Bergische Universität Wuppertal

Fakultät für Mathematik und Naturwissenschaften

Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)

Preprint BUW-IMACM 20/56

H. Asatryan, H. Gottschalk, M. Lippert and M. Rottmann

A Convenient Infinite Dimensional Framework for Generative Adversarial Learning

November 24, 2020

http://www.imacm.uni-wuppertal.de

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

A Convenient Infinite Dimensional Framework for Generative Adversarial Learning

Hayk Asatryan*[1,2,3], Hanno Gottschalk†[1,2,3], Marieke Lippert‡[1] and Matthias Rottmann§[1,2,3]

[1]School of Mathematics and Natural Sciences, University of Wuppertal, Germany
[2]Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)
[3]Interdisciplinary Center for Machine Learning and Data Analytics (IZMD)

November 24, 2020

Abstract. In recent years, generative adversarial networks (GANs) have demonstrated impressive experimental results while there are only a few works that foster statistical learning theory for GANs. In this work, we propose an infinite dimensional theoretical framework for generative adversarial learning. Assuming the class of uniformly bounded k-times α-Hölder differentiable ($C^{k,\alpha}$) and uniformly positive densities, we show that the Rosenblatt transformation induces an optimal generator, which is realizable in the hypothesis space of $C^{k,\alpha}$ generators. With a consistent definition of the hypothesis space of discriminators, we further show that in our framework the Jensen-Shannon (JS) divergence between the distribution induced by the generator from the adversarial learning procedure and the data generating distribution converges to zero. As our convenient framework avoids modeling errors both for generators and discriminators, by the error decomposition for adversarial learning it suffices that the sampling error vanishes in the large sample limit. To prove this, we endow the hypothesis spaces of generators and discriminators with $C^{k,\alpha'}$-topologies, $0 < \alpha' < \alpha$, which render the hypothesis spaces to compact topological spaces such that the uniform law of large numbers can be applied. Under sufficiently strict regularity assumptions on the density of the data generating process, we also provide rates of convergence based on concentration and chaining. To this avail, we first prove subgaussian properties of the empirical process indexed by generators and discriminators. Furthermore, as covering numbers of bounded sets in $C^{k,\alpha}$-Hölder spaces with respect to the L^∞-norm lead to a convergent metric entropy integral if k is sufficiently large, we obtain a finite constant in Dudley's inequality. This, in combination with McDiarmid's inequality, provides explicit rate estimates for the convergence of the GAN learner to the true probability distribution in JS divergence.

Keywords: Generative Adversarial Learning • Inverse Rosenblatt Transformation • Statistical Learning Theory • Chaining • Covering Numbers for Hölder Spaces

MSC 2020: 62G20, 68T05

*asatryan@uni-wuppertal.de
†hanno.gottschalk@uni-wuppertal.de
‡marieke.lippert@uni-wuppertal.de
§rottmann@math.uni-wuppertal.de

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

1 Introduction

Generative learning aims at modelling the distribution of a data generating process and sampling from it. This desire is at least as old as Markov Chain Monte Carlo (MCMC) methods [14] and has born several different types of models and methods such as, to name only a few of them, hidden Markov models (HMM), Gaussian mixture models (GMM), Boltzmann machines (BM), principal component analysis (PCA), autoencoders (AE) and generative adversarial networks (GAN), see [9, 1, 12]. Most of these models either aim at dimensionality reduction (PCA, AE) but do not allow for sampling, or allow for sampling but suffer from the curse of dimensionality (GMM, BM) or allow sampling in high dimension but require an explicit density that has to be constructed or learned beforehand (MCMC). However, this changed with the advent of GANs [12] that can be classified under the more general notion of adversarial learning. Eversince, GANs have demonstrated remarkable capabilities in the domain of generative modeling, in particular, with application to image synthesis [20, 16, 6, 31].

In the typical adversarial learning framework, a generator is supposed to map examples from a noise distribution, say the uniform distribution over a compact set, to examples of a desired distribution, a sample of real / original data of which is available. Mathematically this corresponds to a pushforward or image measure of the noise measure under the generator map. This map is learned in a two player setting: A discriminator is trained to distinguish between the real examples from a data set and the fake ones, i.e., those obtained by the generator. On the other hand, the generator is trained to fool the discriminator. Learning both the discriminator and the generator alternatingly / simultaneously leads to a generator producing ever more realistic examples that approximately follow the distribution the data stems from.

GANs utilize deep neural networks (DNNs) as generators and discriminators. They allow for sampling from a model of the data generating distribution without the necessity of explicitly modelling densities (such as MCMC, GMM), without critical slowing down of learning (MCMC, BM) and barely suffering from the curse of dimensionality.

Many efforts have been made to empirically improve the GAN framework [20, 16, 6] for image synthesis. A remarkable improvement was obtained by the so-called Cycle-GAN framework [31]. Therein, not only the generator but also its inverse is learned (with a corresponding second discriminator). Impressive empirical results are presented, e.g., for transforming horses into zebras and vice versa. Therefore, GANs have also demonstrated to cope with higher resolution image data.

While the loss functions used in many works aim at reducing the Jensen-Shannon (JS) divergence between the true distribution and the class of generated (or parametrized) ones, other loss functions aiming at reducing other distance metrics have also been proposed, e.g., Wasserstein-GANs [2] that aim at reducing the Wasserstein distance. Wasserstein-GANs may in some cases yield better empirical convergence properties, where ordinary GANs suffer from so-called mode collapsing.

In contrast to the rapidly progressing developments on the empirical side, theoretical results on generative adversarial learning[1] remained unexamined until the publication [3]. That work studies the connection between the adversarial principle of generative adversarial learning and the Jensen-Shannon divergence in a framework with a finite dimensional hypothesis space. This includes existence and uniqueness arguments for the optimal generator. While the discriminator is modelled in a rather abstract way, the authors also provide approximation arguments. In

[1]Here we use the term genartive adversarial learning in contrast to genarative adversarial networks if we do not refer to neural networks as models, specifically.

particular, they provide large sample theory for generative adversarial learning as well as a central limit theorem. However this is restricted to finite dimensional parametric models.

Further theoretical work on generative adversarial learning that was recently published includes theory on Wasserstein methods [4], as well as theory on domain shifts quantified by means of an adversarial loss that reduces Jensen-Shannon divergence [26].

