Statistical guarantees for generative models without domination

Nicolas Schreuder NICOLAS.SCHREUDER@ENSAE.FR
Victor-Emmanuel Brunel VICTOR.EMMANUEL.BRUNEL@ENSAE.FR
Arnak S. Dalalyan ARNAK.DALALYAN@ENSAE.FR
Department of Statistics
CREST (UMR CNRS 9194), ENSAE
5, av. Henry Le Chatelier, 91120 Palaiseau, FRANCE

Editors: Vitaly Feldman, Katrina Ligett and Sivan Sabato

Abstract

In this paper, we introduce a convenient framework for studying (adversarial) generative models from a statistical perspective. It consists in modeling the generative device as a smooth transformation of the unit hypercube of a dimension that is much smaller than that of the ambient space and measuring the quality of the generative model by means of an integral probability metric. In the particular case of integral probability metric defined through a smoothness class, we establish a risk bound quantifying the role of various parameters. In particular, it clearly shows the impact of dimension reduction on the error of the generative model.

Keywords: Generative model, risk bound, smoothness class

1. Introduction

The problem of learning generative models has attracted a lot of attention during the last 5 years in machine learning and artificial intelligence. The most prominent example is generating artificial images that look similar to actual photographs, by means of generative adversarial networks. The more general formulation of the problem can be given as a game between the user and the learner. The user samples a set of elements (images of natural scenes, poems, pieces of music, etc.) from a hidden distribution $P^* = P_{\text{user}}$ defined on a hidden (and not so well known) space. The learner receives a noisy and possibly contaminated version of these elements and aims at generating a new set of elements, that are different from those transmitted by the user, but that could have been sampled from the hidden distribution P^*. Note that the revealed elements are usually of very high dimension. However, they may exhibit rich structures such as the harmonic and rhythmic schemes followed by a melody or a poem, or the presence of simple shapes in an image. It is therefore reasonable to assume that these elements can be represented by means of a much lower dimensional latent variable, which is unobserved.

In other words, generative models are used for accomplishing the following task. The user draws n independent samples $\boldsymbol{Y}_1, \dots, \boldsymbol{Y}_n$ from a distribution P_{user} defined on $\mathbb{R}^D$. The learner is given a noisy and contaminated version $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ of this sample. The goal of the learner is to design an algorithm that generates random samples from a distribution P_{learner} which is as close as possible to P_{user}. This can be viewed as a distribution estimation problem with two requirements:

[R1] *It should be easy to sample from* P_{learner}.

[R2] *The way we measure the closeness between P_{learner} and P_{user} for evaluating the error has to admit an interpretation as a sampling error.*

Of course, this formulation is incomplete since it allows to take the uniform distribution over the observed samples as P_{learner}, *i.e.*, $P_{\text{learner}} = \widehat{P}_n = \frac{1}{n}\sum_{i=1}^n \delta_{\boldsymbol{X}_i}$ (the empirical distribution based on the sample $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$). From a generative modeling perspective, $\widehat{P}_n$ is pointless since it does not yield new samples that are different from the previous ones. Hence, generative modeling requires a third distinctive feature:

[R3] *Samples drawn from P_{learner} should be different from those revealed by the user.*

Requirement **R3** is perhaps the hardest to translate into a statistical language. Most prior work focused on the case where both P_{user} and P_{learner}, defined on $\mathbb{R}^D$ equipped with the Borel σ-field, are absolutely continuous with respect to the Lebesgue measure (or another σ-finite measure). This readily implies that the total variation distance between P_{learner} and $\widehat{P}_n$ is equal to 1, which can be considered as a guarantee for P_{learner} to satisfy **R3**.

Positing that P_{user} has a density with respect to the Lebesgue measure, or any other dominating σ-finite measure μ on $\mathbb{R}^D$, is, in general, incompatible with the fact that P_{user} is inherited from a low-dimensional latent variable and supported by a low-dimensional manifold. For instance, in the simple example of $P_{\text{user}} = \mathcal{U}(a\mathbb{S}^{D-1})$, the uniform measure on $a\mathbb{S}^{D-1}$ (the sphere of radius a centered at the origin), there exists no σ-finite measure dominating all the measures $\mathcal{U}(a\mathbb{S}^{D-1})$, for $a > 0$. Very importantly, as a consequence of the restriction to dominated distributions, the available statistical results fail to assess the positive impact of the reduced dimension of the latent space (as compared to the ambient dimension D) on the quality of the generative model.

We propose to circumvent this drawback by restricting the set of candidate generators to those defined as a smooth transformation of the uniform distribution on a low-dimensional hyper-cube. Obviously, the support of these candidate distributions is a path-connected set. Therefore, the empirical distribution $\widehat{P}_n$, as well as any finitely or countably supported distribution is not among these candidates.

The following notation will be used throughout this work. For every positive integer p, we denote by $\mathcal{U}_p$ the uniform distribution on the hyper-cube $[0,1]^p$. For any convex set $\mathcal{X} \subset \mathbb{R}^p$, $\text{Lip}_L(\mathcal{X})$ stands for the set of all Lipschitz-continuous functions defined on $\mathcal{X}$ with a Lipschitz constant less than or equal to L. For a distribution P defined on a measurable space $(E, \mathscr{E})$ and a measurable map $g : E \mapsto F$, where F is another space endowed with a σ-algebra $\mathscr{F}$, we denote by $g\sharp P$ the "push-forward" measure defined by $(g\sharp P)(A) = P\big(g^{-1}(A)\big)$ for all $A \in \mathscr{F}$. For a function $g : \mathcal{X} \to \mathbb{R}$, $\|g\|_\infty = \max_{x\in\mathcal{X}} |g(x)|$ is the supremum norm of g.

The rest of the paper is organised as follows. A brief review of the prior work on generative models is presented in Section 2, while Section 3 provides the formal statement of the problem. In order to convey the main ideas in a simple setting, we analyse the case of noise-free and uncontaminated observations in Section 4. The main results are stated and discussed in Section 5. A summary of the contributions and some avenues for future research are included in Section 6, while Section 7 gathers the proofs of the results stated in previous sections.

2. Related work (and contributions)

The procedures for generative modeling can be split into two groups: prescribed and implicit probabilistic models (Mohamed and Lakshminarayanan, 2016). The former requires an explicit (parametric) specification of the distribution of the observed random variables (*e.g.*, mixture of Gaussian) through a likelihood function, whereas the latter defines a stochastic procedure that directly generates data. The growing complexity of the data makes it harder to design a relevant likelihood function and thus favoured the advent of the latter models. For instance, Generative Adversarial Networks (GANs), perhaps the most well-known generative models based on implicit modeling, enabled groundbreaking advances in the generation of realistic images (Goodfellow et al., 2014; Radford et al., 2015; Goodfellow, 2016; Isola et al., 2017; Zhu et al., 2017; Brock et al., 2018; Karras et al., 2019). In the original GAN framework (Goodfellow et al., 2014) a generator G competes against a discriminator D, both implemented as deep neural networks, in the following zero-sum game: the generator G (resp. the discriminator D) maximizes (resp. minimizes) the objective

$$\Phi(G, D) = \frac{1}{n}\sum_{i=1}^{n} \log D(\boldsymbol{X}_i) + \mathbf{E}_{\tilde{\boldsymbol{X}} \sim G \sharp P_{\boldsymbol{U}}} \log\big(1 - D(\tilde{\boldsymbol{X}})\big), \tag{1}$$

where $P_{\boldsymbol{U}}$ is an easy-to-sample-from noise distribution (*e.g.*, Gaussian or uniform). The goal of the generator is to transform the (low-dimensional) latent variable into artificial data as indistinguishable as possible from the examples drawn from the target distribution. As for the discriminator, the aim is to discriminate between true examples and generated data. See Figure 1 for an illustration of the original GAN model. Informally, the generative model can be thought of as a counterfeiter, trying to produce fake paintings and selling it without detection, while the discriminative model is analogous to art experts, trying to detect the counterfeit paintings. Let us note that here P_{learner} would be the distribution of the generated data, *i.e.*, $G \sharp P_{\boldsymbol{U}}$.

