Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks

Jeyong Lee¹, Hyeok Kyu Kwon¹, Minwoo Chae¹*

¹Department of Industrial and Management Engineering, Pohang University of Science and Technology, Pohang, South Korea.

*Corresponding author(s). E-mail(s): mchae@postech.ac.kr;

Abstract

It is common in nonparametric estimation problems to impose a certain low-dimensional structure on the unknown parameter to avoid the curse of dimensionality. This paper considers a nonparametric distribution estimation problem with a structural assumption under which the target distribution is allowed to be singular with respect to the Lebesgue measure. In particular, we investigate the use of generative adversarial networks (GANs) for estimating the unknown distribution and obtain a convergence rate with respect to the $L^1$-Wasserstein metric. The convergence rate depends only on the underlying structure and noise level. More interestingly, under the same structural assumption, the convergence rate of GAN is strictly faster than the known rate of VAE in the literature. We also obtain a lower bound for the minimax optimal rate, which is conjectured to be sharp at least in some special cases. Although our upper and lower bounds for the minimax optimal rate do not match, the difference is not significant.

Keywords: Convergence rate, deep generative model, generative adversarial networks, nonparametric distribution estimation, singular distribution, Wasserstein distance

1 Introduction

Given $D$-dimensional observations $X_1, \ldots, X_n$ following $P_0$, suppose that we are interested in inferring the underlying distribution $P_0$ or related quantities such as its density function or the manifold on which $P_0$ is supported. The inference of $P_0$ is fundamental in unsupervised learning, and there are numerous inferential methods available in
In this paper, \( X_i \) is modeled as \( X_i = g(Z_i) + \epsilon_i \) for some function \( g : Z \to \mathbb{R}^D \). Here, \( Z_i \) is a latent variable following the known distribution \( P_Z \) supported on \( Z \subset \mathbb{R}^d \), and \( \epsilon_i \) is an error following a normal distribution \( \mathcal{N}(0_D, \sigma^2 I_D) \), where \( 0_D \) and \( I_D \) denote a \( D \)-dimensional vector of zeros and an identity matrix, respectively. The dimension \( d \) of the latent variable \( Z_i \) is typically much smaller than \( D \). This model is often called a (non-linear) factor model in statistical communities (Kundu and Dunson (2014); Yalcin and Amemiya (2001)) and a generative model in machine learning societies (Goodfellow et al., 2014; Kingma & Welling, 2014). Throughout the paper, we use the latter terminology. Accordingly, \( g \) will be referred to as a generator.

Recent advances in deep learning have expanded the use of generative models by modeling \( g \) through deep neural networks (DNN), also known as deep generative models. Two approaches are popularly used for estimating \( g \). Variational autoencoder (VAE; Kingma and Welling (2014); Rezende, Mohamed, and Wierstra (2014)) is perhaps the most well-known algorithm for constructing an estimator \( \hat{g} \) using a likelihood approach. The other approach is generative adversarial networks (GAN). Originally proposed by Goodfellow et al. (2014), GAN has been extended in several directions. One of its extensions considers general integral probability metrics (IPM) as loss functions. Sobolev GAN (Mroueh, Li, Sercu, Raj, & Cheng, 2017), maximum mean discrepancy GAN (Li, Chang, Cheng, Yang, & Póczos, 2017) and Wasserstein GAN (Arjovsky, Chintala, & Bottou, 2017) are important examples in this direction. Another important direction of generalization is the development of novel architectures for generators and discriminators; deep convolutional GAN (Radford, Metz, & Chintala, 2016), progressive GAN (Karras, Aila, Laine, & Lehtinen, 2018) and style GAN (Karras, Laine, & Aila, 2019) are successful ones. In many real applications, GAN often performs better than the likelihood approach in terms of the quality of generated samples.

In spite of the rapid development of GAN, a theoretical understanding of it remains largely unexplored. Specifically, a generative model typically focuses on providing an estimator only for the generator \( g \) and does not yield an explicit estimator for the unknown distribution \( P_0 \). Since the generator is not identifiable, it is crucial to study the convergence rate of the distribution estimator, implicitly defined through \( \hat{g} \). This paper aims to bridge this gap by studying the statistical properties of GAN from the viewpoint of estimating a nonparametric distribution. We investigate the convergence rate of a GAN-based estimator for the underlying distribution concentrated around a low-dimensional structure. Through this analysis, our objective is to provide theoretical insights into why GAN outperforms classical nonparametric methods and likelihood approaches in many applications.

Let \( Q_g \) and \( P_{g,\sigma} \) denote distributions of \( g(Z) \) and \( g(Z) + \epsilon \), respectively, where \( Z \sim P_Z \) and \( \epsilon \sim \mathcal{N}(0_D, \sigma^2 I_D) \) are independent. \( Q_g \) is often called the pushforward measure of \( P_Z \) through the generator \( g \). For the data-generating distribution \( P_0 \), we assume that \( P_0 = P_{g_0,\sigma_0} \) with a true generator \( g_0 \) and \( \sigma_0 \geq 0 \). We further assume that \( g_0 \) possesses a certain low-dimensional structure and \( \sigma_0 \) is small enough so that \( P_0 \) is concentrated around the structure. This assumption on the true distribution...
has been thoroughly investigated by Chae, Kim, Kim, and Lin (2023), inspired by recent articles on structured distribution estimation (Aamari & Levrard, 2019; Divol, 2020; Genovese, Perone-Pacifico, Verdinelli, & Wasserman, 2012a, 2012b; Puchkin & Spokoiny, 2022). Once the true generator $g_0$ possesses a low-dimensional structure that DNN can efficiently capture, deep generative models are highly appropriate for statistical inferences. We consider a composite structure (Horowitz & Mammen, 2007; Juditsky, Lepski, & Tsybakov, 2009) which has recently been studied in deep supervised learning (Bauer & Kohler, 2019; Schmidt-Hieber, 2020). Then, the corresponding distribution $Q_0 = Q_{g_0}$ inherits the structure of $g_0$. Details are described further in Section 3. Although the structural assumption on the distribution through the generator is quite natural, to the best of our knowledge, it has not been studied in the literature except for the work presented in Chae et al. (2023). Similarly to Chae et al. (2023), we adopt this structural assumption in order to develop a statistical theory that explains the benefits of deep generative models and GANs.

Under the above setting, it would be more reasonable to set $Q_0$, rather than $P_0$, as the target distribution to be estimated because $\epsilon$ is a noise. Once an estimator $\hat{g}$ is constructed, one can define an estimator for $Q_0$ as $\hat{Q} = Q_{\hat{g}}$. To evaluate the performance of the estimation, we primarily consider the $L^1$-Wasserstein metric. The metric was originally inspired by the problem of optimal mass transportation (Villani, 2003) and has been widely adopted as an evaluation measure in distribution estimation problems (Chae & Walker, 2019; Nguyen, 2013; Wei & Nguyen, 2022). When $g_0$ possesses a composite structure with parameters $(t_i, \beta_i)_{i=0}^{q}$, see Section 3 for details, we construct a GAN-based estimator that achieves the convergence rate

$$\max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{\beta_i + t_i}} + \sigma_0$$

up to a logarithmic factor; see Theorem 2. Note that the rate does not explicitly depend on the dimensions $D$ and $d$. Under the same assumption, Chae et al. (2023) obtained the rate

$$\max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{\beta_i + t_i}} + \sigma_0$$

up to a logarithmic factor using a VAE-type estimator. Note that our convergence rate is strictly faster than the rate obtained in Chae et al. (2023), which is derived using the sharp probability inequality for likelihood ratios developed by Wong and Shen (1995). Based on this observation, we conjecture that the convergence rate for the VAE-type estimator in Chae et al. (2023) cannot be improved. If this conjecture holds true, our theory will provide valuable insights into the reasons why GAN outperforms VAE.

For the class of structured distributions described above, we also obtain a lower bound

$$\max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{\beta_i + t_i}}$$

3
for the minimax convergence rate; see Theorem 4. When $\sigma_0$ is small enough, this lower bound is only slightly smaller than the rate achieved by a GAN-based estimator. That is, the convergence rate of a GAN-based estimator obtained in this paper is at least very close to the minimax optimal rate. As discussed after Theorem 4, we conjecture that our lower bound cannot be improved in general, and thus, there might be room for improving the upper bound.

Besides the convergence rate with respect to the $L^1$-Wasserstein distance, we also investigate the convergence rate for $d_{\mathcal{F}_0}(\hat{Q}, Q_0)$ with a general integral probability metric $d_{\mathcal{F}_0}$, as defined in (1); see Theorem 3. The $L^1$-Wasserstein distance corresponds to a special case where $\mathcal{F}_0$ is the class of every function with Lipschitz constant bounded by 1. For another example, $\mathcal{F}_0$ can be chosen as an $\alpha$-Hölder class. Additionally, neural network distances are also natural choices for $d_{\mathcal{F}_0}$, where the term neural network distance refers to an integral probability metric $d_{\mathcal{F}_0}$ with $\mathcal{F}_0$ consisting of neural network functions (Arora, Ge, Liang, Ma, & Zhang, 2017; Bai, Ma, & Risteski, 2019; Liu, Bousquet, & Chaudhuri, 2017; Zhang, Liu, Zhou, Xu, & He, 2018).

It would be worthwhile to highlight several technical novelties of this paper compared to existing theories on GANs. A comprehensive overview of related work can be found in Section 1.1.

Firstly, while most existing theories on GANs analyze them from the perspective of nonparametric density estimation, our paper distinguishes itself by focusing on distribution estimation. This allows us to handle both scenarios where the underlying distribution is singular with respect to the Lebesgue measure or possesses a smooth Lebesgue density. In particular, within the framework of existing theory, classical methods such as kernel density estimators and wavelets can achieve the minimax optimal convergence rate. Therefore, their results are insufficient to explain the advantage of GAN compared to classical methods. In this regard, our theory for GAN is particularly beneficial as it provides a framework that can explain the advantages of using GAN for both density estimation and structured distribution estimation problems. There have been recent articles that explore modifications of classical methods for estimating distributions on manifolds (Berenfeld & Hoffmann, 2021; Divol, 2022). However, it remains unclear whether these methods are suitable for the structured distribution estimation problem addressed in the present paper. The structured distribution estimation considered in our paper involves a substantially richer structure than the manifold structure, as discussed in Chae et al. (2023).

Another notable technique in the proof of Theorem 2 lies in the construction of the discriminator class. In the literature, the function class for the discriminator is identical to the function class defining the evaluation metric. In case of the $L^1$-Wasserstein, for example, it is the class $\mathcal{F}_{\text{Lip}}$ of every function with Lipschitz constant bounded by one. In particular, the discriminator class depends solely on the evaluation metric. On the other hand, the discriminator class in our proof depends not only on the evaluation metric but on the generator architecture. Although state-of-the-art GAN architectures such as progressive GAN (Karras et al., 2018) and StyleGAN (Karras et al., 2019) are too complicated to render them theoretically tractable, it is crucial for the success of these procedures that discriminator architectures have similar structures to the generator architectures.
In the proof of Theorem 2, we carefully construct the discriminator class using the generator class. In particular, the discriminator class is constructed so that its complexity, expressed through the metric entropy, is of the same order as that of the generator class. Consequently, the discriminator class becomes a much smaller class than $\mathcal{F}_{\text{Lip}}$, which is the one considered in the literature for obtaining a Wasserstein rate. By reducing the complexity of the discriminator class, we can significantly improve the convergence rate.

