Journal of Machine Learning Research 24 (2023) 1-42 Submitted 9/21; Revised 2/23; Published 3/23

A Likelihood Approach to Nonparametric Estimation of a Singular Distribution Using Deep Generative Models

Minwoo Chae mchae@postech.ac.kr
Department of Industrial and Management Engineering
Pohang University of Science and Technology
Pohang, Gyeongbuk 37673, South Korea

Dongha Kim dongha0718@sungshin.ac.kr
School of Mathematics, Statistics and Data Science
Data Science Center
Sungshin Women's University
Seoul, 02844, South Korea

Yongdai Kim ydkim0903@gmail.com
Department of Statistics
Seoul National University
Seoul, 08826, South Korea

Lizhen Lin lizhen.lin@nd.edu
Department of Applied and Computational Mathematics and Statistics
University of Notre Dame
South Bend, IN 46556, USA

Editor: Daniel Roy

Abstract

We investigate statistical properties of a likelihood approach to nonparametric estimation of a singular distribution using deep generative models. More specically, a deep generative model is used to model high-dimensional data that are assumed to concentrate around some low-dimensional structure. Estimating the distribution supported on this low-dimensional structure, such as a low-dimensional manifold, is challenging due to its singularity with respect to the Lebesgue measure in the ambient space. In the considered model, a usual likelihood approach can fail to estimate the target distribution consistently due to the singularity. We prove that a novel and eective solution exists by perturbing the data with an instance noise, which leads to consistent estimation of the underlying distribution with desirable convergence rates. We also characterize the class of distributions that can be eciently estimated via deep generative models. This class is suciently general to contain various structured distributions such as product distributions, classically smooth distributions and distributions supported on a low-dimensional manifold. Our analysis provides some insights on how deep generative models can avoid the curse of dimensionality for nonparametric distribution estimation. We conduct a thorough simulation study and real data analysis to empirically demonstrate that the proposed data perturbation technique improves the estimation performance signicantly.

Keywords: Data perturbation, deep generative model, distribution on a lower-dimensional manifold, maximum likelihood, singular distribution estimation.

1. Introduction

Suppose that we have observations $X_1; : : : ; X_n$ which are i.i.d. copies of a D-dimensional random vector X following the distribution P. Without any structural assumption, the problem of estimating P or related quantities (e.g. density, support, etc.) with large dimension D is prohibitively dicult, which is widely known as the curse of dimensionality. To avoid the curse of dimensionality, it is natural to assume that the data locate around some lower-dimensional structure which can be captured by the model $X = Y +$, where Y is a random vector possessing a specic low-dimensional structure and is a full-dimensional noise vector with small variance. As an example of low-dimensional structures, one may assume that there exists a low-dimensional manifold on which the probability mass of Y is concentrated. For this model, our primary interests are in estimating Q; the distribution of Y; or related quantities. There is a large literature on estimating the support of Q, i.e., manifold estimation, see, e.g., Ozakin and Gray (2009); Puchkin and Spokoiny (2022); Genovese et al. (2012b,a) and references therein. The problem of estimating Q on the other hand is much less studied and in general a more challenging problem due to the singularity of Q with respect to the Lebesgue measure in the ambient space. Berenfeld and Homann (2019) and Ozakin and Gray (2009) considered kernel density estimators for estimating the (Hausdor) density of Q when the data are assumed to be supported on the image of a submanifold embedded in a higher dimensional space, thus no noise is considered.

In this paper, we consider a special form of $X = Y +$, so-called a probabilistic generative model, which models the observation as $X = f(Z) +$, where Z and are independent random vectors which are not directly observable. The latent variable Z is a d-dimensional random vector drawn from some known distribution P_Z, such as the standard normal or uniform distributions supported on Z, an open subset of R^d, and $f : Z \; ! \; R^D$ is an unknown function which is often called the generator or generating function. The noise vector is assumed to follow the normal distribution $N(0_D; {}^2 I_D)$, where 0_D and I_D denote the D-dimensional zero vector and identity matrix, respectively. We consider the case of $d < D$, in which the distribution of $f(Z)$ is singular with respect to the Lebesgue measure on R^D:

The model $X = f(Z) +$ has been investigated in statistical literature with the name of a nonlinear factor model (Yalcin and Amemiya, 2001). In this paper, we model f using deep neural networks (DNNs), which are known to enjoy universal approximations results (Cybenko, 1989; Hornik et al., 1989, 1990). Accordingly, we adopt the terminology of a deep generative model. In a deep generative model, instead of directly estimating P or Q, one may rst construct an estimator $\hat{f}$ and the resulting distribution of $\hat{f}(Z)$ will serve as an estimator of Q. Although this approach does not provide an explicit estimator of Q, it is easy to draw samples from the estimated distribution.

In recent years, deep generative models have achieved tremendous success for modeling high-dimensional data such as images and videos. Two popular approaches are used in practice to construct an estimator $\hat{f}$. The rst one is likelihood-based. Variational approaches (Kingma and Welling, 2014; Rezende et al., 2014) and EM-based algorithms (Burda et al., 2016; Kim et al., 2020) are two most representative learning methods in this class. The second approach uses the integral probability metrics (IPM; Müller, 1997), often called the adversarial losses in deep learning communities, and constructs an estimator by minimiz-ing these metrics. This approach is widely known as the generative adversarial networks

(GAN), originally developed by Goodfellow et al. (2014) and then generalized in Mroueh et al. (2017); Li et al. (2017) and Arjovsky et al. (2017), to name a few.

In this work, we focus on the likelihood-based approach and study statistical properties of a sieve maximum likelihood estimator (MLE) of deep generative models under the assumption that P is the distribution of $X = f(Z) +$ for some function $f : Z \; ! \; R^D$ and $N(0; {}^2 I_D)$, where 0. The primary goal is to estimate Q, the distribution of $f(Z)$ induced from the distribution of Z via the true generator f. We obtain several important results for this model.

Firstly, we derive a convergence rate of $Q = Q_{\hat f}$ to $Q = Q_f$ in terms of the Wasserstein metric (Villani, 2003), where $\hat f$ is a sieve MLE of f and Q_f denotes the distribution of $f(Z)$, cf. Corollary 4 and Theorem 7. The convergence rate depends on the noise level , intrinsic dimension and smoothness of f; see Section 3 for the denition. More interestingly, Corollary 4 and Theorem 7 do not guarantee the consistency of a sieve MLE for very small . To resolve this issue and improve the convergence rate, we propose a novel method to perturb the data. That is, we obtain a sieve MLE of f based on the perturbed observation $X_i = X_i + {}_i$ where ${}_i$ is an articial noise vector following the distribution $N(0_D; e^2 I_D)$. The perturbation level e will be chosen carefully to provide a desirable convergence rate. Note that X_i always possesses a Lebesgue density p even when = 0: Under general conditions, we derive the convergence rate of a sieve MLE for estimating p with respect to the Hellinger metric, cf. Theorem 3 and Corollary 6. Then, we derive a Wasserstein convergence rate of a sieve MLE of Q based on perturbed observations, cf. Theorem 9. Specically, we attain the convergence rate $e + {}_n$ e up to a logarithmic factor, where ${}_n$ is the Hellinger convergence rate of the sieve MLE of p , and e = + e. Note that ${}_n$ decreases as e increases because p becomes smoother while e increases. Hence, the degree of perturbation e can be determined by minimizing $e + {}_n$ e.

Recently, successful cases of data perturbation for learning deep generative models have been reported in Song and Ermon (2019); Meng et al. (2021). However, theoretical understanding of the data perturbation is still lacking. Our results in this paper can provide a theoretical justication for the success of various data perturbation procedures for deep generative models. Note that most existing theories on deep generative models consider GAN, for which additional noise does not help.

Main results concerning the convergence rates are stated non-asymptotically in the sense that for any xed n 1, we provide sucient conditions under which certain probabilistic inequalities hold. Besides the convergence rate of a sieve MLE, we characterize a class of distributions that can be represented by $f(Z)$ for some f. The class is large enough to include various distributions such as product distributions, classically smooth distributions and distributions supported on a low-dimensional manifold. As an illustrating example, a class of product distributions has the intrinsic dimension 1, and corresponds to the generalized additive model in the regression setting. This kind of structure has not been studied in an unsupervised learning framework. The regularity theory of the optimal transport plays an important role for this characterization.

There are a lot of recent articles studying the statistical properties of the GAN estimator; see Section 1.1 for review. It is a critical limitation of most theoretical studies that they assumed the existence of the smooth Lebesgue density p of the underlying distribution P. They view the GAN in a nonparametric density estimation framework; the convergence rate directly depends on D and the smoothness level of p. Consequently, these results only guarantee that GAN performs as good as classical nonparametric density estimators, and cannot explain why and how it outperforms other methods. Some recent articles reviewed in Section 1.1 go beyond the density estimation framework, but their theories are not exhaustive and possess certain limitations. In this sense, our results about the convergence rates of a sieve MLE with perturbed data are new and important contributions for deep generative models. In contrast, the idea of using perturbed data with the GAN estimator has been shown to be ineective through numerical studies in Section 5, as demonstrated by Figures 10 and 11.

Our convergence rate depends on not only the intrinsic dimension of the manifold $f(Z)$, which is much smaller than D, but also the degree of smoothness of f: Moreover, if f has a low-dimensional composite structure considered as in Horowitz and Mammen (2007), Juditsky et al. (2009), the convergence rate becomes faster. For supervised learning, many studies have shown that DNN can avoid curse of dimensionality when the true regres-sion function has a low-dimensional composite structure (Schmidt-Hieber, 2020; Bauer and Kohler, 2019; Kohler and Langer, 2021) or the support of input variables or covariates concentrate on a low-dimensional manifold (Chen et al., 2019a,b; Schmidt-Hieber, 2019; Nakada and Imaizumi, 2020). Our results are among the rst that have demonstrated that these ne properties of DNN for supervised learning are also valid for unsupervised learning, which is an important advantage of using deep generative models compared to the ones that estimate Q or P directly.

The remainder of this paper is organized as follows. In Section 1.1, we review recently developed theoretical results for GAN. Section 2 introduces a deep generative model. Our main results concerning the convergence rate of a sieve MLE and data perturbation are given in Section 3. Section 4 considers a class of true distributions that can be represented as a true generator. Experimental results and concluding remarks follow in Sections 5 and 6, respectively.

1.1 Related Work

Most works for statistical properties for deep generative models focus on GAN type estimators, which are briey reviewed in this subsection. In a GAN framework, Arora et al. (2017) rstly considered a neural network distance, a special case of IPMs, to measure the discrepancy of an estimator from the true distribution. They noticed that a neural network distance might be so weak that GAN may not consistently estimate the true distribution. Further studies have been conducted by Zhang et al. (2018) and Bai et al. (2019), who provide sucient conditions for a neural network distance to induce the same topology as the Wasserstein metric and KL divergence. In particular, Zhang et al. (2018) obtained convergence rates of GAN estimators with respect to the bounded Lipschitz metric, which however seem to be much slower than the optimal rate. A similar, but slightly dierent approach in studying a neural network distance is given in Liu et al. (2017). This work employs topological properties of neural network distances, hence important structural assumptions such as the smoothness of densities were not considered. Biau et al. (2020) studied asymptotic properties of the original GAN developed by Goodfellow et al. (2014). Rather than considering a neural network distance, they investigated how the approximation of the discriminator can aect the estimation performance with respect to the Jensen{Shannon divergence. However, their analysis is based on the parametric assumption, that is, the number of network parameters is xed as the sample size tends to innity.

There is a dierent line of works that study asymptotic properties of GAN from a nonparametric density estimation point of view. For densities in a Sobolev space, Singh et al. (2018); Liang (2021) derived minimax convergence rates with respect to the Sobolev IPMs which include metrics used in Sobolev (Mroueh et al., 2017), MMD (maximum mean discrepancy; Li et al., 2017) and Wasserstein (Arjovsky et al., 2017) GANs. These results are generalized in Uppal et al. (2019) using Besov IPMs. We would also like to mention Chen et al. (2020), who derived convergence rates with respect to the Hölder IPMs. Although their convergence rate is strictly slower than the minimax rate in Uppal et al. (2019), their results are directly applicable to GANs whose generator and discriminator network architectures are explicitly given. However, all these works are limited to the classical paradigm where the true distribution possesses a smooth Lebesgue density p and the convergence rate depends on the data dimension D, suering from the curse of dimensionality.

There are some recent articles considering the convergence rate of GAN beyond the density estimation framework. To the best of our knowledge, the set-up given in Luise et al. (2020) is the closest to ours. In particular, they assumed that there exists a true generator as in our paper and there is no noise, that is, $P = Q = Q_f$ for some smooth function f. Under this set-up they obtained a convergence rate of GAN for estimating Q with respect to the Sinkhorn divergence (Feydy et al., 2019). Note that although the Sinkhorn divergence metrizes the weak convergence, it is not a standard metric for evaluating the performance of distribution estimation and not comparable with the Wasserstein distance considered in our paper. In particular, their convergence rate directly depends on the regularization parameter dening the Sinkhorn divergence (in their notation), which makes it unclear how tight their convergence rate is. Furthermore, their theory does not incorporate deep neural network structures, hence cannot explain the benet of deep generative models which adapt to various structures such as the composite one. Also, the theory holds only when the smoothness of the true generator exceeds a certain threshold proportional to d. For these reasons, the theory in Luise et al. (2020) has certain limitations.

Schreuder et al. (2021) obtained convergence rates of GAN-based estimators under the assumption that the data-generating distribution is the convolution of $Q = Q_f$ and a general noise distribution, where $f : \mathbb{R}^d \to \mathbb{R}^D$ is a smooth function; hence the data are concentrated around a small neighborhood of a manifold whose dimension is at most d. Rather than assuming the existence of a true generator, Huang et al. (2021) assumed that the support of P is a certain low-dimensional set in $\mathbb{R}^D$ and studied the convergence rate of GAN. In both papers, the convergence rates depend on the intrinsic dimension of the true distribution rather than on the dimension D of the observations. The proofs in these papers rely on the adaptive property of the empirical measure to specic low-dimensional structures, studied in Weed and Bach (2019) and Schreuder (2021). It should be noted that the intrinsic dimension considered in our paper can be smaller compared to the dimensions considered in Schreuder et al. (2021) and Huang et al. (2021).

The analysis of the vanilla GAN in Biau et al. (2020) has been extended to the Wasserstein GAN in Biau et al. (2021). In particular, they considered DNN architectures for both the generator and discriminator classes and proved that the corresponding WGAN estimator can be arbitrarily close to the true distribution in Wasserstein distance with high probability; see Theorem 21 therein. However, their results do not provide specic convergence rate and do not incorporate approximation error of the generator class for specic distribution families.

Finally, we would also like to mention the work by Tang and Yang (2022) who considered the minimax convergence rate for nonparametric distribution estimation under the manifold assumption. Although the structural assumption considered in Tang and Yang (2022) is dierent from ours, they derived the minimax convergence rate for estimating a distribution supported on a submanifold of $\mathbb{R}^D$ with smooth density with respect to the Hausdor measure. In particular, they used a mixture of GAN estimators to achieve the minimax convergence rate. However, it should be emphasized that GAN-based estimators considered in this subsection, including the one in Tang and Yang (2022), is computationally much more intractable than sieve MLEs considered in the present paper.

1.2 Notations and Denitions

For two real numbers a and b, let $a \wedge b$ and $a \vee b$ be the minimum and maximum of a and b, respectively. $\lfloor a \rfloor$ is the largest integer less than or equal to a. The inequality $a \lesssim b$ means that a is less than b up to a constant multiplication. Also, denote $a \asymp b$ if $a \lesssim b$ and $b \lesssim a$. For a vector x, the ℓ^p-norm, $1 \le p \le \infty$, and the number of nonzero elements are represented as $|x|_p$ and $|x|_0$, respectively. Let $B_\epsilon(x)$ be the Euclidean open ball of radius ϵ centered at x. For a vector-valued function f, let $|f|_p$ be the map $x \mapsto |f(x)|_p$. The L^p-norm of a function is denoted $\|\cdot\|_p$, where the domain of a function and dominating measure will be clear in the context. The equality $c = c(A_1, \ldots, A_k)$ means that c depends only on $A_1, \ldots, A_k$. The uppercase letters, such as P and $\hat{P}$, refer to the probability measures corresponding to the densities denoted by the lowercase letters p and $\hat{p}$, respectively, and vice versa. A positive real-valued function f is said to be bounded from above and below if there exist positive constants c_1 and c_2 such that $c_1 \le f(x) \le c_2$ for every x.