In this work we extend the approach of [3] to an infinite dimensional setting where the generators are k-times differentiable α-Hölder ($C^{k,\alpha}$-) functions defined on $[0,1]^d$ (e.g., the input space of images with color intensities in $[0,1]$ for image generation tasks). The discriminator space is chosen consistently, such that optimal discriminators always exist. This enables us to prove the existence of an optimal generator under quite general assumptions on the probability density of the 'true' data generating process on Borel sets of $[0,1]^d$. This convenient, but not very restrictive realizability property of our framework avoids the use of approximations as in [3]. We achieve this proving the Rosenblatt transformation [21] to be in the hypothesis space for the generators consisting of bounded, invertible $C^{k,\alpha}$-functions. In this way, we can estimate the error, measured in terms of the Jensen-Shannon divergence, between the probability distribution from the generative adversarial learning process and the 'true' distribution by the sampling error, i.e., the supremum over the empirical process over the product of the hypothesis spaces of generators and discriminators [24].

Statistical learning theory to a large extent depends on compactness properties of the hypothesis space. In infinite dimensional statistics, where the hypotheses are parametrized by functions in bounded regions of some – say Banach – function space [28, 11, 13], the use of the Banach topology is prohibited by Riesz's theorem which characterizes locally compact Banach spaces (see, e.g., [7, 22]). Therefore, compact embeddings into spaces with weaker topology play a crucial role in infinite dimensional statistics. In our convenient framework, we obtain such embeddings from the embedding of bounded $C^{k,\alpha}$-functions into $C^{k,\alpha'}$-functions for $0 < \alpha' < \alpha < 1$ (see [10]). With the aid of the uniform law of large numbers over compact spaces [8], we can thus conclude that in this setting of adversarial learning, generators rendering the true distribution in the large sample limit can always be learned.

Under stronger regularity assumptions, this statement can also be made quantitative in order to prove explicit rates of convergence. The key observations are that (a) for the embedding of bounded $C^{k,\alpha}$-functions into $L^\infty\left([0,1]^d\right)$ covering numbers are explicitly known and they allow a convergent metric entropy integral in Dudley's inequality, provided the regularity defined by $k+\alpha$ being sufficiently high, and (b) the fact that the empirical process defined by the empirical loss function is a subgaussian process with respect to the $\|\cdot\|_\infty$-norm. Thus explicit estimates for the supremum of the empirical process / the sampling error can be obtained via chaining and concentration (see [28, 17]). These rates, in contrast to [3], do not depend on the dimension of the hypothesis space (which is infinite in our setting). We also give an argument how to adaptively extend the hypothesis space with the sample size in order to eliminate certain assumptions on the regularity of the probability density of the data generating process and achieve almost sure convergence of the generative adversarial learner in JS-distance, also for this case.

Outline. This paper is organized as follows: In Section 2 we state some helpful notions of probability theory and functional analysis including Hölder spaces. Thereafter in Section 3, we introduce the hypothesis class of k-times α-Hödler differentiable generators with the properties outlined above and prove that if the 'true' data generating process has a nonnegative $C^{k,\alpha}$-density function, the optimal generator given by the Rosenblatt transformation is contained in

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

the hypothesis space. Furthermore, we introduce the consistent class of discriminators that also contains the optimal one. In Section 4 we prove the uniform convergence of the empirical loss to its expected value, which implies that the sampling error vanishes in the limit. We also prove the convergence of the probability distribution generated by the empirical risk minimizer in the min-max problem from the adversarial learning setting, to the data generating distribution in Jensen-Shannon divergence. Under stronger regularity assumptions on the 'true' density, we give explicit rate estimates based on covering numbers for the embedding of the hypothesis space into $L^\infty\left([0,1]^d\right)$ in Section 5, which is used to eliminate certain regularity assumptions. We give a summary and outline future research directions in the final Section 6.

2 Preliminaries

In this section we adopt some notions of functional analysis and probability theory, which will be used throughout the rest of the present work. For the sake of completeness, we also give some proofs.

2.1 Hölder Spaces and Their Properties

For an open subset $U \subset \mathbb{R}^{d_1}$ and for a nonnegative integer k (or $k = \infty$), $C^k\left(U, \mathbb{R}^{d_2}\right)$ will stand for the set of all $\mathbb{R}^{d_2}$-valued functions with continuous k-th order derivatives in U, and $C^k\left(\overline{U}, \mathbb{R}^{d_2}\right)$ will stand for the set of all $\mathbb{R}^{d_2}$-valued functions whose k-th derivatives have continuous extensions to $\overline{U}$ (or, equivalently, the k-th derivatives are uniformly continuous on U). For a bounded U, $C^k\left(\overline{U}, \mathbb{R}^{d_2}\right)$ is a Banach space with respect to the norm

$$\|f\|_{C^k\left(\overline{U},\mathbb{R}^{d_2}\right)} \coloneqq \sum_{|n|\leq k} \sup_{x\in U} |D^n f(x)|\,;$$

here $Df(x)$ stands for the Jacobi-Matrix of f at x, hence $D^n f(x)$ is a tensor of level n containing all (mixed) partial derivatives of n-th order. The space of continuous functions C^0 will be denoted by C, as usual. The definition and many of the basic properties of Hölder spaces of real-valued functions, given in [10], extend without difficulty to vector-valued ones:

Definition 2.1. Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let $0 < \alpha \leq 1$, and let k be a nonnegative integer.

1. For $f \in C^k\left(U, \mathbb{R}^{d_2}\right)$ put

$$[f]_{k,\alpha,U} \coloneqq \max_{|n|=k} \sup_{\substack{x,y\in U\\ x\neq y}} \frac{|D^n f(x) - D^n f(y)|}{|x-y|^\alpha}.$$

   Denote

$$C^{k,\alpha}\left(U, \mathbb{R}^{d_2}\right) \coloneqq \left\{ f \in C^k\left(U, \mathbb{R}^{d_2}\right) : [f]_{k,\alpha,U} < \infty \right\}.$$

2. The Hölder space $C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)$ consists of all functions $f \in C^k\left(\overline{U}, \mathbb{R}^{d_2}\right)$ for which the norm

$$\|f\|_{C^{k,\alpha}\left(\overline{U},\mathbb{R}^{d_2}\right)} \coloneqq \|f\|_{C^k\left(\overline{U},\mathbb{R}^{d_2}\right)} + [f]_{k,\alpha,U}$$

   is finite.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

It is easy to see that $C^{k,\alpha}\left(\overline{U},\mathbb{R}^{d_2}\right)$ is a Banach space (see [10, Sec. 4.1]). The members of $C^{k,\alpha}\left(\overline{U},\mathbb{R}^{d_2}\right)$ will be referred to as *k-times α-Hölder differentiable* functions on the set $\overline{U}$ (for a fixed α, we shall sometimes say Hölder differentiable rather than α-Hölder differentiable). Note that

$$C^{k,\alpha_2}\left(\overline{U},\mathbb{R}^{d_2}\right)\subset C^{k,\alpha_1}\left(\overline{U},\mathbb{R}^{d_2}\right)\quad\left(k\geq 0,\ 0<\alpha_1<\alpha_2\leq 1\right);\tag{2.1}$$

moreover, this embedding is compact, i.e., every bounded subset of $C^{k,\alpha_2}\left(\overline{U},\mathbb{R}^{d_2}\right)$ is relatively compact in $C^{k,\alpha_1}\left(\overline{U},\mathbb{R}^{d_2}\right)$ (see, e.g., [10, Lemma 6.33]).