Despite their impressive empirical performance, GANs are notoriously hard to train; Even if some fixes have been proposed (Salimans et al., 2016), several problems are yet to be fully understood and solved (*e.g.*, mode collapse, vanishing gradients, failure to converge). Goodfellow et al. (2014) showed that, when the discriminator is optimal, minimizing (1) with respect to the generator G amounts to minimizing the Jensen-Shannon (JS) divergence between the generated data distribution and the real sample distribution. Arguing that the topology induced by the JS divergence is rather coarse, Arjovsky et al. (2017) proposed to replace this divergence by the Wasserstein-1 distance to stabilize training, leading to the so-called *Wasserstein GAN*. More precisely, the goal of the generator G in this variant is to generate data from a distribution that is as close as possible, w.r.t. the Wasserstein-1 distance, to the empirical distribution of the original data. This leads to the objective

$$\mathsf{W}_1\big(G\sharp P_{\boldsymbol{U}}, \widehat{P}_n\big) = \sup_{f \in \mathrm{Lip}_1(\mathcal{X})} \left| \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{X}_i) - \mathbf{E}_{\tilde{\boldsymbol{X}} \sim G \sharp P_{\boldsymbol{U}}} f(\tilde{\boldsymbol{X}}) \right|.$$

In view of this relation, which follows from the Kantorovitch-Rubinstein duality theorem (Villani, 2008, Theorem 5.9, Remark 6.5), the Wasserstein distance admits a nice interpretation as a sampling error. Replacing the class of Lipschitz functions by an arbitrary functional class $\mathcal{F}$, we obtain general Integral Probability Metrics[1] (IPM): a class of pseudo-metrics on the space of probability

1. The precise definition of an IPM can be found in (3).

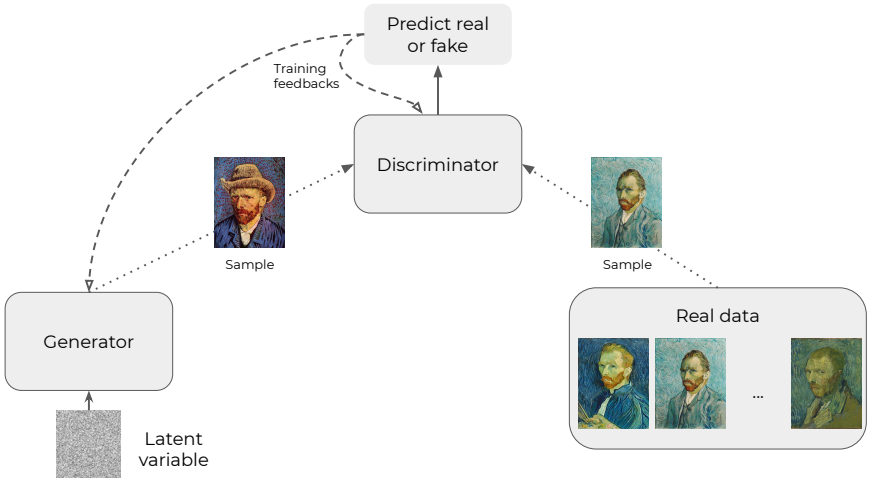

Figure 1: Illustration of the original GAN model on some of Vincent Van Gogh's self-portraits. During the training phase, real data and generated data are fed to the discriminator (dotted arrows) which in turn must predict which data is real and which is fake. Feedback (in the form of gradients of the loss) are then sent to the generator and the discriminator (broken arrows) based on predictions from the latter to update their parameters (through back-propagation in the case of neural networks). Note that the generator does not directly have access to real data.

measures (Müller, 1997). We refer the reader to Liang (2019); Sriperumbudur et al. (2012) for statistical results related to IPM. An IPM can naturally be interpreted as an adversarial loss: to compare two probability distributions, it seeks for the function f^* in $\mathcal{F}$ for which the expectations of $f(\boldsymbol{X})$ under the two distributions have the largest discrepancy. This formalization enables to study a family of pseudo-metrics which encompasses the Wasserstein-1 distance and generalises the Wasserstein GAN problem. In particular, in this work, we will consider IPM indexed by Sobolev-type classes of functions.

Since GANs initially emerged from the deep learning community, the first line of work primarily relied on empirical insights and general mathematical intuitions. Later on, a parallel line of work tackled the GAN problem from the statistical perspectives (Biau et al., 2020b,a; Chen et al., 2020; Liang, 2018; Singh et al., 2018; Luise et al., 2020; Uppal et al., 2019) as well as optimization and algorithmic viewpoints (Liang and Stokes, 2019; Kodali et al., 2017; Pfau and Vinyals, 2016; Nie and Patel, 2020; Nagarajan and Kolter, 2017; Genevay et al., 2018, 2019).

From a statistical perspective, the usual goal is to obtain a bound on the discrepancy between the learned distribution P_{learner} and the true distribution of the data $P^* = P_{\text{user}}$ with respect to a given evaluation metric d. A particularly relevant task is the quantification of the rate of convergence to zero of this discrepancy as the sample size n grows to infinity. Given a family of candidate distributions $\mathcal{P}$, typical bounds are of the form

$$\mathbf{E}_{(X_1,\dots,X_n)\sim P_{\text{obs}}}\left[\mathsf{d}(P_{\text{learner}}, P_{\text{user}})\right] - \inf_{P\in\mathcal{P}} \mathsf{d}(P, P_{\text{user}}) \lesssim n^{-r(\alpha,\beta,d,D)}.$$

for some exponent $r(\alpha, \beta, d, D) > 0$, where the parameter α characterises the *complexity* of the discriminator (*e.g.*, the smoothness of the class $\mathcal{F}$ used in the IPM), β represents the *smoothness* of the generator, d is the intrinsic dimension of the data, (*i.e.*, the dimension of the latent variable $\boldsymbol{U}$) and D is the ambient dimension (*e.g.*, the number of pixels in an image). Since D is typically much larger than d, it is suitable to avoid any dependence on D in the exponent $r(\alpha, \beta, d, D)$.

Chen et al. (2020); Liang (2018); Singh et al. (2018); Uppal et al. (2019) obtained rates depending on the smoothness of the density of the target distribution and (eventually) on the smoothness of the class $\mathcal{F}$ of admissible discriminators. Their rates do depend on the ambient dimension D, leading to the curse of dimensionality phenomenon; they do not account for possible low-dimensionnality of the data. Moreover, the learner distributions proposed in those papers are not necessarily easy-to-sample-from.

Without any smoothness assumptions, Biau et al. (2020a) provide large sample properties of the estimated distribution assuming that all the densities induced by the class of generators are dominated by a fixed known measure on a Borel subset of $\mathbb{R}^D$. When the admissible discriminators are neural networks with a given architecture, Biau et al. (2020b) obtained the parametric rate $n^{-1/2}$.

To our knowledge, Luise et al. (2020) is the only work which establishes statistical guarantees under the assumption that the data generating process is a smooth transformation of a low-dimensional latent distribution. Two key differences with our work is that Luise et al. (2020) measure quality of sampling through the Sinkhorn divergence (while we consider IPMs) and consider smoothness larger than $d/2$. The latter leads to parametric rates of convergence $n^{-1/2}$. Note also that the Sinkhorn divergence, introduced as a compelling computational alternative to the Wasserstein distance (Cuturi, 2013), does not admit a straightforward interpretation as a sampling error.

In this work, we assess the impact of the smoothness of the data generating process and the low-dimensionality of the latent space on the rates of convergence. The rates in the literature either depend on the ambient dimension, which can not explain the effectiveness of GANs, or assume strong smoothness assumption leading to parametric rate. This prevents a fine-grained analysis of the interplay between dimensions and smoothness. In this work we obtain rates which, in terms of dimension, depend only on the intrinsic dimension d of the data and on the smoothness of the data generating process and the admissible discriminators.

3. Problem statement

We are given n points $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ in $\mathbb{R}^D$, that we assume drawn independently from an unknown joint probability distribution $P_{\text{obs}}^{(n)}$. We will make the hypothesis that the data points lie—up to a small noise—on a d-dimensional smooth manifold $\mathcal{M}$ with an intrinsic dimension d much smaller than the ambient dimension D. More precisely, we assume that the $\boldsymbol{X}_i$'s are perturbed versions of n independent copies of a point randomly sampled from a distribution P^* supported on the smooth manifold $\mathcal{M}$. The goal of generative modeling is to design a smooth function

$$g : [0,1]^d \to [0,1]^D$$

such that the image of the uniform distribution $\mathcal{U}_d := \mathcal{U}([0,1]^d)$ by g is close to the target distribution P^*. Of course, this framework requires to make precise what is meant by "smoothness" of the function g and how the closeness of two distributions is measured. Since the goal of the present

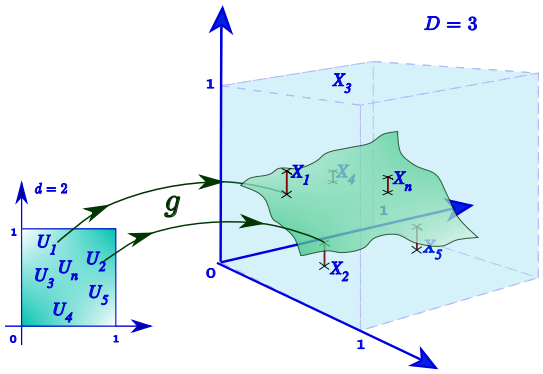

Figure 2: An illustration of **Assumption A**. Most $\boldsymbol{X}_i$'s are close to the manifold defined as the image of $[0,1]^d$ by the smooth map g. A small fraction of the $\boldsymbol{X}_i$'s (such as $\boldsymbol{X}_3$ in this figure) might be at a large distance from $g([0,1]^d)$.

work is to gain a better theoretical understanding of the problem of generative modeling, we assume that the "intrinsic dimension" d is known.

