datasets using Wasserstein GAN, suggesting the effectiveness of such an implicit distribution estimation approach in modern practice.

The current paper studies both the *Adversarial Framework* and *Generative Adversarial Networks* for implicitly learning distributions from a statistical vantage point. As discussed in the paragraphs before, the adversarial framework and GANs is fundamental to statistics, machine learning and econometrics. We intend to answer the following questions:

1. How well do GANs learn a wide range of target distributions (both in nonparametric and parametric cases), under a collection of objective evaluation metrics?
2. How to utilize the adversarial framework to achieve better theoretical guarantee through the lens of regularization?

We discover and isolate a new notion of regularization, which we call *generator-discriminator-pair regularization*, that provides rigorous guidance on balancing the complexities of the generator and discriminator. We emphasize that several curious features of this pair regularization appear to be new to the literature. As a unified theme in the theory, we develop oracle inequalities for analyzing the generative adversarial framework, which could be of independent interest for further theoretical research on GANs.

### 1.1. *Contributions and Organization*

The paper is organized into two main parts: the *Adversarial Framework* and the *Generative Adversarial Networks*.

### Roadmap of Results and Overall Goal

Our overall goal is to provide a complete statistical treatment of the adversarial framework and GANs' mechanism under two important settings: first, the generator and discriminator being the nonparametric classes in the adversarial framework; second, the generator and discriminant being the class parametrized by neural networks as in GANs. We summarize in Table I a roadmap of results for readers to navigate. In Table I, we reserve the following symbols for characteristics of the theorems.

(1.2)
$(\mathcal{G}\dagger)$ : generator $\mathcal{G}$ mis-specified for $\nu$, $\nu \notin \mathcal{G}$
$(\mathcal{F}\ddagger)$ : discriminator $\mathcal{F}$ mis-specified for the metric, $d_{\mathcal{F}} \neq d_{eval}$
$(m*)$ : the result accounts for finite $m$ samples of the generator

The main *technical contributions* are the development of the oracle inequalities for analyzing GANs, and the formulation of the novel generator-discriminator-pair regularization.

TABLE I

ROADMAP OF RESULTS. THE SYMBOLS ARE DEFINED IN (1.2): $(\mathcal{G}\dagger)$ AND $(\mathcal{F}\ddagger)$ TO DENOTE THE MIS-SPECIFICATION FOR THE GENERATOR CLASS AND THE DISCRIMINATOR CLASS RESPECTIVELY, AND $(m*)$ TO INDICATE THE DEPENDENCE ON THE NUMBER OF GENERATOR SAMPLES.

| Goal | Evaluation Metric | Results | | Generator Class $\mathcal{G}$ | Discriminator Class $\mathcal{F}$ | Property |
|---|---|---|---|---|---|---|
| Adversarial Framework (nonparametric) | $d_{\mathcal{F}}$ | Sobolev GAN | minimax optimal (Thm. 1) | Sobolev $W^{\alpha}$ | Sobolev $W^{\beta}$ | |
| | | MMD GAN | upper bound (Thm. 2) | smooth subspace in RKHS | RKHS $\mathcal{H}$ | |
| | | | oracle results (Thm. 3) | any | Sobolev $W^{\beta}$ | $\mathcal{G}\dagger$ |
| Generative Adversarial Networks (parametric) | $d_{TV}$ | leaky-ReLU GANs | upper bound (Thm. 6) | leaky-ReLU | leaky-ReLU | $\mathcal{F}\ddagger$, $m*$ |
| | $d_{TV}, d_{JS}, d_H$ | any GANs | oracle results (Thm. 4 & 5) | neural networks | neural networks | $\mathcal{G}\dagger$, $\mathcal{F}\ddagger$, $m*$ |
| | $d_W$ | Lipschitz GANs | oracle results (Cor. 1) | Lipschitz neural networks | Lipschitz neural networks | $\mathcal{G}\dagger$, $\mathcal{F}\ddagger$, $m*$ |

### Adversarial Framework

One key component of GANs is the adversarial framework: evaluating the performance of the learned distribution by the adversarial loss. Under the adversarial loss $d_{\mathcal{F}_D}(\cdot,\cdot)$ (IPM induced by the specified discriminator $\mathcal{F}_D$), we study the minimax optimal rates for the target density $\nu$ based on $n$-i.i.d. samples. We formulate such adversarial framework following the classic nonparametric literature by considering a wide range of nonparametric target densities $\mu$ and discriminator classes $\mathcal{F}_D$ quantified by their smoothness property. Using a simple oracle inequality, we extend to the case when the generator class $\mathcal{D}_G$ mis-specifies the target density $\nu$, for the procedure

$$\widehat{\mu}_n = \arg\min_{\mu\in\mathcal{D}_G} \max_{f\in\mathcal{F}_D} \mathop{\mathbb{E}}_{Y\sim\mu} f(Y) - \mathop{\mathbb{E}}_{X\sim\widehat{\nu}^n} f(X). \tag{1.3}$$

This procedure is in a general form, not specific to neural networks.

*Our contributions* are: (1) we characterize the minimax optimal rates of the adversarial framework for learning densities for classic nonparametric distribution families, and how to achieve them; (2) we show how the structure of target $\nu$ and that of the class $\mathcal{F}$ affect the minimax rate explicitly, and under what cases fast rates are possible.

**Generative Adversarial Networks**

In practice, GANs are parametrized by neural networks. Built on top of the adversarial framework, we directly analyze the rates for the following parametrized GANs estimator with the generator network $\mathcal{G}$ (parametrized by $\theta$) and discriminator network $\mathcal{F}$ (parametrized by $\omega$)

$$\widehat{\theta}_{m,n} \in \arg\min_{\theta:g_\theta\in\mathcal{G}} \max_{\omega:f_\omega\in\mathcal{F}} \left\{\widehat{\mathbb{E}}_m f_\omega(g_\theta(Z)) - \widehat{\mathbb{E}}_n f_\omega(X)\right\}. \tag{1.4}$$

Here $m$ and $n$ denote the number of the generator samples and target distribution samples. We remark on two key facts about this procedure. First, the distribution estimator is implicit, namely the probability distribution of the random variable push-forwarded by the transformation map $Z \mapsto g_{\widehat{\theta}_{m,n}}(Z)$ with $g_{\widehat{\theta}_{m,n}} : \mathcal{Z} \to \mathcal{X}$. Theory for the implicit distribution estimator (such as GANs) is far less developed in the literature. Second, the objective evaluation metrics we investigate include Jensen-Shannon divergence $d_{JS}$, Total Variation $d_{TV}$, Wasserstein $d_W$ and Hellinger $d_H$ distances, which are mis-specified by the generator $\mathcal{F}$.

*Our contributions* are: (1) we study the parametric rates for the implicit distribution estimator (distribution of $g_{\widehat{\theta}_{m,n}}(Z)$) for the target $\nu$, when both $\mathcal{G}$ and $\mathcal{F}$ are parametrized by general neural networks; (2) We rigorously formulate the complex trade-offs on the choices of the generator $\mathcal{G}$ and the discriminator $\mathcal{F}$ as a *pair regularization*. We evaluate how this new notion of regularization affects the rates for GANs; (3) As a direct application of the general theoretical framework, we showcase how to identify good $\mathcal{G}$ and $\mathcal{F}$ pairs to obtain fast parametric rates using two examples: learning distributions realizable by deep leaky ReLU networks, and learning multivariate Gaussian distributions with two-layer networks. In both cases, the upper rates we obtain provide optimal sample complexity (up to logarithmic factors).

Finally, the paper is organized as follows. Section 2 consists of main nonparametric results and the adversarial framework. Section 3 contains the main parametric results for GAN with neural network generator and discriminator classes, where we introduce the new notion of pair regularization. Further discussions on the generator-discriminator-pair regularization is deferred to Section 4. The main proofs are collected in Section 5, with remaining proofs and supporting lemmas deferred to Appendix A.

### 1.2. *Preliminaries*

We now introduce the preliminary background and notations. In this paper, unless otherwise specified, we restrict the input space to be $\Omega = [0,1]^d$ with dimension $d$ and the base measure to be the Lebesgue measure on $\Omega$. We use $\mu, \nu, \pi$ to denote the distributions, and also reserve $p_\mu(x), p_\nu(x), p_\pi(x)$ for the corresponding density functions w.r.t the Lebesgue measure (the Radon-Nikodym derivative). In other words, for ease of notation we use $\int_\Omega f(x)p_\mu(x)dx = \int_\Omega f d\mu$ to denote the same integration. $\|f\|_q := \left(\int_\Omega |f(x)|^q dx\right)^{1/q}$ denotes the $\ell_q$-norm, for $1 \le q \le \infty$. For a vector $w$, $\|w\|_q$ denotes the vector $\ell_q$-norm. We use the asymptotic notation $A(n) \precsim n^\alpha$, if $\overline{\lim}_{n\to\infty} \frac{\log A(n)}{\log n} \le \alpha$, holding other parameters fixed, similarly $A(n) \succsim n^\alpha$ if $\underline{\lim}_{n\to\infty} \frac{\log A(n)}{\log n} \ge \alpha$. We call $A(n) \asymp n^\alpha$ when $A(n) \precsim n^\alpha$ and $A(n) \succsim n^\alpha$. $[K] := \{1, \ldots, K\}$ refers to the index set, for any $K \in \mathbb{N}_{>0}$. For a vector or a multi-index (possibly infinite dimensional), the subscript $i$ denotes the $i$-th component and $|\gamma| = \sum_i \gamma_i$.

Next, we introduce the function spaces. Let $d$ denotes the dimension. For a multi-index $\gamma \in \mathbb{N}^d_{\ge 0}$, we use $D^{(\gamma)}f$ to denote the $\gamma$-weak derivative for a function $f : \Omega \to \mathbb{R}$. For example, for infinitely smooth $f \in C^\infty(\Omega)$, $D^{(\gamma)}f$ takes the form $D^{(\gamma)}f = \partial^{|\gamma|}f/\partial x_1^{\gamma_1} \ldots \partial x_d^{\gamma_d}$.

Definition 1 (Sobolev space: $\alpha \in \mathbb{N}_{>0}$) *Given a smoothness parameter $\alpha \in \mathbb{N}_{\ge 0}$, $1 \le q \le \infty$, and a radius $r \in \mathbb{R}_{\ge 0}$, the Sobolev space $W^{\alpha,q}(r)$ is defined as*

$$W^{\alpha,q}(r) := \left\{ f \in \Omega \to \mathbb{R} : \Big(\sum_{|\gamma|\le\alpha} \|D^{(\gamma)}f\|_q^q\Big)^{1/q} \le r \right\},$$

*where $\gamma$ is a multi-index and $D^{(\gamma)}$ denotes the $\gamma$-weak derivative.*

For the case $q = 2$, we abbreviate the $W^{\alpha,q=2}(r)$ as $W^{\alpha}(r)$.

To extend the results to distribution on manifolds, we further consider general Reproducing Kernel Hilbert Spaces (RKHS) $\mathcal{H} \subset L^2_\pi$ (with $\pi$ as the base measure) endowed with RKHS norm $\|\cdot\|_{\mathcal{H}}$, and the corresponding positive semidefinite kernel $K(\cdot,\cdot) : \Omega \times \Omega \to \mathbb{R}$. By the Mercer's theorem, one can characterize this RKHS via the following integral operator $\mathcal{T}_\pi : L^2_\pi \to \mathcal{H}$.

DEFINITION 2 (Integral operator of RKHS) *Define the integral operator $\mathcal{T}_\pi : L^2_\pi \to \mathcal{H}$,*

$$\mathcal{T}_\pi f(z) = \int_\Omega K(z,\cdot) f(\cdot) d\pi(\cdot),$$

*and denote the eigenfunctions of this operator by $\psi_i$ and the associated eigenvalues by $t_i, i \in \mathbb{N}_{\geq 0}$ (sorted in non-increasing order), with*

$$\mathcal{T}_\pi \psi_i = t_i \psi_i, \text{ and } \int_\Omega \psi_i \psi_j d\pi = \delta_{ij}.$$

To measure the complexity of functions from a learning theory perspective, we employ the following notion of combinatorial dimension for real-valued function is credited to Pollard [1990]. We will employ this combinatorial dimension as a complexity measure in deriving rates for GANs.

DEFINITION 3 (Pseudo-dimension) *Let $\mathcal{F} = \{f : \Omega \to \mathbb{R}\}$ be a class of functions. The pseudo-dimension of $\mathcal{F}$, denoted by $\mathrm{Pdim}(\mathcal{F})$, is the largest integer $m$ such that $\exists (X_i, y_i) \in \Omega \times \mathbb{R}, i \in [m]$, for any $(b_1, \ldots, b_m) \in \{-1, 1\}^m$ there exists $f \in \mathcal{F}$ such that $\mathrm{sign}(f(X_i) - y_i) = b_i, \forall i \in [m]$.*

Finally, for two functions $f : \mathbb{R}^d \to \mathbb{R}$ and $g : \mathbb{R}^p \to \mathbb{R}^d$, we denote $f \circ g$ to be the composition $f(g(x))$. We use the following notation for the composition of function classes

$$\mathcal{F} \circ \mathcal{G} := \{f \circ g \mid f \in \mathcal{F}, g \in \mathcal{G}\}. \tag{1.5}$$

## 2. THE ADVERSARIAL FRAMEWORK

We start with investigating the adversarial framework including the Wasserstein, Sobolev, and MMD GANs. Recall that the adversarial framework employed by GANs proposes to evaluate the accuracy of learning densities via the adversarial loss specified by the discriminator class. The goal of this section is to study the minimax optimal rates for learning a wide range of distributions, on a host of evaluation metric defined by the adversarial framework. Through the lens of nonparametric statistics, we answer how the structure of the distribution and the choice of the evaluation metric affect the optimal rates, and when fast rates are possible.

### 2.1. *Minimax Optimal Rates*

THEOREM 1 (Minimax optimal rates, Sobolev) *Let $\Omega = [0,1]^d$. Consider the target density $p_\nu(x) \in \mathcal{G} = W^\alpha(r)$ (w.r.t. the Lebesgue measure) in the Sobolev space with smoothness $\alpha \in \mathbb{N}_{\geq 0}$ for some constant $r > 0$, and the evaluation metric that is induced by $\mathcal{F} = W^\beta(1)$ the Sobolev space with smoothness $\beta \in \mathbb{N}_{\geq 0}$. The minimax optimal rate is*

$$\inf_{\widetilde{\nu}_n} \sup_{\nu \in \mathcal{G}} \mathbb{E}\, d_{\mathcal{F}}(\nu, \widetilde{\nu}_n) \asymp n^{-\frac{\alpha+\beta}{2\alpha+d}} \vee n^{-\frac{1}{2}},$$

*where $\widetilde{\nu}_n$ is any estimator for $\nu$ based on $n$ i.i.d. drawn samples $X_1, X_2, \ldots X_n \sim \nu$.*

REMARK 1 The above establishes the minimax optimal rate for Sobolev GAN ($\beta = 1$ for Wasserstein GAN as a special case), with explicit dependence on the smoothness of the density $\alpha$ and that of the evaluation metric $\beta$. First, note there is an interesting transition at $\beta = d/2$ (without depending on $\alpha$): above it the rate is parametric $n^{-1/2}$, and below it the rate is nonparametric. Second, to avoid the curse of dimensionality

in the rates, one needs the sum of smoothness to be proportional to the dimension, i.e. $\alpha + \beta = \Theta(d)$. Note when $\beta$ is large, the rate is indeed faster, however under a weaker evaluation metric. How to choose a good discriminator $\mathcal{F}$ in GANs with provable guarantee under strong evaluation metrics such as $d_{TV}$ will be answered in Theorems 4-6.