For two probability densities p and q, let $d_H(p, q)$ and $K(p, q) = \int \log(p/q) dP$ be the Hellinger distance and KL divergence, respectively. The Wasserstein distance of order $r \in [1, \infty)$ between P and Q is denoted $W_r(P, Q)$ (Villani, 2003). For a function space $\mathcal{F}$, $N(\epsilon, \mathcal{F}, d)$ and $N_{[]}(\epsilon, \mathcal{F}, d)$ denote the covering and bracketing numbers with respect to the (pseudo)-metric d. For $\beta > 0$, let $\mathcal{H}^\beta_M(A)$ be the class of every β-Hölder function $f : A \to \mathbb{R}$ with β-Hölder norm bounded by $M > 0$. Let $\mathcal{H}^\beta(A) = \cup_{M>0} \mathcal{H}^\beta_M(A)$ be the class of every β-Hölder function. If there is no confusion, we simply denote them as $\mathcal{H}^\beta_M$ and $\mathcal{H}^\beta$. For a vector-valued function, $f \in \mathcal{H}^\beta$ refers that each component of f belongs to $\mathcal{H}^\beta$. We refer to Gine and Nickl (2016); van der Vaart and Wellner (1996) for details about these denitions.

2. Deep Generative Models

In this section, we formally dene the model $X = f(Z) + \epsilon$ using a DNN. Let $\mathcal{Z}$ be an open subset of $\mathbb{R}^d$ and $x \mapsto \phi_{\sigma,d}(x)$ be the density of d-fold product measure of the univariate normal distribution $N(0, \sigma^2)$. We often denote $\phi_{\sigma,d}$ as ϕ_σ if there is no confusion. Let $P_{f,\sigma}$ be the distribution of $f(Z) + \epsilon$, where Z and ϵ are independent random vectors distributed as P_Z and $N(0_D, \sigma^2 I_D)$, respectively. Standard uniform or Gaussian distribution is a common choice for P_Z, and some general sub-Gaussian distributions are considered in Luise et al. (2020). For a class $\mathcal{F}$ of functions from $\mathcal{Z}$ to $\mathbb{R}^D$ and two positive numbers $\sigma_{\min} < \sigma_{\max}$, we consider a class of probability distributions

$$\mathcal{P} = \left\{ P_{f,\sigma} : f \in \mathcal{F},\ \sigma \in [\sigma_{\min}, \sigma_{\max}] \right\}. \tag{2.1}$$

Recall that Q_f is the distribution of $f(Z)$, which is often called the pushforward measure of P_Z by the map $f : \mathcal{Z} \to \mathbb{R}^D$. If $\sigma > 0$, $P_{f,\sigma}$ has the Lebesgue density

$$p_{f,\sigma}(x) = \int \phi_\sigma\big(x - f(z)\big)\, dP_Z(z) = \int \phi_\sigma(x - u)\, dQ_f(u). \tag{2.2}$$

The function class $\mathcal{F}$ is modeled via a DNN. We adopt the definitions and notations in Schmidt-Hieber (2020). Let $\rho(x) = x \vee 0$ be the ReLU activation function. For a vector $v = (v_1, \ldots, v_r)^T \in \mathbb{R}^r$, define $\rho_v : \mathbb{R}^r \to \mathbb{R}^r$ as $\rho_v(z) = (\rho(z_1 - v_1), \ldots, \rho(z_r - v_r))^T$ for $z = (z_1, \ldots, z_r)^T$. A neural network with network architecture (L, p) is any function of the form

$$f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad z \mapsto f(z) = W_L \rho_{v_L} W_{L-1} \rho_{v_{L-1}} \cdots W_1 \rho_{v_1} W_0 z, \tag{2.3}$$

where $W_i \in \mathbb{R}^{p_{i+1} \times p_i}$, $v_i \in \mathbb{R}^{p_i}$ and $p = (p_0, \ldots, p_{L+1}) \in \mathbb{N}^{L+2}$. We will consider model (2.1) with the class $\mathcal{F} = \mathcal{F}(L, p, s, K)$, where $\mathcal{F}(L, p, s, K)$ is the collection f of the form (2.3) satisfying

$$\max_{j=0,\ldots,L} |W_j|_\infty \vee |v_j|_\infty \le 1, \qquad \sum_{j=1}^{L} |W_j|_0 + |v_j|_0 \le s, \qquad \| |f|_\infty \|_\infty \le K,$$

$p_0 = d$ and $p_{L+1} = D$. Here, $|W_j|_\infty$ and $|W_j|_0$ denote the maximum-entry norm and the number of nonzero elements of the matrix W_j, respectively.

The statements of main theorems and corollaries in Section 3 are non-asymptotic; they hold for any fixed $n \ge 1$. However, it would be convenient to regard quantities $(\sigma_{\min}, L, p, s)$ as sequences depending on the sample size n, while $(\sigma_{\max}, K)$ remain as fixed constants. In this sense, it would be precise to denote $(\sigma_{\min}, L, p, s)$ and $(\mathcal{F}, \mathcal{P})$ as $(\sigma_{\min,n}, L_n, p_n, s_n)$ and $(\mathcal{F}_n, \mathcal{P}_n)$, respectively. For simplicity, we suppress the subscript when the dependency on n is obvious contextually. Throughout this paper, the model (2.1) with $\mathcal{F} = \mathcal{F}(L, p, s, K)$ will be called a deep generative model with ReLU activation function.

From another viewpoint, the density of the form (2.2) is a mixture of normal distributions. Note that mixtures of normal densities are frequently used in nonparametric statistics to model smooth densities. In particular, an arbitrary smooth density can be approximated by normal mixtures as shown in Ghosal and van der Vaart (2007); Shen et al. (2013). Based on this, it can be shown that a Bayes estimator with a Dirichlet process prior and a sieve MLE achieve the minimax optimal convergence rate up to a logarithmic factor when the true density belongs to a Hölder class. However, the model complexity of normal mixtures required to approximate an arbitrary smooth density, often expressed through the metric entropy, grows rapidly as the dimension D increases which results in slow convergence rates. This large complexity is mainly because the mixing distribution can be of any form. Hence, such a large class of normal mixtures might not be useful for analyzing high-dimensional data. Note that model (2.1) is parametrized by the generator f rather than a mixing distribution. Consequently, the complexity of the model (2.1) can be expressed through the metric entropy of the generator class $\mathcal{F}$, which is detailed in Lemma 1.

3. Convergence Rate of a Sieve MLE

Our main theoretical results are given in this section. We rst present assumptions on the data-generating distribution P. Then, we derive the convergence rate of a sieve MLE for p with respect to the Hellinger distance in the deep generative model. We next obtain the convergence rate of the corresponding sieve MLE of Q under the Wasserstein distance. Our strategy of deriving the convergence rate is as follows. We rst derive a convergence rate of a sieve MLE $\hat{p}$ of p, the Lebesgue density of P; and then recover the corresponding convergence rate of $\hat{Q}$ to Q. However, this strategy only works when is not too small. If is very small, technical diculty arises because the density p peaks around a small neighborhood of f(Z), the likelihood therefore becomes picky and unstable, and a sieve MLE is expected to behave badly. For this case, we propose a novel data perturbation technique to derive the convergence rates for Q under this small regimes.

As mentioned earlier, our main theorems are non-asymptotic in the sense that they hold for any xed n 1. More specically, Theorem 9 is stated with the form of

$$P \;\; W_1(\hat{Q}; Q) > \;_n \;\; _n; \qquad\qquad n \;\; 1 \tag{3.1}$$

for some sequences $_n$ and $_n$ with $_n \;\; _n$. The interpretation of this statement is clear: for any xed n 1, once $_n$ and $_n$ are small enough, the Wasserstein distance between Q and $\hat{Q}$ will be small with high probability. Furthermore, since Q and $\hat{Q}$ are supported on a bounded set, the probabilistic statement (3.1) implies that $EW_1(\hat{Q}; Q) \lesssim \;_n \vee \;_n$ for every n 1. Similar interpretations also hold for assumptions of the Theorems on the noise , that is, for every sample size, there is a sucient condition on the noise for which the probabilistic bound (3.1) holds. Given the non-asymptotic nature of our results, the true data-generating distribution can be interpreted in similar fashions. For any given sample size n 1, the true data-generating distribution is given by a true P induced from the true generator f and some true noise level 2 $[_{min}; _{max}]$ with some appropriate assumptions on $_{min}$ and $_{max}$. The assumptions on $_{min}$ and $_{max}$ may vary with the sample size n.

Note that such non-asymptotic statements and interpretation can be frequently found in modern statistical theory. For example, in a high-dimensional linear regression set-up, the assumption on the dimension and/or the magnitude of the regression coecients may change with the sample size (Buhlmann and van de Geer, 2011; Wainwright, 2019). When the sample size is large, for example, the absolute value of the rst component of may be assumed to be large. For any xed n 1, however, there is one true data-generating distribution with the true parameter satisfying the appropriate assumption. In this set-up, many statistical theories take the form $P(\| \;\; \hat{} \| > \;_n) \;\; _n$, which is quite similar to (3.1).

3.1 Assumption on the True Distribution

Since we consider a deep generative model (2.1), it is natural to assume that $P = P_{f;}$ for some true generator f and 0, or more precisely, P is the convolution of $Q = Q_f$ and $N(0_D; ^2 I_D)$. In particular, we assume that f is a structured function that can be eciently approximated by DNN functions (Yarotsky, 2017; Telgarsky, 2016; Petersen and Voigtlaender, 2018; Ohn and Kim, 2019; Imaizumi and Fukumizu, 2019; Nakada and Imaizumi, 2020). For example, f can belong to a certain class F of smooth composite functions. In Section 4, we will show that the corresponding distribution class $\{Q_f : f \in \mathcal{F}\}$ is large enough to include the classical class of nonparametric smooth densities and densities supported on a lower-dimensional smooth manifolds as special cases.

Note that the generator f is not identiable in general. For example, even for a linear factor model where $f(Z) = AZ$ for a $D \times d$ matrix A, $f(\tilde{Z}) = AZ$ has the same distribution as $f(Z)$. However, the mixing distribution Q is identiable under mild assumptions, e.g. Bruni and Koch (1985).

3.2 A Sieve MLE

Since the parameter space specifying the model (2.1) depends on the sample size n, the model can be regarded as a sieve approximating the true distribution. Then, an estimator can be obtained via a maximum likelihood principle. The corresponding estimator is often called a sieve MLE (Geman and Hwang, 1982). To be specic, let $\ell_n(f;) = \sum_{i=1}^{n} \log p_{f;}(X_i)$ be the log-likelihood function. For a given sequence $_n \downarrow 0$, a sieve MLE is any estimator $(\hat{f}; \hat{}) \in \mathcal{F} \times [_{\min}; _{\max}]$ satisfying

$$\ell_n(\hat{f}; \hat{}) \geq \sup_{f \in \mathcal{F}; \in [_{\min}; _{\max}]} \ell_n(f;) - _n \tag{3.2}$$

and let $\hat{p} = p_{\hat{f};\hat{}}$. We do not abbreviate the subscript n for the rate sequence such as $_n$ and $_n$. The sequence $_n$ allows that strict maximization, which is infeasible in most applications of deep learning, is not necessary. It would be more desirable to consider an estimator which is obtained by a specic algorithm such as the gradient decent method. Unfortunately, it is challenging to study statistical properties of an algorithm-specic estimator in deep learning. To the best of our knowledge, the convergence rate of an algorithm-specic estimator have not been studied in deep learning contexts. We also do not consider algorithmic issues in this paper, and assume that a sieve MLE satisfying (3.2) is available. There are various computational algorithms targeting a sieve MLE in deep generative models, e.g. Burda et al. (2016); Kim et al. (2020).

3.3 Hellinger Convergence Rate of a Sieve MLE of p

Under general conditions, convergence rates of sieve MLEs with respect to the Hellinger metric are well established in Wong and Shen (1995). The key technique to derive convergence rates is to bound the Hellinger bracketing number of the density space for which many techniques are known for various classes of regular functions, see van der Vaart and Wellner (1996). Roughly, the convergence rate $_n$ can be achieved if $\log N_{[]}(; \mathcal{P}; d_H) \lesssim n^2_n$. Metric entropies of deep neural networks are also well-known in recent articles, see Lemma 5 of Schmidt-Hieber (2020). The following lemma provides a relation between the Hellinger bracketing number of $\mathcal{P}$ and the metric entropy of $\mathcal{F}$; which plays a crucial role in deriving the convergence rate of a sieve MLE $\hat{p}$. Below, we do not try to optimize constants which are not essential for deriving convergence rates.

Lemma 1 Let $\mathcal{F}$ be a class of functions from $\mathcal{Z}$ to $\mathbb{R}^D$ such that $\| |f|_1 \|_1 \leq K$ for every $f \in \mathcal{F}$. Let $\mathcal{P} = \{P_{f;} : f \in \mathcal{F}; \in [_{\min}; _{\max}]\}$ with $_{\min} \leq 1$. Then, there exist constants

$c = c(D; K; _{max})$; $c^0 = c^0(D; K; _{max})$ and $= (D)$ such that log

$$N_{[]}(; P; d_H) \;\; \log N \;\; c_{min}^{34}; F; k j \; j_1 k_1 + \log \left(\frac{c^0}{{}_{min}^{D+2} \;\; 4} \right) \qquad (3.3)$$

for every $\;2\;(0;]$.

Remark 2 Note that for a class of general normal location mixtures $\int (x \quad z) dP(z)$ parametrized by the mixing distribution P and scale parameter , the bracketing entropy scales as a polynomial order in ^{1} as $\;!\;\; 0$. Specically, Corollary B1 of Shen et al. (2013) gives an upper bound for the -bracketing entropy of the class $f x \;!\; \int (x \quad z) dP(z) : P([K; K]^D) = 1g$, which is at least of order $O((^{1} \_ \log \;\; ^{1})^D)$. This bound would give a nearly parametric convergence rate of a sieve MLE provided that the model is well-specied and $_{min}$ is bounded away from zero. However, the entropy bound of Shen et al. (2013) grows rapidly as $_{min} \;! \;\; 0$; which is problematic since we are interested in the case that $_{min}$ converges to 0. In contrast, the right hand side of (3.3) depends on $_{min}$ only through a logarithmic function. Hence, the entropy bound (3.3) is much smaller than that of Shen et al. (2013) when $_{min}$ is small, provided that $N(; F; k j \;\; j_1 k_1)$ is of a polynomial order in . If $F = F(L; p; s; 1)$ with $k p k_1 = O(n^a)$ for some constant $a > 0$ and $L = O(\log n)$, for example, $\log N(; F; k j \;\; j_1 k_1)$ is bounded by a multiple of $s f(\log n)^2 + \log \;\; ^1 g$, as shown in Lemma 5 of Schmidt-Hieber (2020). Consequently, $\log N_{[]}(; P; d_H)$ is of order $s f(\log n)^2 + \log \;\; ^1 + \log \;_{min} g$:

Utilizing Lemma 1, the next theorem provides convergence rates of a sieve MLE of p with respect to the Hellinger metric in terms of the entropy bound and approximation error $_{app}$ of the sieve F.

Theorem 3 Let $F; P$ and $= (D)$ be given as in Lemma 1, and $n \;\; 1$. Suppose that $\log N(; F; k j \;\; j_1 k_1) \;\; s f A + 1 \_ \log \;\; ^1 g$ for every $\;> 0$. Assume also that there exists $f \; 2 \; F$ such that $k j f \quad f j_1 k_1 \;\; _{app}$. Furthermore, suppose that $s \;\; 1$, $A \;\; 1$, $_{min} \;\; 1$, $_{app} \;\; 1$ and $\;2\; [_{min}; _{max}]$. Then, a sieve MLE $\hat{p}$ dened through (3.2) satises that

$$P d_H(\hat{p}; p) > \quad 5 e^{\;C_1 n^2_n} + \frac{C_2}{\;\;_n \; n} \qquad (3.4)$$

provided that $_n \;\; ^2 = 6 \;_n$ and $\sqrt{\;} \; 2 \;_n$, where

$$_n = C_3 \left(\sqrt{\frac{s f A + \log(n = _{min}) g}{n}} \; \_ \; \frac{_{app}}{\;} \right);$$

C_1 is an absolute constant, $C_2 = C_2(D)$ and $C_3 = C_3(D; K; _{max})$.

Using Theorem 3, we can derive the convergence rate of a sieve MLE of deep generative models for various f: As an illustrative example, suppose that $f \; 2 \; H_K \;(0; 1)^d$ for some positive constants and K. Since a smooth function can be eciently approximated by DNN, one can obtain a convergence rate as in the following corollary. We omit the proof because it is a special case of Corollary 6 with $q = 0$ and $d = d_0 = t_0$:

Corollary 4 Suppose that $f \; 2 \; H \quad (0;1)^d$, $= n$ and $_{min} = n$ for some $; K > 0$ and 0 . Then, there exists a network architecture $F = F(L; p; s; K)$ (depending only on $(n; d; ; K)$) such that a sieve MLE $\hat{p}$ satises

$$Pd_H(\hat{p}; p) > \quad 5e^{\;C_1 n^2}_{\;n} + \frac{C_2}{\quad} \quad {}_n \quad \frac{}{n}$$

provided that $_n \; ^2 = 6$ and $_n \; \overset{p}{\;} 2, \; _n$ where C_1, $C_2 = C_2(D)$; $= (D)$ are constants in Theorem 3 and $_n = Cn^{\;(\;\;d)=(2+d)}(\log n)^{3=2}$ with $C = C(; ; ; d; D; K; _{max})$.