The special case $\alpha=1$ is of particular interest; the space $C^{k,1}\left(\overline{U},\mathbb{R}^{d_2}\right)$ will be called the Lipschitz space and its members will be referred to as *k-times Lipschitz differentiable* functions on $\overline{U}$. The mean value theorem for vector-valued functions (see, e.g., [23, Th. 9.19]) shows that $C^{1}\left(\overline{U},\mathbb{R}^{d_2}\right)\subset C^{0,1}\left(\overline{U},\mathbb{R}^{d_2}\right)$, hence

$$C^{k}\left(\overline{U},\mathbb{R}^{d_2}\right)\subset C^{k-1,1}\left(\overline{U},\mathbb{R}^{d_2}\right)\quad\left(k\geq 1\right).\tag{2.2}$$

Moreover, the embedding (2.2) is continuous; in fact,

$$\|f\|_{C^{k-1,1}\left(\overline{U},\mathbb{R}^{d_2}\right)}\leq\|f\|_{C^{k}\left(\overline{U},\mathbb{R}^{d_2}\right)}\quad\left(f\in C^{k}\left(\overline{U},\mathbb{R}^{d_2}\right)\right).$$

Let $M_{d_2d_3}(\mathbb{R})$ stand for the set of all $d_2\times d_3$ matrices with real entries. We denote

$$C^{k,\alpha}\left(V,M_{d_2d_3}(\mathbb{R})\right):=\left\{\left(A_{ij}\right)_{d_2\times d_3}:V\to M_{d_2d_3}(\mathbb{R})\,\middle|\,A_{ij}\in C^{k,\alpha}\left(V,\mathbb{R}\right)\text{ for all }i,j\right\}$$

with $V\in\{U,\overline{U}\}$. In the special case $d_2=1$ it is easily seen that the product of Hölder continuous real-valued functions on $\overline{U}$ is again Hölder continuous (see, e.g., [10, Sec. 4.1]), hence the product of k-times Hölder differentiable real-valued functions on $\overline{U}$ is again k-times Hölder differentiable there. Moreover, the multiplication is continuous in $C^{k,\alpha}\left(\overline{U},\mathbb{R}\right)$-norm. Obviously, these assertions remain true for matrix-functions. Moreover, under additional assumptions, the quotient of k-times Hölder differentiable real-valued functions also possesses similar properties:

Lemma 2.2. *The following statements are true:*

(a) If $u,v\in C^{k,\alpha}\left(\overline{U},\mathbb{R}\right)$ and $\inf_{x\in U}|v(x)|>0$, then $\dfrac{u}{v}\in C^{k,\alpha}\left(\overline{U},\mathbb{R}\right)$.

(b) If $u_n\to u, v_n\to v$ in $C^{k,\alpha}\left(\overline{U},\mathbb{R}\right)$-norm and if $\inf_{x\in U,n\in\mathbb{N}}|v_n(x)|>0$, then $\frac{u_n}{v_n}\to\frac{u}{v}$ in $C^{k,\alpha}\left(\overline{U},\mathbb{R}\right)$-norm.

Proof. (a) It is enough to show that $w:=\frac{1}{v}\in C^{k,\alpha}\left(\overline{U},\mathbb{R}\right)$. This is easily seen to be true in the case $k=0$. Now assume the statement is true for $k=m-1$. If $v\in C^{m,\alpha}\left(\overline{U},\mathbb{R}\right)$, then

$$\frac{\partial v}{\partial x_j}\in C^{m-1,\alpha}\left(\overline{U},\mathbb{R}\right)\quad\left(1\leq j\leq d_1\right),$$

which, together with the equality

$$\frac{\partial w}{\partial x_j}=-v^{-2}\cdot\frac{\partial v}{\partial x_j},\tag{2.3}$$

implies that $\frac{\partial w}{\partial x_j}\in C^{m-1,\alpha}\left(\overline{U},\mathbb{R}\right)\ \left(1\leq j\leq d_1\right)$. Thus, $w\in C^{m,\alpha}\left(\overline{U},\mathbb{R}\right)$.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

(b) Similar to the proof of (a), it is enough to show that $\frac{1}{v_n} \to \frac{1}{v}$ in $C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)$-norm. Put $w = \frac{1}{v}, w_n = \frac{1}{v_n}$ $(n = 1, 2, \ldots)$. Again the case $k = 0$ can be easily verified. Assume the statement is true for $k = m - 1$. If $v_n \to v$ in $C^{m,\alpha}\left(\overline{U}, \mathbb{R}\right)$-norm, then $\frac{\partial v_n}{\partial x_j} \to \frac{\partial v}{\partial x_j}$ $(1 \le j \le d_1)$ and $v_n^{-2} \to v^{-2}$ in $C^{m-1,\alpha}\left(\overline{U}, \mathbb{R}\right)$-norm, hence the equality

$$\frac{\partial w_n}{\partial x_j} = -v_n^{-2} \cdot \frac{\partial v_n}{\partial x_j},$$

together with (2.3), yields the convergence $\frac{\partial w_n}{\partial x_j} \to \frac{\partial w}{\partial x_j}$ $(1 \le j \le d_1)$ in $C^{m-1,\alpha}\left(\overline{U}, \mathbb{R}\right)$-norm, which in turn implies that $w_n \to w$ in $C^{m,\alpha}\left(\overline{U}, \mathbb{R}\right)$-norm. □

More interesting is the fact that compositions of k-times Hölder differentiable functions on $\overline{U}$ are k-times Hölder differentiable:

Lemma 2.3. *Let $U_1 \subset \mathbb{R}^{d_1}$ and $U_2 \subset \mathbb{R}^{d_2}$ be bounded open sets, let k be a positive integer, and let $0 < \alpha \le 1$.*

(a) If $f \in C^{k,\alpha}\left(\overline{U}_1, \mathbb{R}^{d_2}\right), f : U_1 \to U_2$ and $g \in C^{k,\alpha}\left(\overline{U}_2, \mathbb{R}^{d_3}\right)$, then $g \circ f \in C^{k,\alpha}\left(\overline{U}_1, \mathbb{R}^{d_3}\right)$.