REMARK 2 (Relations to the literature) The above theorem is an improvement to an earlier draft [Liang, 2017] of this paper, which was the first to formalize nonparametric estimation under the adversarial framework. Admittedly, the improvement for the upper bound is in one line of the original argument, specifically Eqn. (5.1). The minimax lower bound of $n^{-\frac{\alpha+\beta}{2\alpha+d}}$ was first established in this paper (the earlier draft, Liang [2017], page 18-19). In this version we also provide a formal construction for the lower bound of $n^{-\frac{1}{2}}$. We acknowledge an improvement of the upper bound in Liang [2017] was also carried out in a follow-up work [Singh, Uppal, Li, Li, Zaheer, and Póczos, 2018] (see the discussion therein).

One can generalize the above theorem to more general RKHS. The motivation is to accommodate target distributions supported on image manifolds, with similarity better measured by non-linear kernels. It is useful to derive the explicit dependence on the intrinsic dimension of the manifold and the kernel, rather than the ambient dimension $d$. The Sobolev space considered in Thm. 1 can be viewed as a special RKHS when the smoothness index is large enough [De Vito et al., 2019]. In addition, the generalization will enable us to provide theoretical rates for MMD GAN [Arbel et al., 2018, Dziugaite et al., 2015, Li et al., 2015].

In the next theorem, we assume that for all target density $\nu \in \mathcal{G}$ of interest and all $i \in \mathbb{N}_{>0}$, there exists a universal constant on the variance of eigenfunctions in Def. 2,

$$\mathop{\mathbb{E}}_{X\sim\nu} \psi_i(X)^2 \leq C. \tag{2.1}$$

THEOREM 2 (MMD rates, RKHS) *Consider a RKHS $\mathcal{H} \in L^2_\pi$ with base measure $\pi$. Assume that the eigenvalues of the integral operator $\mathcal{T}_\pi$ decay as $t_i \asymp i^{-\kappa}$ for all $i \in \mathbb{N}_{\geq 0}$, with parameter $\kappa \in \mathbb{R}_{>0}$. Consider the evaluation metric $\mathcal{F} = \{f \mid \|f\|_{\mathcal{H}} \leq 1\}$, and the target distribution $\nu$, whose Radon-Nikodym derivative $\frac{d\nu}{d\pi}$ w.r.t the base measure $\pi$ lies in a smooth subset $\mathcal{G} = \{\nu \mid \|\mathcal{T}_\pi^{-(\lambda-1)/2} \frac{d\nu}{d\pi}\|_{\mathcal{H}} \leq r\}$ with smoothness parameter $\lambda \in \mathbb{R}_{>0}$ (for some fixed radius $r > 0$). Under the assumption (2.1), we have*

$$\sup_{\nu\in\mathcal{G}} \mathbb{E}\, d_{\mathcal{F}}\left(\nu, \widetilde{\nu}_n\right) \precsim n^{-\frac{(\lambda+1)\kappa}{2\lambda\kappa+2}} \vee n^{-\frac{1}{2}}.$$

REMARK 3 Remark that the above corollary works with general base measure $\pi$ and domain $\Omega$. Here the target (Radon-Nikodym derivative $\frac{d\nu}{d\pi}$) lies in a subset of the RKHS, with $\lambda$ quantifies its smoothness: the high frequency component decays sufficiently fast. This is a standard formulation studied in the RKHS literature, see Caponnetto and De Vito [2007]. The parameter $\kappa$ describes the intrinsic dimension of the integral operator. When $\kappa > 1$, the intrinsic dimension (trace of $\mathcal{T}_\pi$) is bounded as $\mathrm{Tr}(\mathcal{T}_\pi) = \sum_{i\geq 1} i^{-\kappa} \leq C$, therefore the upper bound reads the parametric rate $n^{-\frac{(\lambda+1)\kappa}{2\lambda\kappa+2}} \vee n^{-\frac{1}{2}} = n^{-\frac{1}{2}}$. When $\kappa < 1$, to obtain $\mathbb{E}\, d_{\mathcal{F}}\left(\nu, \widetilde{\nu}_n\right) \leq \epsilon$, the sample complexity scales

$$n = \epsilon^{2+\frac{2}{\lambda+1}\left(\frac{1}{\kappa}-1\right)}.$$

Therefore the curse of dimensionality only appears in the "effective dimension," described by $1/\kappa - 1$.

The Sobolev space $W^\beta$ can be regarded as a special RKHS with $\kappa = \frac{2\beta}{d}$. In such a case, the generator class $W^\alpha$ can be thought as subset $\mathcal{G}$ in Thm. 2 with $\lambda = \frac{\alpha}{\beta}$. Plug in $\lambda = \frac{\alpha}{\beta}$ and $\kappa = \frac{2\beta}{d}$, the bound in Thm. 2 reduces to $n^{-\frac{\alpha+\beta}{2\alpha+d}}$ which agrees with Thm. 1. Therefore the lower bound in Thm. 1 suggests that the rate for MMD GAN is also sharp, for a particular subclass.

### 2.2. *Oracle Inequality and Regularization*

In this section, we use a simple oracle inequality to show that when the generator class $\mathcal{D}_G$ — typically represented by neural networks — is mis-specified for the target distribution $\nu$, one can still derive oracle results based on the adversarial framework.

Let us recall the notations. Denote $\mathcal{D}_G$ to be class of distributions represented by the generator, and $\mathcal{F}_D$ to be the class of functions realized by the discriminator

$$\mu_n = \arg\min_{\mu\in\mathcal{D}_G} \max_{f\in\mathcal{F}_D} \left\{ \mathop{\mathbb{E}}_{Y\sim\mu} f(Y) - \mathop{\mathbb{E}}_{X\sim\nu_n} f(X) \right\}. \tag{2.2}$$

where $\nu_n$ is some estimate of the density based on $n$ i.i.d. drawn samples from the target distribution $\nu$.

The goal in this section to extend our adversarial framework to obtain upper rates for (2.2). In addition, the oracle inequalities (Lemma 1 and 2) developed will be crucial for model mis-specification, which makes the results of practical relevance.

THEOREM 3 (Mis-specification: nonparametric) *Let $\mathcal{D}_G$ be any generator class. Consider the discriminator metric to be induced by $\mathcal{F}_D = W^\beta(1)$. Consider the target density $p_\nu(x) \in W^\alpha(r)$. With the empirical distribution $\widehat{\nu}^n := \frac{1}{n}\sum_{i=1}^n \delta_{X_i}$ as the plug-in, the GAN estimator*

$$\widehat{\mu}_n \in \arg\min_{\mu\in\mathcal{D}_G} \max_{f\in\mathcal{F}_D} \left\{ \int_\Omega f d\mu - \int_\Omega f d\widehat{\nu}^n \right\},$$

*learns the target density with rate*

$$\mathbb{E}\, d_{\mathcal{F}_D}(\widehat{\mu}_n, \nu) \le \min_{\mu\in\mathcal{D}_G} d_{\mathcal{F}_D}(\mu,\nu) + n^{-\frac{\beta}{d}} \vee \frac{\log n}{\sqrt{n}}.$$

*In contrast, there exists a regularized empirical distribution $\widetilde{\nu}^n$ as the plug-in*

$$\widetilde{\mu}_n \in \arg\min_{\mu\in\mathcal{D}_G} \max_{f\in\mathcal{F}_D} \left\{ \int_\Omega f d\mu - \int_\Omega f d\widetilde{\nu}^n \right\},$$

*where a faster rate is attainable*

$$\mathbb{E}\, d_{\mathcal{F}_D}(\widetilde{\mu}_n, \nu) \le \min_{\mu\in\mathcal{D}_G} d_{\mathcal{F}_D}(\mu,\nu) + n^{-\frac{\alpha+\beta}{2\alpha+d}} \vee \frac{1}{\sqrt{n}}.$$

The proof of the above theorem is based on the following simple oracle inequality Lemma 1. Later, we will generalize the oracle inequality (see Lemma 2) to establish rates when both the generator and discriminator are neural networks, and when one only has finite $m$-samples from the generator. Curiously, a generalization of the oracle inequality gives rise to a curious notion of pair regularization, which we will study in Section 3.

LEMMA 1 (Simple oracle inequality) *Under the condition that $\mathcal{F}_D$ is symmetric class, i.e., $\mathcal{F}_D = -\mathcal{F}_D$, the GAN estimator in (2.2) satisfies*

$$d_{\mathcal{F}_D}(\nu, \mu_n) \le \min_{\mu\in D_G} d_{\mathcal{F}_D}(\mu,\nu) + 2 d_{\mathcal{F}_D}(\nu,\nu_n),$$

*where we refer the first term as the approximation error, and second as the stochastic error.*

REMARK 4 (Regularization) Observe that the rates satisfy $n^{-\frac{\alpha+\beta}{2\alpha+d}} \vee n^{-1/2} \precsim n^{-\frac{\beta}{d}} \vee n^{-1/2}\log n$. Namely, the regularized empirical density as the plug-in for GANs attains a better upper bound. In a high level, the regularized empirical distribution $\widetilde{\nu}^n$ filters out high frequency component of the empirical distribution to enforce regularization. We mention that to obtain a implementable algorithm for the smoothed/regularized empirical distribution $\widetilde{\nu}^n$ in Thm. 3, one may use the following in practice

$$\frac{d\widetilde{\nu}^n}{dx} = \frac{1}{nh_n}\sum_{i\in[n]} K\left(\frac{x-x_i}{h_n}\right),$$

with specific choices of the kernel $K$ and bandwidth $h_n$ as in the nonparametric literature. When using the Gaussian kernel, this so-called "instance noise" technique [Arjovsky and Bottou, 2017, Mescheder et al., 2018, Sønderby et al., 2016] is used in GAN training: each time when evaluating the stochastic gradients for generator and discriminator, sample a mini-batch of data and then perturb them by a Gaussian. Statistically, one may view this data augmentation (or stability to data perturbation) as a form of regularization [Yu, 2013], to prevent the generator from memorizing the empirical data and learning a too complex model. We will show in Section 3 that, specific choice of generator and discriminator pair can also serve the goal of regularization in the parametric regime, in a curious way.

## 3. GENERATIVE ADVERSARIAL NETWORKS

In this section, we consider when both the generator and discriminator are neural networks, and derive rates applicable to GANs used in practice. To be specific, let $\mathcal{F} = \{f_\omega(x) : \mathbb{R}^d \to \mathbb{R}\}$ be the discriminator functions realized by a neural network with parameter $\omega$ describing the weights of the network. Let $\mathcal{G} = \{g_\theta(z) : \mathbb{R}^d \to \mathbb{R}^d\}$ be the generator neural network transformation with weights parameter $\theta$. We keep this parametrization in an abstract form for now as we will first state our general theorems before applying it to specific cases. Consider $Z \sim \pi$ as the random input distribution with distribution $\pi$, and the target distribution $X \sim \nu$. Denote $\mu_\theta$ as the probability distribution of $g_\theta(Z)$. Consider the parametrized GAN estimator used in practice

$$\widehat{\theta}_{m,n} \in \arg\min_{\theta: g_\theta \in \mathcal{G}} \max_{\omega: f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega(g_\theta(Z)) - \widehat{\mathbb{E}}_n f_\omega(X) \right\}, \tag{3.1}$$

where $m$ and $n$ denote the number of the generator samples and target distribution samples.

Let us state the goal of the current section, and connections to the adversarial framework established. So far, we have derived the optimal rates for nonparametric densities under strong evaluation metric such as Wasserstein ($\beta = 1$) or total variation distance ($\beta = 0$). The curse of dimensionality in sample complexity is inevitable unless the density class of interest is sufficiently structured (smooth). Two questions are raised naturally. First, for the structured parametric densities such as the ones parametrized by the generator networks in GANs, are fast parametric rates attainable? Second, can one obtain fast rates under the strong evaluation metric via discriminator networks in GANs, which is mis-specified and differs from the evaluation metric? We will answer both questions, directly for GANs estimator (3.1).

### 3.1. *Generalized Oracle Inequality and Parametric Rate*

First, we will generalize the oracle results to GANs estimator $\widehat{\theta}_{m,n}$ (3.1). Then we will show that the oracle approach, when applied to neural networks as in Thm. 4, sheds light on the choice of *generator-discriminator-pair* as regularization.

LEMMA 2 (Generalized oracle inequality) *Consider the GAN estimator $\widehat{\theta}_{m,n}$ defined in (3.1). Recall the composition in Def. (1.5). Under the condition that $\mathcal{F}$ and $\mathcal{F} \circ \mathcal{G}$ are symmetric, the following oracle inequality holds for any $\mu_\theta$ with $g_\theta \in \mathcal{G}$,*

$$d_{\mathcal{F}}\left(\mu_{\widehat{\theta}_{m,n}}, \nu\right) \le d_{\mathcal{F}}(\mu_\theta, \nu) + 2d_{\mathcal{F}}\left(\widehat{\nu}^n, \nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m, \mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m, \pi).$$

*Here for any distribution $\mu$, we use $\widehat{\mu}^n$ to denote the empirical distribution with $n$ i.i.d. samples from $\mu$.*

The innovative aspects of the above lemma are two-fold. Firstly, the Lemma provides upper bound on the *implicit* distribution estimator $\mu_{\widehat{\theta}_{m,n}}$ (distribution of the random variable $g_{\widehat{\theta}_{m,n}}(Z)$), without knowing the explicit form of the density function in general. Note that we do have direct sampling mechanisms by transforming the random variable $Z$, which is a computational advantage. Secondly, we make explicit the dependence on the number of generator samples $m$, in addition to the number of target samples $n$. The role and complexity of the generator network is made explicit in the bound. It is clear that when $m \to \infty$, the current lemma reduces to Lemma 1. This Lemma made explicit the choice of simulated samples $m$ relative to the real-data samples $n$, and how the classes $\mathcal{G}$ and $\mathcal{F}$ affect the trade-offs.

Next, we apply Lemma 2 to establish parametric rates for densities realized by neural networks, in the following Thm. 4 and 6 (with their corollaries). We emphasize again here that GANs only use a mis-specified discriminator $\mathcal{F}$ parametrized by neural networks with limited capacity. And $d_{\mathcal{F}}$ is *different* from the the objective evaluation metrics such as $d_{TV}, d_H$.