The statement of Corollary 4 is overly simplied to illustrate the role of the dimension, smoothness and noise level in the convergence rate. In particular, the rate gets faster as the noise level increases. This seemingly paradoxical phenomenon occurs because p gets smoother as increases. On the other hand, for a very small value of , for consistent estimation of p it is necessary to have very accurate approximation of f. For this purpose, it is inevitable to increase the number of nonzero network parameters, which leads to an increase in the estimation error. In the set-up of Corollary 4, the number of nonzero network parameters s needed for a suitable degree of approximation is of order $n^{d(2+1)=(2+d)}$ up to a logarithmic factor. Note that the condition $> d$ is equivalent to that $d(2+1)=(2+d)$ is strictly smaller than 1. That is, when $\;d$, too many nonzero coecients are needed to ensure that the approximation error is suciently small. Consequently, Theorem 3 does not even guarantee the consistency. The case for a very small will be handled in Section 3.5 with a novel data perturbation technique. Before that, we assume that is not too small.

When f has a low-dimensional structure, the convergence rate in Corollary 4 can be signicantly improved. We consider the composition structure with low-dimensional smooth component functions as described in Section 3 of Schmidt-Hieber (2020). Specically, we consider a function f of the form

$$f = g_q \; g_{q \;\; 1} \quad g_1 \; g_0 \tag{3.5}$$

with $g_i : (a_i; b_i)^{d_i} \; ! \; (a_{i+1}; b_{i+1})^{d_{i+1}}$. Here, $d_0 = d$ and $d_{q+1} = D$. Denote by $g_i = (g_{i1}; : : : ; g_{id_{i+1}})^T$ the components of g_i and let t_i be the maximal number of variables on which each of the g_{ij} depends. Let $G(q; d; t; ; K)$ be the collection of functions of the form (3.5) satisfying $g_{ij} \; 2 \; H^{\;i}_{\;K} \; (a_i; b_i)^{t_i}$ and $ja_ij \_ jb_ij \quad K$, where $d = (d_0; : : : ; d_{q+1})^T$, $t = (t_0; : : : ; t_q)^T$ and $= (_0; : : : ; _q)^T$. It would be convenient to regard quantities $(q; d; t; ; K)$ as constants. Let

$$_j = {}_j \prod_{l=j+1}^{q} (_l \wedge 1); \qquad j = \underset{j2f0;::::;qg}{\mathrm{argmax}} \frac{t_j}{e_j}; \qquad = {}_j; \; e \qquad t = t_j:$$

We call t and e as the intrinsic dimension and smoothness of f (or of the function class $G(q; d; t; ; K)$), respectively.

Any function f in $G(q; d; t; ; K)$ can be eciently approximated by a DNN as detailed in Lemma 5. The proof can be easily deduced from the proof of Theorem 1 in Schmidt-Hieber (2020). Then, Corollary 6 provides the convergence rates of $\hat{p}$ when f has the composition structure.

Lemma 5 Suppose that f 2 G(q; d; t; ; K). Then, for every 2 (0; 1), there exists a network F = F(L; p; s; K _ 1) with L c_1 log ^{1}, jpj$_1$ c_2 $^{t=}$, s c_3 $^{t=}$ log ^{1} satisfying kjf fj$_1$k$_1$ for some f 2 F, where c_j = c_j(q; d; t; ; K) for j 2 f1; 2; 3g.

Corollary 6 Suppose that f 2 G(q; d; t; ; K), 2 [$_{min}$; $_{max}$], $_{min}$ 1 and

$$\overset{def}{=} \ {}^{2\ 2+t}\ \frac{\text{app}}{n} \qquad 1:$$

Let F = F(L; p; s; K _ 1) with L = [c_1 log $_{app}$ 1], p_0 = = p_{L+1} = [c_2 $_{app}^{=}$], s = [c_3 $_{app}^{=}$ log $_{app}$ t] + 1, where c_j = c_j(q; d; t; ; K), j 3, are constants in Lemma 5. Dene = (D) and as in Theorem 3 with A = c_4(log n) , where c_4 = c_4(q; d; t; ; K) as specied in the proof. If $_n$ ^{2}=6 and $_n$ $_n$ p 2, a sieve MLE p^ satises (3.4).

In Corollary 6, the approximation error $_{app}$ is chosen so that

$$_n \quad \sqrt{\frac{s}{n}} \quad \overline{\text{app}}$$

up to a logarithmic factor. More precisely, if = n and $_{min}$ = n for some 0, we have

$$\underset{n}{=} Cn \ {}^{2+t} (\log n)^{3=2};$$

where C = C(q; d; t; ; K; D; $_{max}$; ;). As one can see, the dimension d in the convergence rate of Corollary 4 is replaced by the intrinsic dimension t. If t is much smaller than d, the improvement from the structural assumption would be signicant.

3.4 Wasserstein Convergence Rate of a Sieve MLE of Q

Since we are primarily interested in estimating Q = Q_f, in this section we consider the problem of estimating Q and utilize the L ^{1}-Wasserstein metric as an evaluation metric. Given a sieve MLE (3.2), an estimator can be easily constructed as Q^ = $Q_{\hat f}$. Note that obtaining an upper bound of W_1(Q^; Q) from d_H(p; p^) is a kind of deconvolution problem. A sharp bound for this problem is established in Section 2.3 of Nguyen (2013) when and ^ are bounded away from zero. For example, with the L^2-Wasserstein metric, a sharp bound W^2_2(Q; Q^) . f log d_H(p; p^)g ^{1} is achievable, see Theorem 2 of Nguyen (2013). Hence, even when d_H(p; p^) decays with a polynomial rate, one can only expect a very slow convergence rate for W_2(Q^; Q); see also Fan (1991) and Meister (2009) for a more formal statistical theory for the deconvolution. Such a logarithmic minimax rate can also be found in a slightly dierent but closely related problem. More specically, Genovese et al. (2012a) considered the problem of estimating the support of the singular distribution Q and obtained a lower bound (log n) ^{1} for the minimax optimal rate under the Hausdor distance, see Theorem 8 therein. The slow minimax rates in the deconvolution and manifold estimation problems are closely related to the super-smoothness of the normal density. Here, a super-smoothness density roughly means that the tail of the Fourier transform of the density decays faster than any inverse polynomial, see Theorem 2 of Nguyen (2013). For a small value of , however, a much faster convergence rate is achievable because is no longer smooth.

Before studying the convergence rate, it would be worth addressing the identifiability issue. Since $p(x) = \int (x \quad u) dQ(u)$, Q can be understood as a mixing distribution for the data distribution P with the normal kernel. In this case, Q is identifiable under very mild conditions, see Bruni and Koch (1985). However, the identifiability does not guarantee an efficient estimation of Q. In some identifiable mixture models, the minimax convergence rate for estimating the mixing distribution can be very slow, see Wei and Nguyen (2022). A stronger identifiability condition is often necessary for obtaining a fast convergence rate of the mixing distribution.

In this subsection, we impose a strong identifiability condition through the reach of a manifold, which is introduced by Federer (1959) and frequently used in manifold estimation contexts. For a set $M \quad \mathbb{R}^D$ and $r > 0$, let $M^r = M \quad B_r(0_D)$ be the r-enlargement of M, where stands for the Minkowski sum. The reach of a closed set M, denoted as $\mathrm{reach}(M)$, is defined as the supremum of r with the property that any point in M^r has a unique Euclidean projection onto M.

In forthcoming Theorem 7, we assume that $\mathrm{reach}(M)$ is bounded below by a positive number, where M is the closure of $f(Z)$. This is one of the most important assumption in manifold estimation literature (Aamari and Levrard, 2019; Divol, 2021; Puchkin and Spokoiny, 2022; Tang and Yang, 2022). Note that even consistent estimation of Q may not be possible if $\mathrm{reach}(M) = 0$, as shown in Berenfeld and Hoffmann (2019).

Theorem 7 Let M be the closure of $f(Z)$. Suppose that $\|jfj_1\|_1 \quad K$ for a constant K. Also, assume that M does not have an interior point in $\mathbb{R}^D$, and $\mathrm{reach}(M) = r$ for some constant $r > 0$. Then, $d_H(p_{f;}; p) \quad 1$ and $\|jfj_1\|_1 \quad K$ imply that $W_1(Q_f; Q) \quad C(+ \sqrt{\log {}^{-1}})$, where $C = C(D; K; r)$.

Theorem 7 guarantees that $W_1(\hat{Q}; Q) \lesssim d_H(\hat{p}; p) +$ up to a logarithmic factor. Since we have already obtained a rate for $d_H(\hat{p}; p)$, it is possible to obtain a Wasserstein convergence rate for estimating Q. For example, when $f \in H_K (0; 1)^d$; Corollary 4 together with Theorem 7 implies that there exists a sieve of deep generative models with which the convergence rate of $W_1(\hat{Q}; Q)$ is $O_p\ n^{(\ d)=(2+d)}(\log n)^{3=2} \quad \sqrt{\log n}$.

Remark 8 Note that Theorem 7 does not require $f(Z)$ to be a topological or smooth manifold. For example, $f(Z)$ can be a union of two manifolds with different dimensions.

3.5 Data Perturbation

When is too small, the convergence rates of $d_H(\hat{p}; p)$ obtained in Corollaries 4 and 6 do not even converge to 0 as the sample size increases: in Corollary 6, for example, when $n^{=t}$, with $< t$. Under these regimes, p peaks around a small neighborhood of $f(Z)$ and the singularity exacerbates, thus a sieve MLE does not behave well. In an extreme case where $= 0$; P itself is a singular measure and likelihood approaches cannot be justified via minimizing the Kullback–Leibler (KL) divergence.

To overcome these difficulties, we consider the perturbed observations $\tilde X_i = X_i + \tilde\epsilon_i$, where $\tilde\epsilon_i \sim N(0_D, \tilde\sigma^2 I_D)$ is an artificial noise vector. Note that $\tilde X_1, \ldots, \tilde X_n$ can be under-stood as i.i.d. observations from the true distribution $\tilde P_* = P_{f_*,\tilde\sigma_*}$, where $\tilde\sigma_*^2 = \sigma_*^2 + \tilde\sigma^2$. Let $(\hat f_{per}, \hat\sigma_{per})$ be a sieve MLE based on the perturbed observation $\tilde X_1, \ldots, \tilde X_n$. Also, define $\hat P_{per} = P_{\hat f_{per},\hat\sigma_{per}}$ and $\hat Q_{per} = Q_{\hat f_{per}}$ accordingly.

Once we use $\hat Q_{per}$ as an estimator for Q_*, we have $W_1(\hat Q_{per}, Q_*) \lesssim \tilde\sigma_* + \tilde\epsilon_n \tilde\sigma_*^{-1}\sqrt{\log \tilde\epsilon_n^{-1}}$ by Theorem 7, where $\tilde\epsilon_n = d_H(\hat p_{per}, \tilde p_*)$. As $\tilde\sigma$ increases, note that $\tilde\epsilon_n$ decreases while $\tilde\sigma_*$ increases. Thus, the convergence rate for $W_1(\hat Q_{per}, Q_*)$ can be optimized by choosing $\tilde\sigma$ accordingly, which is summarized in the following theorem.

Theorem 9 Let $n \ge 1$, $f_* \in \mathcal{G}(q, d, t, \beta, K)$, $\sigma_* \in [\sigma_{min}, \sigma_{max}]$, $\sigma_* = n^{-\alpha}$ and $\sigma_{min} = n^{-\gamma}$ for some $\gamma \ge \alpha \ge 0$. Assume that $Q_*(M) = 1$ and reach$(M) \ge r$, where $r > 0$ and M is the closure of $f_*(\mathcal{Z})$. Then, there exists a network architecture $\mathcal{F} = \mathcal{F}(L, p, s, K)$ (de-pending only on (n, q, d, t, β, K)) such that sieve MLEs $\hat p_{per}$ and $\hat Q_{per}$ based on the perturbed observation $\tilde X_i = X_i + \tilde\epsilon_i$, with $\tilde\epsilon_i \sim N(0_D, n^{-\beta/(\beta+t)} I_D)$, satisfies

$$P\left(W_1(\hat Q_{per}, Q_*) > C_3\left(\sigma_* + \sqrt{\epsilon_n \log(1/\epsilon_n)}\right)\right) \le 5e^{-C_1 n\epsilon_n^2} + \frac{C_2}{n}, \tag{3.6}$$

where

$$\epsilon_n = \begin{cases} C_4 n^{-\frac{\beta - \alpha t}{2\beta+t}}(\log n)^{3/2} & \text{if } \alpha < \beta/\{2(\beta+t)\}, \\ C_5 n^{-\frac{\beta}{2(\beta+t)}}(\log n)^{3/2} & \text{otherwise}; \end{cases}$$

C_1 is an absolute constant, $C_2 = C_2(D)$, $C_3 = C_3(D, K, r)$, C_4 and C_5 depend only on $(q, d, t, \beta, K, D, \sigma_{max}, \alpha, \gamma)$.

To the best of our knowledge, our main result (Theorem 9) is the first theory considering the Wasserstein convergence of $\hat Q$ in a deep generative model with the intrinsic dimension and smoothness of f_*. Most existing theories consider GAN type estimators and have derived convergence rates that depend on either the intrinsic dimension alone or D.

If $\alpha < \beta/\{2(\beta+t)\}$, we have $\epsilon_n \lesssim \sigma_*$ so $\sigma_*\sqrt{\log n}$ in the left hand side of (3.6) is the dominating term. Therefore, regardless of $\alpha < \beta/\{2(\beta+t)\}$, we conclude that

$$W_1(\hat Q_{per}, Q_*) \lesssim n^{-\frac{\beta}{2(\beta+t)}}(\log n)^{3/2} + \sigma_*\sqrt{\log n} \tag{3.7}$$

with high probability. Since $W_1(\hat Q_{per}, Q_*)$ is a bounded random variable, its expectation can also be easily bounded by a multiple of the right hand side of (3.7).

It can be easily deduced from the proof that the data perturbation improves the con-vergence rate only when $\sigma_* \lesssim n^{-\beta/2(\beta+t)}$. Note that the level of perturbation and the network architecture in Theorem 9 depend on the unknown quantities (β, t, α). In other words, our results are non-adaptive to the unknown structure. Hence, the network architec-tures and $\tilde\sigma$ are tuning parameters that should be carefully chosen. To obtain an estimator adaptive to the unknown structure, two approaches are known in the literature for the deep supervised learning. The first one is a penalized likelihood approach such as the lasso and non-convex penalties as considered in Ohn and Kim (2022). Alternatively, Bayesian ap-proaches can be utilized to obtain an adaptive estimator, see Polson and Rockova (2018);

Ohn and Lin (2021). Although these papers studied nonparametric regression, it would be possible to extend their approaches to deep generative models to obtain an adaptive estimator. In practice, there are several heuristic methods to select network architectures (Salimans et al., 2016; Arjovsky et al., 2017; Radford et al., 2016b). The variance e^2 of the additional noise is 1-dimensional, hence it can also be tuned based on the validation error without much diculty; see Section 5 for details.

After the original version of this article was drafted, the rst author investigated the lower bound for the minimax optimal convergence rate with the structural assumption considered in Theorem 9, which is now available in Chae (2022). Specically, he obtained a lower bound $n^{=(2+t\ 2)} + = \sqrt{\ } n$ of the minimax optimal rate. In particular, he provided some rationale for that the rst term $n^{=(2+t\ 2)}$ is sharp. Furthermore, he constructed a GAN type estimator, which achieves the rate $n^{=(2+t)} +$. Therefore, the rate given in Theorem 9 is not optimal. Nonetheless, the dierence is not signicant. Also, the estimator in Chae (2022) is devised for theoretical purposes, and it is not clear to us how to compute it in practice. We would like to emphasize that although likelihood-based approaches are not theoretically optimal, they are popularly used in practice because their computation is much easier than that of GAN.