Finally, we would like to mention that once the statement of Theorem 4 is slightly modified, it might be possible to derive similar lower bounds more easily based on Caffarelli’s regularity theory of optimal transport (Caffarelli, 1990; Urbas, 1988) and minimax theory for density models (Liang, 2021; Niles-Weed & Berthet, 2022; Uppal, Singh, & Póczos, 2019). More specifically, if $P_Z$ is a uniform distribution on a Euclidean ball in $\mathbb{R}^d$ instead of the uniform distribution on the cube $[0, 1]^d$ as in Theorem 4, Caffarelli’s theory provides a useful connection between the density model and generative model, which facilitates an easier proof for the lower bound, see the discussion after Theorem 4 for more details. However, extending this approach to the case where $P_Z = \text{Unif}([0, 1]^d)$ is not straightforward because the uniform convexity of the support of probability measures involved is a key assumption in Caffarelli’s theory. In particular, we may need to construct a sufficiently regular transport map, whose Jacobian determinant is bounded from above and below, from the uniform distribution on a Euclidean ball to the uniform distribution on a cube. Instead of applying the technically involved Caffarelli’s regularity theory, we have chosen to directly construct multiple testing based on generators and apply Fano’s method to obtain the lower bound. We believe this approach is novel and provides a different perspective.

The remainder of the paper is organized as follows. First, we review the literature on the theory of GAN and introduce some notations. Section 2 provides a mathematical set-up, including a brief introduction to DNN and GAN. In Section 3, we discuss the assumption on the true distribution in depth. An upper bound for a convergence rate of a GAN-based estimator and a lower bound of minimax convergence rates are investigated in Sections 4 and 5, respectively. Concluding remarks follow in Section 6. All proofs are provided in Supplement.

1.1 Related statistical theory for GAN

Convergence rates of nonparametric generative models were initially studied in Kundu and Dunson (2014) and Pati, Bhattacharya, and Dunson (2011). Rather than utilizing DNN, they considered a nonparametric Bayesian approach with a Gaussian process prior on the generator function.

Since the development of GAN by Goodfellow et al. (2014), several researchers have studied rates of convergence in deep generative models, particularly focusing on GAN. An earlier version of Liang (2021) was the first one to study the convergence rate under a GAN framework. More specifically, they considered the Sobolev IPMs to evaluate the estimation performance. A similar theory has been developed by Singh et al. (2018), which was later generalized by Uppal et al. (2019) using Besov IPMs. Slightly weaker results were obtained by Chen, Liao, Zha, and Zhao (2020). Although their convergence rates are strictly slower than the minimax optimal rate, they explicitly considered
DNN architectures for the generator and discriminator classes. Convergence rates of
the vanilla GAN with respect to the Jensen–Shannon divergence have recently been
obtained by Belomestny, Moulines, Naumov, Puchkin, and Samsonov (2021).

These works utilized the framework of nonparametric density estimation to under-
stand GAN. They evaluated the performance of GAN using integral probability
metrics, while classical approaches such as the kernel density estimation focused on
other metrics such as the total variation, Hellinger and uniform metrics. Since the
total variation can be viewed as an IPM, some results in the above papers are compa-
rible with that of the classical methods. In these comparable cases, both approaches
achieve the same minimax optimal rate; hence these theories on GAN cannot explain
why deep generative models outperform classical nonparametric methods.

Schreuder, Brunel, and Dalalyan (2021) considered generative models where the
target distribution may not possess a Lebesgue density. They assumed that the true
distribution is the convolution of \(Q_{g_0}\) and a general noise distribution for some function
\(g_0 : [0, 1]^d \rightarrow [0, 1]^D\). While this assumption is similar to ours, it does not explicitly
incorporate the smoothness and composite structure of \(g_0\). As a result, their result
only guarantees that GAN achieves the same rate as the empirical measure. More
recently, Tang and Yang (2023) obtained the minimax rate of distribution estimation
under a submanifold assumption using a mixture of GANs. However, as mentioned
earlier, the composite structure imposed through the generator function in our paper
involves a substantially richer structure than just a manifold structure considered in
Tang and Yang (2023). For instance, the dimension \(t_*\) corresponding to the worst-case
component of a composite function (as defined in (5)) can be much smaller than the
dimension of the manifold on which \(Q_0\) is supported.

1.2 Notations

Maximum and minimum of two real numbers \(a\) and \(b\) are denoted as \(a \lor b\) and \(a \land b\),
respectively. For \(1 \leq p < \infty\), \(\| \cdot \|_p\) denotes the \(\ell^p\)-norm. For a real-valued function \(f\)
and a probability measure \(P\), let \(Pf = \int f(x)dP(x)\). \(\mathbb{E}\) denotes the expectation when
the underlying probability is obvious. Convolution of two probability measures \(P\) and
\(Q\) are denoted \(P * Q\). We write \(c = c(A_1, \ldots, A_k)\) when \(c\) depends only on \(A_1, \ldots, A_k\).
Uppercase letters, such as \(P\) and \(\hat{P}\), refer to probability measures corresponding to
densities denoted by their lowercase counterparts: i.e. \(p\) and \(\hat{p}\), respectively. We write
\(a \preceq b\) if \(a\) is less than \(b\) up to a constant multiplication, where the constant is universal
or at least contextually unimportant. Lastly, \(a \asymp b\) indicates \(a \preceq b\) and \(b \preceq a\).

2 Generative adversarial networks

For a given class \(\mathcal{F}\) of functions from \(\mathbb{R}^D\) to \(\mathbb{R}\), the \(\mathcal{F}\)-IPM (Müller (1997)) between
two probability measures \(P_1\) and \(P_2\) is defined as

\[
\mathcal{F}(P_1, P_2) = \sup_{f \in \mathcal{F}} |P_1f - P_2f|.
\] (1)
For example, if $F = F_{\text{Lip}}$, the class of every function $f : \mathbb{R}^D \to \mathbb{R}$ satisfying $|f(x) - f(y)| \leq |x - y|_2$ for all $x, y \in \mathbb{R}^D$, then the corresponding IPM is the $L^1$-Wasserstein distance by the Kantorovich–Rubinstein duality theorem; see Theorem 1.14 from Villani (2003). Hölder, or more generally Besov, IPMs have been considered in recent articles for evaluating the performance of the density or distribution estimation; see Liang (2021), Uppal et al. (2019), Singh et al. (2018) and Tang and Yang (2023).

Let $G$ be a class of functions from $\mathcal{Z} \subset \mathbb{R}^d$ to $\mathbb{R}^D$, and $F$ be a class of functions from $\mathbb{R}^D$ to $\mathbb{R}$. Two classes $G$ and $F$ are referred to as the generator and discriminator classes, respectively. For given discriminator and generator classes, we define a GAN-based estimator $\hat{g}$ as the minimizer of $d_F(Q_{g}, P_n)$ over $G$, where $P_n$ is the empirical measure based on the $D$-dimensional observations $X_1, \ldots, X_n$. That is, the estimator $\hat{g} \in G$ is such that

$$d_F(Q_{\hat{g}}, P_n) = \inf_{g \in G} d_F(Q_{g}, P_n) + \epsilon_{\text{opt}}. \quad (2)$$

Here, the optimization error $\epsilon_{\text{opt}} \geq 0$ is a prespecified number. An estimator satisfying (2) is of our primary interest. Although the vanilla GAN (Goodfellow et al., 2014) is not of the form (2), the formulation (2) is quite general to include various GANs popularly used in practice (Arjovsky et al., 2017; Li et al., 2017; Mroueh et al., 2017). At the population level, (2) can be viewed as a method to solve the following minimization

$$\minimize_{\hat{g} \in G} d_F(Q_{\hat{g}}, P_0)$$

because one may expect $\mathbb{E}d_F(P_n, P_0) \to 0$, where the expectation is taken with respect to $P_0$. Since the convergence rate for $\mathbb{E}d_F(P_n, P_0)$ might be very slow, however, a careful analysis is necessary. In particular, in Section 4, we will separate the evaluation metric from the $F$-IPM, the one defined through the discriminator class.

In practice, both the generator and discriminator classes are modeled using deep neural networks. To be specific, let $\rho(x) = x \vee 0$ be the ReLU activation function (Glorot, Bordes, & Bengio, 2011). We focus on the ReLU in this paper, but other activation functions can also be used as long as a suitable approximation property holds (Ohn & Kim, 2019). For vectors $v = (v_1, \ldots, v_r)$ and $z = (z_1, \ldots, z_r)$, define $\rho_v(z) = (\rho(z_1 - v_1), \ldots, \rho(z_r - v_r))$. For a nonnegative integer $L$ and $p = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+1}$, a neural network function with the network architecture $(L, p)$ is any function $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ such that

$$z \mapsto f(z) = W_L \rho_{v_L} W_{L-1} \rho_{v_{L-1}} \cdots W_1 \rho_{v_1} W_0 z, \quad (3)$$

where $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$ and $v_i \in \mathbb{R}^{p_i}$. Let $D(L, p, s, F)$ be the collection $f$ from (3) satisfying

$$\max_{j=0, \ldots, L} |W_j|_\infty \vee |\nu_j|_\infty \leq 1, \quad \sum_{j=1}^L |W_j|_0 + |\nu_j|_0 \leq s \quad \text{and} \quad ||f||_\infty \leq F.$$
where $|W_j|_\infty$ and $|W_j|_0$ denote the maximum-entry norm and the number of nonzero elements of the matrix $W_j$, respectively, and $\|f\|_\infty = \|f(z)\|_\infty = \sup_z |f(z)|_\infty$.

When the generator class $\mathcal{G}$ consists of neural network functions, we call the corresponding class $\mathcal{Q} = \{Q_g : g \in \mathcal{G}\}$ as a deep generative model. In this sense, GAN can be viewed as a method for estimating the parameters in deep generative models. In the literature concerning the variational autoencoder, a collection of $P_{g,\sigma}$ is often called a deep generative model as well.

## 3 Assumptions on the true distribution

In this section, we address assumptions on the true distribution $P_0$. As mentioned in the introduction, we assume that $P_0 = P_{g_0,\sigma_0}$ for some function $g_0 : \mathbb{Z} \to \mathbb{R}^D$ and $\sigma_0 \geq 0$. Furthermore, we assume that $g_0$ possesses a structure that allows for efficient estimation of it (or $Q_0 = Q_{g_0}$). Note that it is much more convenient to impose a structure on the generator rather than directly on the density function because there is no constraint on the functional form of the generator.

We suppose that $g_0$ belongs to a class of structured functions. More specifically, we consider a class of composite functions for which deep generative models have benefits. For positive numbers $\beta$ and $K$, let $\mathcal{H}_K^\beta(A)$ be a class of all functions from $A$ to $\mathbb{R}$ with $\beta$-Hölder norm bounded by $K$. See van der Vaart and Wellner (1996) and Giné and Nickl (2016) for the definition of Hölder space. Consider a function $g : \mathbb{R}^d \to \mathbb{R}^D$ as follows:

$$g = h_q \circ h_{q-1} \circ \cdots \circ h_1 \circ h_0 \quad (4)$$

with $h_i : (a_i, b_i)^{d_i} \to (a_{i+1}, b_{i+1})^{d_{i+1}}$ and $h_i = (h_{i,1}, \ldots, h_{i,d_i+1})$. Here, $d_0 = d$ and $d_{q+1} = D$. Let $t_i$ be the maximal number of variables on which each of the $h_{ij}$ depends. Let $\mathcal{G}_0(q, d, t, \beta, K)$ be a collection of functions of the form (4) satisfying $h_{ij} \in \mathcal{H}_K^{\beta_i}((a_l, b_l)^{t_l})$ and $|a_l| \lor |b_l| \leq K$, where $d = (d_0, \ldots, d_q)$, $t = (t_0, \ldots, t_q)$ and $\beta = (\beta_0, \ldots, \beta_{q+1})$. Let

$$\tilde{\beta}_i = \beta_i \prod_{l=i+1}^q (\beta_l \land 1), \quad i_* = \arg\max_{i \in \{0, \ldots, q\}} \frac{t_i}{\tilde{\beta}_i}, \quad \beta_* = \tilde{\beta}_{i_*}, \quad \text{and} \quad t_* = t_{i_*} \quad (5)$$

Note that the decomposition of the form (4) for a given function $g$ may not be unique. For example, the composite function $h_q \circ h_{q-1} \circ \cdots \circ h_1 \circ h_0$ can be seen as a single function $\tilde{h}_0 = (\tilde{h}_0, \ldots, \tilde{h}_0)$ with $\tilde{h}_{0,j} \in \mathcal{H}_K^{\min(\beta_0, \ldots, \beta_{q+1})}((a_0, b_0)^{t_0})$ for a large enough constant $K'$. Also, there might be several maximizers for the map $i \mapsto t_i/\tilde{\beta}_i$. In this case, $i_*$ can be defined as any maximizer.