(b) Let $f_n \in C^{k,\alpha}\left(\overline{U}_1, \mathbb{R}^{d_2}\right), f_n : U_1 \to U_2, g_n \in C^{k,\alpha}\left(\overline{U}_2, \mathbb{R}^{d_3}\right)$ $(n \in \mathbb{N})$, and let $f_n \to f, g_n \to g$ in $C^{k,\alpha}$-norm with f satisfying the condition $f(U_1) \subset U_2$. Then $g_n \circ f_n \to g \circ f$ in $C^{k,\alpha}$-norm.

Proof. (a) First, it is easy to see that if $f \in C^{0,\alpha_1}\left(\overline{U}_1, \mathbb{R}^{d_2}\right)$ and $g \in C^{0,\alpha_2}\left(\overline{U}_2, \mathbb{R}^{d_3}\right)$ for some $\alpha_1, \alpha_2 \in (0, 1]$, then $g \circ f \in C^{0,\alpha_1\alpha_2}\left(\overline{U}_1, \mathbb{R}^{d_3}\right)$. Next, if $k = 1$, then the inclusion $C^1\left(\overline{U}, \mathbb{R}^{d_2}\right) \subset C^{0,1}\left(\overline{U}_1, \mathbb{R}^{d_2}\right)$, together with the chain rule

$$D(g \circ f)(x) = (Dg \circ f)\, Df, \tag{2.4}$$

implies that $g \circ f$ and $D(g \circ f)$ are Hölder continuous on $\overline{U}$, hence $g \circ f \in C^{1,\alpha}\left(\overline{U}_1, \mathbb{R}^{d_3}\right)$.

Now assume that the statement is true for $k = m - 1$ with $m \ge 2$. If $f \in C^{k,\alpha}\left(\overline{U}_1, \mathbb{R}^{d_2}\right)$ and $g \in C^{k,\alpha}\left(\overline{U}_2, \mathbb{R}^{d_3}\right)$, then Df and Dg are $(k-1)$-times Hölder differentiable, hence the equality

$$D^k(g \circ f)(x) = D^{k-1}\left[(Dg \circ f)\, Df\right] \tag{2.5}$$

implies that $D^k(g \circ f)$ is Hölder continuous on $\overline{U}$.

(b) The idea of the proof of this part is similar to that of (a); it is performed by induction on k and uses the equalities (2.4), (2.5) (cf. the proof of the statement (b) of Lemma 2.2). □

In the considerations below we shall assume that $d_1 = d_2 = d$ and, for simplicity, we will write $C^{k,\alpha}\left(\overline{U}\right)$ instead of $C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)$. The following version of the inverse function theorem for Hölder differentiable maps will be needed below.

Theorem 2.4. *Let $U \subset \mathbb{R}^d$ be a bounded open set, let k be a positive integer, and let $0 < \alpha \le 1$.*

(a) If $\varphi : \overline{U} \to \overline{U}$ is a bijective and k-times α-Hölder differentiable function on $\overline{U}$ such that the Jacobian determinant J_φ satisfies the condition

$$\inf_{x \in U} |J_\varphi(x)| > 0, \tag{2.6}$$

then $\varphi^{-1} \in C^{k,\alpha}\left(\overline{U}\right)$ and

$$\left|\varphi^{-1}(y_1) - \varphi^{-1}(y_2)\right| \le \frac{d!\|\varphi\|^{d-1}_{C^1(\overline{U})}}{\inf\limits_{x \in U} |J_\varphi(x)|} |y_1 - y_2| \quad \left(y_1, y_2 \in \overline{U}\right). \tag{2.7}$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

(b) If $\varphi_n : \overline{U} \to \overline{U}$ $(n = 1, 2, \ldots)$ are bijective and k-times α-Hölder differentiable functions on $\overline{U}$ such that

$$\inf_{x \in U, n \in \mathbb{N}} |J_{\varphi_n}(x)| > 0$$

and $\varphi_n \to \varphi$ in $C^{k,\alpha}$-norm, then $\varphi : \overline{U} \to \overline{U}$ is bijective, $\varphi^{-1} \in C^{k,\alpha}\left(\overline{U}\right)$ and $\varphi_n^{-1} \to \varphi^{-1}$ in $C^{k,\alpha}$-norm.

Proof. (a) We first show that $(D\varphi)^{-1}$ is $(k-1)$-times Hölder differentiable on $\overline{U}$. Identifying $D\varphi$ with the Jacobian matrix of φ, we may write $(D\varphi)^{-1}$ in the form

$$(D\varphi)^{-1} = \frac{1}{J_\varphi} \operatorname{adj}(D\varphi),$$

where $\operatorname{adj}(D\varphi)$ is the adjugate matrix of the Jacobian matrix of φ. The entries of $\operatorname{adj}(D\varphi)$ are minors of $D\varphi$ with corresponding sign factors. Since the product of $(k-1)$-times Hölder differentiable functions on $\overline{U}$ is $(k-1)$-times Hölder differentiable, hence so is the matrix-function $\operatorname{adj}(D\varphi)$. In view of (2.6) and Lemma 2.2, we conclude that $(D\varphi)^{-1}$ is $(k-1)$-times Hölder differentiable on $\overline{U}$.

Since $\overline{U}$ is compact and $\varphi : \overline{U} \to \overline{U}$ is a continuous bijection, φ^{-1} is continuous on $\overline{U}$ (see, e.g., [7, Lemma I.5.8]). The inverse function theorem (see, e.g., [23, Th. 9.24] or [27, Th. 3.1]) implies that $\varphi^{-1} \in C^1(U)$. If we differentiate both sides of the relation $\varphi \circ \varphi^{-1} = I$ and apply the chain rule, we obtain

$$D\varphi^{-1} = (D\varphi)^{-1} \circ \varphi^{-1}. \tag{2.8}$$

Therefore $D\varphi^{-1}$ is continuous on $\overline{U}$, i.e., $\varphi^{-1} \in C^1\left(\overline{U}\right)$. Since $(D\varphi)^{-1}$ is $(k-1)$-times Hölder differentiable on $\overline{U}$ and φ^{-1} is continuously differentiable on $\overline{U}$, another application of (2.8) gives $\varphi^{-1} \in C^2\left(\overline{U}\right)$, etc. Thus, we arrive in a finite number of steps at the relation $\varphi^{-1} \in C^k\left(\overline{U}\right)$.

In view of (2.1) and (2.2), we have $\varphi^{-1} \in C^{k-1,\alpha}\left(\overline{U}\right)$. Thus, $(D\varphi)^{-1}$ and φ^{-1} are both $(k-1)$-times Hölder differentiable on $\overline{U}$, hence their composition also possesses that property, according to Lemma 2.3; therefore (2.8) implies that $\varphi^{-1} \in C^{k,\alpha}\left(\overline{U}\right)$.