The following condition will be assumed to be true throughout this work, where $\sigma \geq 0$ and $\varepsilon \in [0,1]$ are fixed yet possibly unknown constants.

Assumption A: There exists a mapping $g^* : [0,1]^d \to [0,1]^D$ (with $d \ll D$), as well as random vectors $\boldsymbol{U}_1, \dots, \boldsymbol{U}_n \in \mathbb{R}^d$ and $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_n \in \mathbb{R}^D$ such that

- $\boldsymbol{U}_i$ are iid uniformly distributed in the hypercube $[0,1]^d$ (denoted by $\boldsymbol{U}_i \overset{\text{iid}}{\sim} \mathcal{U}_d$),
- $\max_{i=1,\dots,n} \mathbf{E}[\|\boldsymbol{\xi}_i\|_2] \leq \sigma$ for some $\sigma < \infty$,
- For some $\mathcal{I} \subset \{1, \dots, n\}$ of cardinality at least $(1-\varepsilon)n$, we have $\boldsymbol{X}_i = g^*(\boldsymbol{U}_i) + \boldsymbol{\xi}_i$ for every $i \in \mathcal{I}$.

The parameters σ and ε, referred to as the noise magnitude and the rate of contamination, are unknown but assumed to be small. The subset $\mathcal{I}$ in the last item of the assumption is the set of inliers. **Assumption A** means that up to some noise, the inliers are drawn from the uniform distribution on the hyper-cube and pushed-forward by g^*. The setting considered here is *adversarial*: the set of inliers and the values of the outliers $\{\boldsymbol{X}_i : i \notin \mathcal{I}\}$ may depend on all the random variables $\boldsymbol{U}_i, \boldsymbol{X}_i, \boldsymbol{\xi}_i$. Furthermore, $\boldsymbol{U}_i$ and $\boldsymbol{\xi}_i$ are not necessarily independent.

In what follows, we set $P^* = g^* \sharp \mathcal{U}_d$ and call it the oracle generator. Let d be a pseudo-metric on the space of all probability measures on $\mathbb{R}^D$. Most relevant examples in the present context are IPMs, but one could also consider the Wasserstein q-distances with $q \geq 1$, the Hellinger distance, the maximum mean discrepancy and so on. For every candidate generator g—a measurable mapping from $[0,1]^d$ to $\mathbb{R}^D$—we define the risk

$$R_{\mathsf{d},P^*}(g) := \mathsf{d}\big(g \sharp \mathcal{U}_d, P^*\big). \tag{2}$$

Our goal is to find a mapping

$$\begin{aligned}\widehat{G} : (\mathbb{R}^D)^n &\to \mathcal{G} \\ (\boldsymbol{X}_1, \dots, \boldsymbol{X}_n) &\mapsto \widehat{g}_n,\end{aligned}$$

such that $R_{\mathsf{d},P^*}(\widehat{g}_n)$ is as small as possible. Note here that $R_{\mathsf{d},P^*}(\widehat{g}_n)$ is a random variable, since $\widehat{g}_n$ is random. Let $\mathcal{G}$ be a set of smooth (at least Lipschitz continuous) functions from $[0,1]^d$ to $\mathbb{R}^D$. We define the generator minimizing the empirical risk, hereafter referred to as the ERM, by

$$\widehat{g}_{n,\mathcal{G}}^{\text{ERM}} \in \arg\min_{g \in \mathcal{G}} \mathsf{d}\big(g\sharp \mathcal{U}_d, \widehat{P}_n\big). \tag{ERM}$$

We assume that the minimum is attained. Our results extend easily to the case in which it is not attained but adds some unnecessary technicalities. Our main result, presented in the next section, provides an upper bound on the risk (2) of the ERM.

To enforce requirement **R2**, we consider distances on the space of probability distributions that can be expressed as integral probability metrics for a class $\mathcal{F}$ of real-valued functions defined on $[0,1]^D$. More precisely, we define an integral probability metrics (IPM) for $\mathcal{F}$ as follows:

$$\mathsf{d}_{\mathcal{F}}(P, Q) = \sup_{f \in \mathcal{F}} |\mathbf{E}_P[f(\boldsymbol{X})] - \mathbf{E}_Q[f(\boldsymbol{X})]|. \tag{3}$$

Classical examples are the total variation and the Wasserstein-1 distances, corresponding respectively to $\mathcal{F} = \{f : \sup_x |f(x)| \leq 1\}$ and $\mathcal{F} = \{f : |f(x) - f(y)| \leq \|x - y\| \text{ for all } x, y\}$.

4. Warming up: guarantees in the noiseless setting for W_1

Let us first consider the noiseless and uncontaminated setting $\sigma = \varepsilon = 0$, corresponding to $P_{\text{obs}} = (P^*)^{\otimes n}$. To convey the main ideas of this work without diving into technicalities, we first consider the case of the Wasserstein W_1-distance. Using arguments that are now standard in learning theory, we get[2]

$$R_{\mathsf{d},P^*}(\widehat{g}_{n,\mathcal{G}}^{\text{ERM}}) \leq \inf_{g \in \mathcal{G}} \mathsf{d}\big(g\sharp \mathcal{U}_d, P^*\big) + 2\,\mathsf{d}\big(\widehat{P}_n, P^*\big). \tag{4}$$

This inequality holds for any pseudo-metric d. It follows from the following chain of inequalities:

$$\begin{aligned} R_{\mathsf{d},P^*}(\widehat{g}_{n,\mathcal{G}}^{\text{ERM}}) &= \mathsf{d}\big(\widehat{g}_{n,\mathcal{G}}^{\text{ERM}}\sharp \mathcal{U}_d, P^*\big) \\ &\leq \mathsf{d}\big(\widehat{g}_{n,\mathcal{G}}^{\text{ERM}}\sharp \mathcal{U}_d, \widehat{P}_n\big) + \mathsf{d}\big(\widehat{P}_n, P^*\big) \\ &\leq \inf_{g \in \mathcal{G}} \mathsf{d}\big(g\sharp \mathcal{U}_d, \widehat{P}_n\big) + \mathsf{d}\big(\widehat{P}_n, P^*\big) \\ &\leq \inf_{g \in \mathcal{G}} \mathsf{d}\big(g\sharp \mathcal{U}_d, P^*\big) + 2\,\mathsf{d}\big(\widehat{P}_n, P^*\big). \end{aligned}$$

Note that if we replace in (ERM) the empirical distribution $\widehat{P}_n$ by another estimator $\tilde{P}_n$ of P^*, then (4) continues to be true with $\tilde{P}_n$ instead of $\widehat{P}_n$ in the right hand side.

2. See (Liang, 2018, Lemma 1) for a similar result.

The inequality (4) provides an upper bound on the risk that is composed of the approximation error $\inf_{g\in\mathcal{G}} \mathsf{d}\big(g\sharp\mathcal{U}_d, P^*\big)$ and the stochastic error $2\,\mathsf{d}\big(\widehat{P}_n, P^*\big)$. While the former is unavoidable, it is not clear how tight the latter is. In particular, the fact that the term $2\,\mathsf{d}\big(\widehat{P}_n, P^*\big)$ measures the distance between the unknown distribution P^* and an approximation of it that does not take into account the specific structure of P^* suggests that it might be possible to get a better upper bound.

This being said, we stick here to inequality (4) and devote the rest of this paper to establishing upper bounds on the stochastic error. To this end, we take advantage of the interplay between the assumptions on P_X and P^* on the one hand, and the set $\mathcal{F}$ defining the IPM $\mathsf{d} = \mathsf{d}_\mathcal{F}$ on the other hand. In the case when both the mapping g^* underlying P^* and the elements of $\mathcal{F}$ are Lipschitz, we get the following result.