The class $\mathcal{G}_0 = \mathcal{G}_0(q, d, t, \beta, K)$ has been extensively studied in recent articles on deep supervised learning to demonstrate the benefits of DNN in estimating a non-parametric function (Bauer & Kohler, 2019; Schmidt-Hieber, 2020). As studied in Chae et al. (2023), a composite structure can naturally be translated to unsupervised learning problems through the distribution class $Q_0 = \{Q_g : g \in \mathcal{G}_0\}$. For example, when $d = D$ and $\mathcal{G}_0$ consists of functions of the form $g(z) = (g_1(z_1), \ldots, g_d(z_d))$, ...
where $z = (z_1, \ldots, z_d)$ and $g_j$, $j = 1, \ldots, d$, is a univariate function, then $t_\ast = 1$ and the corresponding $Q_0$ becomes a class of product distributions. If $G_0$ consists of $\beta$-Hölder functions with $\beta > 1$, the support of $P_Z$ is uniformly convex and its density is bounded from above and below, the corresponding $Q_0$ contains distributions possessing a strictly positive $(\beta-1)$-Hölder density on a bounded and uniformly convex subset of $\mathbb{R}^D$. This fact is based on the well-established regularity theory of optimal transport; see Theorem 12.50 of Villani (2008) for details. It is important to note that the uniform convexity assumption cannot be relaxed within Caffarelli’s regularity theory.

In the literature, structured distribution estimation has predominantly been studied within the framework of manifold structure (Puchkin & Spokoiny, 2022; Tang & Yang, 2023). However, the composite structure introduced through the generator function in our approach incorporates various interesting low-dimensional structures that are not captured by the manifold structure alone. For instance, consider the example mentioned earlier, where $g(z) = (g_1(z_1), \ldots, g_d(z_d))$. In this case, the dimension $t_\ast$, defined as the worst-case component of the composite function (as in (5)), can be much smaller than the dimension $d$ of the manifold on which $g(Z)$ is supported. This highlights the richer and more flexible structure captured by the composite approach compared to the manifold structure.

If $d < D$ and $g_0$ is sufficiently smooth, the distribution $Q_0$ is singular with respect to the Lebesgue measure on $\mathbb{R}^D$. However, the distribution $P_0$ possesses a Lebesgue density provided that $\sigma_0 > 0$. We would like to emphasize that the main theorems in Section 4 hold for all values of $\sigma_0$ in the interval $[0, 1]$. With regard to the noise level $\sigma_0$, it would be worthwhile to discuss two different regimes, as also discussed in Section 3.6 of Chae et al. (2023).

Firstly, consider the case where $\sigma_0$ is a fixed positive constant. In this case, our results do not provide a meaningful convergence rate. The problem of estimating $Q_0$ with additive noise is commonly referred to as the deconvolution problem, and it has been extensively studied in the literature (Fan, 1991; Genovese et al., 2012a; Meister, 2009; Nguyen, 2013) under the assumption of fixed $\sigma_0$. It is worth noting that the estimation problem in this setting is intrinsically very difficult, and this difficulty is often expressed mathematically through the logarithmic minimax rates. While a GAN-based estimator might achieve such a logarithmic convergence rate, we do not pursue its study in the present paper, as our primary focus is on the regime where $\sigma_0$ is small.

In particular, we believe that a theory with such a slow convergence rate would not be suitable for explaining the amazing performance of deep generative models.

A small $\sigma_0$ regime can be mathematically expressed as $\sigma_0 \to 0$ with a suitable rate as $n \to \infty$. In this regime, it is possible to obtain a fast convergence rate for estimating $Q_0$, as guaranteed by our theory. While the data-generating distribution $P_0$ depends on the sample size $n$, the theorems in Section 4 hold for all $n$, ensuring clear interpretation of the results. It is worth noting that such sample size-dependent true distributions have been extensively studied in modern high-dimensional statistics (Bühlmann & van de Geer, 2011; Wainwright, 2019), and our setup can be understood within similar contexts. Although our setup and estimation problems may differ slightly, there have been several recent articles that assume data are concentrated around a small neighborhood of a manifold, and these neighborhoods shrink to the manifold as the sample
size increases; see Puchkin and Spokoiny (2022), Aamari and Levraud (2019), Aamari and Levraud (2018), Divol (2021), Jiao, Shen, Lin, and Huang (2023) and Berenfeld, Rosa, and Rousseau (2022) for relevant discussions in this direction.

4 Convergence rate of a GAN-based estimator

Although a strict minimization of the map \( g \mapsto d_F(Q_g, P_n) \) is computationally intractable, several heuristic approaches are available to approximate the solution to (2). In this section, we investigate the convergence rate of \( \hat{Q} = Q_{\hat{g}} \) under the assumption that the computation of it is possible. A goal is to find a sharp upper bound for \( \mathbb{E} d_{\text{eval}}(\hat{Q}, Q_0) \), where \( d_{\text{eval}} \) is a metric evaluating the performance of the estimation. It is desirable that the rate is independent of \( D \) and \( d \), and it adapts to the structure of \( P_0 \). We consider an arbitrary evaluation metric \( d_{\text{eval}} \) for generality; the \( L_1 \)-Wasserstein distance is of primary interest. Recall that the \( L_p \)-Wasserstein distance between two Borel probability measures \( P \) and \( Q \) on \( \mathbb{R}^D \) is defined as

\[
W_p(P, Q) = \inf_\pi \left( \int |x - y|^p d\pi(x, y) \right)^{1/p}
\]

for \( p \geq 1 \), where the infimum is taken over every coupling \( \pi \) of \( P \) and \( Q \). As mentioned earlier, \( W_1 \) allows the IPM representation \( W_1(P, Q) = d_{\mathcal{F}_{\text{Lip}}}(P, Q) \) by the Kantorovich–Rubinstein duality, which makes it convenient to utilize \( W_1 \) as an evaluation metric in a GAN framework. Although a more general duality theorem is well-known for \( p > 1 \) (Villani (2003), Theorem 1.3), the IPM representation of \( W_p \) is available only for \( p = 1 \).

In literature, the evaluation metric is often identified with \( d_F \), the IPM defined through the discriminator class \( \mathcal{F} \). In this sense, when \( d_{\text{eval}} \) is the \( L_1 \)-Wasserstein metric \( W_1 \), a natural candidate for the discriminator class \( \mathcal{F} \) might be \( \mathcal{F}_{\text{Lip}} \), the class of those functions whose Lipschitz constant is bounded by 1. Indeed, it is the original motivation of the Wasserstein GAN to minimize the objective function

\[
W_1(Q_g, P_n) = d_{\mathcal{F}_{\text{Lip}}}(Q_g, P_n)
\]

over \( g \in \mathcal{G} \). In the original article of the Wasserstein GAN (Arjovsky et al., 2017), (6) is minimized after replacing \( \mathcal{F}_{\text{Lip}} \) by a class \( \mathcal{F} \) of neural network functions. The replacement was only for computational tractability. Although minimizing the map \( g \mapsto d_F(Q_g, P_n) \) with a neural network class \( \mathcal{F} \) is still challenging, several heuristic approaches can be employed to approximate the solution.

Suppose that computing the minimizer of (6), say \( \hat{Q}^W \), is possible. It is natural to ask whether \( \hat{Q}^W \) is a decent estimator, theoretically at least. If the generator class \( \mathcal{G} \) is large enough, for example, \( \hat{Q}^W \) is expected to be close to the empirical measure. Consequently, the convergence rates of \( \hat{Q}^W \) and \( P_n \) would be the same. Schreuder et al. (2021) utilized this idea to prove that \( \hat{Q}^W \) performs at least as good as \( P_n \) does.

The convergence rate of the empirical measure with respect to the Wasserstein metric
is well-known in the literature. Fournier and Guillin (2015) have shown that
\[ E_{W_1}(P_n, P_0) \lesssim \begin{cases} \frac{n^{1/2}}{} & \text{if } D = 1 \\ \frac{n^{1/2} \log n}{D} & \text{if } D > 2. \end{cases} \]
(7)

See also Weed and Bach (2019). The rate (7) becomes slower as \( D \) increases, suffering from the curse of dimensionality. Although \( P_n \) adapts to a certain intrinsic dimension and achieves the minimax rate in some sense (Singh & Póczos, 2018; Weed & Bach, 2019), it is possible to construct a more efficient estimator, particularly when the underlying distribution possesses a smooth structure.

\( \hat{Q}^W \) may perform better than the empirical measure if \( G \) is not too large, but the theoretical analysis of this would be quite challenging. Furthermore, practical estimators might be fundamentally different from \( Q^W \) because it does not take crucial features of state-of-the-art methods into account. For successful GAN approaches, for example, the structures of the generator and discriminator architectures are closely related. In particular, the complexities of the two architectures are similar. On the other hand, the discriminator class \( F = F_{\text{Lip}} \), used in the construction of \( \hat{Q}^W \), has no connection with the generator class. In this sense, it is difficult to view \( \hat{Q}^W \) as a suitable estimator to be theoretically analyzed.

In conclusion, \( d_{\text{eval}} \) is not necessarily identical to \( d_{\mathcal{F}} \) in our analysis. Nonetheless, \( d_{\mathcal{F}} \) should be close to \( d_{\text{eval}} \) in some sense because GAN constructs an estimator by minimizing \( d_{\mathcal{F}}(Q_g, P_n) \) over \( g \in \mathcal{G} \). This is specified as condition (iv) of Theorem 1: \( d_{\mathcal{F}} \) needs to be close to \( d_{\text{eval}} \) only on a relatively small class of distributions.

**Theorem 1** Suppose that \( X_1, \ldots, X_n \) are i.i.d. random vectors following \( P_0 = Q_0 \ast N(0_D, \sigma_0^2 I_D) \) for some distribution \( Q_0 \) (not necessarily of the form \( Q_g \)) and \( \sigma_0 \geq 0 \). For given generator class \( \mathcal{G} \), discriminator class \( \mathcal{F} \) and an estimator \( \hat{Q} = Q_{\hat{g}} \) with \( \hat{g} \in \mathcal{G} \), suppose that
\[
\begin{align*}
(\text{i}) \quad & \inf_{g \in \mathcal{G}} d_{\text{eval}}(Q_g, Q_0) \leq \epsilon_1 \\
(\text{ii}) \quad & d_{\mathcal{F}}(\hat{Q}, P_n) \leq \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(Q_g, P_n) + \epsilon_2 \\
(\text{iii}) \quad & \mathbb{E} d_{\mathcal{F}}(P_n, P_0) \leq \epsilon_3 \\
(\text{iv}) \quad & |d_{\text{eval}}(Q_1, Q_2) - d_{\mathcal{F}}(Q_1, Q_2)| \leq \epsilon_4 \quad \forall Q_1, Q_2 \in \mathcal{Q} \cup \{Q_0\},
\end{align*}
\]
where \( \mathcal{Q} = \{Q_g : g \in \mathcal{G}\} \) and \( \epsilon_j \geq 0 \). Then,
\[
\mathbb{E} d_{\text{eval}}(\hat{Q}, Q_0) \leq 2d_{\mathcal{F}}(P_0, Q_0) + 5\epsilon_1 + \epsilon_2 + 2\epsilon_3 + 3\epsilon_4.
\]

Note that similar inequalities to the statement of Theorem 1 have been explicitly or implicitly considered in the literature to analyze the theoretical properties of GANs (Belomestny et al., 2021; Biau, Sangnier, & Tanielian, 2021; Chen et al., 2020; Schreuder et al., 2021). Theorem 1 is a slight modification of those existing results, with the modification favoring our analysis. The proof of Theorem 1 does not significantly differ from the proofs in the literature.
Two quantities $\epsilon_1$ and $\epsilon_3$ are closely related to the complexity of $\mathcal{G}$ and $\mathcal{F}$, respectively. In particular, $\epsilon_1$ represents an error for approximating $Q_0$ by distributions of the form $Q_g$ over $g \in \mathcal{G}$. The larger the generator class $\mathcal{G}$ is, the smaller the approximation error is (Ohn & Kim, 2019; Telgarsky, 2016; Yarotsky, 2017). Similarly, $\epsilon_3$ gets larger as the complexity of $\mathcal{F}$ increases. Techniques for bounding $\mathbb{E}d_\mathcal{F}(\mathbb{P}_n, P_0)$ are well-known in empirical process theory (Giné & Nickl, 2016; van der Vaart & Wellner, 1996). The second error term $\epsilon_2$ is nothing but the optimization error. The fourth term $\epsilon_4$ is the deviance between the evaluation metric $d_{\text{eval}}$ and $\mathcal{F}$-IPM over $\mathcal{Q} \cup \{Q_0\}$, connecting $d_\mathcal{F}$ and $d_{\text{eval}}$.