To prove (2.7), we need an estimate for the operator norm of an invertible matrix. If $A = (a_{ij})_{d\times d} \in M_{d,d}(\mathbb{R})$ is invertible, then $A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$. The operator norm and the determinant of any matrix $B = (b_{ij})_{d\times d} \in M_{d,d}(\mathbb{R})$ can be easily estimated as

$$|B| \le d \max_{1\le i,j\le d} |b_{ij}| \tag{2.9}$$

and

$$|\det(B)| \le d! \left[\max_{1\le i,j\le d} |b_{ij}|\right]^d, \tag{2.10}$$

respectively. Therefore the absolute value of each entry of the matrix $\operatorname{adj}(A)$ does not exceed

$$(d-1)! \left[\max_{1\le i,j\le d} |a_{ij}|\right]^{d-1},$$

which, together with (2.9), yields

$$\left|A^{-1}\right| \le \frac{d! \left[\max_{1\le i,j\le d} |a_{ij}|\right]^{d-1}}{|\det(A)|}.$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Applying this estimate to $A = (D\varphi)(x)$ and using (2.8), we obtain

$$\left|D\varphi^{-1}(y)\right| \le \frac{d!\|\varphi\|_{C^1(\overline{U})}^{d-1}}{\inf\limits_{x\in U}|J_\varphi(x)|} \quad (y \in U).$$

The latter, combined with the mean value theorem for vector-valued functions, gives (2.7).

(b) The convergence $\|\varphi_n - \varphi\|_{C^{k,\alpha}} \to 0$ implies that $\|\varphi_n - \varphi\|_{C^1} \to 0$, hence

$$C := \frac{d!\sup\limits_{n\in\mathbb{N}}\|\varphi_n\|_{C^1(\overline{U})}^{d-1}}{\inf\limits_{x\in U, n\in\mathbb{N}}|J_{\varphi_n}(x)|} < \infty.$$

(2.7) gives

$$\left|\varphi_n^{-1}(y_1) - \varphi_n^{-1}(y_2)\right| \le C|y_1 - y_2| \quad \left(y_1, y_2 \in \overline{U}; n \in \mathbb{N}\right), \tag{2.11}$$

therefore

$$|x_1 - x_2| \le C|\varphi_n(x_1) - \varphi_n(x_2)| \quad \left(x_1, x_2 \in \overline{U}; n \in \mathbb{N}\right).$$

Letting $n \to \infty$, we obtain the inequality

$$|x_1 - x_2| \le C|\varphi(x_1) - \varphi(x_2)| \quad \left(x_1, x_2 \in \overline{U}\right)$$

which implies the injectivity of φ.

To prove the surjectivity of φ, choose any $y \in \overline{U}$. Since $\varphi_n\left(\overline{U}\right) = \overline{U}$, there exists $x_n \in \overline{U}$ such that $\varphi_n(x_n) = y$, for any $n \in \mathbb{N}$. The compactness of $\overline{U}$ guarantees the existence of a convergent subsequence $\{x_{n_i}\}_{i=1}^{\infty}$. Put $x := \lim\limits_{i\to\infty} x_{n_i}$. The convergence in $C^{k,\alpha}$-norm implies the uniform convergence on $\overline{U}$, hence $\varphi_{n_i}(x_{n_i}) - \varphi(x_{n_i}) \to 0$. Thus,

$$\varphi(x) = \varphi\left(\lim_{i\to\infty} x_{n_i}\right) = \lim_{i\to\infty}\varphi(x_{n_i}) = \lim_{i\to\infty}\left[\varphi(x_{n_i}) - \varphi_{n_i}(x_{n_i}) + y\right] = y.$$

Applying (a) to φ, we see that $\varphi^{-1} \in C^{k,\alpha}\left(\overline{U}\right)$.

Next, we shall prove that $\varphi_n^{-1} \to \varphi^{-1}$ in $C\left(\overline{U}\right)$-norm. To do that, it is enough to show that $\{\varphi_n^{-1}\}_{n=1}^{\infty}$ is relatively compact in $C\left(\overline{U}\right)$ and that φ^{-1} is the only possible accumulation point of $\{\varphi_n^{-1}\}_{n=1}^{\infty}$. Indeed, $\{\varphi_n^{-1}\}_{n=1}^{\infty}$ is uniformly bounded since $\varphi_n^{-1}\left(\overline{U}\right) \subset \overline{U}$ $(n \in \mathbb{N})$. Furthermore, the estimate (2.11) shows that $\{\varphi_n^{-1}\}_{n=1}^{\infty}$ is equicontinuous on $\overline{U}$. Hence the Arzela-Ascoli theorem implies the relative compactness of $\{\varphi_n^{-1}\}_{n=1}^{\infty}$. If ψ is an accumulation point for $\{\varphi_n^{-1}\}_{n=1}^{\infty}$, then there exists a subsequence $\left\{\varphi_{n_i}^{-1}\right\}_{i=1}^{\infty}$ such that $\varphi_{n_i}^{-1} \to \psi$ uniformly on $\overline{U}$. In equalities

$$\varphi_{n_i}^{-1} \circ \varphi_{n_i} = I, \quad \varphi_{n_i} \circ \varphi_{n_i}^{-1} = I$$

letting $i \to \infty$, we easily conclude that $\psi \circ \varphi = I$ and $\varphi \circ \psi = I$, therefore $\psi = \varphi^{-1}$.

Using (2.8), Lemma 2.2 and the established convergence $\varphi_n^{-1} \to \varphi^{-1}$ in $C\left(\overline{U}\right)$-norm, we see that $D\varphi_n^{-1} \to D\varphi^{-1}$ uniformly on $\overline{U}$, hence $\varphi_n^{-1} \to \varphi^{-1}$ in C^1-norm. The latter, together with (2.8) and Lemma 2.2, implies the convergence $\varphi_n^{-1} \to \varphi^{-1}$ in C^2-norm, etc. In a finite number of steps we obtain that $\varphi_n^{-1} \to \varphi^{-1}$ in C^k-norm. The latter, in view of continuity of the embedding (2.2), implies that $\varphi_n^{-1} \to \varphi^{-1}$ in $C^{k-1,1}$-norm. Using this convergence, (2.8) and the statement (b) of Lemma 2.3, we finally conclude that $\varphi_n^{-1} \to \varphi^{-1}$ in $C^{k,\alpha}$-norm. $\square$

An essential component in the proof of our main result (see Theorem 5.3) is the covering of the closed unit ball in the space $C^{k,\alpha}\left(\overline{U}\right)$ by balls of a small radius ε in the space $C\left(\overline{U}\right)$. Below we will give the relevant definitions and formulate the necessary estimate.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Definition 2.5. Let (M, ρ) be a metric space, let $T \subset M$, and let $\varepsilon > 0$.