Theorem 1 *Let Assumption A be fulfilled with $\sigma = \varepsilon = 0$ and $g^* \in \mathrm{Lip}_L([0,1]^d)$ for some $L > 0$. Let $\mathsf{d} = \mathsf{W}_1$ and set $\widehat{g}_n = \widehat{g}^{\mathrm{ERM}}_{n,\mathcal{G}}$. Then, for some universal constant $c > 0$,*

$$\mathbf{E}[R_{\mathsf{W}_1,P^*}(\widehat{g}_n)] \le \inf_{g\in\mathcal{G}} R_{\mathsf{W}_1,P^*}(g) + \frac{cL\sqrt{d}}{n^{1/d}\wedge n^{1/2}}\big(1 + \mathbb{1}_{d=2}\log n\big). \tag{5}$$

The full proof of this result being postponed to Section 7.1, we provide here a sketch of it. In view of (4), it suffices to upper bound $\mathsf{d}(\widehat{P}_n, P) = \mathsf{W}_1(\widehat{P}_n, P^*)$. Since P^* and $\widehat{P}_n$ are the pushforward measures of $\mathcal{U}_d$ and its empirical counterpart by the same Lipschitz mapping, and the composition of two Lipschitz mappings is still Lipschitz, we can upper bound $\mathsf{W}_1(\widehat{P}_n, P^*)$ by $L\mathsf{W}_1(\widehat{P}_{U,n}, \mathcal{U}_d)$. Here, $\widehat{P}_{U,n}$ is the empirical distribution of $\boldsymbol{U}_1, \dots, \boldsymbol{U}_n$ independently sampled from $\mathcal{U}_d$. It is known that, for the Wasserstein-1 distance, there is a universal constant $c > 0$ such that $\mathbf{E}[\mathsf{W}_1(\widehat{P}_{U,n}, \mathcal{U}_d)]$ is upper bounded by the second summand of the right hand side of (5); this fact has been established in the seminal paper Dudley (1969) and later refined and extended by many authors; see Weed and Bach (2019); Singh and Póczos (2018); Lei (2020) and references therein. The version we use here (with an explicit dependence of the constant on the dimension) can be found in Niles-Weed and Rigollet (2019, Prop. 1). This completes the proof.

Some remarks are in order. First, the rate of convergence to zero of the stochastic term, when the sample size goes to infinity, is characterized by the intrinsic dimension only. This rate, $n^{-1/d}$, is much smaller than the naive rate $n^{-1/D}$ provided that the intrinsic dimension is small as compared to D. To the best of our knowledge, despite the embarrassing simplicity of this result, this is the first time that this phenomenon is highlighted in the context of generative modeling.

The second remark concerns the fact that the choice of the set $\mathcal{G}$ in (ERM) impacts only the first term, the approximation error, in the risk bound given by (5). This indicates that inequality (5) might not be tight when $\mathcal{G}$ is a very narrow set. On the positive side, this bound implies that the set $\mathcal{G}$ can be chosen very large, as long as feature **R1** holds and optimisation problem (ERM) is computationally tractable. Finally, one can wonder whether the assumption that g^* is Lipschitz is realistic in some applications. We believe that it is. Indeed, the generator learned by GAN is a Lipschitz function of the input (Seddik et al., 2020) and leads to qualitatively good results. Therefore, it makes perfect sense to assume that g^* is Lipschitz.

5. Main result in the noisy setting for smoothness classes

The rate of convergence obtained in the previous section might be overly pessimistic. Indeed, the Wasserstein distance W_1 might be very weak for many applications: it may be sufficient to take as $\mathcal{F}$ a set which is much smaller than that of the Lipschitz functions. In particular, one can consider the case where $\mathcal{F}$ is a smoothness class with a degree of smoothness strictly larger than one. The main result stated below considers this setting and answers the following three questions:

[Q1] *Can we take advantage of the further smoothness of g^* and that of the functions in $\mathcal{F}$ for improving the risk bound (5)?*

[Q2] *How does the noise magnitude σ impact the risk?*

[Q3] *Can we get meaningful risk bounds if some data points $\boldsymbol{X}_i$ are corrupted?*

To answer these questions, we consider the case of smoothness classes containing all the functions with bounded partial derivatives up to a given order. Let $\mathcal{X} \subset \mathbb{R}^D$ be some compact set, which will be chosen to be $[0,1]^D$ later on in this section. In what follows, for every positive integer α, $C^\alpha(\mathcal{X}, \mathbb{R})$ denotes the set of all α-times continuously differentiable functions. In addition, for a multi-index $\boldsymbol{k} \in \mathbb{N}^D$, we write $\mathtt{D}^{\boldsymbol{k}} f$ for the $\boldsymbol{k}$-th order differential of f. Define the α-smoothness class $\mathcal{W}^\alpha(\mathcal{X}; r)$ over $\mathcal{X}$ with radius $L > 0$ by

$$\mathcal{W}^\alpha(\mathcal{X}; L) := \left\{ f \in C^\alpha(\mathcal{X}, \mathbb{R}) : \max_{|\boldsymbol{k}| \le \alpha} \|\mathtt{D}^{\boldsymbol{k}} f\|_\infty \le L \right\}.$$

Clearly, $\mathcal{W}^1(\mathcal{X}; L)$ is included in the set $\mathrm{Lip}_L(\mathcal{X})$ of Lipschitz-continuous functions. Furthermore, one can check that $\mathcal{W}^1(\mathcal{X}; L)$ is dense in $\mathrm{Lip}_L(\mathcal{X})$.

Theorem 2 *Let Assumption A hold and let the coordinates g_j^* of g^* belong to $\mathcal{W}^\alpha([0,1]^d, L)$ for some $L \ge 1$. Then, if $\mathcal{F} = \mathcal{W}^\alpha([0,1]^D, 1)$ in the definition of the IPM, we have*

$$\mathbf{E}[R_{\mathsf{d}_\mathcal{F}, P^*}(\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}})] \le \inf_{g \in \mathcal{G}} R_{\mathsf{d}_\mathcal{F}, P^*}(g) + L(\sigma + 2\varepsilon) + \frac{cL^\alpha}{n^{\alpha/d} \wedge n^{1/2}} \big(1 + \mathbb{1}_{d=2\alpha} \log n\big). \tag{6}$$

where c is a constant which depends only on α, d, D.

Let us note that this theorem answers the three questions **Q1**-**Q3**. In particular, it shows that if the oracle generator map g^* is α-smooth with $\alpha \le d/2$, and the test function defining the distance $\mathsf{d}_\mathcal{F}$ are α-smooth as well, then the last term of the risk bound of the generator minimizing the empirical risk is of order $n^{-\alpha/d}$. This rate improves with increasing α and reaches the optimal rate $n^{-1/2}$, up to a log factor, when $\alpha = d/2$. It also follows from (6) that the risk of the generator $\widehat{g}_{n,\mathcal{G}}^{\mathrm{ERM}}$ decreases linearly fast in the noise magnitude σ and the contamination rate ε, when these parameters go to zero.

As mentioned earlier, (6) is a consequence of (4) and we do not know whether the latter is tight. However, we can show that the right hand side of (6) is a tight upper bound on the right hand side of (6). More precisely, as stated in the next result, the dependence on σ and ε is tight, while the dependence on n is tight when $\alpha = 1$ or $\alpha > d/2$.

Theorem 3 *Let $\mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)$ be the set of all distributions of n points $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ in $\mathbb{R}^D$ satisfying Assumption A. Let $\mathcal{G}^*$ be a set of functions $g:[0,1]^d\to[0,1]^D$ containing the linear functions. If $\sigma\le 1/2$ and $\mathcal{F}$ contains the projection onto the first axis $\boldsymbol{x}\in[0,1]^D\mapsto x_1\in\mathbb{R}$, then there is a universal constant $c_1>0$ such that*

$$\sup_{g^*\in\mathcal{G}^*}\sup_{P^{(n)}\in\mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)}\mathbf{E}_{P^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n,P^*)]\ge c_1\Big(\sigma+\varepsilon+\frac{1}{n^{1/2}}\Big).$$

If, in addition, $\mathcal{F}$ contains the set of all 1-Lipschitz functions, then

$$\sup_{g^*\in\mathcal{G}^*}\sup_{P^{(n)}\in\mathcal{P}_{n,D}(d,\sigma,\varepsilon)}\mathbf{E}_{\mathcal{P}^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n,P^*)]\ge c_1(\sigma+\varepsilon)+\frac{c_d\big(1+\mathbb{1}_{d=2}\log n\big)}{n^{1/d}\wedge n^{1/2}},$$

where c_1 is a universal constant and c_d is a constant depending on d.

The proof of the theorem is postponed to Section 7.4. Note that it does not establish the tightness of the dependence of the bound in n in the case of smoothness $\alpha\in(1,d/2)$. However, it is very likely that the rate is also optimal in this case as well.

To complete this section, we show that the dependence in ε and σ of the upper bound (6) is tight.

Theorem 4 *Let $\mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)$ be the set of all distributions of n points $(\boldsymbol{X}_1,\dots,\boldsymbol{X}_n)$ in $\mathbb{R}^D$ satisfying Assumption A. Let $\mathcal{G}^*$ be a set of functions $g:[0,1]^d\to[0,1]^D$ containing the affine functions. Assume that $\mathcal{F}$ is a set of functions $f:[0,1]^D\to\mathbb{R}$ bounded by[3] L and containing the projection onto the first axis $\boldsymbol{x}\in[0,1]^D\mapsto x_1\in\mathbb{R}$. Then, if $\sigma\le 1/2$ and $n\ge(6/\varepsilon)\log(20L/\varepsilon)$, we have*

$$\inf_{\widehat{g}_n}\sup_{g^*\in\mathcal{G}^*}\sup_{P^{(n)}\in\mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)}\mathbf{E}[R_{\mathsf{d}_{\mathcal{F}},P^*}(\widehat{g}_n)]\ge 0.1(\sigma+\varepsilon),$$

where the inf is taken over all possible generators $\widehat{g}_n$.