Finally, the term $d_\mathcal{F}(P_0, Q_0)$ in the assertion of Theorem 1 depends primarily on $\sigma_0$. If $\mathcal{F} \subset \mathcal{F}_{\text{Lip}}$, for example, one can easily prove that

$$d_\mathcal{F}(P_0, Q_0) \leq W_1(P_0, Q_0) \leq W_2(P_0, Q_0) \lesssim \sigma_0,$$

where the third inequality holds by well-known formula (Givens & Shortt, 1984). As another example, if $\mathcal{F}$ consists of twice continuously differentiable functions with suitably bounded derivatives, we have

$$|P_0 f - Q_0 f| = \|\mathbb{E}[f(Y) + \epsilon) - f(Y)]\| \approx \|\mathbb{E}[\epsilon^T \nabla f(Y) + \frac{1}{2} \epsilon^T \nabla^2 f(Y) \epsilon]\| \approx \sigma_0^2,$$

for $f \in \mathcal{F}$ where $Y \sim Q_0$ and $\epsilon \sim \mathcal{N}(\mathbf{0}_D, \sigma_0^2 \mathbf{I})$ are independent random vectors. Hence, $d_\mathcal{F}(P_0, Q_0) \lesssim \sigma_0^2$, which gives a better bound than (9) for a small enough $\sigma_0$.

Ignoring the optimization error, suppose for a moment that $\mathcal{G}$ is given and we need to choose a suitable discriminator class to minimize $\epsilon_3 + \epsilon_4$ in Theorem 1. We focus on the case of $d_{\text{eval}} = W_1$. One can easily make $\epsilon_4 = 0$ by taking $\mathcal{F} = \mathcal{F}_{\text{Lip}}$. In this case, however, $\epsilon_3$ would be too large because $\mathbb{E}W_1(\mathbb{P}_n, P_0) \approx n^{-1/2}$ for $D > 2$; cf. Eq. (7). That is, $\mathcal{F}_{\text{Lip}}$ might be too large to be used as a discriminator class. The discriminator class $\mathcal{F}$ should be much smaller than $\mathcal{F}_{\text{Lip}}$ to obtain a fast convergence rate. To achieve this goal, we construct $\mathcal{F}$ so that both $\epsilon_3$ and $\epsilon_4$ are small enough. Such discriminator class can be constructed as, for example,

$$\mathcal{F} = \left\{ f_{Q_1, Q_2} : Q_1, Q_2 \in \mathcal{Q} \cup \{Q_0\} \right\},$$

where $f_{Q_1, Q_2}$ is a (approximate) maximizer of $|Q_1 f - Q_2 f|$ over $f \in \mathcal{F}_{\text{Lip}}$. In this case, $\epsilon_4$ vanishes and the convergence rate of $\hat{Q}$ will be determined solely by $\epsilon_1$, $\epsilon_3$ and $\sigma_0$. Furthermore, the complexity of $\mathcal{F}$ would roughly be the same as that of $\mathcal{G} \times \mathcal{G}$. If the complexity of a function class is expressed through a metric entropy, the logarithmic covering number, the complexities of $\mathcal{G}$ and $\mathcal{F}$ are of the same order. In this case, three quantities $\epsilon_1$, $\epsilon_3$ and $\sigma_0$ can roughly be interpreted as the approximation error, estimation error and noise level, respectively. While we cannot control the noise level $\sigma_0$, both the approximation and estimation errors depend on the complexity of $\mathcal{G}$. Hence, a suitable choice of it is important to achieve a fast convergence rate.
As described in Section 3, we consider a class of composite functions for the true generator. Let

\[ Q_0 = Q_0(q, d, t, \beta, K) = \left\{ Q_g : g \in G_0(q, d, t, \beta, K) \right\}, \]

where \( G_0 = G_0(q, d, t, \beta, K) \) is defined as in Section 3. Although not strictly necessary, it would be convenient to regard quantities \((q, d, t, \beta, K)\) as constants independent of \(n\). In the forthcoming Theorem 2, we obtain a Wasserstein convergence rate of \( \hat{Q} \) under the assumption that \( Q_0 \in G_0 \).

**Theorem 2** Suppose that \( X_1, \ldots, X_n \) are i.i.d. random vectors following \( P_0 = P_{\mathbf{g}, \sigma} = Q_{\mathbf{g}_0, \sigma_0} \) for some \( \sigma_0 \leq 1 \) and \( \mathbf{g}_0 \in G_0(q, d, t, \beta, K) \). Then, there exist a generator class \( G = \mathcal{D}(L, p, s, K \lor 1) \) and a discriminator class \( F \subset \mathcal{F}_{\text{Lip}} \) such that for an estimator \( \hat{Q} \) satisfying (2),

\[
\sup_{Q_0 \in Q_0} \mathbb{E}W_1(\hat{Q}, Q_0) \leq C \left\{ n^{-\frac{\beta_\ast}{2\beta_\ast + \gamma}} (\log n)^{\frac{3\beta_\ast}{2\beta_\ast + \gamma}} + \sigma_0 + \epsilon_{\text{opt}} \right\},
\]

where \( C = C(q, d, t, \beta, K) \).

Theorem 2 only considers a Gaussian additive noise, but the assumption \( P_0 = P_{\mathbf{g}_0, \sigma_0} = Q_{\mathbf{g}_0, \sigma_0} \approx \mathcal{N}(\mathbf{0}_D, \sigma_0^2 I_D) \) can be relaxed in various ways. In the proof of Theorem 2, with regard to the data distribution \( P_0 \), we only need a bound \( d_{\pi}(P_0, Q_0) \lesssim \sigma_0 \) as in (9) and that \( f(X_i) \) is a sub-Gaussian variable for every \( f \in \mathcal{F}_{\text{Lip}} \), with \( f(\mathbf{0}_D) = 0 \), where the sub-Gaussian parameter \( \sigma \) is independent of \( f \). Therefore, for example, the normal distribution \( \mathcal{N}(\mathbf{0}_D, \sigma^2 I_D) \) for the noise distribution can be replaced by any sub-Gaussian distribution with variance \( \sigma^2 \).

In Theorem 2, both the generator class \( G \) and discriminator class \( F \) depend solely on the sample size \( n \) and the parameters \((q, d, t, \beta, K)\), independent of \( Q_0 \) or \( P_0 \). Moreover, from the proof, it can be deduced that the network parameters \((L, p, s)\) of the generator class can be chosen such that \( L \lesssim \log n, |p|_{\infty} \lesssim n^{\gamma}/(2\beta_\ast + \gamma) \) and \( s \lesssim n^{\gamma}/(2\beta_\ast + \gamma) \log n \), where the constants in \( \lesssim \) depend only on \((q, d, t, \beta, K)\).

Ignoring the optimization error \( \epsilon_{\text{opt}} \), the rate (12) consists of the two terms, \( \sigma_0 \) and \( n^{-\beta_\ast}/(2\beta_\ast + \gamma) \), up to a logarithmic factor. If \( \sigma_0 \lesssim n^{-\beta_\ast}/(2\beta_\ast + \gamma) \), it can be absorbed into the polynomial term. Therefore, \( \hat{Q} \) achieves the rate \( n^{-\beta_\ast}/(2\beta_\ast + \gamma) \) when \( \sigma_0 \) is small enough. Note that this rate often appears in nonparametric smooth function estimation, balancing the approximation and estimation errors.

Under the condition given in Theorem 2, Chae et al. (2023) considered a likelihood approach to study the benefit of the deep generative model. More specifically, they obtained a convergence rate of a sieve MLE based on a Gaussian mixture density \( p_{\mathbf{g}, \sigma} \). Note that the density \( p_{\mathbf{g}, \sigma} \) is concentrated around a small neighborhood of a low-dimensional structure induced by \( \mathbf{g} \). As a result, likelihood approaches might be highly unstable due to the singularity issue. To overcome this problem, Chae et al. (2023) considered a sieve MLE based on the perturbed data \( \tilde{X}_i = X_i + \tilde{e}_i \), where \( \tilde{e}_i \) is an artificial noise alleviating the problem caused by the singularity. They proved that a
sieve MLE with an optimal perturbation achieves the Wasserstein rate $n^{-\beta_*/2(\beta_*+t_*)}+\sigma_0$. The rate was obtained based on the perturbed data, thus it was conjectured to be suboptimal. Theorem 2 shows that GAN achieves a strictly faster convergence rate than $n^{-\beta_*/2(\beta_*+t_*)}+\sigma_0$, the one obtained by Chae et al. (2023). Hence, the rate of a likelihood approach in Chae et al. (2023) is sub-optimal.

In some special cases, the convergence rate of a GAN-based estimator obtained from Theorem 2 can be strictly worse than the rate achieved by the empirical measure. For instance, when $\beta_*=1$, $D=d=t_*>2$, and $\sigma_0=\epsilon_{\text{opt}}=0$, the rate (12) simplifies to $n^{-1/(d+2)}$ (up to a logarithmic factor), while the empirical measure $\mathbb{P}_n$ achieves a strictly faster rate of $n^{-1/d}$, as shown in (7). However, by choosing different generator and discriminator classes, it is possible to obtain a GAN-based estimator with a convergence rate equal to that of the empirical measure. Specifically, we can apply Theorem 1 with $F=F_{\text{Lip}}$. In this case, $\epsilon_4=0$ since $d_{\text{eval}}=W_1$. Moreover, by selecting a large enough $G$, i.e. increasing the depth, width, and number of nonzero parameters, we can make $\epsilon_1$ arbitrarily small. As a result, Theorem 1 and Eq. (9) yield the bound

$$\mathbb{E}W_1(\hat{Q},Q_0) \lesssim \sigma_0+\epsilon_{\text{opt}}+\mathbb{E}_W(\mathbb{P}_n,\mathbb{P}_0).$$

Choosing alternative generator and discriminator classes results in an estimator with a convergence rate better than that in Theorem 2. We consider this alternative choice in the statement of Theorem 3.

So far, we have focused on the case $d_{\text{eval}}=W_1$. In the remainder of this section, we consider a general IPM as an evaluation metric. The function space defining the evaluation metric will be denoted $F_0$, hence $d_{\text{eval}}=d_{F_0}$.

**Theorem 3** Suppose that $X_1, \ldots, X_n$ are i.i.d. random vectors following $P_0=P_{g_0,\sigma}$ for some $g_0 \in \mathcal{G}_0(q,d,t,\beta,K)$ and $\sigma_0 \leq 1$. Let $F_0$ be a class of Lipschitz continuous functions from $\mathbb{R}^D$ to $\mathbb{R}$ with Lipschitz constant bounded by a constant $C_1>0$. Then, there exist a generator class $G=D(L,p,s,K\lor 1)$ and a discriminator class $F$ such that $\hat{Q}$ defined as in (2) satisfies

$$\sup_{Q_0 \in \mathcal{Q}_0} \mathbb{E}d_{F_0}(\hat{Q},Q_0) \leq C_2 \left\{ \sigma_0 + \epsilon_{\text{opt}} + n^{-\frac{\beta_*}{2(\beta_*+t_*)}} (\log n)^{\frac{3\beta_*}{2(\beta_*+t_*)}} \right\},$$

where $C_2=C_2(q,d,t,\beta,K,C_1)$. Alternatively, if we take $F=F_0$ and the depth, width and number of nonzero parameters of $G$ are large enough, the estimator satisfies

$$\mathbb{E}d_{F_0}(\hat{Q},Q_0) \leq C_3 \left\{ \sigma_0 + \epsilon_{\text{opt}} + \mathbb{E}d_{F_0}(\mathbb{P}_n,\mathbb{P}_0) \right\},$$

where $C_3=C_3(D,C_1)$. 