It would also be important to study lower bounds specically for likelihood approaches considered in this paper. More specically, one may try to obtain a sharp lower bound for $\sup_Q EW_1(\hat{Q}_e; Q)$, where $\hat{Q}_e$ is a sieve MLE based on the perturbed data $X_i = eX_i + {}_i$ with e_i $N(0_D, e\ I_D)$, and Q ranges over structured distributions considered in Theorem 9. Ideally, we hope

$$\inf_{e0} \sup_{Q} EW_1(\hat{Q}_e; Q) \gtrsim n^{=2(+t)} + ;$$

matching with the upper bound given in Theorem 9. To achieve this goal, we would need two arguments. Each of them is challenging and of independent interest. Firstly, we would need a sharp lower bound for the approximation error of deep neural networks. This would be related to Park et al. (2021), but a far more thorough study is necessary. Another one is regarding the identiability issue; we would need $\|f \quad f\| \lesssim W_1(Q_f; Q_f)$ or a similar inequality, the reverse of $W_1(Q_f; Q_f) \lesssim \|f \quad f\|$. Obtaining such a reverse inequality is known to be challenging; see Nguyen (2013); Wei and Nguyen (2022). Due to these diculties, we do not consider this problem in this paper and leave it as future work.

3.6 Eect of into the Convergence Rate

It is worthwhile to discuss the eect of the noise level into the convergence rate (3.7). Firstly, suppose that is a xed positive constant. Then, the rate (3.7) does not give useful information because the right hand side is not small enough. In fact, estimating Q under an additive noise is known as a deconvolution problem, for which extensive studies have been done in the literature (Fan, 1991; Meister, 2009; Nguyen, 2013). The minimax optimal rate for the Gaussian deconvolution with a xed is very slow, e.g. $(\log n)^{\ 1}$, implying the intrinsic diculty of the estimation problem. Such an intrinsic diculty has also been observed in Genovese et al. (2012a) who considered a slightly dierent problem. Specically, they obtained the minimax optimal rate for estimating the support of Q under the Hausdor distance, see Theorem 8 therein. They assumed that Q is supported on a low-dimensional manifold, but the intrinsic slow rate $(\log n)$ ^{1} was unavoidable. Although their manifold estimation problem is slightly dierent from the deconvolution, they are closely related to each other as discussed in Section 1.1 of Genovese et al. (2012a). Given the inherent challenges of the deconvolution problem, it does not seem possible to achieve a fast convergence rate in estimating Q under xed variance Gaussian noise. In this sense, the constant variance set-up would not be appropriate for studying the amazing performance of deep generative models theoretically.

The rate (3.7) gives meaningful results when is small enough in the sense that converges to zero with a suitable rate as the sample size increases. In this case, data are concentrated in a small neighborhood of a certain low-dimensional structure; hence one may utilize the structural benet to estimate Q eciently. Note that although the set-up is not exactly the same as ours and dierent estimation problems (such as the manifold or regression function) are considered, there are many recent theoretical articles adopting the regime in which data are concentrated around a very small neighborhood of a manifold (Aamari and Levrard, 2018, 2019; Divol, 2021; Jiao et al., 2021; Puchkin and Spokoiny, 2022; Berenfeld et al., 2022); see also Remark 4 of Tang and Yang (2022). In these papers, small neighborhoods depend on the sample size and shrink to a low-dimensional manifold.

Despite the above observation, we wish to emphasize that our results or the probability bounds are again non-asymptotic in nature. That is, for every n, our results hold simultaneously for a range of 's with $2 [_{min}; _{max}]$.

4. Class of True Distributions

Asymptotic properties of a sieve MLE are investigated in the previous sections under the assumption that $P = P_{f;}$ for some f and , that is, P is the convolution of Q_f and $N(0_D; I_D)$. In this section we characterize the class of probability distributions of the form Q_f. In particular, we will show that the class $fQ_f : f\ 2\ Fg$ is quite general to include various structured distributions when f ranges over a certain class F of structured functions. Specically, we will show that various distributions can be represented as Q_f for some function f. Throughout this section, we assume that Z P_Z and Y is a random vector whose distribution Q satises that $Q(Y) = 1$ for Y R^D. A primary goal is to nd a map $f : Z\ !\ R^D$ satisfying $Q = Q_f$. Lu and Lu (2020) considered a similar topic, but they did not consider structures of f such as the smoothness, which are important for obtaining a fast convergence rate.

4.1 Case D = d = 1: 1-dimensional Distributions or Smooth Densities

Suppose that both Y and Z are absolutely continuous real-valued random variables with the cumulative distribution functions F_Y and F_Z, respectively. Then, it is well-known that $F_Y^{1}(F_Z(Z))$ is distributed as Q, where $F_Y^{1}(u) = \inf fy\ 2\ R : F_Y(y) > ug$ is the generalized inverse of F_Y. That is, $Q = Q_f$, where $f = F_Y^{1} F_Z$. Furthermore, it is known that the map f is the unique optimal transport from P_Z to Q with respect to the quadratic cost function, see Section 2.2 of Villani (2003). If Z follows Uniform(0; 1), for example, the smoothness of f is determined by the smoothness of F_Y^{1}. Informally, if the pdf q is -smooth and strictly positive on Z, then F_Y^{1} is $(+1)$-smooth, see Lemma 10 for a formal statement. Note that a smooth 1-dimensional function f can be approximated by DNN eciently. Roughly, if

$f \in \mathcal{H}^\beta$, then for any $\epsilon > 0$, there exists $f^{nn} \in \mathcal{F}(L; p; s; \infty)$ with $L \lesssim \log \epsilon^{-1}$, $|p|_\infty \lesssim \epsilon^{-1/\beta}$ and $s \lesssim \epsilon^{-1/\beta} \log \epsilon^{-1}$ such that $\| |f - f^{nn}|_\infty \|_\infty \le \epsilon$, see Theorem 5 of Schmidt-Hieber (2020).

4.2 Product Distributions

Assume that $D = d$ and $Y = (Y_1; \ldots; Y_D)^T$, where $Y_1; \ldots; Y_D$ are independent random variables. That is, Q is the product probability of $Q_1; \ldots; Q_D$, where Q_j is the distribution of Y_j. If $Z_1; \ldots; Z_D$ are i.i.d. random variables, there exist univariate functions f_j, $j = 1; \ldots; D$, such that Q_j is the distribution of $f_j(Z_j)$, as argued in Section 4.1. Therefore, the map f defined as $f(z) = (f_1(z_1); \ldots; f_D(z_D))^T$ satisfies that $Q = Q_f$. As before, if densities $q_1; \ldots; q_D$ exist and sufficiently smooth, f can be choosen as a smooth function. Specifically, if each $q_j \in \mathcal{H}^\beta$ for every j, one can find $f^{nn} \in \mathcal{F}(L; p; s; \infty)$ with $L \lesssim \log \epsilon^{-1}$, $|p|_\infty \lesssim \epsilon^{-1/\beta}$ and $s \lesssim \epsilon^{-1/\beta} \log \epsilon^{-1}$ such that $\| |f - f^{nn}|_\infty \|_\infty \le \epsilon$. That is, we only need to approximate D many 1-dimensional smooth functions.

4.3 Classical Smooth Densities

Suppose that $D = d$ and Q has the Lebesgue density q. An open set $\Omega \subset \mathbb{R}^r$ is said to be uniformly convex if there exists a twice continuously differentiable function $h : \mathbb{R}^r \to \mathbb{R}$ and a constant $\lambda > 0$ such that $\Omega = \{x \in \mathbb{R}^r : h(x) < 0\}$ and $\nabla^2 h(x) - \lambda I_r$ is positive definite for every $x \in \mathbb{R}^d$, where $\nabla^2 h(x)$ is the Hessian matrix. Note that a uniformly convex set is automatically bounded. The following lemma is a special case of Theorem 12.50 in Villani (2008), originally proven by Caffarelli (1990) and Urbas (1988). As mentioned in Villani (2003), techniques involved in Lemma 10 are really intricate. We refer to page 139 of Villani (2003) for more references about this topic.

Lemma 10 Suppose that (i) $\mathcal{Z}$ and $\mathcal{Y}$ are uniformly convex, (ii) p_Z and q are bounded from above and below on $\mathcal{Z}$ and $\mathcal{Y}$, respectively, and (iii) $q \in \mathcal{H}^\alpha(\mathcal{Y})$ and $p_Z \in \mathcal{H}^\alpha(\mathcal{Z})$ for $\alpha > 0$. Then, there exists a function $f = (f_1; \ldots; f_d) : \mathcal{Z} \to \mathcal{Y}$ such that $Q = Q_f$ and $f \in \mathcal{H}^{\alpha+1}$.

The map f in Lemma 10 is the unique optimal transport from P_Z to Q with respect to the quadratic cost function. For statistical purpose, a map f needs not to be an optimal transport, therefore, conditions on P_Z and Q can be relaxed. For example, note that the uniform distribution on the unit ball $B(0_d)$ has a density which is bounded from above and below, and $B(0_d)$ is uniformly convex. Hence, if Q satisfies the condition in Lemma 10 and there exists a map $h : \mathcal{Z} \to B(0_d)$ such that $h(Z) \sim \mathrm{Uniform}(B(0_d))$, Lemma 10 guarantees the existence of f satisfying $Q = Q_f$. If P_Z is the uniform distribution on the unit cube $(0; 1)^d$, which is a popular choice in practice, such h can be chosen as a smooth function, see Harman and Lacko (2010). Conditions on Q, such as the uniform convexity of $\mathcal{Y}$, can be relaxed in a similar way. Finally, we note that if $f \in \mathcal{H}^{\beta+1}$, there exists $f^{nn} \in \mathcal{F}(L; p; s; \infty)$ with $L \lesssim \log \epsilon^{-1}$, $|p|_\infty \lesssim \epsilon^{-d/(\beta+1)}$ and $s \lesssim \epsilon^{-d/(\beta+1)} \log \epsilon^{-1}$ such that $\| |f - f^{nn}|_\infty \|_\infty \le \epsilon$.

4.4 Distributions on a Manifold

We consider the case where $\mathcal{Y} \subset \mathbb{R}^D$ is a topological manifold with dimension $d \le d$. We start with the case that $\mathcal{Y}$ can be covered by a single chart, that is, there exists a homeomorphism $\varphi : B_1(0_d) \to \mathcal{Y}$. We further assume that $\varphi \in \mathcal{H}^{\beta+1}$ for $\beta > 0$ as a map from $B_1(0_d)$ to $\mathbb{R}^D$, and that $\inf_{x \in B_1(0_d)} |J_\varphi(x)|$ is bounded below by a positive constant, where

$$|J_\varphi(x)| = \sqrt{\det\left(\frac{\partial \varphi}{\partial x}^T \frac{\partial \varphi}{\partial x}\right)}$$

is the Jacobian determinant of φ. Note that a coordinate chart in a smooth manifold is automatically smooth by the definition of a smooth map between manifolds, cf. Lee (2013). Therefore, the ordinary differentiability $\varphi \in H^{+1}$ is an additional condition. This kind of condition is frequently used in literature, see Schmidt-Hieber (2019); Nakada and Imaizumi (2020).

Furthermore, we impose some smooth conditions on the distribution Q. Note that if D is strictly larger than d, the distribution Q cannot possess a Lebesgue density because Y is a null set. We instead consider a density with respect to the Hausdorff measure. Let H_d be the d-dimensional Hausdorff measure in $\mathbb{R}$, which is normalized so that it is the same as the Lebesgue measure if $D = d$. Suppose that Q allows the Radon–Nikodym derivative q with respect to H_d. We further assume that q is bounded from above and below, and that $q \circ \varphi \in H$. Then, by the change of variable formula, the Lebesgue density of $\tilde{Q}$, the distribution of $\varphi^{-1}(Y)$, is given as

$$\tilde{q}(x) = q \circ \varphi(x)|J_\varphi(x)|.$$

Since $|J_\varphi(x)| = 0$ and $\varphi \in H^{+1}$, it is not difficult to see that $|J_\varphi(x)|$ is bounded from above and below, and the map $x \to |J_\varphi(x)|$ belongs to H. Hence, $\tilde{q}$ is bounded from above and below, and belongs to $H(B_1(0_d))$. By Lemma 10, under mild assumptions on P_Z, there exists $g \in H^{+1}(Z)$ such that $\tilde{Q} = Q_g$. Thus, we have $Q = Q_f$, where $f = \varphi \circ g \in H^{+1}$ is a map from Z to $\mathbb{R}^D$. As in Section 4.3, one can choose $f^{nn} \in F(L; p; s; 1)$ with $L \lesssim \log{}^{-1}$, $|p|_\infty \lesssim {}^{-d/(+1)}$ and $s \lesssim {}^{-d/(+1)} \log{}^{-1}$ such that $\|f - f^{nn}\|_\infty \lesssim$.

Now, we illustrate the case of multiple charts. Suppose that a distribution Q is supported on a d-dimensional manifold M that can be covered by J charts (U_j, φ_j), $j = 1, \dots, J$, where $J > 1$. Here, $U_j \subset Y$ are open sets, with homeomorphism $\varphi_j : B_1(0_d) \to U_j$. As before, we further assume that $\varphi_j \in H^{+1}$, $\inf_{x \in B_1(0_d)} |J_{\varphi_j}(x)|$ is bounded below by a positive constant, Q possesses a Hausdorff density that is bounded from above and below, and that $q \circ \varphi_j \in H$. Let $Q_j(\cdot) = Q(\cdot)/Q(U_j)$ be the normalized measure of Q over U_j and denote its corresponding Hausdorff density as q_j. Note that for $y \in U_i \cap U_j$, one has $q_i(y)Q(U_i) = q_j(y)Q(U_j) = q(y)$ because $Q(U_i)Q_i(\cdot)$ and $Q(U_j)Q_j(\cdot)$ agree with Q on $U_i \cap U_j$.

Next we will show that Q can be patched together from Q_j via a partition of unity. Note that a partition of unity of a topological space Y is a set of continuous functions $\{_j : j \in J\}$ from Y to the unit interval $[0, 1]$ such that for every point, $y \in Y$, there is a neighborhood U of y where all but a finite number of the functions are 0, and the sum of all the function values at y is 1, i.e., $\sum_{j \in J} {}_j(y) = 1$. A compact manifold M always admits a finite partition of unity $\{_j : j = 1, \dots, J\}$, $_j(\cdot) : M \to [0, 1]$ such that $\sum_{j=1}^J {}_j(y) = 1$. Furthermore, one can construct $\{_j : j = 1, \dots, J\}$ so that each $_j$ is sufficiently smooth and $_j(y) = 0$ for $y \notin U_j$, see Lemma 3 of Schmidt-Hieber (2019).

Since $q(y) = Q(U_j)q_j(y)$ for each j and $y \in U_j$, one has $q(y) = \sum_{j=1}^{J} Q(U_j)_j(y)q_j(y)$. Let $\tilde{q}_j(y) = c_j{}_j(y)q_j(y)$, where $c_j = [\int {}_j(y)dQ_j(y)]^{-1}$ is the normalizing constant. Then, $q(y) = \sum_{j=1}^{J} {}_j\tilde{q}_j(y)$, where ${}_j = Q(U_j)/c_j$. That is, q is a mixture of $\tilde{q}_j$'s. Since $\tilde{q}_j$ is sufficiently smooth, one can construct $f_j : \tilde{Z} \to \tilde{Y}$ such that Q_j is the distribution of $f_j(\tilde{Z})$ as in the single chart case, where $\tilde{Z}$ is a uniformly convex subset of $\mathbb{R}^d$ and $\tilde{Z}$ follows the uniform distribution on $\tilde{Z}$. Let $Z = (0,1) \times \tilde{Z}$ and P_Z be the product distribution of Uniform$(0,1)$ and the distribution of $\tilde{Z}$. Let $I_1, \dots, I_J$ be disjoint consecutive intervals with lengths ${}_1, \dots, {}_J$ partitioning $(0,1)$, that is, $I_1 = (0, {}_1)$ and $I_j = [\sum_{i=1}^{j-1} {}_i, \sum_{i=1}^{j} {}_i)$ for $j = 2, \dots, J$. Let h_j be the indicator function for the interval I_j. Then, for a random variable Z following Uniform$(0,1)$, we have $P_Z(h_j(Z) = 1) = 1 - P_Z(h_j(Z) = 0) = {}_j$. For $z = (z_1, z_2) \in \mathbb{R}^{d+1}$, define $f(z) = \sum_{j=1}^{J} h_j(z_1)f_j(z_2)$. Then, it is not difficult to see that $Q = Q_f$. Note that each f_j can be efficiently approximated by ReLU network functions as the single chart case. Also, 1-dimensional indicator functions $h_1, \dots, h_J$ can be approximated by piecewise linear functions. Therefore, it is easy to approximate them by shallow ReLU network functions. Finally, the multiplication of h_j and f_j can also be well-approximated by ReLU networks.

Remark 11 Strictly speaking, the regularity of the map f_j is not guaranteed because ${}_j$ is not bounded from below. From the construction of ${}_j$ in Schmidt-Hieber (2019), however, it can be seen that ${}_j$ vanishes only at the boundary of U_j (relative to M). Hence, one may construct a sufficiently regular f_j such that $Q_j \approx Q_{f_j}$. A more rigorous treatment of this topic would be very technical, and we leave it as future work.