The proof of this result is postponed to Section 7.5. If we compare this lower bound with the upper bound of Theorem 2, we see that the linear dependence of the expected risk on the parameters σ and ε is optimal and cannot be improved. This is true for any generator, meaning that the empirical risk minimizer is minimax rate-optimal in terms of σ and ε. We are currently working on establishing similar lower bounds showing the optimality in terms of n as well.

6. Conclusion and outlook

In this work, we introduced a general and nonparametric framework for learning generative models. Given data in a possibly high-dimensional space, we learn their distribution in order to sample new data points that resemble the training ones, while not being identical to those. A key point in our work is to leverage the fact that the distribution of the training samples, up to some noise and adversarial contamination, is supported by a low-dimensional smooth manifold. This allows

3. It can be checked that the same result holds if the functions f satisfy $\max_x f(x)-\min_x f(x)\le L$.

us to alleviate the curse of dimensionality. Such an assumption is very reasonable as it reflects the structural properties of the training samples. For instance, the MNIST dataset (LeCun, 1998) is composed of 28×28 pixels pictures of handwritten digits while the intrinsic dimension of the data is estimated to be around 14 (Costa and Hero, 2004; Levina and Bickel, 2005).

We established risk bounds for the minimizer of the distance between the empirical distribution and admissible generators, where an admissible generator is a smooth function pushing forward a low-dimensional uniform distribution into the high-dimensional sample space. We use Integral Probability Metrics for measuring the discrepancy between the target distribution and our estimate: These metrics, which include the total variation and the Wasserstein-1 distances, mimic the role of a discriminator which would try to discriminate between true samples and the simulated ones.

By proving new bounds on the distance between such distributions and their empirical counterparts, we were able to derive nonasymptotic bounds for the regret of our empirical risk minimizer, with rates of convergence that only depend on the ambient dimension through fixed multiplicative constants. Our new bounds, which are of independent interest, leverage both the smoothness of the distribution of the samples and that of the functions in the IPM class.

We were also able to take into account possible adversarial corruption of the training samples both by noise (*e.g.*, blurry images) and by a small proportion of outliers (*i.e.*, wrong samples in the training set), inducing some error terms that are shown to be unavoidable. To the best of our knowledge, this is the first result assessing the influence of the noise and of the contamination on the error of generative modeling. This constitutes an appealing complement to the recently obtained statistical guarantees (Biau et al., 2020b; Luise et al., 2020).

As a route for future work, we believe that our regret bounds are not minimax optimal in all possible regimes (depending on the smoothness of the generators). Namely, it is not clear that fitting our generator to the empirical distribution $\widehat{P}_n$ yields an optimal method, especially when the smoothness α is less than the half of the dimension d. It might be more judicious to fit the generator to a smoothed version of the empirical distribution $\widehat{P}_n$.

7. Proofs

This section contains the proofs of the main results stated in previous sections. We start by providing the proof of Theorem 1. Then, the proof of Theorem 2 is presented up to the proof of a technical lemma on the composition of smooth functions, postponed to Section 7.3.

7.1. Proof of Theorem 1

To ease notation, we write $\widehat{g}_n$ instead of $\widehat{g}^{\mathrm{ERM}}_{n,\mathcal{G}}$. In view of (4), we have

$$R_{\mathsf{W}_1,P^*}(\widehat{g}_n) \leq \inf_{g\in\mathcal{G}} \mathsf{W}_1(g\sharp\mathcal{U}_d, P^*) + 2\mathsf{W}_1(\widehat{P}_n, P^*).$$

Using the variational formulation of the Wasserstein-1 distance we write

$$\begin{aligned}
\mathsf{W}_1(\widehat{P}_n, P^*) &= \sup_{f \in \mathrm{Lip}_1([0,1]^D)} \Big| \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbf{E}_{\boldsymbol{X} \sim P^*} f(\boldsymbol{X}) \Big| \\
&= \sup_{f \in \mathrm{Lip}_1([0,1]^D)} \Big| \frac{1}{n} \sum_{i=1}^n f \circ g^*(\boldsymbol{U}_i) - \mathbf{E}_{\boldsymbol{U} \sim \mathcal{U}_d} f \circ g^*(\boldsymbol{U}) \Big| \\
&= \sup_{h \in \mathcal{H}_L} \Big(\frac{1}{n} \sum_{i=1}^n h(\boldsymbol{U}_i) - \mathbf{E}_{\boldsymbol{U} \sim \mathcal{U}_d} h(\boldsymbol{U}) \Big)
\end{aligned}$$

where we define the class $\mathcal{H}_L = \{h : [0,1]^d \to \mathbb{R} : h = f \circ g^*, f \in \mathrm{Lip}_1([0,1]^D)\}$. Finally, taking the expectation and noting that $\mathcal{H}_L$ is a subset of the the L-Lipschitz functions on $[0,1]^d$ with values in $\mathbb{R}$, we get

$$\begin{aligned}
\mathbf{E}[\mathsf{W}_1(\widehat{P}_n, P^*)] &\le \mathbf{E}\Big[\sup_{h \in \mathrm{Lip}_L([0,1]^d)} \Big(\frac{1}{n} \sum_{i=1}^n h(\boldsymbol{U}_i) - \mathbf{E}_{\boldsymbol{U} \sim \mathcal{U}_d} h(\boldsymbol{U}) \Big) \Big] \\
&\le L\, \mathbf{E}[\mathsf{W}_1(\widehat{P}_{U,n}, \mathcal{U}_d)] \\
&\le \frac{cL\sqrt{d}}{n^{1/d} \wedge n^{1/2}} \big(1 + \mathbb{1}_{d=2} \log n\big),
\end{aligned}$$

with c a universal constant. The last inequality follows from Niles-Weed and Rigollet (2019, Proposition 1).

7.2. Proof of Theorem 2

In view of (4), we need to establish an upper bound on the expected stochastic error

$$\mathrm{NoisyStochErr}_n = \mathbf{E}[\mathsf{d}_{\mathcal{F}}(P_n, P^*)] = \mathbf{E}\Big[\sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbf{E}[f(g^*(\boldsymbol{U}))] \Big| \Big],$$

where $\boldsymbol{U} \sim \mathcal{U}_d$ and $\mathcal{F} = \mathcal{W}^\alpha([0,1]^D, 1)$. The first step in the proof is a lemma showing the influence of the noise and the corruption on the error $\mathrm{StochErr}_n$.

Lemma 7.1 *If P_X satisfies Assumption A with $\varepsilon \in [0,1]$ and all the functions in $\mathcal{F}$ are bounded by a constant $L_{\mathcal{F}}$ and Lipschitz with constant $L_{\mathcal{F}}$, then*

$$\mathrm{NoisyStochErr}_n \le L_{\mathcal{F}} \sigma + 2 M_{\mathcal{F}} \varepsilon + \mathrm{NoiseFreeStochErr}_n,$$

where

$$\mathrm{NoiseFreeStochErr}_n = \mathbf{E}\Big[\sup_{f \in \mathcal{F}} \Big| \frac{1}{n} \sum_{i=1}^n (f \circ g^*)(\boldsymbol{U}_i) - \mathbf{E}[(f \circ g^*)(\boldsymbol{U})] \Big| \Big],$$

with $\boldsymbol{U}, \boldsymbol{U}_1, \dots, \boldsymbol{U}_n$ iid random vectors drawn from $\mathcal{U}_d$.