THEOREM 4 (GANs upper rate on KL: parametric) *Consider GANs estimator*

$$\widehat{\theta}_{m,n} \in \arg\min_{\theta: g_\theta \in \mathcal{G}} \max_{\omega: f_\omega \in \mathcal{F}, \|f_\omega\|_\infty \le B} \left\{ \widehat{\mathbb{E}}_m f_\omega(g_\theta(Z)) - \widehat{\mathbb{E}}_n f_\omega(X) \right\}. \tag{3.2}$$

*where $B > 0$ is some absolute constant, $m$ and $n$ denote the number of the generator samples and target distribution samples. Recall the pseudo-dimension defined in Def. 3. Then for total variation distance, and Kullback-Leibler divergence, we have*

$$\begin{aligned}
&\mathbb{E}\, d_{TV}^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \le \frac{1}{4}\left[\mathbb{E}\, d_{KL}\left(\nu||\mu_{\widehat{\theta}_{m,n}}\right) + \mathbb{E}\, d_{KL}\left(\mu_{\widehat{\theta}_{m,n}}||\nu\right)\right] \\
&\le \frac{1}{2}\sup_{\theta}\inf_{\omega}\left\|\log\frac{p_\nu}{p_{\mu_\theta}} - f_\omega\right\|_\infty + \frac{B}{4\sqrt{2}}\inf_{\theta}\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty^{1/2} \\
&\quad + C\cdot\sqrt{Pdim(\mathcal{F})\left(\frac{\log m}{m}\vee\frac{\log n}{n}\right)} \vee \sqrt{Pdim(\mathcal{F}\circ\mathcal{G})\frac{\log m}{m}},
\end{aligned} \tag{3.3}$$

*where $C > 0$ is some universal constant independent of $Pdim(\mathcal{F})$, $Pdim(\mathcal{F}\circ\mathcal{G})$ and $m, n$.*

The upper bound in the above theorem on the Jensen-Shannon/Kullback Leibler divergence (and TV distance) consists of three parts: the approximation errors $A_1(\mathcal{F},\mathcal{G},\nu)$, $A_2(\mathcal{G},\nu)$ and the stochastic error $S(\mathcal{F},\mathcal{G},n,m)$,

$$\begin{aligned}
A_1(\mathcal{F},\mathcal{G},\nu) &:= \frac{1}{2}\sup_{\theta}\inf_{\omega}\left\|\log\frac{p_\nu}{p_{\mu_\theta}} - f_\omega\right\|_\infty \\
A_2(\mathcal{G},\nu) &:= \frac{B}{4\sqrt{2}}\inf_{\theta}\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty^{1/2} \\
S_{n,m}(\mathcal{F},\mathcal{G}) &:= \sqrt{\mathrm{Pdim}(\mathcal{F})\left(\frac{\log m}{m}\vee\frac{\log n}{n}\right)} \vee \sqrt{\mathrm{Pdim}(\mathcal{F}\circ\mathcal{G})\frac{\log m}{m}}.
\end{aligned} \tag{3.4}$$

We emphasize that the term $A_1(\mathcal{F},\mathcal{G},\nu)$ is in the $\sup_\theta\inf_\omega$ form, which is crucial and differs from the adversarial idea with the form $\inf_\theta\sup_\omega$. In English, $A_1(\mathcal{F},\mathcal{G},\nu)$ describes how the best discriminator function $f_\omega$ that can express the class of density ratios $p_{\mu_\theta}/p_\nu$, $A_2(\mathcal{G},\nu)$ reflects the the expressiveness of the generator class, and $S_{n,m}(\mathcal{F},\mathcal{G})$ describes the statistical complexity of both the generator and discriminator. In the next section, we will elaborate on the interplay among the two approximation error terms $A_1(\mathcal{F},\mathcal{G},\nu), A_2(\mathcal{G},\nu)$, and the stochastic error term $S_{n,m}(\mathcal{F},\mathcal{G})$.

REMARK 5 To obtain non-trivial rates, the above theorem requires $\mu_\theta$ and $\nu$ to be absolutely continuous, for all $\theta$ of interest. However, this is not essential, as similar results hold qualitatively the same for the non-absolutely continuous case, based on the Hellinger distance. As shown in the next theorem, $-1 \le \frac{\sqrt{p_\nu}-\sqrt{p_{\mu_\theta}}}{\sqrt{p_\nu}+\sqrt{p_{\mu_\theta}}} \le 1$ is well-defined even for non-absolutely continuous distributions $\mu_\theta$ and $\nu$.

THEOREM 5 (GANs upper rate on Hellinger: parametric) *Consider the same GANs estimator $\widehat{\theta}_{m,n}$ as in Thm. 4. The for the Hellinger distance,*

$$d_H(\mu,\nu) := \left(\int\left(\sqrt{p_\mu(x)} - \sqrt{p_\nu(x)}\right)^2 dx\right)^{1/2}, \tag{3.5}$$

*we have*

$$\begin{aligned}
&\mathbb{E}\, d_{TV}^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \le \mathbb{E}\, d_H^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \\
&\le 2\sup_{\theta}\inf_{\omega}\left\|\frac{\sqrt{p_\nu}-\sqrt{p_{\mu_\theta}}}{\sqrt{p_\nu}+\sqrt{p_{\mu_\theta}}} - f_\omega\right\|_\infty + 2B\inf_{\theta}\left\|\frac{\sqrt{p_\nu}-\sqrt{p_{\mu_\theta}}}{\sqrt{p_\nu}+\sqrt{p_{\mu_\theta}}}\right\|_\infty \\
&\quad + C\cdot\sqrt{Pdim(\mathcal{F})\left(\frac{\log m}{m}\vee\frac{\log n}{n}\right)} \vee \sqrt{Pdim(\mathcal{F}\circ\mathcal{G})\frac{\log m}{m}},
\end{aligned} \tag{3.6}$$

*where $C > 0$ is some universal constant.*

Finally, as a corollary of Thm. 4, one can establish similar results for the Wasserstein distance.

COROLLARY 1 *Recall the definitions in (3.4). Assume that $\mathcal{F}$ is with Lipschitz constant $L_{\mathcal{F}}$ and $\mathcal{G}$ with $L_{\mathcal{G}}$. Then for either (1) $Z \sim N(0, I_d)$, or (2) $Z, X$ lie in $[0,1]^d$, we have*

$$\mathbb{E}\, d_W^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \leq C_1 \cdot A_1(\mathcal{F}, \mathcal{G}, \nu) + C_2 \cdot A_2(\mathcal{G}, \nu) + C_3 \cdot S_{n,m}(\mathcal{F}, \mathcal{G})$$

*where $C_1, C_2, C_3 > 0$ are some constants independent of $Pdim(\mathcal{F})$, $Pdim(\mathcal{F} \circ \mathcal{G})$ and $m, n$, but depend on $L_{\mathcal{F}}, L_{\mathcal{G}}$.*

### 3.2. *Generator-Discriminator-Pair Regularization*

In this section, we investigate the new pair regularization, and its trade-offs presented in Thm. 4. One key fact about regularization in GAN is that both the generator and discriminator are choices of "tuning parameters," for users to specify. Therefore, the trade-offs is more complex. For a target distribution of interest, we use the following two thought experiments to explain the intricacies on the interplay between the generator-discriminator-pair.

1. For a fixed generator class $\mathcal{G}$, when the discriminator class $\mathcal{F}$ becomes more complex, it will be easier for the discriminator to tell apart good and bad generators in the TV sense (w.r.t. the target distribution). However, the stochastic error becomes larger as one is learning from a large discriminator model in GANs. This is reflected in the upper bounds obtained in Thm. 4 and 5, shown along the blue dashed arrow direction in Fig. 1.
2. For a fixed discriminator class $\mathcal{F}$, as the generator $\mathcal{G}$ becomes richer, it is capable of expressing densities that are closer to the target distribution. However, at the same time it introduces difficulty for two reasons. First, the generator may create densities that are far away from the target in the TV sense, but being indistinguishable to the discriminator. Second, the stochastic error becomes worse as one is learning from a larger generator model. This is shown by the red dashed arrow direction in Fig. 1.

In general, regularization using the generator-discriminator-pair is more subtle than the conventional bias-variance (or approximation-stochastic error) trade-offs. We visualize such trade-offs in Fig. 1, with $A_1(\mathcal{F}, \mathcal{G}, \nu)$, $A_2(\mathcal{G}, \nu)$ and $S_{n,m}(\mathcal{F}, \mathcal{G})$ defined in (3.4). Here, the tuning parameters lie in a two dimensional domain, rather than in an one dimensional index. For a fixed target $\nu$, as $(\mathcal{G}, \mathcal{F})$ both become richer, $A_2(\mathcal{G}, \mu)$ decreases, $S_{n,m}(\mathcal{F}, \mathcal{G})$ increases, but $A_1(\mathcal{F}, \mathcal{G}, \nu)$ may increase, decrease or stay unchanged. On one hand, one can eliminate some $(\mathcal{G}, \mathcal{F})$ pairs due to notions of dominance on the two dimensional domain. The simple U-shaped picture for bias-variance trade-off no longer exists. On the other hand, by stepping into the two dimensional tuning domain, there are more choices for tuning pairs that potentially give rise to better rates, which we will showcase in Thm. 6.

The following corollary concerns $A_1(\mathcal{F}, \mathcal{G}, \nu)$ and $A_2(\mathcal{G}, \nu)$ through choosing the generator-discriminator-pair, as a step towards understanding the new notion of pair regularization for GANs.

COROLLARY 2 (Choice of generator and discriminator) *Consider the target distribution class $\nu \in \mathcal{D}_R$, and the generator distribution class $\mu_\theta \in \mathcal{D}_G$. With the discriminator chosen as*

$$\mathcal{F}_D := \{\log(p_\nu) - \log(p_{\mu_\theta}) \mid \textit{for all } \nu \in \mathcal{D}_R,\ \mu_\theta \in \mathcal{D}_G\},$$

*then*

$$A_1(\mathcal{F}, \mathcal{G}, \nu) = 0. \tag{3.7}$$

*In addition, if the generator is well-specified in the sense $\mathcal{D}_G \supseteq \mathcal{D}_R$, then*

$$A_2(\mathcal{G}, \nu) = 0. \tag{3.8}$$

*And (3.7) and (3.8) altogether imply $\mathbb{E}\, d_{TV}^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \precsim S_{n,m}(\mathcal{F}, \mathcal{G})$.*

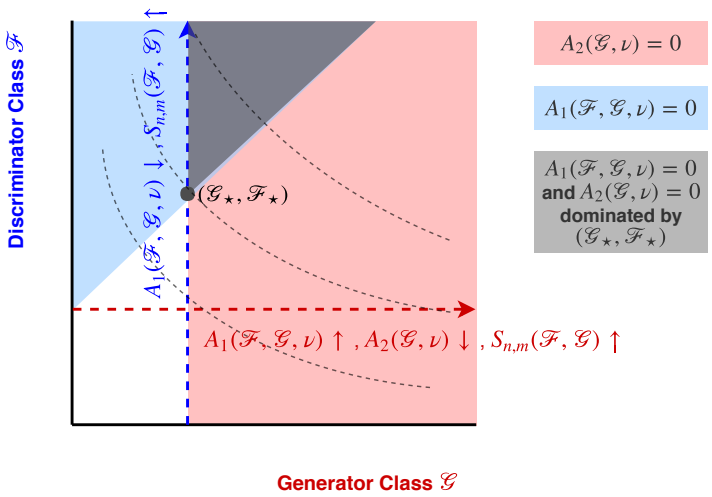


FIGURE 1.— Pair regularization diagram on how well GANs learn densities in TV distance, when tuning with generator $\mathcal{G}$ and discriminator $\mathcal{F}$ pair. The diagram is illustrated based on upper bounds on TV distance, namely $A_1(\mathcal{F},\mathcal{G},\nu)+A_2(\nu,\mathcal{G})+S_{n,m}(\mathcal{F},\mathcal{G})$ in Thm. 4. The red shaded region corresponds to $A_2(\mathcal{G},\nu)=0$ and the blue shaded region is $A_1(\mathcal{F},\mathcal{G},\nu)=0$. The grey dashed line corresponds to the indifference curve for the statistical error $S_{n,m}(\mathcal{F},\mathcal{G})$. One can see that the choice $(\mathcal{G}_\star,\mathcal{F}_\star)$ dominates the other choices in the grey shaded area, and the other choice on the same grey dashed line.

REMARK 6 (Pair regularization and diagram) Let us illustrate the above corollary using Fig. 1. Eqn. (3.7) corresponds to the blue shaded region in the diagram, Eqn. (3.8) represents the red shaded region, and the intersection is highlighted by the grey shaded region. At the intersection, the approximation error $A(\mathcal{F},\mathcal{G},\nu)$ is zero, so all pairs are dominated by the choice $(\mathcal{G}_\star,\mathcal{F}_\star)$ (as other pairs have a larger variance $S_{n,m}(\mathcal{F},\mathcal{G})$). In addition, we argue that $(\mathcal{G}_\star,\mathcal{F}_\star)$ is also the best solution along the indifference curve for $S_{n,m}(\mathcal{F},\mathcal{G})$, denoted by the grey dashed line. To see this, moving $(\mathcal{G}_\star,\mathcal{F}_\star)$ towards the northwest direction on the indifference curve away from $(\mathcal{G}_\star,\mathcal{F}_\star)$, $A_1, S_{m,n}$ stay unchanged, but $A_2(\mathcal{G}_\star,\nu)\le A_2(\mathcal{G}',\nu)$. Moving $(\mathcal{G}',\mathcal{F}')$ towards the southeast direction, $A_2, S_{m,n}$ stay the same, but $A_1(\mathcal{G}_\star,\mathcal{F}_\star,\nu)\le A_1(\mathcal{G}',\mathcal{F}',\nu)$. Similarly, one can argue that all pairs above the indifference curve is dominated by $(\mathcal{G}_\star,\mathcal{F}_\star)$.

We acknowledge that the diagram is illustrated using an upper bound on the TV distance, however, qualitatively, similar phenomenon extends to $\mathbb{E}\, d_{TV}^2\left(\nu,\mu_{\widehat{\theta}_{m,n}}\right)$ and $\mathbb{E}\, d_H^2\left(\nu,\mu_{\widehat{\theta}_{m,n}}\right)$ (see the first paragraph in Section 3.2). We defer the further discussion on the pair regularization versus classic regularization to Section 4.

### 3.3. *Applications: Leaky ReLU Networks*

We showcase how to apply our theory and regularization insight to GANs used in practice in this section. We consider two special cases of leaky ReLU generator and discriminator, to make explicit the rates for estimating parametric densities. The main tools are Thm. 4 and the pair regularization. The goal of this section is to show for good choice of $(\mathcal{G},\mathcal{F})$, near optimal sample complexity is attainable. Admittedly, we do not aim to identify the optimal pair of $(\mathcal{G}_\star,\mathcal{F}_\star)$ over the entire two dimensional turning domain. In fact, such optimization can be hard. The reason is, to characterize the implicit density of $g_{\widehat{\theta}_{m,n}}(Z)$ given by neural networks transformations, and how it approximates general nonparametric target density $\nu$ is a challenging future work outside the statistical goal of the current paper.

Let's introduce the neural networks parameter space. The *generator* $x=g_\theta(z):\mathbb{R}^d\to\mathbb{R}^d$ is parametrized

by a Multi-Layer Perceptron (MLP):

$$\begin{aligned} h_0 &= z, \\ h_l &= \sigma_a(W_l h_{l-1} + b_l),\ 0 < l < L \\ x &= W_L h_{L-1} + b_L, \end{aligned}$$

where $h_l$ denotes the output of hidden units, and $x$ is the transformed final output of the MLP. Here the activation is leaky ReLU

$$\sigma_a(t) = \max\{t, at\},\ \text{for some fixed } 0 < a \leq 1. \tag{3.9}$$

Denote the parameter space for the generator weights as

$$\theta \in \Theta(d, L) := \{\theta = (W_l \in \mathbb{R}^{d\times d}, b_l \in \mathbb{R}^d, 1 \leq l \leq L) \mid \text{rank}(W_l) = d, \forall 1 \leq l \leq L\}.$$

We require the $W_l$ to be full rank so that the generator transformation $g_\theta$ is invertible. One can verify, when the input distribution $Z \sim U([0,1]^d)$ is uniform, the class of densities realizable by $g_\theta(Z)$, for $\theta \in \Theta(d, L)$ has the following closed form,

$$\log\left(p_{\mu_\theta}(x)\right) = c_1 \sum_{l=1}^{L-1} \sum_{i=1}^{d} \mathbb{1}_{m_{li}(x)\geq 0} + c_0(\theta), \tag{3.10}$$

with some proper choice of $c_1, c_0(\theta)$. Here $m_{li}(x)$ is the function computed by the $i$-th hidden unit in the $l$-th layer of a certain MLP[1], with dual leaky ReLU activation (defined in next paragraph) and weights properly chosen as a function of $\theta$. For details, see derivation (5.7) and (5.9). Remark that from the closed form expression, as the depth grows (as a function of $n$), the generator is capable of expressing increasingly complex distributions. Clearly from the expression, one can see that for any $\theta, \theta' \in \Theta(d, L)$, $\mu_\theta$ and $\mu_{\theta'}$ are *absolutely continuous*.