14
Note that the rate in (14) is slower than that in (15) if
\[ \mathbb{E}d_{\mathcal{F}_0}(P_n, P_0) \lesssim n^{-\frac{\beta_i}{\beta_i + t_i}} \log n^{\frac{3\pi}{\beta_i + t_i}}. \]

It is unclear whether it is possible to construct an estimator which achieves the rate of the minimum of two rates in (14) and (15). This interesting problem is left as a topic for future research.

When \( \mathcal{F}_0 \) consists of neural networks, \( \mathcal{F}_0 \)-IPM is often called a neural network distance. Although it is not a standard choice, neural network distances can serve as an evaluation metric. In particular, convergence in a neural network distance guarantees a weak convergence under mild assumptions (Zhang et al., 2018). If \( \mathcal{F}_0 = \mathcal{D}(L_0, p_0, \infty) \), then it is not difficult to see that \( \mathbb{E}d_{\mathcal{F}_0}(P_n, P_0) \lesssim \sqrt{s_0/n} \) up to a logarithmic factor. This can be proved using a well-known empirical process theory combined with metric entropy of deep neural networks (See Lemma 5 from Schmidt-Hieber (2020)). Therefore, if \( s_0 \gg n^{1/2(2\beta_0 + t_0)} \), the right hand side of (14) provides a strictly faster rate than (15).

Another important class of metrics is a H"{o}lder IPM. When \( \mathcal{F}_0 = \mathcal{H}_\alpha([-K, K]^D) \) for some \( \alpha > 0 \), it is well-known that
\[
\mathbb{E}d_{\mathcal{F}_0}(P_n, P_0) \lesssim \begin{cases} 
    n^{-\alpha/D} & \text{if } \alpha < D/2 \\
    n^{-1/2} \log n & \text{if } \alpha = D/2 \\
    n^{-1/2} & \text{if } \alpha > D/2,
\end{cases}
\]
see Schreuder (2021), for example. Similar bounds can be obtained for more general Besov IPMs. Hence, when \( \alpha/D < \beta_0/(2\beta_0 + t_0) < 1/2 \), the rate provided by the right-hand side of (14) is strictly faster than that of (15).

5 Lower bound of the minimax risk

In this section, we study a lower bound for the minimax optimal rate, particularly focusing on the case \( d_{\text{eval}} = W_1 \). With \( P_Z \) the uniform distribution on \([0,1]^d\), we investigate the minimax optimal rate for the distribution class \( Q_0 = \{Q_g : g \in \mathcal{G}_0\} \), where \( \mathcal{G}_0 = \mathcal{G}_0(q, d, t, \beta, K) \). Our analysis is focused on the regime where \( t_i \leq \min\{d_0, \ldots, d_0+1\} \) and \( \beta_i \geq 1 \) for all \( i \). Beyond this regime, obtaining a lower bound using our proof technique becomes challenging. Note that \( \beta_i = \beta_i \) for all \( i \) in this regime.

For given \( \mathcal{G}_0 \) and \( \sigma_0 \geq 0 \), the minimax risk is defined as
\[ \mathcal{M}(\mathcal{G}_0, \sigma_0) = \inf_{\mathcal{Q}} \sup_{\mathcal{Q}_0 \in \mathcal{G}_0} \mathbb{E}W_1(\mathcal{Q}, Q_0), \]
where the infimum ranges over all possible estimators. Although the exact value of \( \mathcal{M}(\mathcal{G}_0, \sigma_0) \) is rarely available in nonparametric problems, several techniques are known in the literature to obtain a lower bound of it. We refer to Tsybakov (2008) and Wainwright (2019) for a comprehensive review. We will utilize a general technique known as Fano’s method to obtain a lower bound.
Theorem 4 Suppose that \(d \leq D\), \(\sigma_0 \geq 0\), \(t_i \leq \min\{d_0, \ldots, d_{q+1}\}\), \(\beta_i \geq 1\) for all \(i\), \(P_Z\) is the uniform distribution on \([0, 1]^d\) and \(\mathcal{G}_0 = \mathcal{G}_0(q, d, t, \beta, K)\). If \(K\) is large enough (depending on \(\beta\) and \(d\)), there exists a constant \(C > 0\) such that

\[
\mathcal{M}(\mathcal{G}_0, \sigma_0) \geq C \max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{2\beta_i + t_i - 2}}. \tag{16}
\]

Note that the lower bound (16) does not depend on \(\sigma_0\). With a direct application of Le Cam’s method, one can easily show that \(\mathcal{M}(\mathcal{G}_0, \sigma_0) \gtrsim \sigma_0 / \sqrt{n}\), hence

\[
\mathcal{M}(\mathcal{G}_0, \sigma_0) \geq C \left( \max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{2\beta_i + t_i - 2}} + \frac{\sigma_0}{\sqrt{n}} \right).
\]

Since we are particularly interested in the small \(\sigma_0\) regime (i.e. nearly singular cases), our discussion below focuses on the case where \(\sigma_0\) is small enough.

Firstly, note that the rate in the right hand side of (16) can be strictly larger than \(n^{-\frac{\beta_i}{2\beta_i + t_i - 2}}\), where \(t_i^*\) and \(\beta_i^*\) are defined in (5). If \((\beta_i^*, t_i^*) = (1, 2)\) and \((\beta_i, t_i) = (1.6, 3)\), for example, then \(t_i^* / \beta_i^* > t_i / \beta_i\) and \(\beta_i^*/(2\beta_i^* + t_i^* - 2) < \beta_i/(2\beta_i + t_i - 2)\). However, the rate in (16) cannot be larger than \(n^{-\frac{\beta_i}{2\beta_i + t_i}}\), which is the convergence rate (up to a logarithmic factor) in Theorem 2. To provide a more convenient comparison of the upper and lower bounds of the convergence rate, we can express the bounds as follows

Upper bound in Theorem 2: \(\max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{2\beta_i + t_i}}\).

Lower bound in Theorem 4: \(\max_{i \in \{0, \ldots, q\}} n^{-\frac{\beta_i}{2\beta_i + t_i - 2}}\).

Therefore, the lower bound is only slightly smaller than the upper bound, indicating that the convergence rate of a GAN-based estimator is very close to the minimax optimal rate.

Regarding the difference between the upper and lower bounds, we conjecture that the lower bound is sharp in at least some special cases, and thus cannot be improved in general. In particular, when \(q = 0\) and \(t_0 = d = D\), we believe that the lower bound in Theorem 4 is sharp. This conjecture is based on the results presented in Uppal et al. (2019) and Liang (2021). They considered GAN for nonparametric density estimation, i.e. \(D = d\) in their framework. For example, Theorem 4 in Liang (2021) guarantees that, for \(D = d \geq 2\) and \(\sigma_0 = 0\),

\[
\inf_{\mathcal{Q}} \sup_{Q_0 \in \mathcal{Q}_0} \mathrm{EW}_1(\hat{Q}, Q_0) \asymp n^{-\frac{\beta^* + 1}{2\beta^* + d}}, \tag{17}
\]

where

\[
\mathcal{Q}_0 = \left\{ Q : q \in \mathcal{H}_1^{\beta^*}([0, 1]^d) \right\}.
\]
More precisely, they considered Sobolev classes rather than Hölder classes. We also refer to Niles-Weed and Berthet (2022) for similar results but using different proof techniques. Interestingly, when $D = d$, there is a close connection between the density model $\hat{Q}_0$ and the generative model $Q_0 = \{Q_g: g \in H^\beta_K([0,1]^d)\}$. This connection is based on Caffarelli's regularity theory of optimal transport, often referred to as the Brenier map. Roughly speaking, for a certain $\beta'$-Hölder density $q$, there exists a $(\beta' + 1)$-Hölder function $g$ such that $Q = Q_g$. Therefore, a density model consisting of densities with this property can be viewed as a sub-model of the generative model with $\beta = \beta' + 1$. It is noteworthy that the rate (17) is the same as the right-hand side of (16) with $q = 0$, $d = D = t_0$, and $\beta_0 = \beta = \beta' + 1$. This is why we conjecture that the lower bound (16) cannot be improved in general. It is important to note that this argument is a conjecture because Caffarelli’s regularity theory requires the uniform convexity of the domain and co-domain of $g$, but $[0,1]^d$ is not uniformly convex. For rigorous statements, counterexamples, and historical background on this field, we refer readers to Chapter 12 of Villani (2008). Additionally, Cordero-Erausquin and Figalli (2019) provides some recent advancements in this area.

The minimax optimal rate in Tang and Yang (2023) is consistent with the lower bound (16) in some special cases with $q = 0$. However, the proof techniques presented in the existing literature are not directly applicable to the structured distribution estimation problem considered in our paper. The techniques employed in Uppal et al. (2019) and Liang (2021) for both upper and lower bounds rely on wavelet thresholding. It is unclear how to extend these techniques to our case with $q > 0$ and a singular $Q_0$. Additionally, Tang and Yang (2023) also relies on wavelet thresholding for estimating a distribution on a manifold, which is similar to the techniques used in Uppal et al. (2019) and Liang (2021), making them not fully applicable to our specific estimation problem. Although Tang and Yang (2023) incorporates an additional step of estimating the charts of the manifold, these techniques do not directly address our estimation problem.

6 Conclusion

Under a structural assumption on the generator, we investigated a convergence rate of a GAN-based estimator and a lower bound of the minimax optimal rate. Notably, the rate is faster than that obtained by likelihood approaches. In practice, however, the computation of GAN incorporates a challenging minimax optimization problem and our understanding of it remains largely unexplored. For example, it is unclear where a practical estimator constructed via a stochastic gradient algorithm converges to (Hsieh, Mertikopoulos, & Cevher, 2021; Mescheder, Geiger, & Nowozin, 2018). The discriminator constructed in the proof of Theorem 2 is even further away from the one used in practice. Our theory only guarantees that there exists a discriminator class $\mathcal{F}$ which yields an estimator whose convergence rate is close to the minimax optimal rate. Regardless, our theory plays an important role in further advancing GAN theory.

We conclude the paper with some possible directions for future work. One of the most important tasks is to reduce the current gap between the upper and lower bounds of the convergence rate. As discussed in Section 5, it would be crucial to construct an
estimator that achieves the lower bound in Theorem 4. After an early version of this paper was drafted on arXiv, Stéphanovitch, Aamari, and Levrard (2023) studied a similar problem, particularly focusing on the special case where \( q = 0 \) and \( t_0 = d \). They obtained a minimax optimal estimator, but its construction relies on wavelet features rather than DNNs. Furthermore, their proof techniques cannot be extended to more general cases where \( q > 0 \). Techniques from the literature concerning the estimation of optimal transport maps might be employed to address this problem, as explored in works such as Deb, Ghosal, and Sen (2021), Hütter and Rigollet (2021), Divol, Niles-Weed, and Pouladian (2022), Manole, Balakrishnan, Niles-Weed, and Wasserman (2021) and Pouladian and Niles-Weed (2021). The problem of estimating optimal transport maps appears to be closely related to our set-up, and the rate (4) can be found in this literature. Investigating whether a GAN-based estimator can achieve the minimax rate when \( d_{\text{eval}} = W_1 \) is another important research problem. In particular, it would be valuable to explore whether the discriminator and generator classes modeled by deep neural networks can attain the minimax rate when \( d_{\text{eval}} = W_1 \). Finally, based on the approximation property of the convolutional neural networks (CNN) architectures (Kohler, Krzyzak, & Walter, 2020; Yarotsky, 2021), studying the benefit of CNN-based GAN would be an intriguing problem.