5. Numerical Experiments

In this section, we empirically demonstrate that the data perturbation method proposed in Section 3.4 plays an important role to improve the performance of a sieve MLE of deep generative models. In addition, we illustrate that deep generative models can detect low-dimensional structures well. Numerical studies are carried out by analyzing various synthetic and real data sets and comparisons are made between our estimators and others such as the MLE of a linear factor model, GAN and Wasserstein GAN.

5.1 Synthetic and Real Data Sets

5.1.1 Synthetic Data

For simulation study, we firstly consider distributions on 1-dimensional manifolds. Specifically, we generate data from the model $X = f(Z) + \epsilon$ with $D = 2$ and $\sigma = 0$, where Z is a univariate random variable following Uniform$(0,1)$. For the true generator $f = (f_1, f_2)$, we consider the following three functions:

$$
\begin{aligned}
&\text{Case 1.}\ f_1(z) = 6(z - 0.5); && f_2(z) = 0.5(z - 2)z(z + 2)\\
&\text{Case 2.}\ f_1(z) = 2\cos(2z); && f_2(z) = 2\sin(2z)\\
&\text{Case 3.}\ \begin{array}{l} f_1(z) = 2\cos(2z) + 1;\ f_2(z) = 2\sin(2z) + 0.4 \ \text{if } z > 0.5\\ f_1(z) = 2\cos(2z) - 1;\ f_2(z) = 2\sin(2z) - 0.4 \ \text{otherwise.} \end{array}
\end{aligned}
\tag{5.1}
$$

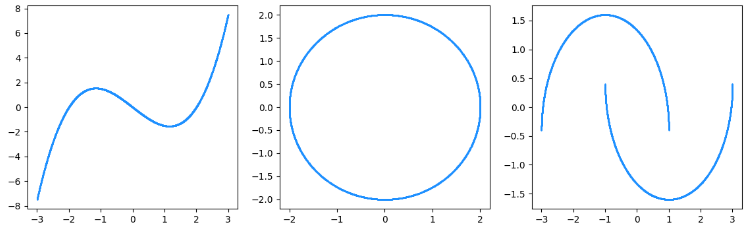

Figure 1: Supports of Q for the three synthetic data sets in (5.1).

The supports of Q for the three cases are depicted in Figure 1. The generator of Case 2 leads the uniform distribution on a circle. Note that a circle cannot be covered by a single chart. Also, for Case 3, the true generator is discontinuous. In this case, the support of Q is the union of two disjoint 1-dimensional manifolds.

We next consider two more distributions, a distribution on the Swiss roll (Marsland, 2015) and the uniform distribution on the sphere, which are supported on 2-dimensional manifolds with the ambient space R^3. The distribution on the Swiss roll is the distribution of $f(Z)$, where Z follows the uniform distribution on $(0;1)^2$ and the true generator $f = (f_1; f_2; f_3) : (0;1)^2$! R^3 is dened as

$$t_1 = 1{:}5(1 + 2z_1); \qquad t_2 = 21z_2;$$
$$f_1(z_1; z_2) = t_1 \cos(t_1); \qquad f_2(z_1; z_2) = t_2; \qquad f_3(z_1; z_2) = t_1 \sin(t_1):$$

Similar to the circle, the sphere cannot be covered by a single chart. In all the experiments, the sample sizes of validation and test data are set to be 3,000, while the training sample size varies.

5.1.2 Big Five Personality Traits Data Set

The big ve personality traits data set (Big-ve; Goldberg (1990)) consists of answers for 50 questions, with the ve-level Likert scale (1 to 5) from 1,015,342 respondents. This data set has been frequently analyzed in literature with linear factor models, see Ohn and Kim (2021) and references therein. We only use the data of the 874,434 respondents who answer to all questions completely. Each variable is rescaled to take values from 1 to 1. We randomly draw 20,000 samples from the entire data, 10,000 of which are used as validation data and the others as test data. The remains are used as training data.

5.1.3 MNIST and Omniglot Data Sets

We analyze two well-known image data sets, MNIST and Omniglot. MNIST data set (LeCun et al., 1998) contains handwritten digit images of 28 28 pixel sizes and has a training data set consisting of 60,000 images and a test data set of 10,000 images. We randomly sample 10,000 images from the training data set and use them as validation data. Omniglot (Lake et al., 2015) data set consists of various character images of 28 28 pixel sizes taken from 50 dierent alphabets. It has 24,345 training samples and 8,070 test samples. As before, we split the training data set into two subsets, each of which has 20,000 and 4,345 samples, respectively, and use one for training data and the other for validation data.

5.2 Learning Algorithm to Obtain the MLE

Assume that the generator $f = f$ is parametrized by . With a slight abuse of notation, let $p_{;}$ = $p_{f;}$, that is,

$$p_{;}(x) = \int (x \quad f(z))\, dP_Z(z):$$

Mostly, the log-likelihood is computationally intractable. Alternatively, one can maximize a lower bound of the log-likelihood by use of a family of variational distributions using methods of variational inference (Jordan et al., 1999). The most well-known algorithm is the variational autoencoder (VAE; Kingma and Welling, 2014; Rezende et al., 2014) and the lower bound used in VAE is often called the ELBO (evidence lower bound).

Various alternative lower bounds of the log-likelihood that are tighter than the ELBO but still computationally tractable, have been proposed afterwards, see Burda et al. (2016); Cremer et al. (2017); Kingma et al. (2016); Rezende and Mohamed (2015); Salimans et al. (2015); Snderby et al. (2016). Among these, the importance weighted autoencoders (IWAE, Burda et al., 2016) is an important variant of the VAE. Recently, it is shown that IWAE can be understood as an EM algorithm to obtain the MLE, see Dieng and Paisley (2019); Kim et al. (2020). Thus, we use the IWAE algorithm to obtain a sieve MLE. Specically, let z ! $q(z\,j\,x)$ be a variational density parametrized by . A popular choice for $q(\ j\,x)$ is the density of $N((x); (x))$, where x ! (x) and x ! (x) are DNN functions with network parameters . For given i.i.d. samples $Z_1; : : : ; Z_K$ from $q(jx)$, let

$$\hat{L}^{IWAE}(; ; ; x) := \log \left(\frac{1}{K} \sum_{k=1}^{K} \frac{p_{;}(x; Z_k)}{q(Z_k jx)} \right);$$

where $p_{;}(x; z) = p_Z(z)(x \quad f(z))$ and K is a given positive integer. Then, IWAE simultaneously estimates ; and by maximizing $\sum_{i=1}^{n} \hat{L}^{IWAE}(; ; ; X_i)$: We set $K = 10$ throughout our experiments.

5.3 Implementation Details

5.3.1 Data Perturbation

The model is trained after perturbing the training data by an articial noise $N(0_D; e^2 I_D)$. For each data set, we consider various values of e.

5.3.2 Architectures

For analyzing ve synthetic and Big-ve data sets, we consider DNN architectures with the leaky ReLU activation function (Xu et al., 2015). For the variational distribution $q(\ j\,x)$, we use the multivariate normal distribution $N((x); (x))$, where (x) is a diagonal matrix. Both the mean and variance are modelled by DNNs. For synthetic data, we set $L = 2$, $d = 10$, $p = (d; 200; 200; D)$ for f, and $L = 2$, $p = (D; 200; 200; d)$ for and . For the Big-ve data set, we set $L = 3$, $d = 5$, $p = (d; 200; 200; 200; D)$ for f, and $L = 3$, $p = (D; 200; 200; 200; d)$ for and .

For analyzing two image data, we use a deep convolutional neural network (Radford et al., 2016a) with $L = 6$ and the ReLU activation function for modeling f. Also, convolutional neural networks with $L = 6$ and the leaky ReLU activation function are used to build model architectures for and . For the both data sets, we set $d = 40$.

5.3.3 Optimization

We train deep generative models using the Adam optimization algorithm (Kingma and Ba, 2015) with a mini-batch size of 100. The learning rate is xed as 10^{3} for synthetic and Big-ve data, and $3\ 10^{4}$ for two image data.

5.3.4 Sparse Learning Framework

For learning sparse generative models, we adopt the pruning algorithm proposed by Han et al. (2015). Firstly, a non-sparse model is trained with a pre-specied maximum number of training epochs, 200 in our experiments, and then the number of training epochs which minimizes the IWAE loss on the validation data is chosen. Next, the model is pruned by zeroing out small weights. Specically, 25% of small weights are replaced by zero. We then re-train the model keeping the zero weights unchanged. This procedure is repeated one more time to make 50% of the total weights become zero in the nal model.

5.4 Performance Comparisons

The performance of a given estimator $\hat{Q}$ is evaluated by the Wasserstein distance $W_1(\hat{Q}; Q)$ estimated on test data as follows. Let $\hat{Q}_M$ be the empirical measure based on the M i.i.d. samples from $\hat{Q}$. Note that it is easy to generate samples from $\hat{Q}$ via the estimated generator. Similarly, let Q_M be the empirical measure based on the M observations in test data. Then, $W_1(\hat{Q}; Q)$ can be estimated by $W_1(\hat{Q}_M; Q_M)$. In general, $W_1(\hat{Q}_M; Q_M)$ can be computed via a linear programming. We use a more stable algorithm developed by Cuturi (2013). We call $W_1(Q_M; \hat{Q}_M)$ the estimated W_1 distance.

5.4.1 Results for Synthetic Data

For the three 1-dimensional synthetic data sets, various training sample sizes ranging from 100 to 50,000 are considered . For each case, we obtain a sieve MLE for three times with random initialization and report the average based on the three sieve MLEs. Firstly, we trace the estimated variance $\hat{}^2$: Figure 2 draws the values of $j\hat{}\ \ ej{=}e$ as the sample size increases, where $e^2 = {}^2 + e^2 = e^2$. It seems that $j\hat{}\ \ ej{=}e\ !\ 0$ as n increases regardless of the value of e^2, which suggests that sieve MLEs perform reasonably well.

The estimated W_1 distances for various training sample sizes are shown in Figure 3. It is interesting to see that the estimated W_1 distance of a sieve MLE does not converge to 0 when e^2 is either too small or too large, which well corresponds to Theorem 7. Figure 4 provides the curves of the estimated W_1 distances over the degree of perturbation (i.e. e) with the training sample size being xed at $n = 50; 000$: As can be seen, the estimated W_1 distance is minimized at an intermediate value of e in all three cases, which again conrms

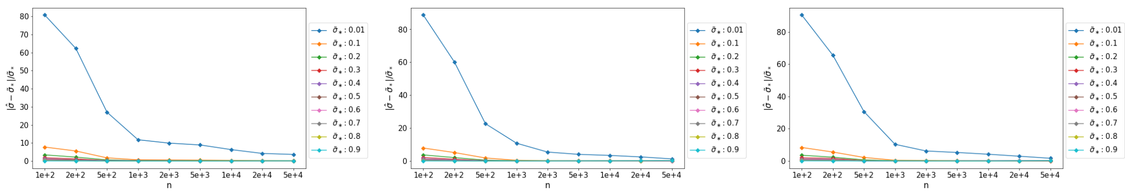

Figure 2: Values of j^ ej=e for various e and n for the three 1-dimensional synthetic data sets.

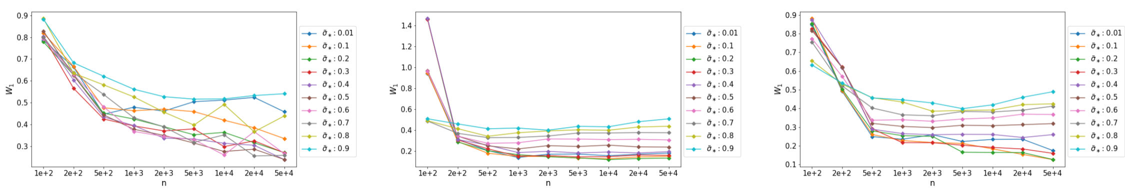

Figure 3: The estimated W_1 distance over the sample size with various values of e for the three 1-dimensional synthetic data sets.

the validity of our theoretical results. Figure 5 presents generated samples from Q̂ estimated with n = 50; 000 and the optimal choice of e that minimizes the estimated W_1 distance.

Similar phenomena can be found for the Swiss roll and sphere models. That is, the estimated W_1 distance is minimized at an intermediate value of e. Generated samples from Q̂ with n = 50; 000 and the optimal choice of e are plotted over the support of Q in Figure 6.

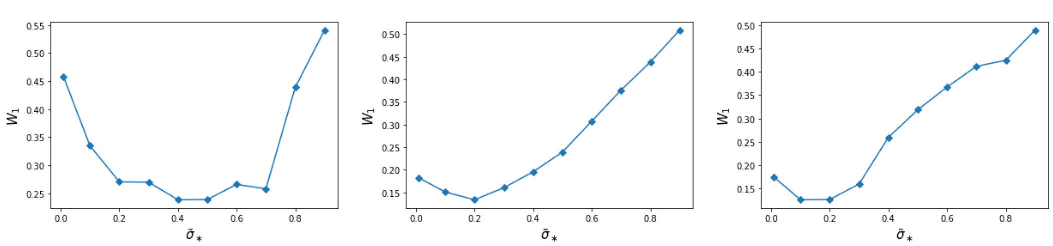

Figure 4: The estimated W_1 distance over e with the training sample size being xed at n = 50; 000 for the three 1-dimensional synthetic data sets .

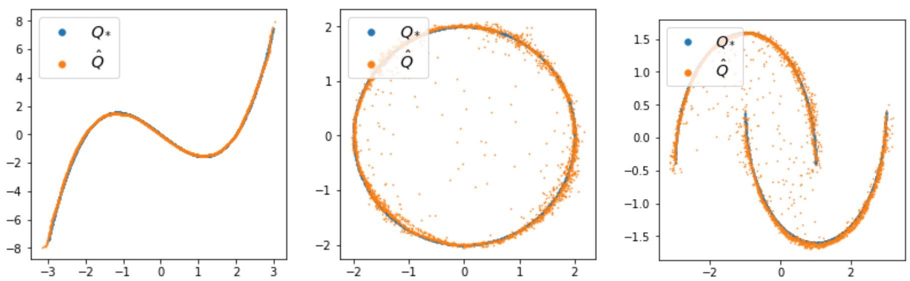

Figure 5: Generated samples from $\hat{Q}$ for the three 1-dimensional synthetic data sets.

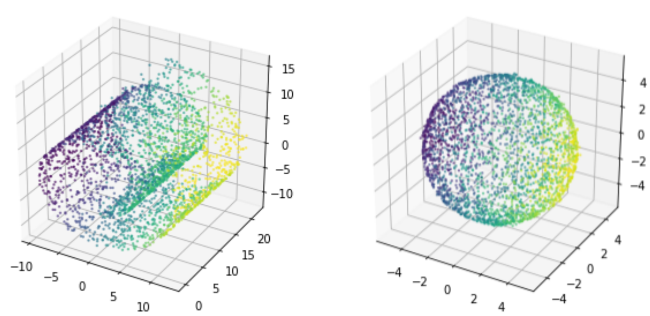

Figure 6: Generated samples from $\hat{Q}$ for the two 2-dimensional synthetic data sets.

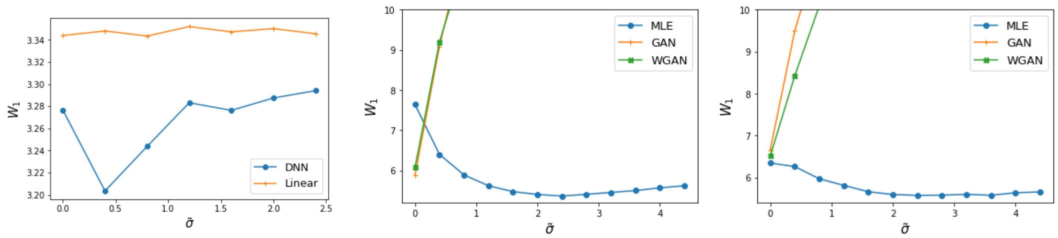

Figure 7: The estimated W_1 distance over e for Big-ve (left), MNIST (middle) and Omniglot (right) data.

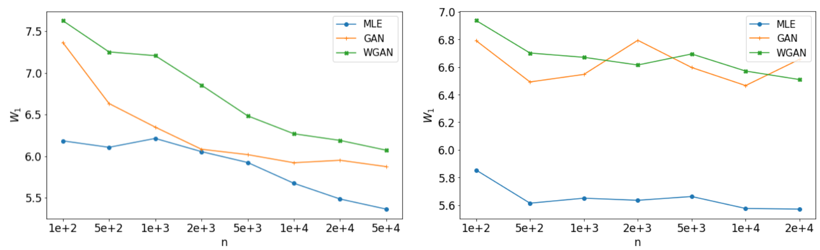

Figure 8: The estimated W_1 distance over the sample size for MNIST (lefg) and Omniglot (right) data. An optimal e is chosen for sieve MLEs based on the validation errir, and no data perturbation is applied for GAN and WGAN.