The *discriminator* $f_\omega(x) : \mathbb{R}^d \to \mathbb{R}$ is parametrized by a feedforward neural network with activation functions include dual leaky ReLU activation

$$\sigma_a^\star(t) := \min\{t, at\},\ \text{for } a \geq 1, \tag{3.11}$$

and threshold activation $\sigma_\infty^\star(t) := \mathbb{1}_{t\leq 0}$. The structure a feedforward network is that hidden units are grouped in a sequence of $L$ layers (the depth of the network), where a node is in layer $1 \leq l \leq L$, if it has a predecessor in layer $l-1$ and no predecessor in any layer $l' \geq l$. Computation of the final output unit proceeds layer-by-layer: at any layer $l < L$, each hidden unit $u$ receives an input in the form of a linear combination $\widetilde{x}_u' w_u + b_u$, and then outputs $\sigma_a(\widetilde{x}_u' w_u + b_u)$, where the vector $\widetilde{x}_u$ collects the output of all the units with a directed edge into $u$ (i.e., from prior layers). $\omega$ denotes all the weights in such feedforward network.

THEOREM 6 (Leaky-ReLU generator and discriminator, uniform as input) *Consider a multi-layer perceptron generator $g_\theta : \mathbb{R}^d \to \mathbb{R}^d$, $\theta \in \Theta(d, L)$ with depth $L$ and width $d$, using leaky ReLU $\sigma_a(\cdot)$ activation (3.9) with any $0 < a \leq 1$. Consider the class of realizable densities, i.e., $X \sim \nu$ enjoys the same distribution as $g_{\theta_*}(Z)$ with some $\theta_* \in \Theta(d, L)$ and $Z \sim U([0,1]^d)$. Choose the discriminator $f_\omega : \mathbb{R}^d \to \mathbb{R}$ to be a feedforward neural network (architecture shown in Fig. 2) with depth $L+2$, using dual leaky ReLU $\sigma_{1/a}^\star(\cdot)$ (3.11) and threshold activations (only at the final layer), with parameter $\omega \in \Omega(d, L)$ defined in (5.11).*

*Then, the GAN estimator $\mu_{\widehat{\theta}_{m,n}}$ defined in (3.2), satisfies the following parametric rates for the total variation distance,*

$$\mathbb{E}\, d_{TV}^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \precsim \sqrt{d^2 L^2 \log(dL)\left(\frac{\log m}{m} \vee \frac{\log n}{n}\right)}.$$

[1]The architecture and weights depend on the generator network $g_\theta$, with depth $L$ and $d$ hidden units in each layer.

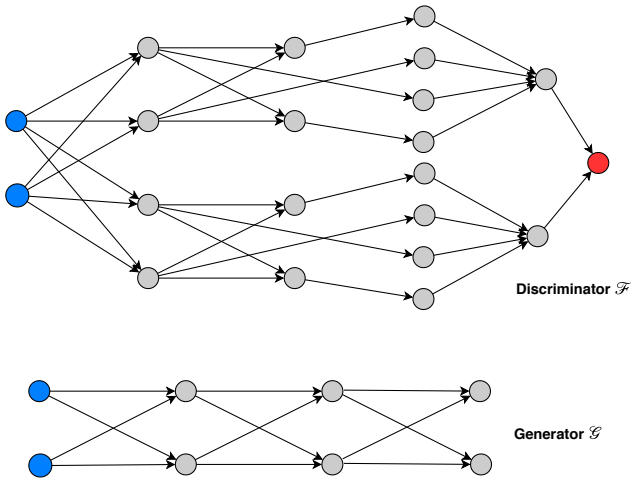


FIGURE 2.— Illustration of discriminator $\mathcal{F}$ (feed-forward network) and generator $\mathcal{G}$ (multi-layer perceptron) in Theorem 6, for $L = 3$.

REMARK 7 (Relations to literature) The above theorem is built on top of Thm. 4 and Cor. 2. Remark here we use the neural networks' architecture as pair regularization. Remark that in our setting, we can allow for *very deep* ReLU neural network with $L \precsim \sqrt{n \wedge m / \log(n \vee m)}$, with generator's width being as small as the dimension $d$.

Investigations on the parametric rates for GANs have been considered in Bai, Ma, and Risteski [2018], based on spectral norm-based capacity controls as regularization of networks, i.e. $\forall l \in [L], \|W_l\|_{\text{op}}, \|W_l^{-1}\|_{\text{op}} \leq C$. The approach they are taking is to establish multiplicative equivalence on $d_{\mathcal{F}}(\mu, \nu) \asymp d_W(\mu, \nu)$ for $\mu, \nu \in \mathcal{G}$ restricted to the generator class.

In contrast, we make use of the oracle inequality approach developed in an early version of the current paper [Liang, 2017], and the notion of pair regularization. We study through the angle of pseudo-dimensions, without requiring that the spectral radius of each $W_l, W_l^{-1}$ is bounded. This has two advantages. First, the generator class can express a wider range of densities, as we only require that $W_l$ has full rank. Second, we make explicit the dependence of the depth of the neural networks $L$ in the rate. In addition, we were able to get a better polynomial dependence on both the dimension $d$ and the depth $L$, in the error.

Finally, as a sanity check, we show that GANs can also achieve the correct dimension dependence in sample complexity ($n = O(d^2 \log d)$) when estimating multivariate Gaussian with unknown mean and covariance (where from information-theoretic lower bound we need at least $n = O(d^2)$ samples). This is to showcase that with the power of pair regularization, GANs can obtain provable guarantee in classic statistical realms.

COROLLARY 3 (Multivariate Gaussian estimation, isotropic Gaussian as input) *Consider $\nu \sim N(b_*, \Sigma_*)$ to be a multivariate Gaussian in $\mathbb{R}^d$. Consider a linear generator (neural network with no hidden layer) with input distribution $N(0, I_p)$ ($p \geq d$), and the discriminator to be a one hidden layer neural network with quadratic activation $\sigma(t) = t^2$, the GAN estimator $\mu_{\widehat{\theta}_{m,n}}$ defined in (3.2), satisfies the following rates,*

$$\mathbb{E}\, d_{TV}^2\left(\nu, \mu_{\widehat{\theta}_{m,n}}\right) \precsim \sqrt{\frac{d^2 \log d}{n} + \frac{(pd + d^2)\log(p + d)}{m}}.$$

## 4. CONCLUSION AND DISCUSSION

We further discuss on the following question: even overlooking computation, what is the advantage of GANs compared to classic nonparametric density estimation, and the classic parametric models. We use the diagram as in Fig. 1 to point out some conclusions (based on Thm. 1-5 obtained in this paper) and some conjectures.

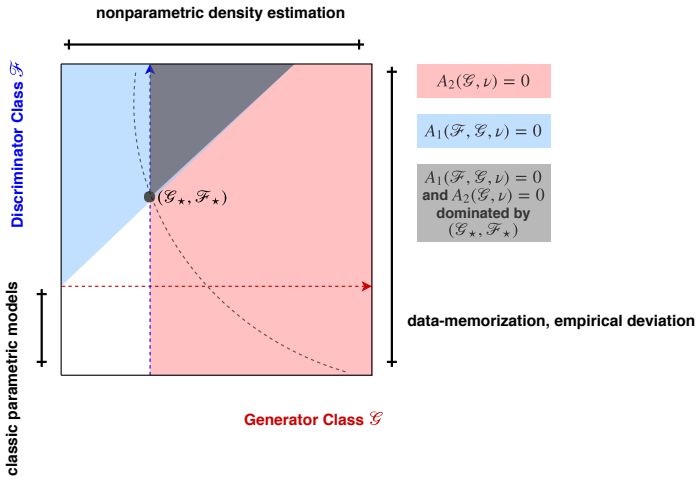


FIGURE 3.— Diagram for generator-discriminator-pair regularization.

1. Classic parametric models: can be viewed as the left interval (along y-axis) in Fig. 3, where the generator class $\mathcal{G}$ is simple and limited. The discriminator can be viewed as assessing how well we are estimating the finite parameters, which relates to how well we are learning densities in the parametric class. More advanced discriminator won't help. The pair regularization effectively reduces to one dimensional tuning on the discriminator: what is a good loss function on the parameter set.
2. Classic nonparametric density estimation: can be viewed as the top interval (along x-axis) in Fig. 3. Here the $d_{\mathcal{F}} = d_{TV}$, and by tuning the generator class $\mathcal{G}$ (using sieves, kernels, etc.), one can achieve the optimal rates when the target density lies in a certain nonparametric class. The minimax theory for the adversarial framework (Thm. 1) informs us, when the target is truly nonparametric, tuning with the generator class is optimal: there is no theoretical gain in utilizing the generator-discriminator-pair to tune. Though, with simpler evaluation metrics, one can obtain faster rates, shown in Thm. 1.
3. Empirical density, or data memorization: can be viewed as the right interval (along y-axis) in Fig. 3. Here the generator class is flexible enough to memorize the training data, and one should try to avoid this by means of regularization (Thm. 3).
4. For a certain target density $\nu$ (in between parametric and nonparametric for many realistic cases), tuning with the generator and discriminator pair $(\mathcal{G}, \mathcal{F})$ as illustrated in Fig. 3 could potentially do better than both that in the parametric and nonparametric tuning domains. We *conjecture* that *tuning with the generator-discriminator-pair* $(\mathcal{G}_\star, \mathcal{F}_\star)$ could potentially explain the empirical success of GANs on the statistical side, as one has the choice of flexibly tuning the generator and discriminator pair with deep neural networks, in the two dimensional domain balancing $A_1(\mathcal{F}, \mathcal{G}, \nu), A_2(\mathcal{G}, \nu), S_{n,m}(\mathcal{F}, \mathcal{G})$ simultaneously.

Admittedly, to fully understand such phenomenon in pair-regularization, one may need to re-think the class of distributions of interest, and what constitutes "low complexity/structured" class rather than the "smoothness" used in the nonparametric literature. In this paper, we only consider the statistical problem of how well GANs learn density, assuming the optimization, say (3.2), can be done to sufficient accuracy. Admittedly, computation of GANs is a considerably harder question [Arbel, Sutherland, Bińkowski, and Gretton, 2018, Daskalakis, Ilyas, Syrgkanis, and Zeng, 2017, Liang and Stokes, 2018, Lucic, Kurach, Michalski, Gelly, and Bousquet, 2017, Mescheder, Nowozin, and Geiger, 2017], which we leave as future work.

## 5. PROOF OF MAIN RESULTS

### 5.1. *Oracle Inequalities*

We now develop the oracle inequalities, which are the main innovative tool for analyzing the rates for GANs. We remark that these are deterministic inequalities that hold generally, which could be of independent interest for further research on GANs.

PROOF OF LEMMA 1: For any $\mu \in \mu_G$, we know that due to the optimality of GAN in (2.2),

$$d_{\mathcal{F}_D}(\mu, \nu_n) - d_{\mathcal{F}_D}(\mu_n, \nu_n) \geq 0.$$

Due to the triangle inequality of IPM, we have

$$\begin{aligned} d_{\mathcal{F}_D}(\mu_n, \nu) &\leq d_{\mathcal{F}_D}(\mu_n, \nu_n) + d_{\mathcal{F}_D}(\nu_n, \nu) \\ &\leq d_{\mathcal{F}_D}(\mu, \nu_n) + d_{\mathcal{F}_D}(\nu_n, \nu) \quad \text{(optimality of } \nu_n) \\ &\leq d_{\mathcal{F}_D}(\mu, \nu) + d_{\mathcal{F}_D}(\nu, \nu_n) + d_{\mathcal{F}_D}(\nu_n, \nu). \end{aligned}$$

Now take $\mu = \arg\min_{\mu \in \mu_G} d_{\mathcal{F}_D}(\mu, \nu)$, and recall that $\mathcal{F}_D$ is symmetric around 0, we have

$$d_{\mathcal{F}_D}(\mu_n, \nu) \leq \min_{\mu \in \mu_G} d_{\mathcal{F}_D}(\mu, \nu) + 2 d_{\mathcal{F}_D}(\nu, \nu_n).$$

*Q.E.D.*

PROOF OF LEMMA 2: For ease of notation, we abbreviate $\widehat{\theta}_{m,n}$ as $\widehat{\theta}$ in this proof when there is no confusion. Recall GANs estimator (3.1), and the definition of $d_{\mathcal{F}}\left(\mu_{\widehat{\theta}_{m,n}}, \nu\right)$, we have

$$\begin{aligned} & d_{\mathcal{F}}\left(\mu_{\widehat{\theta}_{m,n}}, \nu\right) = \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_{\widehat{\theta}}(Z) - \mathbb{E}\, f_\omega(X) \right\} \\ & \leq \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_{\widehat{\theta}}(Z) - \widehat{\mathbb{E}}_n f_\omega(X) \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_n f_\omega(X) - \mathbb{E}\, f_\omega(X) \right\} \\ & \leq \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}}(Z) - \widehat{\mathbb{E}}_n f_\omega(X) \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_{\widehat{\theta}}(Z) - \widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}}(Z) \right\} \\ & \qquad + \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_n f_\omega(X) - \mathbb{E}\, f_\omega(X) \right\}. \end{aligned}$$

Here the first inequality we insert the quantity $\widehat{\mathbb{E}}_n f_\omega(X)$, and the second we insert the quantity $\widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}}(Z)$ to the first term. For any $\theta$ such that $g_\theta \in \mathcal{G}$, we recall the optimality condition of GANs estimator

$$\sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}_{m,n}}(Z) - \widehat{\mathbb{E}}_n f_\omega(X) \right\} \leq \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega \circ g_\theta(Z) - \widehat{\mathbb{E}}_n f_\omega(X) \right\},$$

then one can proceed with (for any $\theta$ with $g_\theta \in \mathcal{G}$)

$$
\begin{aligned}
& d_{\mathcal{F}}\left(\mu_{\widehat{\theta}_{m,n}}, \nu\right) \\
& \leq \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega \circ g_\theta(Z) - \widehat{\mathbb{E}}_n f_\omega(X) \right\} \quad \text{(optimality of } \widehat{\theta}_{m,n}\text{)} \\
& \quad + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_{\widehat{\theta}}(Z) - \widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}}(Z) \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_n f_\omega(X) - \mathbb{E}\, f_\omega(X) \right\} \\
& \leq \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega \circ g_\theta(Z) - \mathbb{E}\, f_\omega \circ g_\theta(Z) \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_\theta(Z) - \mathbb{E}\, f_\omega(X) \right\} \\
& \quad + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega(X) - \widehat{\mathbb{E}}_n f_\omega(X) \right\} \quad \text{(insert } \mathbb{E}\, f_\omega \circ g_\theta(Z) \text{ and } \mathbb{E}\, f_\omega(X)\text{)} \\
& \quad + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}[f_\omega \circ g_{\widehat{\theta}}(Z)] - \widehat{\mathbb{E}}_m [f_\omega \circ g_{\widehat{\theta}}(Z)] \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_n [f_\omega(X) - \mathbb{E}\, f_\omega(X) \right\} \\
& \leq 2 \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_n f_\omega(X) - \mathbb{E}\, f_\omega(X) \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \widehat{\mathbb{E}}_m f_\omega \circ g_\theta(Z) - \mathbb{E}\, f_\omega \circ g_\theta(Z) \right\} \\
& \quad + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_{\widehat{\theta}}(Z) - \widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}}(Z) \right\} + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_\theta(Z) - \mathbb{E}\, f_\omega(X) \right\}
\end{aligned}
$$

where the last step uses the fact that $f_\omega \in \mathcal{F}$ then $-f_\omega \in \mathcal{F}$. As the above holds for any $\theta$ such that $g_\theta \in \mathcal{G}$, we know then (by moving the last term to the LHS)

$$
\begin{aligned}
& d_{\mathcal{F}}\left(\mu_{\widehat{\theta}_{m,n}}, \nu\right) - d_{\mathcal{F}}(\mu_\theta, \nu) \\
& \leq 2 d_{\mathcal{F}}\left(\widehat{\nu}^n, \nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m, \mu_\theta\right) + \sup_{f_\omega \in \mathcal{F}} \left\{ \mathbb{E}\, f_\omega \circ g_{\widehat{\theta}}(Z) - \widehat{\mathbb{E}}_m f_\omega \circ g_{\widehat{\theta}}(Z) \right\} \\
& \leq 2 d_{\mathcal{F}}\left(\widehat{\nu}^n, \nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m, \mu_\theta\right) + \sup_{f_\omega \in \mathcal{F}, g_\theta \in \mathcal{G}} \left\{ \mathbb{E}\, f_\omega \circ g_\theta(Z) - \widehat{\mathbb{E}}_m f_\omega \circ g_\theta(Z) \right\} \\
& \leq 2 d_{\mathcal{F}}\left(\widehat{\nu}^n, \nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m, \mu_\theta\right) + d_{\mathcal{F} \circ \mathcal{G}}(\widehat{\pi}^m, \pi).
\end{aligned}
$$

Here the second to the last step is by the fact that $g_{\widehat{\theta}} \in \mathcal{G}$. *Q.E.D.*

### 5.2. *Minimax Optimal Rates*

We start with an equivalent definition of the Sobolev space for $W^{\alpha,q}(r)$ for $q = 2$ is through the coefficients of the Fourier series. The following is also called the Sobolev ellipsoid. The definition (for $q = 2$) naturally extends to non-integer $\alpha \in \mathbb{R}_{>0}$ through the Bessel potential. Denote $\mathbf{F}[f](\xi)$ denotes the Fourier transform of $f(x)$, and $\mathbf{F}^{-1}$ as its inverse.