(i) A subset N of M is called an *ε-net* (or ε-covering) for T if

$$T \subset \bigcup_{x \in N} B[x, \varepsilon]$$

(here $B[x, \varepsilon] \coloneqq \{y \in M : \rho(y, x) \le \varepsilon\}$ is the closed ball of radius ε and centered at x).

(ii) The *covering number* of T, denoted by $N(T, \rho, \varepsilon)$, is the minimum cardinality of an ε-net for T. The function $\varepsilon \mapsto \log N(T, \rho, \varepsilon)$ is the *metric entropy* of T.

Recall that a subset of a metric space is called *totally bounded* if it has a finite ε-net for every $\varepsilon > 0$. According to Hausdorff's theorem (see, e.g., [7, Lemma I.6.15]), a subset of a metric space is totally bounded iff it is Cauchy-precompact, i.e., every sequence of its points has a Cauchy subsequence. In particular, all the covering numbers of a compact set are finite.

Theorem 2.6. *If $U \subset \mathbb{R}^{d_1}$ is bounded, open and convex, then*

$$\log N\left(B_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)}[0,1], \|\cdot\|_{\infty}, \varepsilon\right) \le \frac{C_1 \left[\lambda^{(d_1)}\left(U^1\right)\right]^{d_2}}{\varepsilon^{\frac{d_1 d_2}{\alpha+k}}}, \tag{2.12}$$

where

$$U^1 \coloneqq \left\{x \in \mathbb{R}^{d_1} : \operatorname{dist}(x, U) < 1\right\}$$

with dist *being the euclidean distance, and C_1 is a constant depending only on d and $\alpha + k$.*

In the case $d_2 = 1$ the estimate (2.12) is proved in [28, Th. 2.7.1]; the general case follows without difficulty since $C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right) = \left[C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)\right]^{d_2}$.

2.2 Measure-Theoretic and Probabilistic Preliminaries

We recall the definition of the pushforward (image) measure (see, e.g., [5, Sec 3.6]).

Definition 2.7. Let $(\Omega, \mathscr{A})$ and $(\Omega', \mathscr{A}')$ be two measurable spaces, let $\mu : \mathscr{A} \to [0, \infty]$ be a positive measure on $\mathscr{A}$, and let $\varphi : \Omega \to \Omega'$ be a $(\mathscr{A}, \mathscr{A}')$-measurable function, i.e.,

$$\varphi^{-1}(\mathscr{A}') \subset \mathscr{A}.$$

The pushforward measure $\varphi_* \mu$ (also called the image of the measure μ under the mapping φ) is defined by the formula

$$\varphi_* \mu(A') \coloneqq \mu\left(\varphi^{-1}(A')\right) \quad (A' \in \mathscr{A}').$$

To measure the dissimilarity between two probability distributions, we will use the Kullback-Leibler and the Jensen-Shannon divergences (see [15, 3]). Later these divergences will be helpful to evaluate a learned measure imitating the given one.

Definition 2.8. Let ν and μ be two equivalent (i.e., absolutely continuous with respect to one another) probability measures on the measurable space $(\Omega, \mathscr{A})$. The *Kullback-Leibler divergence* (also called the relative entropy) $d_{KL}(\nu || \mu)$ is defined by the formula

$$d_{KL}(\nu || \mu) \coloneqq \int_{\Omega} \log \frac{d\nu}{d\mu} d\nu,$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

provided that the integral on the right-hand side exists (here $\frac{d\nu}{d\mu}$ is the Radon-Nikodym derivative).

Jensen's integral inequality shows that under the considered assumptions the integral $\int \log \frac{d\nu}{d\mu} d\nu$ exists (in the extended sense) and it is nonnegative. Moreover, $d_{KL}(\nu||\mu)$ vanishes iff $\nu = \mu$.

Since the Kullback-Leibler divergence is not symmetric in general, we will use another measure of dissimilarity between two probability distributions, called the Jensen-Shannon divergence.

Definition 2.9. Under the assumptions of the previous definition, the *Jensen-Shannon divergence* is defined by the formula

$$d_{JS}(\nu,\mu) := \frac{1}{2}\left[d_{KL}\left(\mu \left\| \frac{\mu+\nu}{2}\right.\right) + d_{KL}\left(\nu \left\| \frac{\mu+\nu}{2}\right.\right)\right].$$

If for a measure κ the Radon-Nikodym derivatives $m = \frac{d\nu}{d\kappa}$ and $n = \frac{d\mu}{d\kappa}$ exist, the Kullback-Leibler divergence can be written as

$$d_{KL}(\nu||\mu) = \int m \log \frac{m}{n} d\kappa =: d_{KL}(m||n).$$

Furthermore, the Jensen-Shannon-divergence can be written as

$$d_{JS}(\nu||\mu) = \frac{1}{2}\left[d_{KL}\left(m \left\| \frac{m+n}{2}\right.\right) + d_{KL}\left(n \left\| \frac{m+n}{2}\right.\right)\right] =: d_{JS}(m||n).$$

Below we will consider only measures which have a density. Throughout this work, $\lambda^{(d)}$ will denote the Lebesgue measure on $\mathbb{R}^d$.

Next, we recall the definition of the conditional density (see [25, Sec. 2.7]).

Let $Y_j : \Omega \to \mathbb{R}^{d_j}$ $(j = 1, 2)$ be random variables such that the pair $Y := (Y_1, Y_2)$ has a density $f_Y(y_1, y_2)$. Then Y_j has a marginal density given by the formula

$$f_{Y_j}(y_j) = \int_{\mathbb{R}^{d_{3-j}}} f_Y(y_1, y_2)\, dy_{3-j} \quad \left(y_j \in \mathbb{R}^{d_j}; j = 1, 2\right).$$

We define the conditional density $f_{Y_1|Y_2}(\cdot|y_2) = f_{Y_1|Y_2=y_2}(\cdot)$ by the formula

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \frac{f_Y(y_1, y_2)}{f_{Y_2}(y_2)} & \text{if} \quad f_{Y_2}(y_2) > 0, \\ 0 \quad \text{if} \quad f_{Y_2}(y_2) = 0. \end{cases} \tag{2.13}$$

The conditional (cumulative) distribution function $F_{Y_1|Y_2}(\cdot|y_2) = F_{Y_1|Y_2=y_2}(\cdot)$ is defined by the formula[2]

$$F_{Y_1|Y_2}(y_1|y_2) = \int_{-\infty}^{y_1} f_{Y_1|Y_2}(s|y_2)\, ds \quad \left(y_1 \in \mathbb{R}^{d_1}\right).$$

To estimate the sampling error in generative adversarial learning, some results on maxima of so-called subgaussian random processes will be needed. We give the necessary definitions.