5.4.2 Results for Big-five Data Set

The Big-ve data set is trained with various values of e; and the estimated W_1 distances over various values of e are depicted in the left panel of Figure 7. Again, it is clear that the estimated W_1 distance is minimized at an intermediate value of e: In addition, we provide the results of the MLE of a sparse linear factor model for comparison, which has been considered in literature for analysing the Big-ve data set, see Ohn and Kim (2021). A deep generative model is signicantly better than a sparse linear factor model, which indicates that nonlinear factor models are necessary for practical data analysis.

5.4.3 Results for MNIST and Omniglot Data Sets

The results about the estimated W_1 distance for various e are shown in the middle and right panels of Figure 7. Again, we observe that the estimated W_1 distance is minimized at an intermediate value of e: On the other hand, the data perturbation does not work at all for GAN and Wasserstein GAN. Moreover, a sieve MLE with proper data perturbation outperforms GAN and Wasserstein GAN for the both image data sets, as detailed in Figure 8.

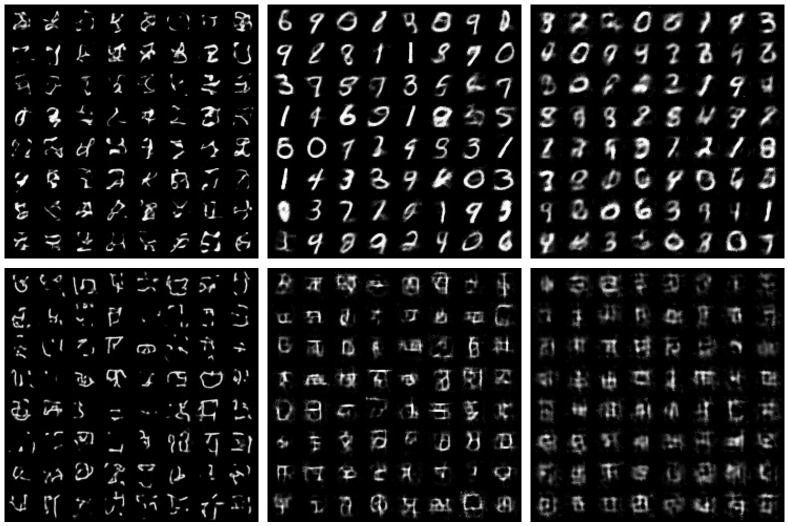

Figure 9: Randomly generated images from a sieve MLE $\hat{Q}$ for MNIST (upper) and Omniglot (lower). We considered three values of e, 0.0, 2.0 and 4.0 from left to right.

Figure 9 presents randomly generated images from sieve MLEs $\hat{Q}$ for MNIST and Omniglot data sets with three values of e, 0.0, 2.0 and 4.0. It is obvious that $e = 2{:}0$ gives the best results for the both data, which implies that the estimated W_1 distance is positively related to the cleanness of corresponding synthetic images. Randomly generated images of GAN and Wasserstein GAN learned with data perturbation for MNIST and Omniglot are given in Figures 10 and 11, respectively, which again conrms that data perturbation is not helpful for GAN and Wasserstien GAN to generate synthetic images.

5.5 Meta-learning for Low-dimensional Composite Structures

In Section 3.2, we have proved that a sieve MLE of deep generative models can capture a low-dimensional composition structure well. Using this exibility of a sieve MLE, we can learn a low-dimensional composite structure from a sieve MLE as follows. For example, suppose that f possesses a generalized additive model (GAM) structure such as

$$f_j(z) = g_{j1}(z_1) + \quad + g_{jd}(z_d)$$

for $j = 1; : : : ; D$: Then, we can estimate the component functions $g_{jl}; l = 1; : : : ; d$ by minimizing

$$\sum_{i=1}^{N} \hat{f}_j(z_i) \quad g_{j1}(z_{i1}) + \quad + g_{jd}(z_{id}) \quad {}_2$$

under certain regularity conditions, where z_i's are independently generated samples from P_Z:

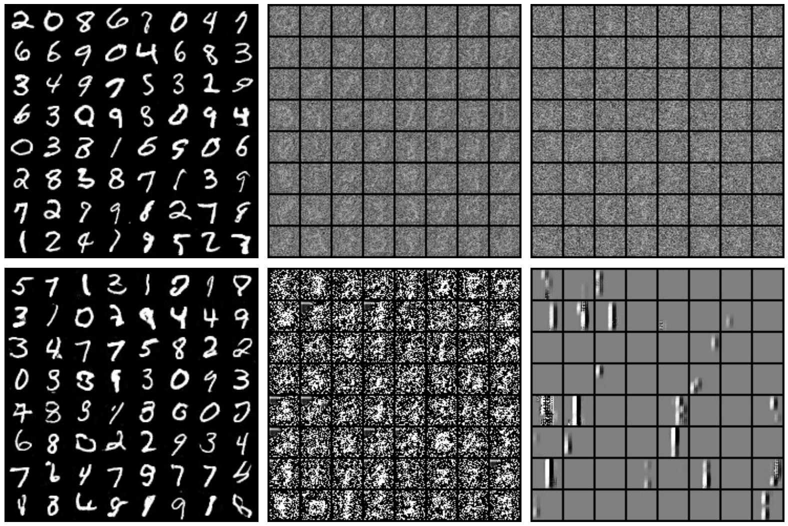

Figure 10: Randomly generated images by GAN (upper) and WGAN (lower) estimators for MNIST. We consider three values of e, 0.0, 2.0 and 4.0 from left to right.

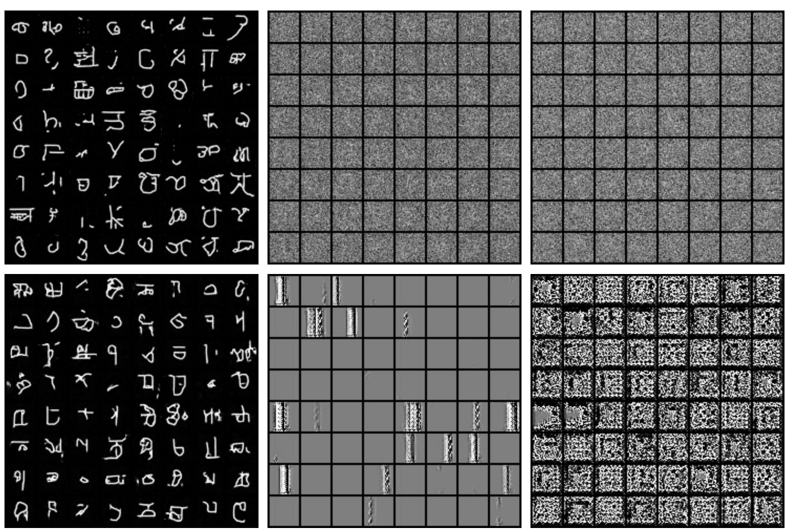

Figure 11: Randomly generated images by GAN (upper) and WGAN (lower) estimators for Omniglot. We consider three values of e, 0.0, 2.0 and 4.0 from left to right.

We investigate the above meta-modeling approach by simulation. We generate data of size 50,000 from the following two generative models:

Model 1: GAM

$$
\begin{aligned}
z &= (z_1; z_2; z_3) \quad N(0; I_3) \\
f_1(z) &= \quad 2{:}3 + \frac{1}{0{:}7 + \exp(0{:}3 \quad 2z_1)} + 0{:}3z_2^2 \quad f_2(z) \\
&= 0{:}9 + 0{:}8z_1 \qquad 0{:}1z_1^3 + \log(z_2^2 + 1{:}5) \quad 0{:}4z_3^2 \\
f_3(z) &= 1{:}8 + \frac{3{:}5}{2z_2^2 + z_2 + 4} \quad 0{:}2\exp(z_3) \\
f_4(z) &= 1{:}2z_1 \quad 0{:}1z_2^3 + 0{:}05z_3^4 \\
f_5(z) &= 3 + 0{:}5\log(2{:}5 + \exp(z_1)) \quad 0{:}2\exp(z_3 + 0{:}2)
\end{aligned}
$$

Model 2: Non-additive model

$$
\begin{aligned}
z &= (z_1; z_2; z_3) \quad N(0; I_3) \\
f_1(z) &= \frac{5z_3}{3{:}7 + \exp(\quad 2z_1 + 0{:}4z_2)} \\
f_2(z) &= 0{:}9 \quad 0{:}1z_1 \quad 0{:}2z_1(z_2 \quad 0{:}1)^2 + 0{:}15z_1z_2z_3 \\
f_3(z) &= \log(2 + (z_1 \quad z_2)^2) \quad 0{:}2z_1\exp(0{:}2 \ z_3) \\
f_4(z) &= 1{:}5 \quad 0{:}3z_1^2 + 0{:}07z_1z_2z_3 \\
f_5(z) &= \frac{3z_1 \quad 1{:}2}{z_2^2 + 2z_2 + 3{:}3} + 0{:}5\log(1 + (z_1 \quad 0{:}1)^2 + z_2^2z_3^2)
\end{aligned}
$$

We estimated the components of the GAM from a sieve MLE of the deep generative model by the proposed meta-modeling and compare the estimated W_1 distances of the original sieve MLE and the estimated GAM in Figure 12. The orginal sieve MLE outperforms the GAM for the two simulation models but the dierence of the estimated W_1 distances is smaller for the rst model where the true model is a GAM than the second model, which indicates that the sieve MLE captures the underlying low-dimensional composite structure well.

For the Big-ve data set, the upper left panel of Figure 13 compares the estimated W_1 distances of three estimates, (sieve) MLEs of the linear and deep generative models and the estimated GAM obtained by the meta-learning. The GAM improves over the linear model but is slightly inferior to the deep generative model. The ve estimated component functions for f_{14}; a randomly selected coordinate, are drawn in Figure 13. Some of them clearly show non-linearity, which partly explains why the performance of the deep generative model is much better than the linear factor model.

6. Discussion

In this work, we consider the estimation of a distribution of high-dimensional data based on a deep generative model which includes the estimation of classical smooth densities and

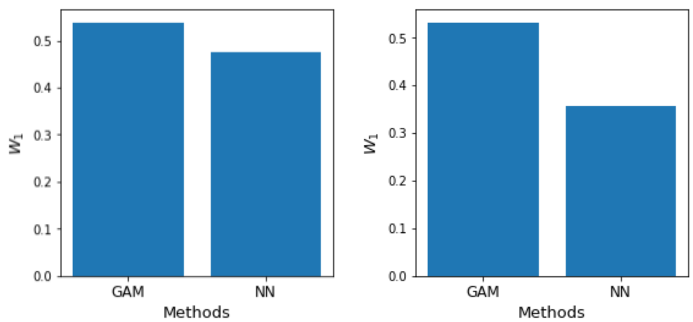

Figure 12: Estimated W_1 distances of a sieve MLE and the estimated GAM for Model 1 (left) and Model 2 (right)

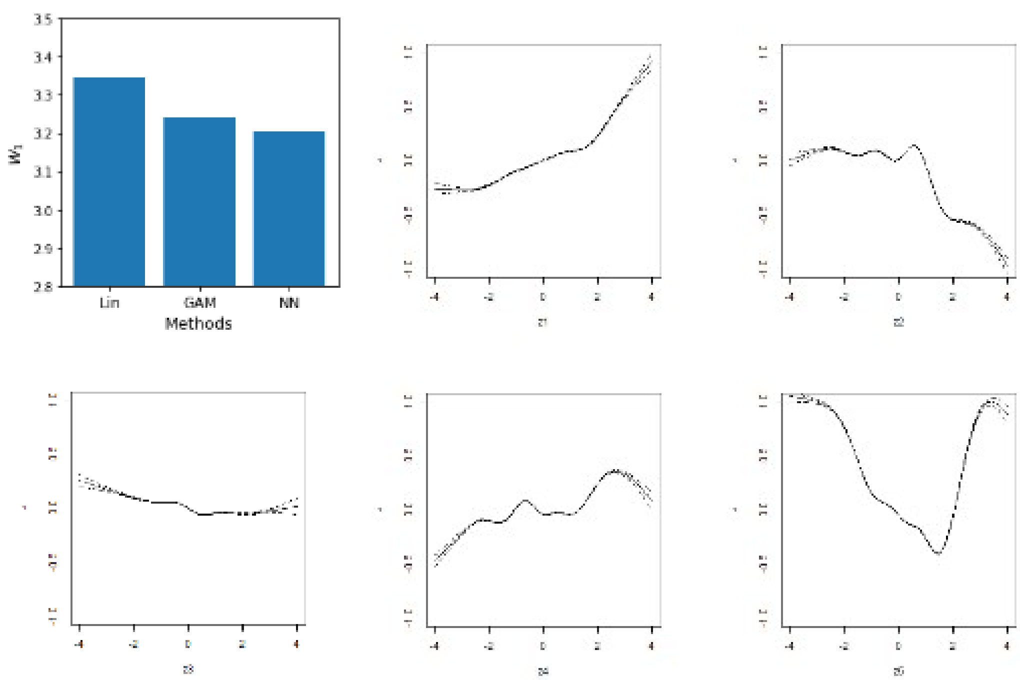

Figure 13: The estimated W_1 distances of (sieve) MLEs of the linear model, deep generative model and the estimated GAM (upper left) and the ve estimated component functions of a randomly selected corrdinate (i.e. $\hat{f}_{14}$) of the GAM for the Big-ve data set

	Manifold	Noise level	Upper bound	Lower bound
G1	C^2	$\|\|_1 . 1$	$n^{2=(2+d)}$	$n^{2=(2+d)}$ (log
G2	C^2	$N(0_D; I_D)$	$(\log n)^{1=2}$	$n)^{1}$
P	C^2	$\|\|_1 . n^{2=(3d+8)}\|\|_1 .$	$n^{2=d}$ _ $(^2=n)^{2=(d+4)}$ n	$(^2=n)^{2=(d+4)}$
A	C	$n^{1=d}\|\|_1 . n^{2=d}$	$^{=d}$ _	$n^{=d}$ _ $(=n)^{=(d+)}$ $n^{2=d}$
D	C^2		$n^{2=d}$	

Table 1: Convergence rates of the manifold estimators with respect to the Hausdor distance from existing papers: Genovese et al. (2012b) (G1), Genovese et al. (2012a) (G2), Puchkin and Spokoiny (2022) (P), Aamari and Levrard (2019) (A), Divol (2021) (D). C in the second column refers that M is a dierentiable manifold of order . For Genovese et al. (2012b) and Aamari and Levrard (2019), it is assumed that is perpendicular to the manifold, see Genovese et al. (2012b) for details.

distributions supported on lower-dimensional manifolds as special cases. The case when Q is supported on a smooth manifold M with dim(M) = d, is the most interesting and challenging case. For this model, one may be interested in estimating the manifold or the support of M itself. One can easily construct an estimator for M by $\hat{M} \triangleq \hat{f}(Z)$ based on an estimator $\hat{f}$. The performance of $\hat{M}$ might be evaluated through a convergence rate with respect to the Hausdor metric. Some existing results on convergence rates are summarized in Table 1 with assumptions on the underlying manifold and noise level. All these papers assume that the reach of the underlying manifold is bounded below by a positive constant. Technical assumptions from dierent papers may vary, but none of these papers explicitly consider the regularity of q, the density with respect to the volume measure. In particular, Genovese et al. (2012b) assumed that the error vector is perpendicular to the manifold which is somewhat a strong condition. In Genovese et al. (2012a), the perpendicular error is replaced by standard Gaussian error leading to a slow convergence rate. This slow rate is standard in a deconvolution problem with a supersmooth Gaussian kernel. The other three papers considered bounded errors which decay to zero with suitable rates. If the noise level is suciently small and M 2 C^2, the minimax convergence rate would be $n^{2=d}$. It would be interesting to investigate whether an estimator $\hat{M}$ constructed from a deep generative model can achieve this rate. More generally, it would be worthwhile to study the manifold estimation problem through the lens of deep generative models.

We have some interesting observations from the results of analysis of the two image data sets in Section 5. While GAN and WGAN generate clearer images than a sieve MLE, the performance of a sieve MLE in terms of the evaluation metric $W_1(\hat{Q}_M; Q_M)$ is better than both, if a suitable degree of perturbation is applied. Surprisingly, opposite results are obtained if FID (Frechet Inception distance; Heusel et al. (2017)) is used as a measure of performance. Note that FID is an approximation of L^2-Wasserstein distance in the feature space of Inception model (Szegedy et al., 2016), and it is one of the most popularly used performance measures in image generation problems. The obtained FID values are 2.76, 4.19 and 9.58 for GAN, WGAN and sieve MLE with the optimal e, respectively. That is, both GAN and WGAN are signicantly better than a sieve MLE in terms of FID. At this point, we are not aware of any reason why two performance measures, $W_1(\hat{Q}_M, Q_M)$ and FID, yield opposite results, which we leave as a future work.