**Acknowledgments**

The author would like to thank Lizhen Lin for her valuable comments and discussions on an earlier version of this paper. This work was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA2101-03.

**References**

Aamari, E., & Levrard, C. (2018). Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. *Discrete Comput. Geom.*, 59(4), 923–971.

Aamari, E., & Levrard, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation. *Ann. Statist.*, 47(1), 177–204.

Arjovsky, M., Chintala, S., Bottou, L. (2017). Wasserstein generative adversarial networks. *Proc. ICML* (pp. 214–223).

Arora, S., Ge, R., Liang, Y., Ma, T., Zhang, Y. (2017). Generalization and equilibrium in generative adversarial nets (GANs). *Proc. ICML* (pp. 224–232).

Bai, Y., Ma, T., Risteski, A. (2019). Approximability of discriminators implies diversity in GANs. *Proc. ICLR* (pp. 1–10).
Bauer, B., & Kohler, M. (2019). On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *Ann. Statist.*, 47(4), 2261–2285.

Belomestny, D., Moulines, E., Naumov, A., Puchkin, N., Samsonov, S. (2021). Rates of convergence for density estimation with GANs. *ArXiv:2102.00199*, .

Berenfeld, C., & Hoffmann, M. (2021). Density estimation on an unknown submanifold. *Electron. J. Stat.*, 15, 2179–2223.

Berenfeld, C., Rosa, P., Rousseau, J. (2022). Estimating a density near an unknown manifold: a Bayesian nonparametric approach. *ArXiv:2205.15717*, .

Biau, G., Sangnier, M., Tanielian, U. (2021). Some theoretical insights into Wasserstein GANs. *J. Mach. Learn. Res.*, 22(1), 5287–5331.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Caffarelli, L.A. (1990). Interior $W^{2,p}$ estimates for solutions of the Monge–Ampère equation. *Ann. of Math.*, 131(1), 135–150.

Chae, M., & Walker, S.G. (2019). Bayesian consistency for a nonparametric stationary Markov model. *Bernoulli*, 25(2), 877–901.

Chae, M., Kim, D., Kim, Y., Lin, L. (2023). A likelihood approach to nonparametric estimation of a singular distribution using deep generative models. *J. Mach. Learn. Res.*, 24(77), 1–42.

Chen, M., Liao, W., Zha, H., Zhao, T. (2020). Statistical guarantees of generative adversarial networks for distribution estimation. *ArXiv:2002.03938*, .

Cordero-Erausquin, D., & Figalli, A. (2019). Regularity of monotone transport maps between unbounded domains. *Discrete Contin. Dyn. Syst.*, 39(12), 7101-7112.

Deb, N., Ghosal, P., Sen, B. (2021). Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. *Proc. NeurIPS* (Vol. 34, pp. 29736–29753).
Divol, V. (2020). Minimax adaptive estimation in manifold inference. ArXiv:2001.04896,

Divol, V. (2021). Minimax adaptive estimation in manifold inference. Electron. J. Stat., 15(2), 5888–5932,

Divol, V. (2022). Measure estimation on manifolds: an optimal transport approach. Probab. Theory Related Fields, 183(1), 581–647,

Divol, V., Niles-Weed, J., Pooladian, A.-A. (2022). Optimal transport map estimation in general function spaces. ArXiv:2212.03722,

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. Ann. Statist., 19(3), 1257–1272,

Fournier, N., & Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. Probab. Theory Related Fields, 162(3-4), 707–738,

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L. (2012a). Manifold estimation and singular deconvolution under Hausdorff loss. Ann. Statist., 40(2), 941–963,

Genovese, C.R., Perone-Pacifico, M., Verdinelli, I., Wasserman, L. (2012b). Minimax manifold estimation. J. Mach. Learn. Res., 13(1), 1263–1291,

Ghosal, S., & van der Vaart, A. (2017). Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press.

Giné, E., & Nickl, R. (2016). Mathematical foundations of infinite-dimensional statistical models. Cambridge University Press.

Givens, C.R., & Shortt, R.M. (1984). A class of Wasserstein metrics for probability distributions. The Michigan Mathematical Journal, 31(2), 231–240,

Glorot, X., Bordes, A., Bengio, Y. (2011). Deep sparse rectifier neural networks. Proc. AISTATS (pp. 315–323).
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. *Proc. NIPS* (pp. 2672–2680).

Hastie, T., Tibshirani, R., Friedman, J.H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Horowitz, J.L., & Mammen, E. (2007). Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. *Ann. Statist.*, 35(6), 2589–2619.

Hsieh, Y.-P., Mertikopoulos, P., Cevher, V. (2021). The limits of min-max optimization algorithms: Convergence to spurious non-critical sets. *Proc. ICML* (pp. 4337–4348).

Hütter, J.-C., & Rigollet, P. (2021). Minimax estimation of smooth optimal transport maps. *Ann. Statist.*, 49, 1166–1194.

Jiao, Y., Shen, G., Lin, Y., Huang, J. (2023). Deep nonparametric regression on approximate manifolds: Nonasymptotic error bounds with polynomial prefactors. *Ann. Statist.*, 49(2), 691 – 716.

Juditsky, A.B., Lepski, O.V., Tsybakov, A.B. (2009). Nonparametric estimation of composite functions. *Ann. Statist.*, 37(3), 1360–1404.

Karras, T., Aila, T., Laine, S., Lehtinen, J. (2018). Progressive growing of GANs for improved quality, stability, and variation. *Proc. ICLR* (pp. 1–26).

Karras, T., Laine, S., Aila, T. (2019). A style-based generator architecture for generative adversarial networks. *Proc. CVPR* (pp. 4401–4410).

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational Bayes. *Proc. ICLR* (pp. 1–14).

Kohler, M., Krzyzak, A., Walter, B. (2020). On the rate of convergence of image classifiers based on convolutional neural networks. *ArXiv:2003.01526*, .

Kundu, S., & Dunson, D.B. (2014). Latent factor models for density estimation. *Biometrika*, 101(3), 641–654.

Li, C.-L., Chang, W.-C., Cheng, Y., Yang, Y., Póczos, B. (2017). MMD GAN: Towards deeper understanding of moment matching network. *Proc. NIPS* (pp. 21.
Liang, T. (2021). How well generative adversarial networks learn distributions. *J. Mach. Learn. Res.*, 22(228), 1–41.

Liu, S., Bousquet, O., Chaudhuri, K. (2017). Approximation and convergence properties of generative adversarial learning. *Proc. NIPS* (pp. 5545–5553).

Manole, T., Balakrishnan, S., Niles-Weed, J., Wasserman, L. (2021). Plugin estimation of smooth optimal transport maps. *ArXiv:2107.12364*,

Meister, A. (2009). *Deconvolution Problems in Nonparametric Statistics*. Springer, New York.

Mescheder, L., Geiger, A., Nowozin, S. (2018). Which training methods for GANs do actually converge? *Proc. ICML* (pp. 3481–3490).

Mroueh, Y., Li, C.-L., Sercu, T., Raj, A., Cheng, Y. (2017). Sobolev GAN. *ArXiv:1711.04894*,

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Adv. in Appl. Probab.*, 29(2), 429–443.

Nguyen, X. (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Statist.*, 41(1), 370–400.

Niles-Weed, J., & Berthet, Q. (2022). Minimax estimation of smooth densities in Wasserstein distance. *Ann. Statist.*, 50(3), 1519–1540.

Ohn, I., & Kim, Y. (2019). Smooth function approximation by deep neural networks with general activation functions. *Entropy*, 21(7), 627.

Pati, D., Bhattacharya, A., Dunson, D.B. (2011). Posterior convergence rates in non-linear latent variable models. *ArXiv:1109.5000*,

Pooladian, A.-A., & Niles-Weed, J. (2021). Entropic estimation of optimal transport maps. *ArXiv:2109.12004*,

22
Puchkin, N., & Spokoiny, V.G. (2022). Structure-adaptive manifold estimation. *J. Mach. Learn. Res.*, 23, 1–62.

Radford, A., Metz, L., Chintala, S. (2016). Unsupervised representation learning with deep convolutional generative adversarial networks. *Proc. ICLR* (pp. 1–16).

Rezende, D.J., Mohamed, S., Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. *Proc. ICML* (pp. 1278–1286).

Schmidt-Hieber, J. (2020). Nonparametric regression using deep neural networks with ReLU activation function. *Ann. Statist.*, 48(4), 1875–1897.

Schreuder, N. (2021). Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *Math. Methods Statist.*, 29, 76–86.

Schreuder, N., Brunel, V.-E., Dalalyan, A. (2021). Statistical guarantees for generative models without domination. *Proc. Algorithmic Learning Theory* (pp. 1051–1071).

Singh, S., & Póczos, B. (2018). Minimax distribution estimation in Wasserstein distance. *ArXiv:1802.08855*,.

Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M., Póczos, B. (2018). Nonparametric density estimation with adversarial losses. *Proc. NeurIPS* (pp. 10246–10257).

Stéphanovitch, A., Aamari, E., Levrard, C. (2023). Wasserstein generative adversarial networks are minimax optimal distribution estimators. *ArXiv:2311.18613*,.

Tang, R., & Yang, Y. (2023). Minimax rate of distribution estimation on unknown submanifolds under adversarial losses. *Ann. Statist.*, 51(3), 1282–1308.

Telgarsky, M. (2016). Benefits of depth in neural networks. *Proc. COLT* (pp. 1517–1539).

Tsybakov, A.B. (2008). *Introduction to Nonparametric Estimation*. Springer, New York.

Uppal, A., Singh, S., Póczos, B. (2019). Nonparametric density estimation and convergence of GANs under Besov IPM losses. *Proc. NeurIPS* (pp. 9089–9100).

Urbas, J.I. (1988). Regularity of generalized solutions of Monge–Ampère equations. *Math. Z.*, 197(3), 365–393.
van der Vaart, A.W., & Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. Springer.

Villani, C. (2003). *Topics in Optimal Transportation*. American Mathematical Society.

Villani, C. (2008). *Optimal Transport: Old and New*. Springer.

Wainwright, M.J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press.

Weed, J., & Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernoulli*, 25(4A), 2620–2648.

Wei, Y., & Nguyen, X. (2022). Convergence of de Finetti’s mixing measure in latent structure models for observed exchangeable sequences. *Ann. Statist.*, 50(4), 1859–1889.

Wong, W.H., & Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, 23(2), 339–362.

Yalçın, İ., & Amemiya, Y. (2001). Nonlinear factor analysis as a statistical method. *Statist. Sci.*, 16(3), 275–294.

Yarotsky, D. (2017). Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94, 103–114.

Yarotsky, D. (2021). Universal approximations of invariant maps by neural networks. *Constr. Approx.*, 1–68.