[2] For $a_j = \left(a_j^1, \ldots, a_j^d\right)$ $(j = 1, 2)$ we put $\int_{a_1}^{a_2} = \int_{a_1^1}^{a_2^1} \cdots \int_{a_1^d}^{a_2^d}$ for brevity.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Definition 2.10. A random variable X is called σ^2-subgaussian if $\mathbb{E}\left[|X|\right] < \infty$ and

$$\mathbb{E}\left[e^{\lambda[X-\mathbb{E}[X]]}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}} \quad (\lambda \geq 0)$$

(the constant σ^2 is called a variance proxy).

The following lemma shows that the class of subgaussian random variables is quite large (we refer to [13] for a proof).

Lemma 2.11 (Hoeffding)**.** *If $a \leq X \leq b$ (a.s.) with some $a, b \in \mathbb{R}$, then X is $\frac{(b-a)^2}{4}$-subgaussian.*

Definition 2.12. A real valued random process $\{X_t\}_{t\in T}$ on a metric space (T, ρ) is called subgaussian if $\mathbb{E}\left[X_t\right] = 0 \ \ (t \in T)$ and the increments $X_t - X_s$ are $\left[\rho\left(t, s\right)\right]^2$-subgaussian.

Theorem 2.13 (Dudley)**.** *If $\{\pm X_t\}_{t\in T}$ are continuous subgaussian processes on the compact metric space (T, ρ), then the estimate*

$$\mathbb{E}\left[\sup_{t\in T}|X_t|\right] \leq 12 \int\limits_0^\infty \sqrt{\log N\left(T, \rho, \varepsilon\right)} d\varepsilon \tag{2.14}$$

holds.

The estimate (2.14) is often referred to as the *entropy bound.* The proof of (2.14) can be found in [13, Sec. 5.3].

The following concentration inequality due to C. McDiarmid (see [17]) will make it easy to pass from estimates for the expectation values of the maxima of random processes to estimates for the maxima of those processes themselves.

Theorem 2.14 (McDiarmid)**.** *Let $X = (X_1, X_2, \ldots, X_n)$ be a family of independent random variables with X_k taking values in a set A_k for each k. Suppose that $f : \prod\limits_{k=1}^n A_k \to \mathbb{R}$ is a function with the $(c_1, \ldots, c_n)$-bounded differences property: for each $k = 1, 2, \ldots, n$ and for any vectors $x, y \in \prod\limits_{k=1}^n A_k$ that differ only in the k-th coordinate, the inequality*

$$f\left(x\right) - f\left(y\right) \leq c_k$$

holds. Then

$$\mathbb{P}\left(f\left(X\right) - \mathbb{E}\left[f\left(X\right)\right] \geq t\right) \leq e^{-\frac{2t^2}{\sum\limits_{i=1}^n c_i^2}} \quad (t \geq 0)\,. \tag{2.15}$$

3 Generators and Discriminators

Let μ be a given probability measure on the measurable space $\left([0,1]^d, \mathscr{B}\left([0,1]^d\right)\right)$ with $\mathscr{B}\left([0,1]^d\right)$ the Borel-σ-algebra, and let $\lambda^{(d)}$ denote the restriction of the d-dimensional Lebesgue measure to $\left([0,1]^d, \mathscr{B}\left([0,1]^d\right)\right)$. The main objective of this section is to introduce a certain subset of $\left\{f : [0,1]^d \to [0,1]^d\right\}$, called the hypothesis space of generators, such that the pushforward measure $\mu_\varphi := \varphi_*\lambda^{(d)}$ has a density with regularity similar to those of the original measure μ, for each generator φ. Moreover, such a hypothesis space of generators possesses some useful compactness properties which will be stated below. To distinguish between the original and generated measures, we introduce a consistent hypothesis space of discriminators.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

3.1 The Set of Generators

Below we will mainly consider the Hölder space $C^{k,\alpha}\left(\overline{U}\right)$ with $U = (0,1)^d$. Put $\Omega = [0,1]^d$ for brevity.

Definition 3.1. Let K be a positive constant. The set of all $k \geq 1$ times α-Hölder differentiable bijective functions $\varphi : \Omega \to \Omega$, satisfying the conditions $\|\varphi\|_{C^{k,\alpha}} \leq K$ and $|J_\varphi| \geq \frac{1}{K}$, will be denoted by $\mathscr{H}_{G,K}$. The members of $\mathscr{H}_{G,K}$ will be called *admissible generators*, and the set $\mathscr{H}_{G,K}$ will be called the *hypothesis space of generators.*

Proposition 3.2. *$\mathscr{H}_{G,K}$ is a closed subset of the space $C^{k,\alpha}(\Omega)$. For each $\varphi \in \mathscr{H}_{G,K}$, the pushforward measure $\mu_\varphi = \varphi_*\lambda^{(d)}$ is absolutely continuous with respect to $\lambda^{(d)}$, and its density f_φ is a $(k-1)$-times α-Hölder differentiable function on Ω, satisfying the condition $\min_{x\in\Omega} f_\varphi(x) > 0$. The density f_φ is given by the formula*

$$f_\varphi = |J_{\varphi^{-1}}|\,. \tag{3.1}$$

Proof. The statement (b) of Theorem 2.4 shows that $\mathscr{H}_{G,K}$ is closed in $C^{k,\alpha}(\Omega)$. Using (2.10), we may estimate

$$|J_\varphi| \leq d! \max_{x\in\Omega}\left[\max_{1\leq i,j\leq d}\left|\frac{\partial \varphi_i(x)}{\partial x_j}\right|\right]^d \leq d!K^d, \tag{3.2}$$

for every $\varphi \in \mathscr{H}_{G,K}$. The change-of-variables formula shows that

$$\varphi_*\lambda^{(d)}(A) = \lambda^{(d)}\left(\varphi^{-1}(A)\right) = \int\limits_{\varphi^{-1}(A)} d\lambda^{(d)} = \int\limits_A |J_{\varphi^{-1}}|\, d\lambda^{(d)} = \int\limits_A \frac{1}{|J_\varphi \circ \varphi^{-1}|} d\lambda^{(d)},$$

for any $\varphi \in \mathscr{H}_{G,K}$ and $A \in \mathscr{B}\left([0,1]^d\right)$; hence each $\varphi \in \mathscr{H}_{G,K}$ generates a measure $\mu_\varphi = \varphi_*\lambda^{(d)}$ with a density

$$f_\varphi = |J_{\varphi^{-1}}| = \frac{1}{|J_\varphi \circ \varphi^{-1}|} \geq \frac{1}{d!K^d}.$$

□

Corollary 3.3. *$\mathscr{H}_{G,K}$ is compact in $C^{k,\alpha'}(\Omega)$ for $0 < \alpha' < \alpha$.*

Proof. Indeed, $\mathscr{H}_{G,K}$ is a bounded subset of the space $C^{k,\alpha}(\Omega)$ and the embedding (2.1) is compact, hence $\mathscr{H}_{G,K}$ is relatively compact in $C^{k,\alpha'}(\Omega)$. Since $\mathscr{H}_{G,K}$ is closed in $C^{k,\alpha'}(\Omega)$, it is compact. □

Our major requirement on the original measure is the following

Assumption 3.4. *μ is absolutely continuous with respect to $\lambda^{(d)}$ and the Radon-Nikodym derivative $f_\mu \coloneqq \frac{d\mu}{d\lambda^{(d)}}$ satisfies the condition*

$$\varepsilon = \min_{x\in\Omega} f_\mu(x) > 0.$$

Moreover, $f_\mu \in C^{k,\alpha}(\Omega)$ for some $\alpha \in (0,1]$ and $k \geq 1$.