DEFINITION 4 *For $\alpha \in \mathbb{R}_{>0}$, the Sobolev space $W^{\alpha,2}(r)$ definition extends to non-integer $\alpha$,*

$$
W^\alpha(r) := \left\{ f \in \Omega \to \mathbb{R} : \left\| \mathbf{F}^{-1} \left[ (1 + |\xi|^2)^{\frac{\alpha}{2}} \mathbf{F}[f](\xi) \right] \right\|_2 \leq r \right\}.
$$

DEFINITION 5 (Sobolev ellipsoid) *Let $\theta = \{\theta_\xi, \xi = (\xi_1, \ldots, \xi_d) \in \mathbb{N}^d\}$ collects the coefficients of the Fourier series, define*

$$
\Theta^\alpha(r) := \left\{ \theta \in \mathbb{N}^d \to \mathbb{R} : \sum_{\xi \in \mathbb{N}^d} (1 + \sum_{i=1}^d \xi_i^2)^\alpha \theta_\xi^2 \leq r^2 \right\}.
$$

It is clear that $\Theta^\alpha(r)$ (frequency domain) is an equivalent representation of $W^\alpha(r)$ (spatial domain, Def. 4) in $L^2(\mathbb{N}^d)$ for trigonometric Fourier series. For more details on Sobolev spaces, we refer the readers to Nemirovski [2000], Nickl and Pötscher [2007], Tsybakov [2009].

PROOF OF THEOREM 1: The proof consists of three main parts, the upper bound and the nonparametric minimax lower bound, and the parametric lower bound. In the proof, for simplicity, we only consider $\alpha, \beta \in \mathbb{N}_{\geq 0}$. Extensions to the $\mathbb{R}_{\geq 0}$ follows the same proof idea.

**Step 1: upper bound**

Recall that the base measure $\pi(x)$ to be a uniform measure on $[0,1]^d$ (Lebesgue measure). For the density $\nu(x)$ w.r.t. the Lebesgue measure, we can represent it in the Fourier trigonometric series form

$$p_\nu(x) = \sum_{\xi\in\mathbb{N}^d} \theta_\xi(\nu)\psi_\xi(x), \quad \theta(\nu)\in\mathbb{N}^d \text{ denotes the coefficients of } \nu$$

with the tensorized basis $\psi_\xi(x) = \prod_{i=1}^d \psi_{\xi_i}(x_i)$ . We construct the following estimator $p_{\widetilde{\nu}_n}$, with a cut-off parameter $M$ to be determined later,

$$p_{\widetilde{\nu}_n}(x) := \sum_{\xi\in\mathbb{N}^d} \widetilde{\theta}_\xi(\nu)\psi_\xi(x),$$

where based on i.i.d. samples $X^{(1)}, X^{(2)}, \dots X^{(n)} \sim \nu$

$$\widetilde{\theta}_\xi(\nu) := \begin{cases} \frac{1}{n}\sum_{j=1}^n \prod_{i=1}^d \psi_{\xi_i}(X_i^{(j)}), & \text{for } \xi \text{ satisfies } \|\xi\|_\infty \le M \\ 0, & \text{otherwise} \end{cases}.$$

Note $\widetilde{\nu}_n$ filters out all the high frequency (less smooth) components, when the multi-index $\xi$ has largest coordinate larger than $M$. Similarly, expand the discriminator function $f\in\mathcal{F}$ in the same Fourier basis,

$$f(x) = \sum_{\xi\in\mathbb{N}^d} \theta_\xi(f)\psi_\xi(x).$$

Recall the Sobolev ball Def. 5, for any $\nu(x)\in W^\alpha(r)$, we have for the estimator $\widetilde{\nu}_n$

$$\begin{aligned}
\mathbb{E}\, d_{\mathcal{F}}(\nu,\widetilde{\nu}_n) &= \mathbb{E}\sup_{f\in\mathcal{F}} \int_\Omega f(x)\left(p_\nu(x) - p_{\widetilde{\nu}_n}(x)\right)dx \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{\xi\in\mathbb{N}^d} \theta_\xi(f)\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right) \\
&= \mathbb{E}\sup_{f\in\mathcal{F}} \left\{ \sum_{\xi\in[M]^d} \theta_\xi(f)\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right) + \sum_{\xi\in\mathbb{N}^d\setminus[M]^d} \theta_\xi(f)\theta_\xi(\nu) \right\} \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{\xi\in[M]^d} \theta_\xi(f)\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right) + \mathbb{E}\sup_{f\in\mathcal{F}} \sum_{\xi\in\mathbb{N}^d\setminus[M]^d} \theta_\xi(f)\theta_\xi(\nu).
\end{aligned}$$

For the truncated first term, we know

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}} \sum_{\xi\in[M]^d} \theta_\xi(f)\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right) \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}} \left\{ \sum_{\xi\in[M]^d} (1+\|\xi\|_2^2)^\beta \theta_\xi^2(f) \right\}^{1/2} \left\{ \sum_{\xi\in[M]^d} (1+\|\xi\|_2^2)^{-\beta}\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right)^2 \right\}^{1/2} \\
&\le \mathbb{E}\left\{ \sum_{\xi\in[M]^d} (1+\|\xi\|_2^2)^{-\beta}\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right)^2 \right\}^{1/2} \quad (\text{as } \sup_{f\in\mathcal{F}} \sum_{\xi\in[M]^d} (1+\|\xi\|_2^2)^\beta \theta_\xi^2(f) \le 1) \qquad (5.1) \\
&\le \left\{ \sum_{\xi\in[M]^d} (1+\|\xi\|_2^2)^{-\beta}\,\mathbb{E}\left(\widetilde{\theta}_\xi(\nu) - \theta_\xi(\nu)\right)^2 \right\}^{1/2} \quad (\text{Jensen's inequality}) \qquad (5.2) \\
&\le \sqrt{C_{d,\beta}\frac{M^{d-2\beta}\vee 1}{n}}
\end{aligned}$$

where the last line $\mathbb{E}\left(\widetilde{\theta}_\xi(\nu)-\theta_\xi(\nu)\right)^2 \le \frac{1}{n}\mathbb{E}_{X\sim\nu}\psi_\xi^2(X) \le \frac{1}{n}$ for trigonometric series for any multi-index $\xi$. In addition, simple calculus shows that

$$\sum_{\xi\in[M]^d}(1+\|\xi\|_2^2)^{-\beta} \le C'_{d,\beta}\int_0^{\sqrt{d}M}\frac{r^{d-1}}{(1+r^2)^\beta}dr \le C_{d,\beta}\left(M^{d-2\beta}\vee 1\right).$$

For the second term, the following inequality holds

$$\begin{aligned}
&\mathbb{E}\sup_{f\in\mathcal{F}}\sum_{\xi\in\mathbb{N}^d\setminus[M]^d}\theta_\xi(f)\theta_\xi(g) \le \mathbb{E}\sup_{f\in\mathcal{F}}\left\{\sum_{\xi\in[M]^d}\theta_\xi^2(f)\right\}^{1/2}\cdot\left\{\sum_{\xi\in[M]^d}\theta_\xi^2(g)\right\}^{1/2}\\
&\le \sup_{f\in\mathcal{F}}\left\{(1+M^2)^{-\beta}\sum_{\xi\in[M]^d}(1+\|\xi\|_2^2)^\beta\theta_\xi^2(f)\right\}^{1/2}\left\{(1+M^2)^{-\alpha}\sum_{\xi\in[M]^d}(1+\|\xi\|_2^2)^\alpha\theta_\xi^2(g)\right\}^{1/2}\\
&\le r\sqrt{\frac{1}{M^{2(\alpha+\beta)}}}.
\end{aligned}$$

Combining two terms, we have for any $\nu\in\mathcal{G}$, with the optimal choice of $M \asymp n^{\frac{1}{2\alpha+d}}$

$$\begin{aligned}
\sup_{\nu\in\mathcal{G}}\mathbb{E}\, d_{\mathcal{F}}(\nu,\widetilde{\nu}_n) &\le \inf_{M\in\mathbb{N}}\left\{\sqrt{C\frac{M^{d-2\beta}\vee 1}{n}}+r\sqrt{\frac{1}{M^{2(\alpha+\beta)}}}\right\}\\
&\precsim n^{-\frac{\alpha+\beta}{2\alpha+d}}\vee n^{-\frac{1}{2}}.
\end{aligned}\tag{5.3}$$

Let us now establish the lower bound. Again we consider the $\Omega=[0,1]^d$ as the domain, which is the same as in the upper bound.

**Step 2: nonparametric lower bound**

The main idea behind the proof is to reduce the estimation problem to a multiple hypothesis testing problem that is at least as hard. In this proof, it turns out the Hölder space $W^{\alpha,\infty}$ — which is a subspace of the Sobolev space $W^\alpha$ — suffices for the minimax lower bound.

First, we need to construct multiple hypothesis $\nu$'s that are valid densities in $W^{\alpha,\infty}(1)$. Specify a kernel function $K(u)=(a_1\exp(-\frac{1}{1-4u^2})-a_2)I(|u|<1/2), u\in\mathbb{R}$ for some small fixed $a_1,a_2>0$ to ensure that $K(x)\in W^{\alpha\vee\beta,\infty}(1)$, and $\int K(u)du=0$. Let $m$ be a parameter (that depends on the sample size $n$) to be determined later, and denote $h_m=1/m$. Define the hypothesis class to be (of cardinality $2^{m^d}$)

$$\begin{aligned}
\Omega_\alpha &= \left\{g_w(x)=1+\sum_{\xi\in[m]^d}w_\xi h_m^\alpha\varphi_\xi(x), w\in\{0,1\}^{m^d}\right\},\\
\Lambda_\beta &= \left\{f_v(x)=\sum_{\xi\in[m]^d}v_\xi h_m^\beta\varphi_\xi(x), v\in\{-1,1\}^{m^d}\right\},
\end{aligned}$$

where

$$\varphi_\xi(x)=\prod_{i=1}^d K\left(\frac{x_i-\frac{\xi_i-1/2}{m}}{h_m}\right),\quad \text{with } h_m=1/m.$$

Let us verify (1) $\Omega_\alpha\subset W^{\alpha,\infty}(r)$ for some $r$, and that each element in the hypothesis set is a valid density; (2) $\Lambda_\beta\subset W^{\beta,\infty}(1)$. To start, for any multi-index $\gamma$ such that $|\gamma|\le\alpha$,

$$\|D^{(\gamma)}g_w\|_\infty \le \sup_{\xi\in[m]^d}h_m^\alpha\|D^{(\gamma)}\varphi_\xi\|_\infty = h_m^{\alpha-|\gamma|}\|D^{(\gamma)}K(u)\|_\infty \le h_m^{\alpha-|\gamma|}\le 1.$$

Similarly for $\forall \gamma, |\gamma| \leq \beta$, we know

$$\|D^{(\gamma)} f_v(x)\|_\infty \leq h_m^{\beta-|\gamma|} \leq 1.$$

We also need to bound $\|g_w\|_\infty$, for any $w$

$$\|g_w\|_\infty \leq 1 + h_m^\alpha \sup_{\xi \in [m]^d} \|\varphi_\xi(x)\|_\infty \leq 1 + h_m^\alpha \leq 1 + 1/100, \tag{5.4}$$

as long as $m$ is large enough. So far we have shown $\Omega_\alpha \subset W^{\alpha,\infty}(r)$ and $\Lambda_\beta \subset W^{\beta,\infty}(1)$. Last, we can check $g_\omega$ is a proper density as we know $g_w(x) \geq 0$, and

$$\int \varphi_\xi(x)dx = \prod_{i=1}^d \int K\left(\frac{x_i - \frac{\xi_i - 1/2}{m}}{h_m}\right) dx_i = 0,$$
$$\int g_\omega(x)dx = 1 + \sum_{\xi \in [m]^d} w_\xi h_m^\alpha \int \varphi_\xi(x)dx = 1.$$

To select hypothesis within $\Omega_\alpha$ are hard to distinguish based on finite samples, we use the Varshamov-Gilbert construction in conjunction with Fano's inequality (we use the version in Lemma 9). The technicality is to construct multiple hypothesis that are separated w.r.t. the adversarial loss, then show that the hypothesis are close in statistical sense. Let's use the construction credited to Varshamov-Gilbert (Lemma 2.9 in Tsybakov [2009]): we know that there exists a subset $\{w^{(0)}, \ldots, w^{(H)}\} \subset \{0,1\}^h$ such that $w^{(0)} = (0, \ldots, 0)$,

$$\rho(w^{(j)}, w^{(k)}) \geq \frac{h}{8}, \ \forall\, j, k \in [H], \ j \neq k,$$
$$\log H \geq \frac{h}{8} \log 2,$$

where $\rho(w, w')$ denotes the Hamming distance between $w$ and $w'$ on the hypercube. In our case $h = m^d$. For the loss function, any $w, w' \in \{w^0, \ldots, w^H\}$

$$\begin{aligned}
d_{\mathcal{F}}(g_w, g_{w'}) &:= \sup_{f \in W^\beta(1)} \int f(x) g_w(x)dx - \int f(x) g_{w'}(x)dx \\
&\geq \sup_{f \in W^{\beta,\infty}(1)} \int f(x) g_w(x)dx - \int f(x) g_{w'}(x)dx \\
&\geq \sup_{f \in \Lambda_\beta} \int f(x) \left(g_w(x) - g_{w'}(x)\right) dx \\
&= \sup_{v \in \{-1,+1\}^{m^d}} h_m^{\alpha+\beta} \sum_{\xi \in [m]^d} v_\xi (w_\xi - w'_\xi) \int \varphi_\xi^2(x)dx \\
&= h_m^{\alpha+\beta+d} \sum_{\xi \in [m]^d} I(w_\xi \neq w'_\xi) \int \prod_{i=1}^d K^2(u_i)\, du \\
&\geq c_{a_1,a_2} h_m^{\alpha+\beta+d} \rho(w, w') \geq c_{a_1,a_2} \frac{m^d}{8} h_m^{\alpha+\beta+d} \asymp h_m^{\alpha+\beta}.
\end{aligned}$$