Zhang, P., Liu, Q., Zhou, D., Xu, T., He, X. (2018). On the discrimination-generalization tradeoff in GANs. *Proc. ICLR* (pp. 1–26).
A Proof of Theorem 1

Choose \( g_\ast \in \mathcal{G} \) such that

\[
d_{\text{eval}}(Q_\ast, Q_0) \leq \inf_{g \in \mathcal{G}} d_{\text{eval}}(Q_g, Q_0) + \epsilon_1 \leq 2\epsilon_1, \tag{18}\]

where \( Q_\ast = Q_{g_\ast} \). Then,

\[
d_{\text{eval}}(\hat{Q}, Q_0) \leq d_{\text{eval}}(\hat{Q}, Q_\ast) + d_{\text{eval}}(Q_\ast, Q_0) \leq d_{\mathcal{F}}(\hat{Q}, P_n) + d_{\mathcal{F}}(P_n, Q_\ast) + 2\epsilon_1 + \epsilon_4 \tag{iv}
\]

\[
\leq \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(Q_g, P_n) + d_{\mathcal{F}}(P_n, Q_\ast) + 2\epsilon_1 + \epsilon_2 + \epsilon_4 \tag{ii}
\]

\[
\leq \inf_{g \in \mathcal{G}} d_{\mathcal{F}}(Q_g, Q_0) + d_{\mathcal{F}}(P_n, Q_0) + d_{\mathcal{F}}(P_n, P_0) + d_{\mathcal{F}}(P_0, Q_0) + d_{\mathcal{F}}(Q_0, Q_\ast) + 2\epsilon_1 + \epsilon_2 + \epsilon_4 \tag{iv}
\]

\[
\leq \inf_{g \in \mathcal{G}} d_{\text{eval}}(Q_g, Q_0) + d_{\mathcal{F}}(P_n, Q_0) + d_{\mathcal{F}}(P_n, P_0) + d_{\mathcal{F}}(P_0, Q_0) + d_{\mathcal{F}}(Q_0, Q_\ast) + 2\epsilon_1 + \epsilon_2 + 2\epsilon_4 \tag{i}
\]

\[
\leq d_{\mathcal{F}}(P_n, Q_0) + d_{\mathcal{F}}(P_n, P_0) + d_{\mathcal{F}}(P_0, Q_0) + d_{\mathcal{F}}(Q_0, Q_\ast) + 3\epsilon_1 + \epsilon_2 + 2\epsilon_4 \leq 2d_{\mathcal{F}}(P_n, P_0) + 2d_{\mathcal{F}}(P_0, Q_0) + d_{\mathcal{F}}(Q_0, Q_\ast) + 3\epsilon_1 + \epsilon_2 + 2\epsilon_4 \tag{iv}
\]

By taking the expectation, we complete the proof. \( \square \)
B Proof of Theorem 2

We will construct a generator class \( G \) and a discriminator \( F \) satisfying condition (8) of Theorem 1 with \( d_{eval} = W_1 \). By the construction of the estimator \( \hat{Q} \), condition (8)-(ii) is automatically satisfied with \( \epsilon_2 = \epsilon_{opt} \) for any \( G \) and \( F \).

Let \( \delta > 0 \) be given. Lemma 3.5 from Chae et al. (2023) implies that there exists \( g_* \in \mathcal{D}(L, p, s, K \lor 1) \), with \( L \leq c_1 \log \delta^{-1} \), \( |p|_{\infty} \leq c_1 \delta^{-t_*/\beta_*} \), \( s \leq c_1 \delta^{-t_*/\beta_*} \log \delta^{-1} \) for some constant \( c_1 = c_1(q, d, t, \beta, K) \), such that \( \|g_* - g_0\|_{\infty} < \delta \).

Let \( g_* \in \mathcal{G} \). Then, by the Kantorovich–Rubinstein duality (see Theorem 1.14 in Villani (2003)),

\[
W_1(Q_*, Q_0) = \sup_{f \in \mathcal{F}_{lip}} |Q_* f - Q_0 f| \\
\leq \sup_{f \in \mathcal{F}_{lip}} \int \left| f(g_*(z)) - f(g_0(z)) \right| dP_Z(z) \\
\leq \int \left| g_*(z) - g_0(z) \right| dP_Z(z) \leq \sqrt{D} \|g_* - g_0\|_{\infty} \leq \sqrt{D} \delta.
\]

Hence, condition (8)-(i) holds with \( \epsilon_1 = \sqrt{D} \delta \).

Let \( \epsilon > 0 \) be given. For two Borel probability measures \( Q_1 \) and \( Q_2 \) on \( \mathbb{R}^D \), one can choose \( f_{Q_1, Q_2} \in \mathcal{F}_{lip} \) such that \( f_{Q_1, Q_2}(0_D) = 0 \) and

\[
W_1(Q_1, Q_2) = \sup_{f \in \mathcal{F}_{lip}} |Q_1 f - Q_2 f| \leq |Q_1 f_{Q_1, Q_2} - Q_2 f_{Q_1, Q_2}| + \epsilon.
\]

Then, by the Lipschitz continuity,

\[
\sup_{|x|_{\infty} \leq K} |f_{Q_1, Q_2}(x)| \leq \sup_{|x|_{\infty} \leq K} |x|_2 = \sqrt{DK}.
\]

Let \( \{g_1, \ldots, g_N\} \) be an \( \epsilon \)-cover of \( \mathcal{G} \cup \{g_0\} \) with respect to \( \|\cdot\|_{P_Z, 2} \) and

\[
\mathcal{F} = \left\{ f_{jk} : 1 \leq j, k \leq N \right\},
\]

where

\[
\|g\|_{P_Z, p} = \left( \int |g(z)|_p^pdP_Z(z) \right)^{1/p}
\]

and \( f_{jk} = f_{Q_{g_j}, Q_{g_k}} \). Since \( \|g - \bar{g}\|_{P_Z, 2} \leq \sqrt{D} \|g - \bar{g}\|_{\infty} \) for every \( g, \bar{g} \in \mathcal{G} \cup \{g_0\} \) and

\[
\log N(\epsilon, \mathcal{G}, \|\cdot\|_{\infty})
\]
\begin{equation}
\leq (s + 1) \left\{ \log 2 + \log \epsilon^{-1} + \log(L + 1) + 2 \sum_{l=0}^{L+1} \log(p_l + 1) \right\}
\end{equation}

by Lemma 5 of Schmidt-Hieber (2020), the number $N$ can be bounded as

\begin{equation}
\log N \leq \log \left( N(\epsilon/\sqrt{D}, G, \| \cdot \|_\infty) + 1 \right)
\leq c_2 s \left( \log D + \log \epsilon^{-1} + L \log \delta^{-1} \right)
\leq c_3 d^{-\epsilon/\beta_*} \log \delta^{-1} \left\{ \log \epsilon^{-1} + (\log \delta^{-1})^2 \right\},
\end{equation}

where $c_2 = c_2(t_*, \beta_*)$ and $c_3 = c_3(c_1, c_2, D)$. Here, $N(\epsilon, G, \| \cdot \|_\infty)$ denotes the covering number of $G$ with respect to $\| \cdot \|_\infty$.

Next, we will prove that condition (8)-(iv) is satisfied with $\epsilon_4 = 5\epsilon$. Note that $d_F \leq W_1$ by the construction. For $g_j, \bar{g} \in G \cup \{g_0\}$, we can choose $g_j$ and $\bar{g}$ such that $\|g - g_j\|_{p,2} \leq \epsilon$ and $\|\bar{g} - g_0\|_{p,2} \leq \epsilon$. Then,

\begin{equation}
W_1(Q_{\bar{g}}, Q_\bar{g}) \leq W_1(Q_{g_j}, Q_{g_j}) + W_1(Q_{g_j}, Q_{g_0}) + W_1(Q_{g_0}, Q_{\bar{g}})
\leq W_1(Q_{g_j}, Q_{g_0}) + d_F(Q_{g_j}, Q_{g_0}) + W_1(Q_{g_0}, Q_{\bar{g}}) + \epsilon.
\end{equation}

Note that

\begin{equation}
W_1(Q_{g_j}, Q_{g_0}) = \sup_{f \in F_{\cup g}} \left| \int f(g(z))dP_{\bar{g}}(z) - \int f(g_j(z))dP_{g_0}(z) \right|
\leq \int |g(z) - g_j(z)|_2 dP_{g_0}(z) \leq \|g - g_j\|_{p,2} \leq \epsilon.
\end{equation}

Similarly, $W_1(Q_{g_j}, Q_{g_0}) \leq \epsilon$, and therefore,

\begin{equation}
d_F(Q_{g_j}, Q_{g_0}) \leq d_F(Q_{g_j}, Q_{g_0}) + d_F(Q_{g_0}, Q_{\bar{g}}) + d_F(Q_{\bar{g}}, Q_{g_0})
\leq d_F(Q_{g_j}, Q_{g_0}) + 2\epsilon.
\end{equation}

Hence, the right hand side of (20) is bounded by $d_F(Q_{g_0}, Q_{\bar{g}}) + 5\epsilon$. That is, condition (8)-(iv) holds with $\epsilon_4 = 5\epsilon$.

Next, note that $P_n$ is the empirical measure based on i.i.d. samples from $P_0$. Let $Y$ and $\epsilon$ be independent random vectors following $Q_0$ and $N(0, \sigma_0^2 I_D)$, respectively. For any $f \in F$, by the Lipschitz continuity,

\begin{equation}
|f(Y + \epsilon)| \leq |Y + \epsilon|_2 \leq |Y|_2 + |\epsilon|_2.
\end{equation}

Since $Y$ is bounded almost surely and $\sigma_0 \leq 1$, $f(Y + \epsilon)$ is a sub-Gaussian random variable with the sub-Gaussian parameter $\sigma = \sigma(K, D)$. By the Hoeffding’s inequality,

\begin{equation}
P_0 \left( \|P_n f - P_0 f\|_2 > t \right) \leq 2 \exp \left[ -\frac{nt^2}{2\sigma^2} \right].
\end{equation}
for every $f \in \mathcal{F}$ and $t \geq 0$; see Proposition 2.5 from Wainwright (2019) for Hoeffding’s inequality for unbounded sub-Gaussian random variables. Since $\mathcal{F}$ is a finite set with the cardinality $N^2$,

$$P_0\left(\sup_{f \in \mathcal{F}} |\mathbb{P}_n f - P_0 f| > t\right) \leq 2N^2 \exp\left[\frac{-nt^2}{2\sigma^2}\right].$$

If $t \geq 2\sigma \sqrt{(\log(2N^2))/n}$, the right hand side is bounded by $e^{-nt^2/(4\sigma^2)}$. Therefore,

$$\mathbb{E}d_F(P_n, P_0) = \int_0^\infty P_0(d_F(P_n, P_0) > t) dt \leq 2\sigma \sqrt{\frac{\log(2N^2)}{n}} + \int_0^\infty \exp\left[-\frac{nt^2}{4\sigma^2}\right] dt \leq 2\sigma \sqrt{\frac{\log(2N^2)}{n}} + \sigma \sqrt{\frac{\pi}{n}},$$

and condition (8)-(iii) is also satisfied with $\epsilon_3$ equal to the right hand side of the last display. Note that

$$d_F(P_0, Q_0) \leq W_1(P_0, Q_0) \leq W_2(P_0, Q_0) \leq \sqrt{D}\sigma_0,$$

where the last inequality holds because $P_0$ is the convolution of $Q_0$ and $\mathcal{N}(\mathbf{0}_D, \sigma_0^2 \mathbf{I}_D)$.

By Theorem 1, we have

$$\mathbb{E}W_1(\hat{Q}, Q_0) \leq 2\sqrt{D}\sigma_0 + 5\sqrt{D}\delta + \epsilon_{\text{opt}} + 4\sigma \sqrt{\frac{\log(2N^2)}{n}} + 2\sigma \sqrt{\frac{\pi}{n}} + 10\epsilon \leq c_4 \left\{ \epsilon_{\text{opt}} + \sigma_0 + \delta + \sqrt{\frac{\log N}{n}} + \epsilon \right\},$$

where $c_4 = c_4(\sigma, D)$. Combining with (19), we have

$$\mathbb{E}W_1(\check{Q}, Q_0) \leq c_5 \left\{ \epsilon_{\text{opt}} + \sigma_0 + \delta + \frac{\sqrt{\log \delta^{-1}}(\sqrt{\log \epsilon^{-1}} + \log \delta^{-1})}{\sqrt{n}\delta^{1/2}\beta} + \epsilon \right\},$$

where $c_5 = c_5(c_3, c_4)$. The proof is complete if we take

$$\delta = n^{-\beta_*/(2\beta_* + t_*)}(\log n)^{\frac{3\beta_*}{3\beta_* + t_*}}$$

and $\epsilon = n^{-\log n}$. \qed
C Proof of Theorem 3

The proof of the first assertion is the same as that of Theorem 2. The only difference is that some constants in the proof depend on the Lipschitz constant $C_1$.