Proof The triangle inequality yields

$$\text{StochErr}_n \leq \mathbf{E}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(\boldsymbol{X}_i)-(f\circ g^*)(\boldsymbol{U}_i)\right\}\right|\right] + \text{NoiseFreeStochErr}_n.$$

Let us define $\boldsymbol{Y}_i = g^*(\boldsymbol{U}_i) + \boldsymbol{\xi}_i$ for $i = 1,\dots,n$. The third item of Assumption A implies that $\boldsymbol{Y}_i = \boldsymbol{X}_i$ for $i \in \mathcal{I}$. For $i \notin \mathcal{I}$, we have $|f(\boldsymbol{X}_i) - f(\boldsymbol{Y}_i)| \leq 2M_{\mathcal{F}}$. Therefore, the first term in the right hand side of the last display can be further bounded as follows:

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(\boldsymbol{X}_i)-(f\circ g^*)(\boldsymbol{U}_i)\right\}\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n}\left\{f(\boldsymbol{Y}_i)-(f\circ g^*)(\boldsymbol{U}_i)\right\}\right| + \frac{2M_{\mathcal{F}}(n-n_{\mathcal{I}})}{n}\\ &\leq \frac{L_{\mathcal{F}}}{n}\sum_{i=1}^{n}\left\|\boldsymbol{Y}_i - g^*(\boldsymbol{U}_i)\right\| + 2M_{\mathcal{F}}\,\varepsilon\\ &= \frac{L_{\mathcal{F}}}{n}\sum_{i=1}^{n}\left\|\boldsymbol{\xi}_i\right\| + 2M_{\mathcal{F}}\,\varepsilon.\end{aligned}$$

To get the claimed result, it suffices to take the expectation of both sides of the last display. ■

The next step consists in upper bounding the stochastic error in the noise free case. If we use the notation $\mathcal{F}\circ g^* = \{f\circ g^* : f\in\mathcal{F}\}$, the noise free stochastic error can be written as

$$\text{NoiseFreeStochErr}_n = \mathbf{E}[\mathsf{d}_{\mathcal{F}\circ g^*}(\widehat{P}_{U,n},\mathcal{U}_d)]. \tag{7}$$

We see that the problem is reduced to that of evaluating the distance between the uniform distribution and the empirical distribution of n independent random points uniformly distributed on the unit hypercube. In order to upper bound this distance, we first show that the class $\mathcal{F}\circ g^*$, under the assumptions of Theorem 2, is included in a smoothness class of order α. The precise statement is the following.

Lemma 7.2 *Let $g : [0,1]^d \to [0,1]^D$ and $h : [0,1]^D \to [-1,1]$ two mappings such that $g \in \mathcal{W}^\alpha([0,1]^d, L)$ and $h \in \mathcal{W}^\alpha([0,1]^D, 1)$ for some $\alpha \in \mathbb{N}^*$ and some $L \geq 1$. Then, there exists a constant $C = C(D,d,\alpha)$ such that*

$$|\mathrm{D}^{\boldsymbol{k}}(h\circ g)(\boldsymbol{x})| \leq CL^\alpha, \quad \forall \boldsymbol{x}\in[0,1]^d,$$

for every multi-index $\boldsymbol{k} = (k_1,\dots,k_d) \in \mathbb{N}^d$ such that $|\boldsymbol{k}| \leq \alpha$.

This lemma, in conjunction with (7) and the assumption $g^* \in \mathcal{W}^\alpha([0,1]^d, L)$, implies that

$$\begin{aligned}\text{NoiseFreeStochErr}_n &\leq \mathbf{E}[\mathsf{d}_{\mathcal{W}^\alpha([0,1]^d, CL^\alpha)}(\widehat{P}_{U,n},\mathcal{U}_d)]\\ &= CL^\alpha\mathbf{E}[\mathsf{d}_{\mathcal{W}^\alpha([0,1]^d,1)}(\widehat{P}_{U,n},\mathcal{U}_d)].\end{aligned}$$

The last step is to use Schreuder (2020, Theorem 4), which provides the inequality

$$\mathbf{E}[\mathsf{d}_{\mathcal{W}^\alpha([0,1]^d, CL^\alpha)}(\widehat{P}_{U,n},\mathcal{U}_d)] \leq \tilde{C}L^\alpha n^{-(\alpha\wedge d/2)/d}(1+\mathbb{1}_{\alpha=d/2}\log n).$$

This completes the proof of the theorem.

7.3. Image of a smoothness class by a smooth function

Proof of Lemma 7.2 The proof relies on Fraenkel (1978, Formula B) providing an explicit formula for derivatives of composite functions: for any multi-index $\boldsymbol{k}$ such that $1 \leq |\boldsymbol{k}| \leq \alpha$ and for any $\boldsymbol{x} \in [0,1]^d$,

$$\mathrm{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x}) = \boldsymbol{k}! \sum_{\boldsymbol{a}:1\leq|\boldsymbol{a}|\leq|\boldsymbol{k}|} \frac{(\mathrm{D}^{\boldsymbol{a}}h)(g(\boldsymbol{x}))}{\boldsymbol{a}!} Q_{\boldsymbol{k},\boldsymbol{a}}(g;\boldsymbol{x}),$$

where $Q_{\boldsymbol{k},\boldsymbol{a}}(g;\cdot)$ is a homogeneous polynomial of degree $|\boldsymbol{a}|$ in derivatives of $g_1,\dots,g_D$. Since the partial derivatives of h of any order up to α are bounded by one, we infer from the last display that

$$\left|\mathrm{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x})\right| = \boldsymbol{k}! \sum_{\boldsymbol{a}:1\leq|\boldsymbol{a}|\leq|\boldsymbol{k}|} \frac{1}{\boldsymbol{a}!} \left|Q_{\boldsymbol{k},\boldsymbol{a}}(g;\boldsymbol{x})\right|. \tag{8}$$

We can give an explicit expression of $Q_{\boldsymbol{k},\boldsymbol{a}}$ using the following notation. Let r be the cardinality of the set $\{\boldsymbol{\beta} \in \mathbb{N}^d \,|\, 0 < \boldsymbol{\beta} \leq \boldsymbol{\gamma}\}$ and $\boldsymbol{\beta}(1),\dots,\boldsymbol{\beta}(r)$ be its elements somehow enumerated. Define, for $\boldsymbol{\gamma} \in \mathbb{N}^d$ and for $a \in \mathbb{N}$, the set of multi-indices

$$R(\boldsymbol{\gamma}, a) = \left\{ \boldsymbol{\rho} \in \mathbb{N}^r \,\Big|\, \sum_{j=1}^{r} \rho_j \boldsymbol{\beta}(j) = \boldsymbol{\gamma}, |\boldsymbol{\rho}| = a \right\},$$

and, for any $v : \mathbb{R}^d \to \mathbb{R}$, the polynomials

$$P_{\boldsymbol{\gamma}}(a, v; \boldsymbol{x}) = \sum_{\boldsymbol{\rho} \in R(\gamma,a)} \frac{a!}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{(\mathrm{D}^{\boldsymbol{\beta}(j)} v(\boldsymbol{x}))^{\rho_j}}{\boldsymbol{\beta}(j)!}. \tag{9}$$

The functions $Q_{\boldsymbol{k},\boldsymbol{a}}$ in (8) are given by

$$Q_{\boldsymbol{k},\boldsymbol{a}}(g;\boldsymbol{x}) = \sum_{\boldsymbol{\gamma}(1)+\dots+\boldsymbol{\gamma}(D)=\boldsymbol{k}} \prod_{m=1}^{D} P_{\boldsymbol{\gamma}(m)}(a_m, g_m; \boldsymbol{x}).$$

Since, according to the conditions of the lemma, all the partial derivatives of g appearing in (9) for $v = g_m$ are bounded by $L \geq 1$, we have

$$\left|P_{\boldsymbol{\gamma}(m)}(a_m, g_m; \boldsymbol{x})\right| \leq \sum_{\boldsymbol{\rho} \in R(\boldsymbol{\gamma}(m), a_m)} L^{|\boldsymbol{\rho}|} \frac{a_m!}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{1}{\boldsymbol{\beta}(j)!}.$$

Since $|\boldsymbol{\rho}| \leq a_m$ and $|\boldsymbol{a}| \leq |\boldsymbol{k}| \leq \alpha$, this leads to

$$\left|Q_{\boldsymbol{k},\boldsymbol{a}}(g;\boldsymbol{x})\right| \leq L^{\alpha} \boldsymbol{a}! \sum_{\boldsymbol{\gamma}(1)+\dots+\boldsymbol{\gamma}(D)=\boldsymbol{k}} \prod_{m=1}^{D} \left(\sum_{\boldsymbol{\rho} \in R(\boldsymbol{\gamma}(m), a_m)} \frac{1}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{1}{\boldsymbol{\beta}(j)!} \right).$$

Combining this inequality with (8), we arrive at

$$|\mathrm{D}^{\boldsymbol{k}}(h \circ g)(\boldsymbol{x})| \leq L^{\alpha} \boldsymbol{k}! \sum_{1\leq|\boldsymbol{a}|\leq|\boldsymbol{k}|} \sum_{\boldsymbol{\gamma}(1)+\dots+\boldsymbol{\gamma}(D)=\boldsymbol{a}} \prod_{m=1}^{D} \left(\sum_{\boldsymbol{\rho} \in R(\boldsymbol{\gamma}(m), a_m)} \frac{1}{\boldsymbol{\rho}!} \prod_{j=1}^{r} \frac{1}{\boldsymbol{\beta}(r)!} \right).$$

Denoting by $C(D, d, \alpha)$ the maximum of the right hand side over all multi-indices $\boldsymbol{k}$ such that $|\boldsymbol{k}| \leq \alpha$, we get the claim of the lemma. ■

7.4. Proof of the lower bounds in Theorem 3

Since the bound we wish to prove does not depend on the dimension, we assume without loss of generality that $D = d$. First, we start by considering the case $\sigma + \varepsilon \geq 2/n^{1/2}$.