Now let's show that based $n$ i.i.d. data generated from density $g_w(x)$, it is hard to distinguish the hypothesis. Note that for $|t| < 1/50$, $\log(1+t) \geq t - t^2$. Recall (5.4) we know $\|(g_w(x) - g_0(x))/g_w(x)\|_\infty \leq$

$\frac{1/100}{1-1/100} \leq 1/50$. Therefore

$$\begin{aligned} d_{KL}\left(P_{w^{(j)}}^{\otimes n}, P_{w^{(0)}}^{\otimes n}\right) &= n \cdot d_{KL}\left(P_{w^{(j)}}, P_{w^{(0)}}\right) \\ &= n \int -\log\left(1 + \frac{g_0 - g_{w^{(j)}}}{g_{w^{(j)}}}\right) g_{w^{(j)}} dx \\ &\leq n \int \frac{(g_0 - g_{w^{(j)}})^2}{g_{w^{(j)}}} dx \leq 1.01n \sum_{\xi \in [m]^d} \int h_m^{2\alpha} \varphi_\xi^2(x) dx \\ &\leq 1.01n \sum_{\xi \in [m]^d} \int h_m^{2\alpha+d} \prod_{i=1}^{d} K^2(u_i)\, du \precsim n h_m^{2\alpha+d} m^d. \end{aligned}$$

Therefore if we choose $m \asymp n^{-\frac{1}{2\alpha+d}}$, we know

$$\frac{1}{H} \sum_{j=1}^{H} D_{\mathrm{KL}}(P_{w^{(j)}}^{\otimes n}, P_{w^{(0)}}^{\otimes n}) \leq c \log H = c' m^d.$$

Using the Fano's inequality, the lower bound for density estimation is of the order $h_m^{\alpha+\beta} = n^{-\frac{\alpha+\beta}{2\alpha+d}}$, as

$$\begin{aligned} \inf_{\widetilde{\nu}_n} \sup_{\nu \in W^\alpha(r)} \mathbb{E}\, d_{\mathcal{F}}(\widetilde{\nu}_n, \nu) &\geq \inf_{\hat{g}} \sup_{g \in W^{\alpha,\infty}(r)} \mathbb{E} \sup_{f \in W^{\beta,\infty}(1)} \int f(x)\left(\hat{g}(x) - g(x)\right) dx \\ &\geq \inf_{\hat{w}} \sup_{w \in \{w^{(0)}, \ldots, w^{(H)}\}} \mathbb{E}\, d_{\mathcal{F}}(g_{\hat{w}}, g_w) \\ &\geq c h_m^{\alpha+\beta} \cdot \inf_{\hat{w}} \sup_{w \in \{w^{(0)}, \ldots, w^{(H)}\}} P_w\left(d_{\mathcal{F}}(g_{\hat{w}}, g_w) \geq c h_m^{\alpha+\beta}\right) \\ &\geq c h_m^{\alpha+\beta} \frac{\sqrt{H}}{1+\sqrt{H}} \left(1 - 2c' - \sqrt{\frac{2c'}{\log H}}\right) \quad \text{(Lemma 9)} \\ &\geq c n^{-\frac{\alpha+\beta}{2\alpha+d}}. \end{aligned}$$

**Step 3: parametric lower bound**

The parametric rate lower bound $n^{-1/2}$ can be obtained by the following reduction to a two point hypothesis testing problem. Consider the uniform measure $p_{\nu_0}(x) = 1$ for $x \in [0,1]^d$, and

$$p_{\nu_1}(x) = 1 + \frac{1}{\sqrt{n}} K\left(x(1) - \frac{1}{2}\right).$$

One can verify both $\nu_0, \mu_1$ are valid densities on $[0,1]^d$ with

$$d_{\chi^2}(\nu_1^{\otimes n}, \nu_0^{\otimes n}) = (1 + d_{\chi^2}(\nu_1, \nu_0))^n - 1 = (1 + c/n)^n - 1 \leq e^c - 1$$

where the last line uses the fact

$$d_{\chi^2}(\nu_1, \nu_0) = \frac{1}{n} \int_{-1/2}^{1/2} K^2(u) du \leq \frac{c}{n}. \tag{5.5}$$

Therefore, we know by Pinsker's inequality

$$d_{TV}(\nu_1^{\otimes n}, \nu_0^{\otimes n}) \leq \sqrt{d_{\chi^2}(\nu_1^{\otimes n}, \nu_0^{\otimes n})/2} \leq \sqrt{(e^c - 1)/2}.$$

Recall the fact that the kernel $K(u) \in C^\infty$ with all bounded derivatives in the domain $[-1/2, 1/2]$. Therefore, we know $p_{\nu_1}(x) - p_{\nu_0}(x) = \frac{1}{\sqrt{n}} K(x(1) - \frac{1}{2}) \in W^\infty(r/\sqrt{n})$ with some absolute constant $r > 0$. Now it is clear

that $p_{\nu_0}, p_{\nu_1} \in W^\infty(r) \subset W^\alpha(r')$ for any $\alpha > 0$, with some proper constant $r' > 1 + r/\sqrt{n}$ independent of $n$. Hence, by the Le Cam's method (Lemma 4 in Cai et al. [2015]), for any $\widetilde{\nu}_n$

$$\begin{aligned}\sup_{\nu \in W^\alpha(r)} \mathbb{E}\, d_{\mathcal{F}}\left(\widetilde{\nu}_n, \nu\right) &\geq \sup_{\nu \in \{\nu_0, \nu_1\}} \mathbb{E}\, d_{\mathcal{F}}\left(\widetilde{\nu}_n, \nu\right) \\ &\geq c \cdot d_{\mathcal{F}}(\nu_0, \nu_1)(1 - d_{TV}(\nu_1^{\otimes n}, \nu_0^{\otimes n})) \\ &\geq c' \cdot d_{W^\infty(1)}(\nu_0, \nu_1) = c' r^{-1} \frac{1}{\sqrt{n}} \cdot \int_{-1/2}^{1/2} K^2(u) du \geq c'' \cdot \frac{1}{\sqrt{n}}\end{aligned}$$

where the last step is by choosing the discriminator function $f(x) = r^{-1}\sqrt{n}[p_{\nu_0}(x) - p_{\nu_1}(x)]$ with $f \in \mathcal{F} \in W^\infty(1) \subset W^\beta(1)$. *Q.E.D.*

### 5.3. *Rates for Neural Networks*

Proof of Theorem 4: The proof consists of three steps. Remark in this proof, we wrote $\int$ as $\int_\Omega$ as there won't be confusion.

**Step 1: $A_1(\mathcal{F}, \mathcal{G}, \nu)$ approximation term**

For any distribution $g_{\widehat{\theta}_{m,n}}(Z)$ (we abbreviate $\widehat{\theta}_{m,n}$ as $\widehat{\theta}$ in this proof), by Pinsker's inequality (Lemma 8),

$$d_{TV}^2\left(\nu, \mu_{\widehat{\theta}}\right) \leq \frac{1}{2} d_{KL}\left(\nu || \mu_{\widehat{\theta}}\right).$$

The above implies that for any $X \sim \nu$

$$\begin{aligned}4 d_{TV}^2\left(\nu, \mu_{\widehat{\theta}}\right) &\leq d_{KL}\left(\nu || \mu_{\widehat{\theta}}\right) + d_{KL}\left(\mu_{\widehat{\theta}} || \nu\right) \\ &= \int \log \frac{p_\nu(x)}{p_{\mu_{\widehat{\theta}}}(x)} \left(p_\nu(x) - p_{\mu_{\widehat{\theta}}}(x)\right) dx \quad \text{(for any } f_\omega \in \mathcal{F}) \\ &= \int \left(\log \frac{p_\nu(x)}{p_{\mu_{\widehat{\theta}}}(x)} - f_\omega(x)\right) \left(p_\nu(x) - p_{\mu_{\widehat{\theta}}}(x)\right) dx + \int f_\omega(x) \left(p_\nu(x) - p_{\mu_{\widehat{\theta}}}(x)\right) dx \\ &\leq \int \left(\log \frac{p_\nu(x)}{p_{\mu_{\widehat{\theta}}}(x)} - f_\omega(x)\right) \left(p_\nu(x) - p_{\mu_{\widehat{\theta}}}(x)\right) dx + d_{\mathcal{F}}\left(\mu_{\widehat{\theta}}, \nu\right) \\ &\leq \left\| \log \frac{p_\nu(x)}{p_{\mu_{\widehat{\theta}}}(x)} - f_\omega(x) \right\|_\infty \left\| p_\nu - p_{\mu_{\widehat{\theta}}} \right\|_1 + d_{\mathcal{F}}\left(\mu_{\widehat{\theta}}, \nu\right) \\ &\leq 2 \left\| \log \frac{p_\nu}{p_{\mu_{\widehat{\theta}}}} - f_\omega \right\|_\infty + d_{\mathcal{F}}\left(\mu_{\widehat{\theta}}, \nu\right)\end{aligned}$$

where the last line is due to the fact that $\mu_{\widehat{\theta}}, \nu(x)$ are both proper densities, so $\left\|p_\nu(x) - p_{\mu_{\widehat{\theta}}}(x)\right\|_1 \leq 2$. Take $f_\omega$ to be the one minimize the first term on RHS, we have

$$4 d_{TV}^2\left(\nu, \mu_{\widehat{\theta}}\right) \leq 2 \inf_{f_\omega \in \mathcal{F}} \left\| \log \frac{p_\nu}{p_{\mu_{\widehat{\theta}}}} - f_\omega \right\|_\infty + d_{\mathcal{F}}\left(\mu_{\widehat{\theta}}, \nu\right).$$

**Step 2: oracle inequality and $A_2(\mathcal{G},\nu)$ approximation term**

Now, let's apply the oracle approach developed in Lemma 2 to $d_{\mathcal{F}}\left(\mu_{\widehat{\theta}},\nu\right)$. For any $\theta$ such that $g_\theta \in G$, we know

$$\begin{aligned}
d_{\mathcal{F}}\left(\mu_{\widehat{\theta}},\nu\right) &\leq d_{\mathcal{F}}(\mu_\theta,\nu) + 2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m,\mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi) \\
&\leq B\cdot d_{TV}(\mu,\nu) + 2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m,\mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi) \\
&\leq B\cdot\sqrt{\frac{1}{4}\left[d_{KL}\left(\mu_\theta||\nu\right) + d_{KL}\left(\nu||\mu_\theta\right)\right]} \\
&\quad + 2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m,\mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi) \\
&\leq B\sqrt{\frac{1}{4}\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty \|p_{\mu_\theta}-p_\nu\|_1} + 2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m,\mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi)
\end{aligned}$$

where second line uses the fact that for any $f\in\mathcal{F}$, $\|f\|_\infty \leq B$.

**Step 3: the stochastic term $S_{m,n}(\mathcal{F},\mathcal{G})$ by empirical processes**

Assemble the bounds, we have for any $\theta$

$$\begin{aligned}
4d_{TV}^2\left(\nu,\mu_{\widehat{\theta}_{m,n}}\right) \leq 2\inf_\omega\left\|\log\frac{p_\nu}{p_{\mu_{\widehat{\theta}}}} - f_\omega\right\|_\infty + B\sqrt{\frac{1}{2}\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty} \\
+ 2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m,\mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi)
\end{aligned}$$

Therefore by choosing $\theta_\star$ minimizes $\left\|\log\frac{\mu_\theta}{\nu}\right\|_\infty$ over the generator class

$$\begin{aligned}
\mathbb{E}\, d_{TV}^2\left(\nu,\mu_{\widehat{\theta}_{m,n}}\right) &\leq \frac{1}{2}\mathbb{E}\left\{\inf_\omega\left\|\log\frac{p_\nu}{p_{\mu_{\widehat{\theta}}}} - f_\omega\right\|_\infty\right\} + \frac{B}{4\sqrt{2}}\sqrt{\inf_\theta\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty} \\
&\quad + \mathbb{E}\left\{2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_{\theta_\star}^m,\mu_{\theta_\star}\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi)\right\} \\
&\leq \frac{1}{2}\sup_\theta\inf_\omega\left\|\log\frac{p_\nu}{p_{\mu_\theta}} - f_\omega\right\|_\infty + \frac{B}{4\sqrt{2}}\inf_\theta\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty^{1/2} \\
&\quad + \mathbb{E}\left\{2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_{\theta_\star}^m,\mu_{\theta_\star}\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi)\right\}.
\end{aligned}$$

Recall the symmetrization in Lemma 3,

$$\begin{aligned}
&\mathbb{E}\left\{2d_{\mathcal{F}}\left(\widehat{\nu}^n,\nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_{\theta_\star}^m,\mu_{\theta_\star}\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m,\pi)\right\} \\
&\leq 4\,\mathbb{E}\,\mathcal{R}_n\left(\mathcal{F}\right) + 2\,\mathbb{E}\,\mathcal{R}_m\left(\mathcal{F}\right) + 2\,\mathbb{E}\,\mathcal{R}_m\left(\mathcal{F}\circ\mathcal{G}\right) \\
&\leq C\sqrt{\mathrm{Pdim}(\mathcal{F})\left(\frac{\log m}{m}\vee\frac{\log n}{n}\right)} + C\sqrt{\mathrm{Pdim}(\mathcal{F}\circ\mathcal{G})\frac{\log m}{m}},
\end{aligned}$$

where the last step uses the relationship between Rademacher complexity and pseudo-dimension, shown in Lemma 6. *Q.E.D.*

Proof of Theorem 5: Due to Le Cam's inequality (Lemma 2.3 in Tsybakov [2009]), we know

$$\begin{aligned}
d_{TV}^2\left(\nu,\mu_{\widehat{\theta}_{m,n}}\right) \leq d_H^2\left(\nu,\mu_{\widehat{\theta}_{m,n}}\right) &= \int\left(\sqrt{p_\nu(x)} - \sqrt{p_{\mu_{\widehat{\theta}}}(x)}\right)^2 dx \\
&= \int\frac{\sqrt{p_\nu(x)} - \sqrt{p_{\mu_{\widehat{\theta}}}(x)}}{\sqrt{p_\nu(x)} + \sqrt{p_{\mu_{\widehat{\theta}}}(x)}}\left(p_\nu(x) - p_{\mu_{\widehat{\theta}}}(x)\right)dx \quad \text{for any } f_\omega\in\mathcal{F} \\
&\leq 2\left\|\frac{\sqrt{p_\nu} - \sqrt{p_{\mu_{\widehat{\theta}}}}}{\sqrt{p_\nu} + \sqrt{p_{\mu_{\widehat{\theta}}}}} - f_\omega\right\|_\infty + d_{\mathcal{F}}\left(\mu_{\widehat{\theta}},\nu\right).
\end{aligned}$$

Due to the oracle inequality Lemma 2, one has for any $\theta$

$$d_{\mathcal{F}}\left(\mu_{\widehat{\theta}}, \nu\right) \leq d_{\mathcal{F}}(\mu_\theta, \nu) + 2d_{\mathcal{F}}\left(\widehat{\nu}^n, \nu\right) + d_{\mathcal{F}}\left(\widehat{\mu}_\theta^m, \mu_\theta\right) + d_{\mathcal{F}\circ\mathcal{G}}(\widehat{\pi}^m, \pi).$$

For the first term, we can further upper bound,

$$\begin{aligned} d_{\mathcal{F}}(\mu_\theta, \nu) &\leq B \cdot d_{TV}(\mu_\theta, \nu) \leq B \cdot \sqrt{d_H\left(\mu_\theta, \nu\right)} \\ &\leq 2B \cdot \left\| \frac{\sqrt{p_\nu} - \sqrt{p_{\mu_\theta}}}{\sqrt{p_\nu} + \sqrt{p_{\mu_\theta}}} \right\|_\infty \end{aligned}$$

where the last line follows because

$$\begin{aligned} \sqrt{d_H\left(\mu_\theta, \nu\right)} &= \sqrt{\int \left( \frac{\sqrt{p_\nu(x)} - \sqrt{p_{\mu_{\widehat{\theta}}}(x)}}{\sqrt{p_\nu(x)} + \sqrt{p_{\mu_{\widehat{\theta}}}(x)}} \right)^2 \left( \sqrt{p_\nu(x)} + \sqrt{p_{\mu_\theta}(x)} \right)^2 dx} \\ &\leq \left\| \frac{\sqrt{p_\nu} - \sqrt{p_{\mu_\theta}}}{\sqrt{p_\nu} + \sqrt{p_{\mu_\theta}}} \right\|_\infty \sqrt{\int 2\left(p_\nu(x) + p_{\mu_\theta}(x)\right) dx}. \end{aligned}$$

The rest of the proof follows exactly the same as in Thm. 4. *Q.E.D.*

Proof of Theorem 6: The proof proceeds in three steps.