Acknowledgments

The authors are very grateful to the Editor, the Associate Editor and the reviewers for their valuable comments which have led to substantial improvement in the paper. MC was supported by Samsung Science and Technology Foundation under Project Number SSTF-BA2101-03. DK was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1G1A1010894). YK was supported by the NRF grant (No. 2020R1A2C3A01003550) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2022-0-00184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics) funded by MSIT. LL would like to acknowledge the generous support of NSF grants DMS CAREER 1654579 and DMS 2113642.

Appendix A. Proofs

A.1 Proof of Lemma 1

For $f_1, f_2 \in \mathcal{F}$ with $\| |f_1 - f_2|_1 \|_\infty \le \delta_1$, we have that

$$
\begin{aligned}
p_{f_1,\sigma}(x) - p_{f_2,\sigma}(x) &= \int \phi_\sigma(x - f_1(z)) \left\{1 - \frac{\phi_\sigma(x - f_2(z))}{\phi_\sigma(x - f_1(z))}\right\} dP_Z(z) = \\
&\int \phi_\sigma(x - f_1(z)) \left\{1 - \exp\left(\frac{|x - f_1(z)|_2^2 - |x - f_2(z)|_2^2}{2\sigma^2}\right)\right\} dP_Z(z) \le \int \phi_\sigma(x - \\
&f_1(z)) \frac{|x - f_2(z)|_2^2 - |x - f_1(z)|_2^2}{2\sigma^2} dP_Z(z) \\
&= \int \phi_\sigma(x - f_1(z)) \frac{|f_2(z)|_2^2 - |f_1(z)|_2^2 - 2x^T(f_2(z) - f_1(z))}{2\sigma^2} dP_Z(z) \\
&\le \int \phi_\sigma(x - f_1(z)) \frac{KD\delta_1 + \sqrt{D}|x|_2\delta_1}{\sigma^2} dP_Z(z),
\end{aligned}
$$

where the last inequality holds because $\left| |f_1(z)|_2^2 - |f_2(z)|_2^2 \right| \le 2KD\delta_1$ and $|x^T(f_1(z) - f_2(z))| \le \sqrt{D}|x|_2\delta_1$. Since $|x|_2 \le |x - f(z)|_2 + |f(z)|_2 \le 1 + |x - f(z)|_2^2 + \sqrt{D}K$ and $|x|_2^2\phi_\sigma(x)/(2\sigma^2) \le (2\pi\sigma^2)^{-D/2}/e$, the last display is further bounded by

$$
\begin{aligned}
&\delta_1 \int \phi_\sigma(x - f_1(z)) \left(\frac{2KD + \sqrt{D}}{\sigma^2} + \frac{\sqrt{D}|x - f_1(z)|_2^2}{\sigma^2}\right) dP_Z(z) \\
&\le \delta_1 (2\pi\sigma^2)^{-D/2} \left(\frac{2KD + \sqrt{D}}{\sigma^2} + \frac{2\sqrt{D}}{e}\right).
\end{aligned}
\tag{A.1}
$$

Also, for $\sigma_1, \sigma_2 \in [\sigma_{min}, \sigma_{max}]$ with $|\sigma_1 - \sigma_2| \le \eta_2$, it holds that $|\sigma_1^{-2} - \sigma_2^{-2}| \le \sigma_1^{-2}\sigma_2^{-2}(\sigma_1 + \sigma_2)\eta_2$ and $|\log(\sigma_2/\sigma_1)| \le \eta_2/(\sigma_1 \wedge \sigma_2)$. Hence

$$
\begin{aligned}
p_{f,\sigma_1}(x) - p_{f,\sigma_2}(x) &= \int \phi_{\sigma_1}(x - f(z)) \left[1 - \left(\frac{\sigma_1}{\sigma_2}\right)^D \exp\left(\frac{|x - f(z)|_2^2}{2} \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) \right) \right] dP_Z(z) \\
&\le \int \phi_{\sigma_1}(x - f(z)) \left(\frac{|x - f(z)|_2^2}{2} \left| \frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right| + D \left| \log \frac{\sigma_1}{\sigma_2} \right| \right) dP_Z(z) \qquad \text{(A.2)} \\
&\le \eta_2 \int \phi_{\sigma_1}(x - f(z)) \left(\frac{(\sigma_1 + \sigma_2)|x - f(z)|_2^2}{2\sigma_1^2 \sigma_2^2} + \frac{D}{\sigma_1 \wedge \sigma_2} \right) dP_Z(z) \\
&\le \eta_2 (2\pi\sigma_1^2)^{-D/2} \eta_2 \left(\frac{\sigma_1 + \sigma_2}{e\,\sigma_2^2} + \frac{D}{\sigma_1 \wedge \sigma_2} \right) D
\end{aligned}
$$

Let $\epsilon > 0$ be given. Let $\{f_1, \dots, f_{N_1}\}$ and $\{\sigma_1, \dots, \sigma_{N_2}\}$ be η_1-covering of $\mathcal{F}$ and η_2-covering of $[\sigma_{min}, \sigma_{max}]$, respectively. By (A.1) and (A.2), there exist constants $c_1 = c_1(D, K)$ and $c_2 = c(D)$ such that $\eta_1 = c_1 \epsilon \sigma_{min}^{D+2}$ and $\eta_2 = c_2 \epsilon \sigma_{min}^{D+1}$ implies that $\{p_{f_i,\sigma_j} : i = 1, \dots, N_1; j = 1, \dots, N_2\}$ forms an $\epsilon/2$-covering of $\mathcal{P}$ with respect to $\|\cdot\|_\infty$. For each (i, j), define l_{ij} and u_{ij} as

$$
l_{ij}(x) = \max\{p_{f_i,\sigma_j}(x) - \epsilon/2, 0\} \quad \text{and} \quad u_{ij}(x) = \min\{p_{f_i,\sigma_j}(x) + \epsilon/2, H(x)\},
$$

where $H(x) = \sup_{p \in \mathcal{P}} p(x)$ is an envelop function of $\mathcal{P}$. Note that

$$
\begin{aligned}
H(x) &\le (2\pi\sigma_{min}^2)^{-D/2} \sup_{|y|_1 \le K} \exp\left(-\frac{|x - y|_2^2}{2\sigma_{max}^2} \right) \\
&\le (2\pi\sigma_{min}^2)^{-D/2} \exp\left(-\frac{|x|_2^2 - 2K^2 D}{4\sigma_{max}^2} \right) = 2^{D/2} \left(\frac{\sigma_{max}}{\sigma_{min}} \right)^D e^{K^2 D/2} \phi_{\sqrt{2}\sigma_{max}}(x);
\end{aligned}
$$

where the second inequality holds because $|x - y|_2^2 \ge |x|_2^2/2 - |y|_2^2 \ge |x|_2^2/2 - K^2 D$. Since $\int_{|x|_1 > t} \phi_\sigma(x) dx \le D e^{-t^2/(2\sigma^2)}$, we have that $\int_{|x|_1 > B} H(x) dx \le \epsilon$, where

$$
B = 2\sigma_{max} \left\{ \log\frac{1}{\epsilon} + D \log\frac{\sigma_{max}}{\sigma_{min}} + \frac{D \log 2}{2} + \frac{K^2 D}{2} + \log D \right\}^{1/2}.
$$

It follows that

$$
\int \{u_{ij}(x) - l_{ij}(x)\} dx \le \epsilon \int_{|x|_1 \le B} dx + \int_{|x|_1 > B} H(x) dx \le \epsilon\{(2B)^D + 1\} \overset{\text{def}}{=} \epsilon'.
$$

Since $d_H^2(u_{ij}, l_{ij}) \le \|u_{ij} - l_{ij}\|_1$, we have that

$$
N_{[]}(\sqrt{\epsilon'}, \mathcal{P}, d_H) \le N_{[]}(\epsilon', \mathcal{P}, \|\cdot\|_1) \le N_1 N_2 \lesssim \frac{\sigma_{max} - \sigma_{min}}{\eta_2} N(\eta_1, \mathcal{F}, \||\cdot|_1\|_\infty).
$$

Since $(\log \frac{1}{\epsilon})^{D/2} \le 1/\sqrt{\epsilon}$ for every small enough ϵ, once δ is small enough, say $\delta \le \delta_0$ for some $\delta_0 = \delta_0(D)$, it holds that $\epsilon' \ge c_3 \delta^4 \{\log(\sigma_{max}/\sigma_{min})\}^{-D}$, where $c_3 = c_3(D, K, \sigma_{max})$. Hence,

$$
\eta_1 \ge \frac{c_1 c_3 \sigma_{min}^{D+3} \delta^4}{\sigma_{min} \{\log(\sigma_{max}/\sigma_{min})\}^D}.
$$

Since $_{\min}$ 1, $_{\min}\mathrm{flog}(_{\max}={}_{\min})g^{D}$ is bounded by a constant which depends only on $_{\max}$ and D, so $_1$ is bounded below by c_4 $^{D+3\;4}_{\min}$, where $c_4 = c_4(D;K;{}_{\max})$. A similar lower bound can be obtained for $_2$, which completes the proof.

A.2 Proof of Theorem 3

We will apply Theorem 4 of Wong and Shen (1995) with $= 0+$. Choose four absolute constants $c_1;:::;c_4$ as in their Theorem 1. These constants can be chosen so that $c_1 = 1=3$ and $c_3 > 2$. Dene c and c^0 as in the statement of Lemma 1.

For every $2\ (0;c_3]$,

$$\log N_{[]}(=c_3;P;d_H)\quad 4(s+1)\log{}^{1} + sA + (D+3)(s+1)\log{}_{\min} + c_5^1 s$$

by Lemma 1, where $c_5 = c_5(c;c^0;c_3)$. Hence,

$$Z\ ^{p}2\ \overline{q\ \log N_{[]}(=c_3;P;d_H)}\ d$$
$$\tfrac{}{}2\ ^8$$
$$p\ \overline{2\ q\ sA + (D+3)(s+1)\log{}_{\min}^{1} + c_5 s +}\quad p\ 2\ \overline{p\ 4(s+1)}\ \overline{r\ \log\frac{2^8}{2}}$$

for every $^p2\ \ c_3 = {}^p2$. For $= {}_n = c_6\ ^p n\ ^1 sfA^- + \log(\overline{n={}_{\min})g}$ with a large enough constant $c_6 = c_6(c_4;c_5;D)$, the last display is bounded by $c_4 n^{1=2}\,{}^2_n$ for every n, so Eq. (3.1) of Wong and Shen (1995) is satised. Note that Eq. (3.1) of Wong and Shen (1995) still holds if c_6 is replaced by any constant larger than c_6.

It is well-known (see Example B.12 of Ghosal and van der Vaart (2017)) that

$$K(p;p_{f;})\quad \int K\ N\ \ f(z);{}^2;N\ \ f(z);{}^2 dP_Z(z) = \int \frac{jf(z)\ \ f(z)j_2^2}{2\ ^2} dP_Z(z)\quad \frac{D\ \ ^2_{\mathrm{app}}}{2\ \ ^2} \stackrel{\mathrm{def}}{=} {}_n.$$

Also, it is easy to see that

$$\int {}^2(x)dx = \int \log\frac{(x)}{(x\ \ y)}\ \ \frac{jyj_2^4 + 4jx^T yj^2}{4\ ^2}(x)dx\quad \frac{jyj_2^4}{4^2} + \frac{jyj_2^2}{}\int \frac{jxj_2^2}{^2}(x)dx:$$

Combining this with Example B.12, (B.17) and Exercise B.8 of Ghosal and van der Vaart (2017), we have that

$$\int \log\frac{p(\ \ x)}{p_{f;}(x)}{}^2 dP(x)$$
$$\iint \log\frac{(x\ \ f(z))}{(x\ \ f(z))}{}^2 (x\ \ f(z))dx\,dP_Z(z) + 4K(p;p_{f;})$$
$$D\ \frac{^2_{\mathrm{app}}\ ^4}{4\ \ ^2} + D_{\mathrm{app}}\ ^2 \int \frac{j^2 xj_2^2}{^2}(x)dx + \frac{2D_{\mathrm{app}}}{}\quad c_7\frac{^2_{\mathrm{app}}}{^2} = {}_n;\quad \frac{^2}{^2}\ \ \mathrm{def}$$

where $c_7 = c_7(D)$. (Note that both $_n$ and $_n$ need not depend on n. We use the notations $_n$ and $_n$ for the notational consistency with Theorem 4 of Wong and Shen (1995)). Let $_n = _n$ _ p $\overline{12_n}$. Then, Theorem 4 of Wong and Shen (1995) implies that

$$P d_H(\hat{p}; p) > 5e^{c_2 n^2_n} + n^2 5e^{c_2 n^2_n} + \frac{12_n}{n_n} = 5e^{c_2 n^2_n} + \frac{D}{n}$$ By re-defining $_n$ $\frac{2c_7^2}{}$

constants, the proof is complete.

A.3 Proof of Corollary 6

By Lemma 5 of Schmidt-Hieber (2020), we have

$$\log N(; F; kj\ j_1 k_1) \quad s f c_4 (\log n)^2 + \log\ ^1 g$$

for every > 0, where $c_4 = c_4(q; d; t; ; K)$. By applying Lemma 5 and Theorem 3 with A $= c_4(\log n)^2$, we have the conclusion.

A.4 Proof of Theorem 7

For any constant $c_0 = c_0(D; K; r)$, if + p $\overline{\log\ ^1}$ c_0, the assertion of Theorem 7 holds trivially by taking a large enough constant $C = C(D; K; r)$. Therefore, it suces to prove the assertion of Theorem 7 when and p $\overline{\log\ ^1}$ are suciently small.

For given 2 (0; 1], suppose that $d_H(p_{f;}; p)$ and $kjf j_1 k_1$ K. Throughout this proof, $P_{f;}$ and Q_f will be denoted as P and Q, respectively. Let Y; Y; ; be independent random vectors, with the underlying probability such that Y Q, Y Q, $N(0_D; ^2 I_D)$, $N(0_D; ^2 I_D)$.

Since

$$\int_{jxj_1 > t = D} (x) dx \quad \int_p (x) dx \quad De^{t^2 = (2D^2)}{}_{jxj_2 > t}$$

for any $t > 0$, we have $\int_{jxj_2 > t} (x) dx$ with $t = (2D^2 \log(D=))^{1=2}$. Hence, 1 P M_t

$$= Y + 2 M_t \quad (jj_2 > t) :$$

Since $jP(B) \quad P(B)j \quad d_H(P; P)$ for every Borel set B, see Eq. (8) of Gibbs and Su (2002), we have that $P(M^t)$ 1 2.

We will next prove that 2t, which is the main part of the proof. For this, we assume on the contrary that $> 2t$ which we will show lead to a contraction. Firstly, if $> r=2$, then 1 $P([\ K \quad t; K + t]^D)$ is bounded below by a constant that depends on K; D and r, which contradicts to $P(M_t)$ 1 2 provided that t and are smaller than a certain threshold depending only on K; D and r. (Note that t and are suciently small as assumed at the beginning of the proof.) If 2 [2t; r=2], then we claim that for every x 2 R^D, there exists y 2 R^D such that $jx \quad yj_2$ and $B_{=2}(y) \setminus M^t = ;$. Let $(x; M) = \inf fjx \quad x j_2 : x\ 2\ M g$. The proof of the claim is divided into three cases.

(Case 1) $(x; M)\ _0^c$: Obviously, one can choose $y = x$.

(Case 2) $(x; M)$ 2 (0;): Let x_0 be the unique Euclidean projection of x onto M, and $x_t = x_0 + t(x \quad x_0)$. Dene two continuous functions $d_0(t) = jx_t \quad x_0 j_2$ and $d(t) = (x_t; M)$.