For the second assertion, we utilize Theorem 1 with $F = F_0$. Since $d_{eval} = d_F$, we have $\epsilon_t = 0$. Also, for a large enough $G$, i.e., large depth, width and number of nonzero parameters, $\epsilon_1$ can be set to be an arbitrarily small number. Since $F$ consists of Lipschitz continuous function, $d_F(P_0, Q_0) \lesssim \sigma_0$. It follows by Theorem 1 that $\mathbb{E}d_{F_0}(Q, Q_0) \lesssim \sigma_0 + \epsilon_{opt} + \mathbb{E}d_{F_0}(\mathbb{P}_n, P_0)$.

D Proof of Theorem 4

The proof is divided into several cases. For cases with $q = 0$, we write $\beta_0$ as $\beta$ for simplicity.

Case 1: $q = 0$ and $t_0 = d = D$

In this case, $\beta_0 = \beta$, $t_0 = d = 0$ and $G_0 = \mathcal{H}_K^\beta([0,1]^d) \times \cdots \times \mathcal{H}_K^\beta([0,1]^d)$. Our proof relies on Fano’s method for which we refer to Chapter 15 from Wainwright (2019).

Let $\phi : \mathbb{R} \to [0, \infty)$ be a fixed function satisfying that

(i) $\phi$ is $[\beta + 1]$-times continuously differentiable on $\mathbb{R}$,

(ii) $\phi$ is unimodal and symmetric about 1/2, and

(iii) $\phi(z) > 0$ if and only if $z \in (0,1)$.

where $[x]$ denotes the largest integer less than or equal to $x$. Figure 1 shows an illustration of $\phi$ and related functions. For a positive integer $m = m_n$, with $m_n \uparrow \infty$ as $n \to \infty$, let $z_j = j/m$, $I_j = [z_j, z_{j+1})$ for $j = 0, \ldots, m - 1$, $J = \{0,1,\ldots,m-1\}^d$ and $\phi_j(z) = \phi(m(z - z_j))$. For a multi-index $j = (j_1, \ldots, j_d) \in J$ and $\alpha = (\alpha_j)_{j \in J} \in \{-1, +1\}^{|J|}$, define $g_\alpha : [0,1]^d \to \mathbb{R}^d$ as

$$g_\alpha(z) = \left( z_1 + \frac{c_1}{m^{\beta}} \sum_{j \in J} \alpha_j \phi_j(z_1) \cdot \phi_j(z_2, \ldots, z_d) \right),$$

where $c_1 = c_1(\phi, d)$ is a small enough constant described below. Then, it is easy to check that $g_\alpha$ is a one-to-one function from $[0,1]^d$ onto itself, and $g_\alpha \in \mathcal{H}_K^\beta([0,1]^d) \times \cdots \times \mathcal{H}_K^\beta([0,1]^d)$ for large enough $K = K(\beta, c_1)$.

Let $Z = (Z_1, \ldots, Z_d)$ be a uniform random variable on $(0,1)^d$. Then, by the change of variables formula, the Lebesgue density $q_\alpha$ of $Y = g_\alpha(Z)$ is given as

$$q_\alpha(y) = \left| \frac{\partial z}{\partial y} \right| = \left( 1 + \frac{c_1}{m^{\beta}} \sum_{j \in J} \alpha_j \phi'_j(z_1) \phi_j(y_2) \cdot \phi_j(y_d) \right)^{-1}$$

for $y \in [0,1]^d$, where $\phi'$ denotes the derivative of $\phi$. Here, $z_1 = z_1(y_1, \ldots, y_d)$ is implicitly defined.
We first find an upper bound of $K(q_{\alpha}, q_{\alpha'})$ for $\alpha, \alpha' \in \{-1, +1\}^{|J|}$, where $K(p, q) = \int \log \frac{p}{q} dP$ is the Kullback–Leibler divergence. Since $\beta \geq 1$, $q_{\alpha}$ is bounded from above and below for a small enough $c_1$. Also, $g_{\alpha}(C_j) = C_j$, where $C_j = I_{j_1} \times \cdots \times I_{j_d}$.

Therefore, we have

$$|q_{\alpha}(y) - q_{\alpha'}(y)| \lesssim 1 \frac{1}{q_{\alpha}(y)} - 1 \frac{1}{q_{\alpha'}(y)} | \leq 2 \frac{c_1}{m^{\beta - 1}} \|\phi'\|_{\infty} | \phi|^{d-1}.$$  

Since the ratio $q_{\alpha} / q_{\alpha'}$ is bounded from above and below, we can use a well-known inequality $K(q_{\alpha}, q_{\alpha'}) \lesssim d_H^2(q_{\alpha}, q_{\alpha'})$, where $d_H$ denotes the Hellinger distance; see Lemma B.2 from Ghosal and van der Vaart (2017). Since $|\sqrt{q_{\alpha}} - \sqrt{q_{\alpha'}}| \lesssim |q_{\alpha} - q_{\alpha'}|$, we have

$$K(q_{\alpha}, q_{\alpha'}) \lesssim \int_{[0,1]^d} |q_{\alpha}(y) - q_{\alpha'}(y)|^2 dy \lesssim \frac{c_1^2 \|\phi'\|_{\infty}^2 \|\phi\|^{2(d-1)}}{m^{2(\beta-1)}}.$$

Next, we derive a lower bound for $W_1(q_{\alpha}, q_{\alpha'})$. Suppose that $\alpha_j \neq \alpha'_j$ for some $j \in J$. Then, the excess mass of $Q_{\alpha}$ over $Q_{\alpha'}$ on $C_j$ is

$$\int_{y \in C_j; q_{\alpha}(y) > q_{\alpha'}(y)} \left\{ q_{\alpha}(y) - q_{\alpha'}(y) \right\} dy$$

$$= \frac{1}{2} \left( \int_{C_j} |q_{\alpha}(y) - q_{\alpha'}(y)| dy \right) \geq \frac{1}{2} \int_{C_j} \left| \frac{1}{q_{\alpha}(y)} - \frac{1}{q_{\alpha'}(y)} \right| dy$$

$$\geq \frac{2c_1}{m^3} \int_{C_j} |\phi'_{j_1}(z_1)\phi_{j_2}(z_2)\cdots\phi_{j_d}(z_d)| \left| \frac{\partial y}{\partial z} \right| dz$$

$$= \frac{2c_1}{m^3} \int_{C_j} |\phi'_{j_1}(z_1)\phi_{j_2}(z_2)\cdots\phi_{j_d}(z_d)| \left| \frac{\partial y}{\partial z} \right| dz$$

$$\geq \frac{c_1}{m^{(\beta-1)+d}} \int_{(0,1)^d} |\phi'(z_1)\phi(z_2)\cdots\phi(z_d)| dz.$$
In virtue of Corollary 1.16 from Villani (2003), with a (unique) optimal transport plan between \( Q_\alpha \) and \( Q_{\alpha'} \), some portion \( \gamma \in (0, 1) \) of this excess mass must be transported at least the distance of \( c_2/m \), where constants \( \gamma \) and \( c_2 \) can be chosen so that they depend only on \( d \) and \( \phi \). Hence, for some constant \( c_3 = c_3(\phi, d) \),

\[
W_1(q_\alpha, q_{\alpha'}) \geq \frac{c_1 c_3}{m^{d+1}} H(\alpha, \alpha'),
\]

where \( H(\alpha, \alpha') = \sum_{j \in J} I(\alpha_j \neq \alpha'_j) \) denotes the Hamming distance between \( \alpha \) and \( \alpha' \).

With the Hamming distance on \( \{-1, +1\}^{|J|} \), it is well-known (e.g. see page 124 of Wainwright (2019)) that there is a \(|J|/4\)-packing \( \mathcal{A} \) of \( \{-1, +1\}^{|J|} \) whose cardinality is at least \( e^{|J|/16} \). Let \( P_\alpha \) be the convolution of \( Q_\alpha \) and \( N(0_d, \sigma_0^2 I_d) \). Then, \( K(p_\alpha, p_{\alpha'}) \leq K(q_\alpha, q_{\alpha'}) \) by Lemma B.11 of Ghosal and van der Vaart (2017). By Fano’s method (Proposition 15.12 from Wainwright (2019)), we have

\[
\mathfrak{m}(G_0, \sigma_0) \gtrsim \frac{c_1 c_3}{m^2} \left\{ 1 - \frac{nc^2 \beta(\phi, d)m^{-2(\beta-1)} + \log 2}{m^d/16} \right\}.
\]

If \( n \gg m^{d+2(\beta-1)} \) and \( c_1 \) is small enough, we have the desired result.

**Case 2: \( q = 0 \) and \( t_0 = d < D \)**

Define a subset \( G_1 \) of \( G_0 = \mathcal{H}_K^\beta([0, 1]^d) \times \cdots \times \mathcal{H}_K^\beta([0, 1]^d) \) as

\[
G_1 = \left\{ g \in G_0 : g_{d+1}(z) = \cdots = g_D(z) = 0 \right\},
\]

where \( g(\cdot) = (g_1(\cdot), \ldots, g_D(\cdot)) \). The problem of obtaining a lower bound of the minimax risk \( \mathfrak{m}(G_1, \sigma_0) \) reduces to Case 1, hence \( \mathfrak{m}(G_1, \sigma_0) \) is bounded below by a multiple of \( n^{-\beta/(2\beta+d-2)} \). Since \( G_1 \subset G_0 \), we have \( \mathfrak{m}(G_0, \sigma_0) \geq \mathfrak{m}(G_1, \sigma_0) \).

**Case 3: \( q = 0 \) and \( t_0 < d \leq D \)**

Similarly to Case 2, define a subset \( G_2 \) of \( G_0 \) as

\[
G_2 = \left\{ g \in G_0 : g(z) = (g_1(z_{1:t_0}), \ldots, g_{t_0}(z_{1:t_0}), 0, \ldots, 0) \right\},
\]

for some \( g_j : [0, 1]^{t_0} \to \mathbb{R}, j = 1, \ldots, t_0 \). Then, the problem reduces to Case 1 with \( d \) replaced by \( t_0 \). Hence, we obtain a desired lower bound \( \mathfrak{m}(G_0, \sigma_0) \geq \mathfrak{m}(G_2, \sigma_0) \gtrsim n^{-\beta/(2\beta+t_0-2)} \).

**Case 4: General \( q \)**

For \( G_0 = G_0(q, d, t, \beta, K) \), fix \( i_0 \in \{0, \ldots, q\} \). We consider a subset \( G_3 \) of \( G_0 \) consisting of functions of the form \( g = h_q \circ h_{i_0-1} \circ \cdots \circ h_1 \circ h_0 \), where each \( h_i : [a_i, b_i]^{d_i} \to [a_{i+1}, b_{i+1}]^{d_{i+1}} \) satisfies the following properties:
1. For $i < i_0$, $h_i(x) = (x_1, x_{d_i+1}, 0, d_i+1-d_i, i+1)$.
2. For $i > i_0$, $h_i(x) = (x_1, 0, d_{i+1}-d_i, 0)$.
3. $h_{i_0}(x) = (h_{i_01}(x_1), . . . , h_{i_0t_{i_0}}(x_1), 0, . . . , 0)$ for some function $h_{i_0j} \in \mathcal{H}_i^{\beta_0}(\alpha_{i_0}, \beta_{i_0})$.

Since $t_{i_0} \leq \min\{d_0, . . . , d_p+1\}$, we have

$$g(z) = (h_{i_01}(z_1), . . . , h_{i_0t_{i_0}}(z_1), 0, . . . , 0).$$

Again, the problem reduces to Case 1 with $(d, \beta)$ replaced by $(t_{i_0}, \beta_{i_0})$. Therefore, $2\mathcal{R}(G_0, \sigma_0) \geq 2\mathcal{R}(G_3, \sigma_0) \gtrsim n^{-\beta_{i_0}/(2\beta_{i_0}+t_{i_0}-2)}$. Since this inequality holds for all $i_0 \in \{0, \ldots, g\}$, the assertion of the theorem follows. $\Box$

32