**Step 1: recursive formula of generator density**

Consider the generator network realized by a multi-layer perceptron:

$$\begin{aligned} h_1 &= \sigma(W_1 z + b_1) \\ &\cdots \\ h_l &= \sigma(W_l h_{l-1} + b_l) \\ &\cdots \\ x &= W_L h_{L-1} + b_L \ . \end{aligned}$$

Denote the parameter space of interest

$$\theta \in \Theta(d, L) := \{(W_l \in \mathbb{R}^{d\times d}, b_l \in \mathbb{R}^d, 1 \leq l \leq L) \mid \operatorname{rank}(W_l) = d, \forall 1 \leq l \leq L\}. \tag{5.6}$$

Let us denote the density function of of the random variable $h_\ell$ to be $p(h_\ell)$. Consider the density evolution from layer $l-1$ to layer $l$ (basic change of variables with Jacobian $\partial h_l / \partial h_{l-1}$)

$$\begin{aligned} \log p(h_l) &= \log p(h_{l-1}) - \log |\det\left(\frac{\partial h_l}{\partial h_{l-1}}\right)| \\ &= \log p(h_{l-1}) - \log |\det W_l| - \sum_{i=1}^d \log |\sigma'(\sigma^{-1}(h_l(i)))|. \end{aligned}$$

Recursively apply the above equality to track the density of $X$, we have

$$\begin{aligned} \log p_{\mu_\theta}(x) &= \log p(h_{L-1}) - \log |\det W_L|, \quad \text{where } h_{L-1} = W_L^{-1}(x - b_L) \\ &= \log p(h_{L-2}) - \sum_{j=L-1}^{L} \log |\det W_j| - \sum_{i=1}^d \log |\sigma'(\sigma^{-1}(h_{L-1}(i)))|, \\ &\cdots \qquad \text{where } h_{L-2} = W_{L-1}^{-1}(\sigma^{-1}(h_{L-1}) - b_{L-1}) \\ &= \log p_\mu(z) - \sum_{j=1}^{L} \log |\det W_j| - \sum_{j=1}^{L-1} \sum_{i=1}^d \log |\sigma'(\sigma^{-1}(h_j(i)))|, \\ &\qquad \text{where } z = W_1^{-1}(\sigma^{-1}(h_1) - b_1). \end{aligned}$$

Now consider $\mu(z) = 1$ to be the uniform measure on $z \in [0,1]^d$. Consider leaky ReLU activation $\sigma(t) = \max(t, at)$ for $0 < a \leq 1$, then $\sigma^{-1}(t) = \min(t, t/a)$, and $\log|\sigma'(t)| = \log(a) \cdot 1_{t\leq 0}$.

Let's consider the realizable case when $\log p_\nu(x) = \log p_{\mu_{\theta_*}}(x)$ for some $\theta_* \in \Theta(d, L)$. Denote $m_l := \sigma^{-1}(h_{L-l})$, for any $1 \leq l \leq L-1$. Then it follows that

$$m_1 = \sigma^{-1}(W_L^{-1}x - W_L^{-1}b_L) \tag{5.7}$$

$$m_l = \sigma^{-1}(W_{L-l+1}^{-1}m_{l-1} - W_{L-l+1}^{-1}b_{L-l+1}), \quad 1 \leq l \leq L-1. \tag{5.8}$$

Therefore, the density can be written out explicitly,

$$\log p_{\mu_\theta}(x) = -\sum_{j=1}^{L}\log|\det W_j| - \sum_{j=1}^{L-1}\sum_{i=1}^{d}\log\sigma'(m_{L-j}(i)) \tag{5.9}$$

$$= -\sum_{j=1}^{L}\log|\det W_j| - \sum_{j=1}^{L-1}\sum_{i=1}^{d}\log\sigma'(m_j(i)) \tag{5.10}$$

In addition, we know that for any $\theta$ and $\theta_*$, $\mu_\theta$ and $\mu_{\theta_*}$ (namely $\nu$) are absolutely continuous to each other, as $\mu_\theta(x) > 0$ for any $x \in [0,1]^d$.

**Step 2: construction of discriminator networks**

Now consider a discriminator network which follows

$$\begin{aligned} m_1 &= \sigma^{-1}(V_1x + c_1) \\ &\cdots \\ m_{L-1} &= \sigma^{-1}(V_{L-1}m_{L-2} + c_{L-1}) \\ h_\omega(x) &= \sum_{j=1}^{L-1}\sum_{i=1}^{d}\log(1/a)1_{m_j(i)\leq 0} + c_L \ . \end{aligned}$$

Here the parameter set is,

$$\omega \in \Omega(d, L) := \{(V_l \in \mathbb{R}^{d\times d}, c_l \in \mathbb{R}^d, c_L \in \mathbb{R}, 1 \leq l \leq L-1) \mid \mathrm{rank}(V_l) = d, \forall 1 \leq l \leq L-1\}. \tag{5.11}$$

Choose the discriminator function $w = (w_1, w_2)$ where $w_1, w_2 \in \Omega(d, L)$

$$f_\omega(x) = h_{\omega_1}(x) - h_{\omega_2}(x).$$

Then we can verify that Cor. 2 follows. Recall the upper bound in Theorem 6, we can see that for the choice of generator and discriminator

$$\begin{aligned} \frac{1}{2}\sup_\theta\inf_\omega\left\|\log\frac{p_\nu}{p_{\mu_\theta}} - f_\omega\right\|_\infty &= 0 \\ \frac{B}{4\sqrt{2}}\inf_\theta\left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty^{1/2} &= 0 \end{aligned}$$

as $\log v(x)$ can be realized by $\log p_{\mu_{\theta_*}}(x)$, and that for any $\theta \in \Theta(d, L)$, there exist an $\omega \in \Omega(d, L)$ such that

$$f_\omega(x) = \log p_\nu(x) - \log p_{\mu_\theta}(x).$$

**Step 3: complexity bound**

Recall the result in Bartlett, Harvey, Liaw, and Mehrabian [2017] on the Vapnik-Chervonenkis dimension of feed-forward neural networks (See Lemma 7 with degree at most 1 and number of pieces $p + 1 = 2$), we

know for leaky-ReLU neural networks $\mathcal{F}$ and $\mathcal{F} \circ \mathcal{G}$ respectively by simple counting

$$\begin{aligned} \text{for network } \mathcal{F}: \quad & \text{number of weights } W_{\mathcal{F}} \leq 2(d^2 L + 2dL) + 2, \\ & \text{number of units } U_{\mathcal{F}} \leq 4dL, \\ & \text{depth } L_{\mathcal{F}} \leq L + 2 \ ; \\ \text{for network } \mathcal{F} \circ \mathcal{G}: \quad & \text{number of weights } W_{\mathcal{F}\circ\mathcal{G}} \leq W_{\mathcal{F}} + d^2 L \\ & \text{number of units } U_{\mathcal{F}\circ\mathcal{G}} \leq U_{\mathcal{F}} + dL, \\ & \text{depth } L_{\mathcal{F}\circ\mathcal{G}} \leq L_{\mathcal{F}} + L \ . \end{aligned}$$

Therefore, we have the following upper bound on VC-dimension,

$$\begin{aligned} \text{Pdim}(\mathcal{F}) \asymp \text{VCdim}(\mathcal{F}) &\leq C \cdot L_{\mathcal{F}} W_{\mathcal{F}} \log U_{\mathcal{F}} = C d^2 L^2 \log(dL), \\ \text{Pdim}(\mathcal{F} \circ \mathcal{G}) \asymp \text{VCdim}(\mathcal{F} \circ \mathcal{G}) &\leq C \cdot L_{\mathcal{F}\circ\mathcal{G}} W_{\mathcal{F}\circ\mathcal{G}} \log U_{\mathcal{F}\circ\mathcal{G}} \leq C' d^2 L^2 \log(dL). \end{aligned}$$

Finally, by Cor. 2, we have the result proved. *Q.E.D.*

## REFERENCES


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

Michael Arbel, Dougal J Sutherland, Mikołaj Bińkowski, and Arthur Gretton. On gradient regularizers for mmd gans. *arXiv preprint arXiv:1805.11565*, 2018.

Martin Arjovsky and Léon Bottou. Towards principled methods for training generative adversarial networks. *arXiv preprint arXiv:1701.04862*, 2017.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017.

Sanjeev Arora and Yi Zhang. Do gans actually learn the distribution? an empirical study. *arXiv preprint arXiv:1706.08224*, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (gans). *arXiv preprint arXiv:1703.00573*, 2017.

Susan Athey, Guido W Imbens, Jonas Metzger, and Evan M Munro. Using wasserstein generative adversarial networks for the design of monte carlo simulations. Technical report, National Bureau of Economic Research, 2019.

Kerry Back and David P. Brown. Implied Probabilities in GMM Estimators. *Econometrica*, 61(4):971–975, 1993. ISSN 0012-9682. .

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in gans. *arXiv preprint arXiv:1806.10586*, 2018.

Peter L. Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian. Nearly-tight vc-dimension bounds for piecewise linear neural networks. In *Proceedings of the 22nd Annual Conference on Learning Theory (COLT 2017)*, 2017.

Lawrence D Brown. Fundamentals of statistical exponential families: with applications in statistical decision theory. Ims, 1986.

T. Tony Cai, Tengyuan Liang, and Harrison H. Zhou. Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional gaussian distributions. *Journal of Multivariate Analysis*, 137:161 – 172, 2015. ISSN 0047-259X. . URL http://www.sciencedirect.com/science/article/pii/S0047259X1500038X.

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

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training gans with optimism. *arXiv preprint arXiv:1711.00141*, 2017.

Ernesto De Vito, Nicole Mücke, and Lorenzo Rosasco. Reproducing kernel Hilbert spaces on manifolds: Sobolev and Diffusion spaces. *arXiv:1905.10913 [cs, math, stat]*, May 2019.

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

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *arXiv preprint arXiv:1809.09953, Econometrica, forthcoming*, August 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

Lars Peter Hansen. Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, 50(4):1029–1054, 1982. ISSN 0012-9682. .

Guido W Imbens, Phillip Johnson, and Richard H Spady. Information theoretic approaches to inference in moment condition models. Technical report, National Bureau of Economic Research, 1995.

Yujia Li, Kevin Swersky, and Rich Zemel. Generative moment matching networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1718–1727, 2015.

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

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

Shuang Liu and Kamalika Chaudhuri. The inductive bias of restricted f-gans. *arXiv preprint arXiv:1809.04542*, 2018.

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. *arXiv preprint arXiv:1705.08991*, 2017.

Mario Lucic, Karol Kurach, Marcin Michalski, Sylvain Gelly, and Olivier Bousquet. Are gans created equal? a large-scale study. *arXiv preprint arXiv:1711.10337*, 2017.

Daniel McFadden. A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica*, 57(5):995–1026, 1989. ISSN 0012-9682. .

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. The numerics of gans. *arXiv preprint arXiv:1705.10461*, 2017.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. Which training methods for gans do actually converge? In *International Conference on Machine Learning*, pages 3478–3487, 2018.

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

Arkadi Nemirovski. Topics in non-parametric. *Ecole d'Eté de Probabilités de Saint-Flour*, 28:85, 2000.

Richard Nickl and Benedikt M Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20(2):177–199, 2007.

Ariel Pakes and David Pollard. Simulation and the Asymptotics of Optimization Estimators. *Econometrica*, 57(5):1027–1057, 1989. ISSN 0012-9682. .

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

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation under adversarial losses. *arXiv preprint arXiv:1805.08836*, 2018.

Casper Kaae Sønderby, Jose Caballero, Lucas Theis, Wenzhe Shi, and Ferenc Huszár. Amortised map inference for image super-resolution. *arXiv preprint arXiv:1610.04490*, 2016.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.

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

Ramon van Handel. Probability in high dimension. Technical report, PRINCETON UNIV NJ, 2014.

Larry Wassermann. *All of nonparametric statistics*. Springer Science+ Business Media, New York, 2006.

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

Bin Yu. Stability. *Bernoulli*, 19(4):1484–1500, 2013.