Note that $d_0(t) = d(t)$ for all t 2 [0; 1]. Otherwise, $jx_t \quad zj_2 < jx_t \quad x_0j_2$ for some t 2 [0; 1] and z 2 Mnx_0. Since x_t lies in the line segment with end points x and x_0,

$$jx \quad x_0j_2 = jx \quad x_tj_2 + jx_t \quad x_0j_2 > jx \quad x_tj_2 + jx_t \quad zj_2 \; jx \quad zj_2;$$

and thus, x_0 cannot be the unique projection of x onto M. Note also that $d(t) = d_0(t)$ for all t 2 [1; 1 + =jx x_0j_2]. Otherwise, ft 2 [1; 1 + =jx x_0j_2] : $d(t) < d_0(t)$g is a non-empty set with the inmum t_0, and it is not dicult to see that x_{t^0} has at least two Euclidean projection onto M. Let y = $x_{1+=jx \ x \ j}$. Then, we have jy xj_2 = and (y; M) = jy x_0j_2 = jx x_0j_2 + . Since t =2, we have $B_{=2}(y) \setminus M$ = ; t

(Case 3) (x; M) = 0: Since B(x) is not contained in M for any > 0, one can choose x^0 2 B(x)nM. If is small enough, by Case 2, there exists y^0 such that $jx^0 \quad y^0j_2$ and $B_{=2}(y^0) \setminus M^t$ = ;. Note that jx y^0j_2 jx x^0j_2 + $jx^0 \quad y^0j_2$ + . One can take y as any limit point of y as ! 0.

By the claim, we have

$$= x \quad 2 \; B_{=2}(y \quad x) \qquad Y + 2 M \quad j Y$$

for every x 2 R^D. Since jy xj_2 , the right hand side is bounded below by a positive constant, say c, that depends only on D. It follows that P (M t) = (Y + 2 M) 1 t c, which contradicts P (M) 1 t2 for small enough . This completes the proof of 2t. Note that the '$_1$-diameter of [K; K]^D is 2KD, $W_1 \quad W_2$ and W_1 is bounded by a multiple of the total variation, see Theorem 4 of Gibbs and Su (2002). Also, it is easy to see that $W_2(P; Q)$ and $W_2(P; Q)$. Hence,

$$W_1(Q; Q) \; W_2(Q; P) + W_1(P; P) + W_2(P; Q) \quad + KDkp \qquad pk_1 + :$$

Since kp pk_1 $2d_H(p; p)$ and 2t, the proof is complete.

A.5 Proof of Theorem 9

Let p = $p_{f;e}$, where e = + n $^{=f2(+t)g}$. Also, let e = log e= log n, that is, e = n e. Then, by Corollary 6, (3.4) holds with

$$_n = Cn^{\frac{et}{2+t}} (\log n)^{3=2};$$

where C = C(q; d; t; ; K; D; $_{max}$;).

Firstly, suppose that < =f2(+ t)g. In this case, < e 2, so

$$\frac{\log 2}{\log n} e < :$$

Hence, $_n$ can be re-written as

$$_n = C^0 n^{\frac{t}{2+t}} (\log n)^{3=2}$$

with an adjusted constant $C^0 = C^0(q; d; t; ; K; D; _{max}; ;)$ satisfying 2 $^{t=(2+t)}$C $C^0 < C$.

Similarly, if $=f2(+ t)g$, we have

$$\frac{}{t)} \qquad \frac{\log 2}{\log n}\ e \qquad \frac{2(+}{2(+ t)}:$$

Hence, $_n$ can be re-written as

$$_n = C^{00} n^{\frac{}{2(+t)}} (\log n)^{3=2}$$

with $C_{00} = C^{00}(q; d; t; ; K; D; _{max}; ;)$.

Finally, Theorem 7 gives the desired result with re-dened constants.

References

Eddie Aamari and Clement Levrard. Stability and minimax optimality of tangential Delaunay complexes for manifold reconstruction. Discrete Comput. Geom., 59(4):923{971, 2018.

Eddie Aamari and Clement Levrard. Nonasymptotic rates for manifold, tangent space and curvature estimation. Ann. Statist., 47(1):177{204, 2019.

Martin Arjovsky, Soumith Chintala, and Leon Bottou. Wasserstein generative adversarial networks. In Proc. International Conference on Machine Learning, pages 214{223, 2017.

Sanjeev Arora, Rong Ge, Yingyu Liang, Tengyu Ma, and Yi Zhang. Generalization and equilibrium in generative adversarial nets (GANs). In Proc. International Conference on Machine Learning, pages 224{232, 2017.

Yu Bai, Tengyu Ma, and Andrej Risteski. Approximability of discriminators implies diversity in GANs. In Proc. International Conference on Learning Representations, pages 1{10, 2019.

Benedikt Bauer and Michael Kohler. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. Ann. Statist., 47(4):2261{2285, 2019.

Clement Berenfeld and Marc Homann. Density estimation on an unknown submanifold. ArXiv:1910.08477, 2019.

Clement Berenfeld, Paul Rosa, and Judith Rousseau. Estimating a density near an unknown manifold: a Bayesian nonparametric approach. ArXiv:2205.15717, 2022.

Gerard Biau, Benoˆt Cadre, Maxime Sangnier, and Ugo Tanielian. Some theoretical properties of GANs. Ann. Statist., 48(3):1539{1566, 2020.

Gerard Biau, Maxime Sangnier, and Ugo Tanielian. Some theoretical insights into Wasserstein GANs. J. Mach. Learn. Res., 22(119):1{45, 2021.

Carlo Bruni and Giorgio Koch. Identiability of continuous mixtures of unknown Gaussian distributions. Ann. Probab., 13(4):1341{1357, 1985.

Peter Bühlmann and Sara van de Geer. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer, 2011.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In Proc. International Conference on Learning Representations, pages 1{14, 2016.

Luis A Caarelli. Interior $W^{2;p}$ estimates for solutions of the Monge{Ampere equation. Ann. of Math., 131(1):135{150, 1990.

Minwoo Chae. Rates of convergence for nonparametric estimation of singular distributions using generative adversarial networks. ArXiv:2202.02890, 2022.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Ecient approximation of deep ReLU networks for functions on low dimensional manifolds. In Proc. Neural Information Processing Systems, pages 8174{8184, 2019a.

Minshuo Chen, Haoming Jiang, Wenjing Liao, and Tuo Zhao. Nonparametric regression on low-dimensional manifolds using deep ReLU networks. ArXiv:1908.01842, 2019b.

Minshuo Chen, Wenjing Liao, Hongyuan Zha, and Tuo Zhao. Statistical guarantees of generative adversarial networks for distribution estimation. ArXiv:2002.03938, 2020.

Chris Cremer, Quaid Morris, and David Duvenaud. Reinterpreting importance-weighted autoencoders. In Proc. International Conference on Learning Representations, 2017.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In Proc. Neural Information Processing Systems, pages 2292{2300, 2013.

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

Adji B Dieng and John Paisley. Reweighted expectation maximization. ArXiv:1906.05850, 2019.

Vincent Divol. Minimax adaptive estimation in manifold inference. Electron. J. Stat., 15 (2):5888{5932, 2021.

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

Herbert Federer. Curvature measures. Trans. Amer. Math. Soc., 93(3):418{491, 1959.

Jean Feydy, Thibault Sejourne, Francois-Xavier Vialard, Shun-ichi Amari, Alain Trouve, and Gabriel Peyre. Interpolating between optimal transport and mmd using sinkhorn divergences. In Proc. International Conference on Articial Intelligence and Statistics, pages 2681{2690. PMLR, 2019.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. Ann. Statist., 10(2):401{414, 1982.

Christopher R Genovese, Marco Perone-Pacico, Isabella Verdinelli, and Larry Wasserman. Manifold estimation and singular deconvolution under Hausdor loss. Ann. Statist., 40 (2):941{963, 2012a.

Christopher R Genovese, Marco Perone-Pacico, Isabella Verdinelli, and Larry Wasserman. Minimax manifold estimation. J. Mach. Learn. Res., 13(1):1263{1291, 2012b.

Subhashis Ghosal and Aad van der Vaart. Fundamentals of Nonparametric Bayesian Inference. Cambridge University Press, 2017.

Subhashis Ghosal and Aad W van der Vaart. Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann. Statist., 35(2):697{723, 2007.

Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. Int. Stat. Rev., 70(3):419{435, 2002.

Evarist Gine and Richard Nickl. Mathematical Foundations of Innite-Dimensional Statistical Models. Cambridge University Press, 2016.

Lewis R Goldberg. An alternative \description of personality": the big-ve factor structure. Journal of Personality and Social Psychology, 59(6):1216, 1990.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Proc. Neural Information Processing Systems, pages 2672{2680, 2014.

Song Han, Je Pool, John Tran, and William Dally. Learning both weights and connections for ecient neural network. In Proc. Neural Information Processing Systems, pages 1135{ 1143, 2015.

Radoslav Harman and Vladimr Lacko. On decompositional algorithms for uniform sampling from n-spheres and n-balls. J. Multivariate Anal., 101(10):2297{2304, 2010.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In Proc. Neural Information Processing Systems, pages 6629{6640, 2017.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. Neural Networks, 2(5):359{366, 1989.

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. Neural Networks, 3(5):551{560, 1990.

Joel L Horowitz and Enno Mammen. Rate-optimal estimation for a general class of nonparametric regression models with unknown link functions. Ann. Statist., 35(6):2589{2619, 2007.

Jian Huang, Yuling Jiao, Zhen Li, Shiao Liu, Yang Wang, and Yunfei Yang. An error analysis of generative adversarial networks for learning distributions. ArXiv:2105.13010, 2021.

Masaaki Imaizumi and Kenji Fukumizu. Deep neural networks learn non-smooth functions eectively. In Proc. International Conference on Articial Intelligence and Statistics, pages 869{878, 2019.

Yuling Jiao, Guohao Shen, Yuanyuan Lin, and Jian Huang. Deep nonparametric regression on approximately low-dimensional manifolds. ArXiv:2104.06708, 2021.

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. Mach. Learn., 37(2):183{233, 1999.

Anatoli B Juditsky, Oleg V Lepski, and Alexandre B Tsybakov. Nonparametric estimation of composite functions. Ann. Statist., 37(3):1360{1404, 2009.

Dongha Kim, Jaesung Hwang, and Yongdai Kim. On casting importance weighted autoencoder to an EM algorithm to learn deep generative models. In Proc. International Conference on Articial Intelligence and Statistics, pages 2153{2163. PMLR, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Proc. International Conference on Learning Representations, 2015.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In Proc. International Conference on Learning Representations, pages 1{14, 2014.

Durk P Kingma, Tim Salimans, Rafal Jozefowicz, Xi Chen, Ilya Sutskever, and Max Welling. Improved variational inference with inverse autoregressive ow. In Proc. Neural Information Processing Systems, pages 4743{4751, 2016.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected very deep neural network regression estimates. To appear in Ann. Statist., 2021.

Brenden M Lake, Ruslan Salakhutdinov, and Joshua B Tenenbaum. Human-level concept learning through probabilistic program induction. Science, 350(6266):1332{1338, 2015.

Yann LeCun, Leon Bottou, Yoshua Bengio, and Patrick Haner. Gradient-based learning applied to document recognition. Proc. IEEE, 86(11):2278{2324, 1998.

John M Lee. Introduction to Smooth Manifolds. Springer, New York, 2nd edition, 2013.

Chun-Liang Li, Wei-Cheng Chang, Yu Cheng, Yiming Yang, and Barnabas Pozos. MMD GAN: Towards deeper understanding of moment matching network. In Proc. Neural Information Processing Systems, pages 2203{2213, 2017.

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

Shuang Liu, Olivier Bousquet, and Kamalika Chaudhuri. Approximation and convergence properties of generative adversarial learning. In Proc. Neural Information Processing Systems, pages 5545{5553, 2017.

Yulong Lu and Jianfeng Lu. A universal approximation theorem of deep neural networks for expressing probability distributions. Proc. Neural Information Processing Systems, pages 1{12, 2020.

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

Stephen Marsland. Machine Learning: An Algorithmic Perspective. CRC press, 2015.

Alexander Meister. Deconvolution Problems in Nonparametric Statistics. Springer, New York, 2009.

Chenlin Meng, Jiaming Song, Yang Song, Shengjia Zhao, and Stefano Ermon. Improved autoregressive modeling with distribution smoothing. ArXiv:2103.15089, 2021.

Youssef Mroueh, Chun-Liang Li, Tom Sercu, Anant Raj, and Yu Cheng. Sobolev GAN. ArXiv:1711.04894, 2017.

Alfred M■ller. Integral probability metrics and their generating classes of functions. Adv. in Appl. Probab., 29(2):429{443, 1997.

Ryumei Nakada and Masaaki Imaizumi. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. J. Mach. Learn. Res., 21(174):1{38, 2020.

XuanLong Nguyen. Convergence of latent mixing measures in nite and innite mixture models. Ann. Statist., 41(1):370{400, 2013.

Ilsang Ohn and Yongdai Kim. Smooth function approximation by deep neural networks with general activation functions. Entropy, 21(7):627, 2019.

Ilsang Ohn and Yongdai Kim. Posterior consistency of factor dimensionality in high-dimensional sparse factor models. To appear in Bayesian Anal., 2021.

Ilsang Ohn and Yongdai Kim. Nonconvex sparse regularization for deep neural networks and its optimality. Neural Comput., 34(2):476{517, 2022.

Ilsang Ohn and Lizhen Lin. Adaptive variational Bayes: Optimality, computation and applications. ArXiv:2109.03204, 2021.

Arkadas Ozakin and Alexander Gray. Submanifold density estimation. Proc. Neural Information Processing Systems, 22:1375{1382, 2009.

Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. In Proc. International Conference on Learning Representations, pages 1{29, 2021.

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

Nicholas G Polson and Veronika Rockova. Posterior concentration for sparse deep learning. In Proc. Neural Information Processing Systems, volume 31, 2018.

Nikita Puchkin and Vladimir G Spokoiny. Structure-adaptive manifold estimation. J. Mach. Learn. Res., 23:1{62, 2022.

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

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In Proc. International Conference on Learning Representations, pages 1{16, 2016b.

Danilo Rezende and Shakir Mohamed. Variational inference with normalizing ows. In Proc. International Conference on Machine Learning, pages 1530{1538. PMLR, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In Proc. International Conference on Machine Learning, pages 1278{1286, 2014.

Tim Salimans, Diederik Kingma, and Max Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In Proc. International Conference on Machine Learning, pages 1218{1226, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In Proc. Neural Information Pro-cessing Systems, volume 29, 2016.

Johannes Schmidt-Hieber. Deep ReLU network approximation of functions on a manifold. ArXiv:1908.00695, 2019.

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

Nicolas Schreuder. Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. Math. Methods Statist., 29:76{86, 2021.

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

Weining Shen, Surya T Tokdar, and Subhashis Ghosal. Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. Biometrika, 100(3):623{640, 2013.

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabas Pozos. Nonparametric density estimation with adversarial losses. In Proc. Neural Information Processing Systems, pages 10246{10257, 2018.

Casper Kaae Snderby, Tapani Raiko, Lars Maale, Sren Kaae Snderby, and Ole Winther. Ladder variational autoencoders. Proc. Neural Information Processing Systems, 29:3738{3746, 2016.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. ArXiv:1907.05600, 2019.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In Proc. Conference on Computer Vision and Pattern Recognition, pages 2818{2826, 2016.

Rong Tang and Yun Yang. Minimax rate of distribution estimation on unknown submanifold under adversarial losses. ArXiv:2202.09030, 2022.

Matus Telgarsky. Benets of depth in neural networks. In Proc. Conference on Learning Theory, pages 1517{1539, 2016.

Ananya Uppal, Shashank Singh, and Barnabas Poczos. Nonparametric density estimation and convergence of GANs under Besov IPM losses. In Proc. Neural Information Processing Systems, pages 9089{9100, 2019.

John IE Urbas. Regularity of generalized solutions of Monge{Ampere equations. Math. Z., 197(3):365{393, 1988.

Aad W. van der Vaart and Jon A. Wellner. Weak Convergence and Empirical Processes. Springer, 1996.

Cedric Villani. Topics in Optimal Transportation. American Mathematical Society, 2003.

Cedric Villani. Optimal Transport: Old and New. Springer, 2008.

Martin J Wainwright. High-Dimensional Statistics: A Non-Asymptotic Viewpoint. Cambridge University Press, 2019.

Jonathan Weed and Francis Bach. Sharp asymptotic and nite-sample rates of convergence of empirical measures in Wasserstein distance. Bernoulli, 25(4A):2620{2648, 2019.

Yun Wei and XuanLong Nguyen. Convergence of de Finetti's mixing measure in latent structure models for observed exchangeable sequences. To appear in Ann. Statist., 2022.

Wing Hung Wong and Xiaotong Shen. Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Ann. Statist., 23(2):339{362, 1995.

Bing Xu, Naiyan Wang, Tianqi Chen, and Mu Li. Empirical evaluation of rectied activations in convolutional network. ArXiv:1505.00853, 2015.

Ilker Yalcin and Yasuo Amemiya. Nonlinear factor analysis as a statistical method. Statist. Sci., 16(3):275{294, 2001.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. Neural Networks, 94:103{114, 2017.

Pengchuan Zhang, Qiang Liu, Dengyong Zhou, Tao Xu, and Xiaodong He. On the discrimination-generalization tradeo in GANs. In Proc. International Conference on Learning Representations, pages 1{26, 2018.