Let us define $g^*(\boldsymbol{x}) = (2\boldsymbol{x} + 1)/4$. This function is clearly 1-Lipschitz. Let ξ_1 be a random variable drawn from the uniform in $[0, 1]$ distribution. We define $P_0^{(n)}$ to be the distribution of i.i.d. vectors $\boldsymbol{X}_1, \dots, \boldsymbol{X}_n$ such that $\boldsymbol{X}_i \overset{\text{dist}}{\sim} g^*(\boldsymbol{U}) + \sigma\xi_1$ for $i = 1, \dots, n\varepsilon$ and $\boldsymbol{X}_i = (1, \dots, 1)^\top$ for $i > n\varepsilon$. Then, it is clear that $P_0^{(n)} \in \mathcal{P}_{n,D}\big(d, \sigma, (1-\varepsilon)n\big)$ and

$$\begin{aligned}
\mathbf{E}_{P_0^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)] &= \mathbf{E}_{P_0^{(n)}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(\boldsymbol{X}_i) - \mathbf{E}[(f\circ g^*)(\boldsymbol{U})]\right|\right] \\
&\geq \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n X_{i,1} - \mathbf{E}[g^*(\boldsymbol{U})_1]\right|\right] \\
&= \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n \big(X_{i,1} - \mathbf{E}[X_{i,1}]\big) + \varepsilon + 0.5\sigma - \varepsilon\mathbf{E}[g^*(\boldsymbol{U})_1]\right|\right] \\
&= \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n \big(X_{i,1} - \mathbf{E}[X_{i,1}]\big) + 0.5(\sigma+\varepsilon)\right|\right]. \qquad (10)
\end{aligned}$$

The first inequality above follows by replacing the sup over $\mathcal{F}$ by the corresponding expression evaluated at the representer $f_0(\boldsymbol{x}) = x_1$. The third line above follows from $\mathbf{E}[X_{i,1}] = \mathbf{E}[g^*(\boldsymbol{U})_1] + 0.5\sigma$ if $i \leq n\varepsilon$ whereas $\mathbf{E}[X_{i,1}] = 1$ if $i > n\varepsilon$. The last line is a consequence of $\mathbf{E}[g(\boldsymbol{U})_1] = 0.5$. Combining the above lower bound with the triangle inequality, we arrive at

$$\begin{aligned}
\mathbf{E}_{P_0^{(n)}}[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)] &\geq 0.5(\sigma+\varepsilon) - \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n \big(X_{i,1} - \mathbf{E}[X_{i,1}]\big)\right|\right] \\
&\geq 0.5(\sigma+\varepsilon) - \left(\mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n \big(X_{i,1} - \mathbf{E}[X_{i,1}]\big)\right|^2\right]\right)^{1/2} \\
&\geq 0.5(\sigma+\varepsilon) - 0.5/\sqrt{n} \\
&\geq \big(\sigma + \varepsilon + 1/\sqrt{n}\big)/6.
\end{aligned}$$

To get the second line above, we used that the first-order moment is bounded by the second-order moment. In the third line, we used that the variance of the sum of independent random variables is the sum of variances and that the variance of a random variables taking its values in $[0, 1]$ is always $\leq 1/4$. Finally, the last line is derived from the assumption $\sigma + \varepsilon \geq 2/\sqrt{n}$.

We now turn to the case $\sigma + \varepsilon \leq 2/\sqrt{n}$. In this case, we use the same distribution $P_0^{(n)}$ as in the previous case but we choose $\sigma = \varepsilon = 0$. From (10) we derive that

$$\begin{aligned}
\mathbf{E}_{P_0^{(n)}}\big[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)\big] &\geq \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^n \big(X_{i,1} - \mathbf{E}[X_{i,1}]\big)\right|\right] \\
&\geq 0.5\,\mathbf{E}_{U_i \overset{\text{iid}}{\sim} \mathcal{U}_1}\left[\left|\frac{1}{n}\sum_{i=1}^n (U_i - 0.5)\right|\right] \geq 0.105/\sqrt{n}
\end{aligned}$$

In view of the assumption $\sigma + \varepsilon \leq 2/\sqrt{n}$, this leads to

$$\begin{aligned}\mathbf{E}_{P_0^{(n)}}\left[\mathsf{d}_{\mathcal{F}}(\widehat{P}_n, P^*)\right] &\geq \mathbf{E}_{P_0^{(n)}}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(X_{i,1} - \mathbf{E}[X_{i,1}]\right)\right|\right] \\ &\geq 0.5\,\mathbf{E}_{U_i \overset{\text{iid}}{\sim} \mathcal{U}_1}\left[\left|\frac{1}{n}\sum_{i=1}^{n}(U_i - 0.5)\right|\right] \geq 0.035(\sigma + \varepsilon + 1/\sqrt{n}),\end{aligned}$$

which completes the proof of the first inequality of the theorem. For the second inequality, it suffices to combine the first inequality with the lower bound established in the seminal paper Dudley (1969).

7.5. Proof of the lower bound in Theorem 4

We split the proof of Theorem 4 into two propositions: The first one shows the tightness of the dependence on the contamination rate whereas the second one establishes the tightness of the dependence on the noise-level.

Proposition 1 (Tightness wrt to the contamination rate) *Under the assumptions of Theorem 4,*

$$\inf_{\widehat{g}_n}\sup_{g^*}\sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)} \mathbf{E}[R_{\mathsf{d}_{\mathcal{F}},P^*}(\widehat{g}_n)] \geq \varepsilon/3.$$

Proof It can be easily checked that the supremum of the expected risk over $\mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)$ is always not smaller than the supremum of the same quantity over $\mathcal{P}_{n,1}(d,0,\varepsilon,g^*)$. To ease notation, we write $\mathcal{P}_n(d,\varepsilon,g^*) = \mathcal{P}_{n,1}(d,0,\varepsilon,g^*)$ and also set $\mu = \mathcal{U}_d$.

Step 1: Reduction to Huber contamination model. Note that the set of admissible data distributions $\mathcal{P}_{n,D}(d,\varepsilon,g^*)$ comprises the data distributions from Huber's deterministic contamination model (Bateni and Dalalyan, 2020, Section 2.2), namely data distributions such that a (deterministic) proportion $(1-\varepsilon)$ of the data is distributed according to a reference distribution P^* while the remaining proportion ε is independently drawn from another distribution Q. Therefore, denoting by $\mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)$ such distributions, it holds, for any estimator $\widehat{g}_n$ and generator g^*,

$$\begin{aligned}\sup_{P^{(n)} \in \mathcal{P}_n(d,\varepsilon,g^*)} \mathbf{E}[R_{\mathsf{d}_{\mathcal{F}},P^*}(\widehat{g}_n)] &= \sup_{P^{(n)} \in \mathcal{P}_n(d,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu, \widehat{g}_n\sharp\mu)] \\ &\geq \sup_{P^{(n)} \in \mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu, \widehat{g}_n\sharp\mu)].\end{aligned}$$

Furthermore, let us denote by $\mathcal{P}_D^{\mathrm{HC}}(d,\varepsilon,g^*)$ the set of data distributions such that there is a distribution Q defined on the same space as a reference distribution $P^* = g^*\sharp\mu$ such that the observations $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_n$ are independent and drawn from the mixture distribution $(1-\varepsilon)P^* + \varepsilon Q$. In view of $\sup_{g,g'} d_{\mathcal{F}}(g\sharp\mu, g'\sharp\mu) \leq L$ and (Bateni and Dalalyan, 2020, Proposition 1), for any estimator $\widehat{g}_n$ and generator g^*, we have

$$\sup_{P^{(n)} \in \mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu, \widehat{g}_n\sharp\mu)] \geq \sup_{P^{(n)} \in \mathcal{P}_n^{\mathrm{HC}}(d,\varepsilon/2,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^*\sharp\mu, \widehat{g}_n\sharp\mu)] - e^{-n\varepsilon/6}L.$$

The second step consists in lower bounding the risk in the Huber contamination model using an argument based on two simple hypotheses.