## APPENDIX A: REMAINING PROOFS

### A.1. *Other Theorems and Corollaries*

PROOF OF THEOREM 2: The proof logic of this corollary follows similarly as in Theorem 1. We need to adapt the proof to the density ratio w.r.t. the general base measure $\pi$. Express $f \in \mathcal{F}$ under the eigenfunctions

$$f(x) = \sum_{i\in\mathbb{N}} f_i \psi_i(x), \text{ with } \sum_i t_i^{-1} f_i^2 \le 1$$

where $t_i \asymp i^{-\kappa}$ and $f_i = \int f\psi_i d\pi$ are the coefficients. Consider the series representation of the target density $\frac{d\nu}{d\pi}$ w.r.t. the base measure $\pi$

$$\frac{d\nu}{d\pi}(x) = \sum_{i\in\mathbb{N}} \nu_i \psi_i(x), \text{ then}$$

$$\|\mathcal{T}_\pi^{-(\lambda-1)/2} \frac{d\nu}{d\pi}\|_{\mathcal{H}} \le r \text{ is equivalent to } \sum_i t_i^{-\lambda} \nu_i^2 \le r^2.$$

Define density

$$\frac{d\widetilde{\nu}_n}{d\pi}(x) := \sum_{i\in\mathbb{N}} \widetilde{\nu}_i \psi_i(x),$$

where based on i.i.d. samples $X^{(1)}, X^{(2)}, \dots X^{(n)} \sim \nu$

$$\widetilde{\nu}_i := \begin{cases} \frac{1}{n}\sum_{j=1}^n \psi_i(X^{(j)}), & \text{for } i \le M \\ 0, & \text{otherwise} \end{cases}.$$

Follow the sample logic as in the proof of Theorem 1, we have for any $\nu(x) \in \mathcal{G}$, with the optimal choice of $M \asymp n^{\frac{1}{\lambda\kappa+1}}$, the following holds

$$\begin{aligned}
\mathbb{E}\, d_{\mathcal{F}}(\nu, \widetilde{\nu}_n) &= \mathbb{E} \int f(d\nu - d\widetilde{\nu}_n) \\
&= \mathbb{E} \int f \left( \frac{d\nu}{d\pi} - \frac{d\widetilde{\nu}_n}{d\pi} \right) d\pi \\
&\le \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{i\le M} f_i \left(\widetilde{\nu}_i - \nu_i\right) + \mathbb{E} \sup_{f\in\mathcal{F}} \sum_{i>M} f_i \nu_i. \\
&\le \sqrt{\sum_{i\le M} t_i^{-1} f_i^2} \sqrt{\sum_{i\le M} t_i\, \mathbb{E}(\widetilde{\nu}_i - \nu_i)^2} + C r t_M^{\frac{\lambda+1}{2}} \\
&\le \inf_{M\in\mathbb{N}} \left\{ \sqrt{C \frac{M^{1-\kappa} \vee 1}{n}} + Cr \sqrt{\frac{1}{M^{\kappa(\lambda+1)}}} \right\} \\
&\precsim n^{-\frac{(\lambda+1)\kappa}{2\lambda\kappa+2}} \vee n^{-\frac{1}{2}}.
\end{aligned}$$

*Q.E.D.*

PROOF OF THEOREM 3: By the entropy integral Lemma 3, if $\mathcal{F}_D$ consists of $L$-Lipschitz functions (Wasserstein GAN) on $\mathbb{R}^d$, $d \ge 2$, plug in the $\ell_\infty$-covering number bound for Lipschitz functions,

$$\begin{aligned}
&\log \mathcal{N}(\epsilon, \mathcal{F}_D, \|\cdot\|_\infty) \le C \left(\frac{L}{\epsilon}\right)^d, \\
&\mathbb{E}\, d_{\mathcal{F}_D}(\nu, \widehat{\nu}^n) \le 2 \inf_{0<\delta<1/2} \left( 4\delta + \frac{8\sqrt{2}}{\sqrt{n}} \int_\delta^{1/2} \sqrt{\log \mathcal{N}(\epsilon, \mathcal{F}_D, \|\cdot\|_\infty)} d\epsilon \right) \\
&\quad \le 16 \left( \frac{4\sqrt{2C}}{d-2} \right)^{\frac{2}{d}} L n^{-\frac{1}{d}} = \mathcal{O}\left( \left( \frac{C}{d^2 n} \right)^{-\frac{1}{d}} \right).
\end{aligned}$$

This matches the best known bound as in Canas and Rosasco [2012] (Section 2.1.1).

Let's consider when $\mathcal{F}_D$ denotes Sobolev space $W^{\beta,2}$ on $\mathbb{R}^d$. Recall the entropy number estimate for $W^{\beta,2}$ [Nickl and Pötscher, 2007], we have

$$\begin{aligned}
&\log \mathcal{N}(\epsilon, \mathcal{F}_D, \|\cdot\|_\infty) \le C \left(\frac{1}{\epsilon}\right)^{\frac{d}{\beta}\vee 2}, \\
&\mathbb{E}\, d_{\mathcal{F}_D}(\nu, \widehat{\nu}^n) \le \mathcal{O}\left( n^{-\frac{\beta}{d}} + \frac{\log n}{\sqrt{n}} \right).
\end{aligned}$$

Remark in addition that the parametric rate $\frac{1}{\sqrt{n}}$ is inevitable, which can be easily seen from the Sudakov minoration,

$$\mathop{\mathbb{E}}_{\epsilon} \sup_{f\in\mathcal{F}_D} \frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i) \geq \sup_{\epsilon} \frac{\epsilon}{2}\sqrt{\frac{\log \mathcal{M}(\epsilon, \mathcal{F}_D, \|\cdot\|_n)}{n}} \geq \frac{1}{\sqrt{n}}.$$

For the smoothed/regularized density, one can apply Lemma 1 and Theorem 1 to obtain the claimed result. *Q.E.D.*

Proof of Corollary 1: Now let's consider Wasserstein distance. Consider in addition the Lipschitz constants of $\mathcal{F}$ to be $L_{\mathcal{F}}$, and $\mathcal{G}$ to be $L_{\mathcal{G}}$, namely

$$|f_\omega(x) - f_\omega(x')| \leq L_{\mathcal{F}}\|x - x'\|$$
$$\|g_\theta(z) - g_\theta(z')\| \leq L_{\mathcal{G}}\|z - z'\|$$

Consider first the case when $Z \sim N(0, I_d)$ (unbounded). Then for any $f \in Lip(1)$, we know

$$f(g_\theta(z)) \in Lip(L_{\mathcal{G}}). \tag{A.1}$$

In other words, $f \circ g_\theta(Z)$ are $L_{\mathcal{G}}^2$ sub-Gaussian (Lemma 8), therefore

$$d_W^2\left(\nu, \mu_{\widehat{\theta}}\right) \leq 2L_{\mathcal{G}}^2 \cdot d_{KL}\left(\nu||\mu_{\widehat{\theta}}\right)$$

and

$$\begin{aligned} d_{\mathcal{F}}\left(\nu, \mu_\theta\right) &\leq L_{\mathcal{F}} \cdot d_W\left(\nu, \mu_\theta\right) \\ &\leq \sqrt{2}L_{\mathcal{F}}L_{\mathcal{G}}\sqrt{d_{KL}\left(\nu||\mu_\theta\right)}. \end{aligned}$$

Follow the analysis with as in the TV distance, we have

$$\begin{aligned} \mathbb{E}\, d_W^2\left(\nu, \mu_{\widehat{\theta}}\right) \leq L_{\mathcal{G}}^2 \sup_\theta \inf_\omega \left\|\log\frac{p_\nu}{p_{\mu_\theta}} - f_\omega\right\|_\infty + L_{\mathcal{G}}^3 L_{\mathcal{F}} \inf_\theta \left\|\log\frac{p_{\mu_\theta}}{p_\nu}\right\|_\infty^{1/2} \\ + C\sqrt{\mathrm{Pdim}(\mathcal{F})\left(\frac{\log m}{m} \vee \frac{\log n}{n}\right)} + C\sqrt{\mathrm{Pdim}(\mathcal{F}\circ\mathcal{G})\frac{\log m}{m}}. \end{aligned}$$

Consider then the case when $z, x \in [0,1]^d$ is bounded, we know

$$\|g_\theta(z) - g_\theta(z')\| \leq L_{\mathcal{G}}\sqrt{d} \tag{A.2}$$

Therefore $\|g_\theta(z)\| \leq M + L_{\mathcal{G}}\sqrt{d}$, and the support of $g_\theta(Z)$ and $X$ lies in $R := M + (L_{\mathcal{G}} + 1)\sqrt{d}$. Hence

$$\mathbb{E}\, d_W^2\left(\nu, \mu_{\widehat{\theta}}\right) \leq R^2\, \mathbb{E}\, d_{TV}^2\left(\nu, \mu_{\widehat{\theta}}\right).$$

The last line is because for any $f(x)$ that has Lipchitz constant 1 with $f(0) = 0$, it is true that $f(x)$ is bounded in a bounded domain with radius $R$. And such an centering of $f$ is without loss of generality since $\sup_{f:f\in Lip(1)}\int f(\mu - \nu)dx = \sup_{f-f(0):f\in Lip(1)}\int (f - f(0))(\mu - \nu)dx$, for probability measures $\mu, \nu$.

*Q.E.D.*

Proof of Corollary 3: Suppose $\log p_\nu(x) = -\frac{1}{2}(x - b_*)'\Sigma_*^{-1}(x - b_*) + \frac{1}{2}\log\det(\Sigma_*^{-1}) - \frac{d}{2}\log(2\pi)$. And the generator class is depth-one NN, with weights $\theta = (W, b)$, $X = WZ + b$, then $\log p_{\mu_\theta}(x) = -\frac{1}{2}(x - b)'(WW')^{-1}(x - b) + \frac{1}{2}\log\det((WW')^{-1}) - \frac{d}{2}\log(2\pi)$.

For the discriminator, if one is allowed to use $\sigma(t) = t^2$, then one can have $O(d)$ units in discriminator network with depth 2, so that the two approximation error term is zero (Note one can also realize with ReLU activation in a bounded domain, using saw construction, as in Yarotsky [2017]). By Lemma 7 with degree at most 2, $\mathrm{VCdim}(\mathcal{F}) \precsim d^2\log d$, $\mathrm{VCdim}(\mathcal{F}\circ\mathcal{G}) \precsim (pd + d^2)\log(p + d)$. Therefore $\mathbb{E}\, d_{TV}^2\left(g_\theta(Z), X\right) \leq C\left(\frac{d^2\log d}{n\wedge m} + \frac{(pd+d^2)\log(p+d)}{m}\right)^{1/2}$. *Q.E.D.*

## A.2. *Supporting Lemmas*

Let's define the empirical Rademacher complexity

$$\mathcal{R}_n(\mathcal{F}) := \mathop{\mathbb{E}}_{\epsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i). \tag{A.3}$$

Lemma 3 (Symmetrization and entropy integral) *For $\widehat{\nu}^n(x) = \frac{1}{n}\sum_{i=1}^{n}\delta_{X_i}(x)$, then*

$$\mathbb{E}\, d_{\mathcal{F}}(\nu, \widehat{\nu}^n) \leq 2\,\mathbb{E}\,\mathcal{R}_n(\mathcal{F}). \tag{A.4}$$

*Assuming $\sup_{f\in\mathcal{F}}\|f\|_\infty \leq 1$, one has the standard entropy integral bound,*

$$\mathbb{E}\, d_{\mathcal{F}}(\nu, \widehat{\nu}^n) \leq 2\,\mathbb{E}\inf_{0<\delta<1/2}\left(4\delta + \frac{8\sqrt{2}}{\sqrt{n}}\int_\delta^{1/2}\sqrt{\log\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_n)}d\epsilon\right),$$

*where $\|f\|_n := \sqrt{1/n\sum_{i=1}^{n}f(X_i)^2}$ is the empirical $\ell_2$-metric on data $\{X_i\}_{i=1}^n$. Furthermore, because $\|f\|_n \leq \max_i|f(X_i)|$, and therefore $\mathcal{N}(\epsilon, \mathcal{F}, \|\cdot\|_n) \leq \mathcal{N}(\epsilon, \mathcal{F}|_{X_1\ldots,X_n}, \infty)$ and so the upper bound in the conclusions also holds with $\mathcal{N}(\epsilon, \mathcal{F}|_{X_1\ldots,X_n}, \infty)$.*

PROOF: We use the Dudley entropy integral, a standard result in empirical process theory. For the first inequality, it is easy to connect to standard symmetrization technique,

$$\mathbb{E}\, d_{\mathcal{F}}(\nu, \widehat{\nu}^n) \leq \mathop{\mathbb{E}}_{X, X'} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i) - f(X_i') \leq 2 \mathop{\mathbb{E}}_{X} \mathop{\mathbb{E}}_{\epsilon} \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(X_i).$$

*Q.E.D.*

The next two results, Theorems 12.2 and 14.1 in Anthony and Bartlett [2009], show that the metric entropy may be bounded in terms of the pseudo-dimension and that the latter is bounded by the Vapnik-Chervonenkis (VC) dimension.

LEMMA 4 *Assume for all $f \in \mathcal{F}$, $\|f\|_\infty \leq M$. Denote the pseudo-dimension of $\mathcal{F}$ as $\mathrm{Pdim}(\mathcal{F})$, then for $n \geq \mathrm{Pdim}(\mathcal{F})$, we have for any $\epsilon$ and any $X_1, \ldots, X_n$,*

$$\mathcal{N}(\epsilon, \mathcal{F}|_{X_1, \ldots, X_n}, \infty) \leq \left( \frac{2eM \cdot n}{\epsilon \cdot \mathrm{Pdim}(\mathcal{F})} \right)^{\mathrm{Pdim}(\mathcal{F})}.$$

LEMMA 5 *If $\mathcal{F}$ is the class of functions generated by a neural network with a fixed architecture and fixed activation functions, then*

$$\mathrm{Pdim}(\mathcal{F}) \leq \mathrm{VCdim}(\tilde{\mathcal{F}})$$

*where $\tilde{\mathcal{F}}$ has only one extra input unit and one extra computation unit compared to $\mathcal{F}$.*

LEMMA 6 (Rademacher complexity and Pseudo-dimension) *Under the condition $\max_i |f(X_i)| \leq B$, then for any $n \geq \mathrm{Pdim}(\mathcal{F})$,*

$$\mathcal{R}_n(\mathcal{F}) \leq C \cdot B \sqrt{\frac{\mathrm{Pdim}(\mathcal{F}) \log n}{n}}$$

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

PROOF: The proof is a direct application of the Dudley entropy integral in Lemma 3 and the covering number bound by pseudo-dimension in Lemma 4. See A.2.2 in Farrell, Liang, and Misra [2020] for details. *Q.E.D.*

LEMMA 7 (Theorem 6 in Bartlett et al. [2017], Vapnik-Chervonenkis dimension) *Consider function class computed by a feed-forward neural network architecture with $W$ parameters and $U$ computation units arranged in $L$ layer. Suppose that all non-output units have piecewise-polynomial activation functions with $p+1$ pieces and degree no more than $d$, and the output unit has the identity function as its activation function. Then the VC-dimension and pseudo-dimension is upper bounded*

$$\mathrm{VCdim}(\mathcal{F}), \mathrm{Pdim}(\mathcal{F}) \leq C \cdot \left( LW \log(pU) + L^2 W \log d \right),$$

*with some universal constants $C > 0$. The same result holds for pseudo-dimension $\mathrm{Pdim}(\mathcal{F})$.*

LEMMA 8 (van Handel [2014], special case of Theorem 4.8 and Example 4.9) *For any two random variables $g_\theta(Z), X \in \mathbb{R}^d$ with $g_\theta(Z) \sim \mu_\theta$ and $X \sim \nu$, Pinsker's inequality asserts that*

$$2 d_{TV}^2(\mu_\theta, \nu) \leq d_{KL}(\mu_\theta || \nu).$$

*Assume in addition that $Z \sim N(0, I_d)$ to be isotropic Gaussian and for all $\theta$, $\|g_\theta(z) - g_\theta(z')\| \leq L\|z - z'\|$ is $L$-Lipschitz. Then for any $X \sim \nu$ we know*

$$d_W^2(\mu_\theta, \nu) \leq 2L^2 d_{KL}(\nu || \mu_\theta).$$

PROOF: Consider any 1-Lipchitz function $f : \mathbb{R}^d \to \mathbb{R}$, then $f \circ g_\theta$ is $L$-Lipschitz, which implies $f \circ g_\theta$ is $L^2$-subGaussian due to Gaussian concentration Theorem 3.25 in van Handel [2014]. Therefore we know $f(g_\theta(Z))$ is $L^2$-subGaussian for any $f$ that is 1-Lipchitz, together with Theorem 4.8 in van Handel [2014], the proof completes. *Q.E.D.*

LEMMA 9 (Theorem 2.5 in Tsybakov [2009]) *Let $d(\cdot, \cdot)$ be a metric on $\Theta$. Assume that $H \geq 2$ and suppose $\Theta$ contains $\theta_0, \theta_1, \ldots, \theta_H$ such that:*

1. *$d(\theta_j, \theta_k) \geq 2s > 0$, for all $j, k \in [H]$ and $j \neq k$.*
2. *$\frac{1}{H} \sum_{j=1}^{H} d_{KL}(P_j, P_0) \leq c \log H$ with $0 < c < 1/8$ and $P_j = P_{\theta_j}$ for $j \in [H]$.*

*Then for any estimator $\hat{\theta}$,*

$$\sup_{\theta \in \Theta} P_\theta(d(\hat{\theta}, \theta) \geq s) \geq \frac{\sqrt{H}}{1 + \sqrt{H}} \left( 1 - 2c - \sqrt{\frac{2c}{\log H}} \right) > 0.$$