Step 2: Construction of hypotheses. Let us define the generators $g_1^*, g_2^* : [0,1]^d \to [0,1]$ as

$$g_1^*(\boldsymbol{u}) = (1-\varepsilon)u_1 \quad \text{and} \quad g_2^*(\boldsymbol{u}) = (1-\varepsilon)u_1 + \varepsilon, \quad \text{for } \boldsymbol{u} = (u_1, \dots, u_d) \in [0,1]^d.$$

For contamination distributions $Q_1 := \mathcal{U}([1-\varepsilon, 1])$ and $Q_2 := \mathcal{U}([0, \varepsilon])$, define the data generating distributions

$$P_1^{(n)} = [(1-\varepsilon)g_1^* \sharp \mu + \varepsilon Q_1]^{\otimes n} \text{ and } P_2^{(n)} = [(1-\varepsilon)g_2^* \sharp \mu + \varepsilon Q_2]^{\otimes n}.$$

One can easily check that $P_1^{(n)} = P_2^{(n)} = \mathcal{U}([0,1])^{\otimes n}$ and $P_j^{(n)} \in \mathcal{P}_n^{\mathrm{HC}}(d, \varepsilon, g_j^*)$ for $j = 1, 2$. Using the fact that the maximum is larger than the arithmetic mean, in conjunction with the triangular inequality, we obtain

$$\begin{aligned}
\sup_{g^*} \sup_{P^{(n)} \in \mathcal{P}_n^{\mathrm{HC}}(d,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] &\geq \frac{1}{2}\left[\mathbf{E}_{P_1^{(n)}} \mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, \widehat{g}_n \sharp \mu) + \mathbf{E}_{P_2^{(n)}} \mathsf{d}_{\mathcal{F}}(g_2^* \sharp \mu, \widehat{g}_n \sharp \mu)\right] \\
&= \frac{1}{2}\mathbf{E}_{P_0^{(n)}}\left[\mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, \widehat{g}_n \sharp \mu) + \mathsf{d}_{\mathcal{F}}(g_2^* \sharp \mu, \widehat{g}_n \sharp \mu)\right] \\
&\geq \frac{1}{2}\mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, g_2^* \sharp \mu) \geq \varepsilon/2.
\end{aligned}$$

The last inequality comes from choosing the representer $f(\boldsymbol{u}) = u_1$ from $\mathcal{F}$.

Conclusion. Combining the previous two steps, we get

$$\sup_{P^{(n)} \in \mathcal{P}_n^{\mathrm{HDC}}(d,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] \geq (1/4)\varepsilon - e^{-n\varepsilon/6}L.$$

Choosing $n \geq (6/\varepsilon)\log(20L/\varepsilon)$, we get the claim of the proposition. ■

Proposition 2 (Tightness wrt to the noise level) *Under the assumptions of Theorem 4, we have*

$$\inf_{\widehat{g}_n \in \mathcal{G}} \sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_{n,D}(d,\sigma,\varepsilon,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] \geq \sigma/2.$$

Proof Once again, without loss of generality we assume that $D = 1$, $\varepsilon = 0$ and drop the dependence of different quantities on these two parameters. Recall that $\mu = \mathcal{U}_d$. Let us define the generators $g_j^* : [0,1]^d \to [0,1]^D$, $j = 1, 2$, by

$$g_1^*(\boldsymbol{u}) \equiv 0 \qquad \text{and} \qquad g_2^*(\boldsymbol{u}) \equiv \sigma, \qquad \text{for } \boldsymbol{u} = (u_1, \dots, u_d) \in [0,1]^d.$$

These functions allow us to define the data generating distributions

$$P_1^{(n)} = [g_1^* \sharp \mu * \delta_\sigma]^{\otimes n} \qquad \text{and} \qquad P_2^{(n)} = [g_2^* \sharp \mu * \delta_0]^{\otimes n}.$$

One can easily check that $P_1^{(n)} = P_2^{(n)} = \delta_\sigma^{\otimes n}$, which belongs to $\mathcal{P}_n(d, \sigma, g_1^*) \cap \mathcal{P}_n(d, \sigma, g_2^*)$. Furthermore, $g_j^* \in \mathcal{G}^*$ for $j = 1, 2$ since the latter contains all the affine functions. Using the same arguments as in the proof of the previous proposition, we arrive at

$$\begin{aligned}
\sup_{g^* \in \mathcal{G}^*} \sup_{P^{(n)} \in \mathcal{P}_n(d,\sigma,g^*)} \mathbf{E}[\mathsf{d}_{\mathcal{F}}(g^* \sharp \mu, \widehat{g}_n \sharp \mu)] &\geq \frac{1}{2}\mathbf{E}_{P_1^{(n)}}\left[\mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, \widehat{g}_n \sharp \mu) + \mathsf{d}_{\mathcal{F}}(g_2^* \sharp \mu, \widehat{g}_n \sharp \mu)\right] \\
&\geq \frac{1}{2}\mathsf{d}_{\mathcal{F}}(g_1^* \sharp \mu, g_2^* \sharp \mu) \geq \sigma/2.
\end{aligned}$$

This completes the proof of the proposition. ■

To get the claim of Theorem 4, it suffices to combine the claims of the last two propositions with the fact that $(0.2\varepsilon \vee 0.5\sigma) \geq 0.1(\varepsilon + \sigma)$.

Acknowledgments

This work was partially supported by the grant Investissements d'Avenir (ANR11-IDEX-0003/Labex Ecodec/ANR-11-LABX-0047).

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of Machine Learning Research*, volume 70, pages 214–223. PMLR, 2017.

Amir-Hossein Bateni and Arnak S. Dalalyan. Confidence regions and minimax rates in outlier-robust estimation on the probability simplex. *Electron. J. Statist.*, 14(2):2653–2677, 2020. doi: 10.1214/20-EJS1731. URL https://doi.org/10.1214/20-EJS1731.

Gérard Biau, Benoît Cadre, Maxime Sangnier, and Ugo Tanielian. Some theoretical properties of GANs. *Annals of Statistics*, 48(3):1539–1566, 2020a.

Gérard Biau, Maxime Sangnier, and Ugo Tanielian. Some theoretical insights into Wasserstein GANs. *arXiv preprint arXiv:2006.02682*, 2020b.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations*, 2018.

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

Jose A Costa and Alfred O Hero. Learning intrinsic dimension and intrinsic entropy of high-dimensional datasets. In *2004 12th European Signal Processing Conference*, pages 369–372. IEEE, 2004.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013.

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

L.E. Fraenkel. Formulae for high derivatives of composite functions. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 83(2), pages 159–165. Cambridge University Press, 1978.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. In *International Conference on Artificial Intelligence and Statistics*, pages 1608–1617, 2018.

Aude Genevay, Lénaic Chizat, Francis Bach, Marco Cuturi, and Gabriel Peyré. Sample complexity of Sinkhorn divergences. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1574–1583. PMLR, 2019.

Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

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

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

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

Yann LeCun. The MNIST database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Jing Lei. Convergence and concentration of empirical measures under Wasserstein distance in unbounded functional spaces. *Bernoulli*, 26(1):767–798, 2020.

Elizaveta Levina and Peter J Bickel. Maximum likelihood estimation of intrinsic dimension. In *Advances in neural information processing systems*, pages 777–784, 2005.

Tengyuan Liang. On how well generative adversarial networks learn densities: Nonparametric and parametric results. *arXiv preprint arXiv:1811.03179*, 2018.

Tengyuan Liang. Estimating certain integral probability metric (IPM) is as hard as estimating under the IPM. *arXiv preprint arXiv:1911.00730*, 2019.

Tengyuan Liang and James Stokes. Interaction matters: A note on non-asymptotic local convergence of generative adversarial networks. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 907–915, 2019.

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

Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *arXiv preprint arXiv:1610.03483*, 2016.

Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

Vaishnavh Nagarajan and J Zico Kolter. Gradient descent GAN optimization is locally stable. In *Advances in neural information processing systems*, pages 5585–5595, 2017.

Weili Nie and Ankit B Patel. Towards a better understanding and regularization of GAN training dynamics. In *Uncertainty in Artificial Intelligence*, pages 281–291. PMLR, 2020.

Jonathan Niles-Weed and Philippe Rigollet. Estimation of Wasserstein distances in the spiked transport model. *arXiv preprint arXiv:1909.07513*, 2019.

David Pfau and Oriol Vinyals. Connecting generative adversarial networks and actor-critic methods. *arXiv preprint arXiv:1610.01945*, 2016.

Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in neural information processing systems*, pages 2234–2242, 2016.

Nicolas Schreuder. Bounding the expectation of the supremum of empirical processes indexed by Hölder classes. *arXiv preprint arXiv:2003.13530*, 2020.

Mohamed El Amine Seddik, Cosme Louart, Mohamed Tamaazousti, and Romain Couillet. Random matrix theory proves that deep learning representations of GAN-data behave as gaussian mixtures. *arXiv preprint arXiv:2001.08370*, 2020.

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

Shashank Singh, Ananya Uppal, Boyue Li, Chun-Liang Li, Manzil Zaheer, and Barnabás Póczos. Nonparametric density estimation with adversarial losses. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 10246–10257. Curran Associates Inc., 2018.

Bharath K Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

A. Uppal, S. Singh, and B. Poczos. Nonparametric density estimation: Convergence rates for GANs under Besov IPM losses. In *Advances in Neural Information Processing Systems 32*, pages 9089–9100. Curran Associates, Inc., 2019.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

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

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