Remark 3.5. While the above assumptions seem restrictive, for any given μ one can easily define an approximate problem, such that Assumption 3.4 is fulfilled. To this avail, $[0,1]^d$ is embedded in $\mathbb{R}^d$, convolved with $N(0,\epsilon^2\mathbb{1})$ and then re-projected to $[0,1]^d$ applying the $\mod 1$-mapping in all dimensions. On the side of the input data, this corresponds to the mapping $Y \mapsto (Y+N) \mod 1$, where $N \sim N(0,\epsilon^2\mathbb{1})$ is noise. This does not only allow to construct an explicit lower bound for the density of the modified measure, but also an explicit computation of the $C^{k,\alpha}$-norm and thereby a computable choice for K.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Let $Y = (Y_1, \ldots, Y_d)$ be a random vector with distribution μ satisfying Assumption 3.4. Consider the *Rosenblatt Transformation* ψ, given by the formula

$$\psi(y_1, y_2, \ldots, y_d) = \left(F_{Y_1}(y_1), F_{Y_2|Y_1}(y_2|y_1), \ldots, F_{Y_d|(Y_1,\ldots,Y_{d-1})}(y_d\,|(y_1, \ldots, y_{d-1}))\right)$$

for every $(y_1, \ldots, y_d) \in \Omega$ (see [21]). Since $0 \le F_{Y_1} \le 1$ and $0 \le F_{Y_j|(Y_1,\ldots,Y_{j-1})} \le 1$ $(2 \le j \le d)$, this implies $\psi(\Omega) \subset \Omega$.

The following theorem proves that μ is *realizable* in $\mathscr{H}_{G,K}$:

Theorem 3.6. *Under Assumption 3.4, the Rosenblatt Transformation ψ has the following properties:*

(i) ψ is k-times α-Hölder differentiable, i.e., $\psi \in C^{k,\alpha}(\Omega)$.

(ii) ψ is bijective and the inverse map ψ^{-1} is given by the formula

$$\psi^{-1}(x_1, x_2, \ldots, x_d) = (y_1, y_2, \ldots, y_d)$$

where[3]

$$y_1 = F_{Y_1}^{-1}(x_1), y_2 = F_{Y_2|Y_1}^{-1}(x_2|y_1), \ldots, y_d = F_{Y_d|(Y_1,\ldots,Y_{d-1})}^{-1}(x_d|\,(y_1, \ldots, y_{d-1}))\,.$$

(iii) The Jacobian determinant of ψ coincides with the density f_μ.

Proof. In view of Assumption 3.4 and Lemma 2.2, the densities

$$f_{Y_1}(\cdot), f_{Y_2|Y_1}(\cdot|y_1), \ldots, f_{Y_d|(Y_1,\ldots,Y_{d-1})}(\cdot\,|(y_1, \ldots, y_{d-1}))$$

are k-times Hölder differentiable. Clearly, the integration preserves the Hölder continuity. Therefore the distribution functions

$$F_{Y_1}(\cdot), F_{Y_2|Y_1}(\cdot|y_1), \ldots, F_{Y_d|(Y_1,\ldots,Y_{d-1})}(\cdot\,|(y_1, \ldots, y_{d-1}))$$

are k-times Hölder differentiable, too. Thus, $\psi \in C^{k,\alpha}(\Omega)$.

In order to prove the bijectivity of ψ, we choose an arbitrary $x = (x_1, \ldots, x_d) \in \Omega$ and show that the equation $\psi(y) = x$ has a unique solution. Put

$$\begin{aligned} \psi_1(y_1) &\coloneqq F_{Y_1}(y_1) \quad (y_1 \in [0,1])\,, \\ \psi_j(y_1, \ldots, y_j) &\coloneqq F_{Y_j|(Y_1,\ldots,Y_{j-1})}(y_j|\,(y_1, \ldots, y_{j-1})) \quad (y_1, \ldots, y_j \in [0,1];\, 2 \le j \le d) \end{aligned}$$

and rewrite the definition of ψ in the form

$$\psi(y_1, y_2, \ldots, y_d) = (\psi_1(y_1), \ldots, \psi_d(y_1, \ldots, y_d))\,.$$

Assumption 3.4 ensures that

$$\frac{\partial \psi_1(y_1)}{\partial y_1} = f_{Y_1}(y_1) > 0 \quad (y_1 \in (0,1))\,, \tag{3.3}$$

$$\frac{\partial \psi_j(y_1, \ldots, y_j)}{\partial y_j} = f_{Y_j|(Y_1,\ldots,Y_{j-1})}(y_j|\,(y_1, \ldots, y_{j-1})) > 0 \;\; (y_1, \ldots, y_j \in (0,1)\,;\, 2 \le j \le d)\,, \tag{3.4}$$

[3]For a non-injective distribution function F we put $F^{-1}(y) \coloneqq \inf\{x \in \mathbb{R} : F(x) \ge y\}$.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

hence $\psi_j(y_1, \ldots, y_j)$ is strictly increasing with respect to y_j on the segment $[0,1]$, for $1 \le j \le d$. Moreover, $\psi_j|_{y_j=0} = 0$, $\psi_j|_{y_j=1} = 1$. Therefore the equation $\psi_1(y_1) = x_1$ has a unique solution which we denote by b_1. Next, the equation $\psi_2(b_1, y_2) = x_2$ has a unique solution which we denote by b_2, etc. Thus, we see in a finite number of steps that the equation $\psi(y) = x$ has a unique solution $y = (b_1, b_2, \ldots, b_d)$ where

$$b_1 = F_{Y_1}^{-1}(x_1), b_2 = F_{Y_2|Y_1}^{-1}(x_2|b_1), \ldots, b_d = F_{Y_d|(Y_1,\ldots,Y_{d-1})}^{-1}(x_d|(b_1, \ldots, b_{d-1})).$$

To compute the Jacobian determinant J_ψ, observe that

$$\frac{\partial \psi_i}{\partial y_j} = 0 \quad (1 \le i < j \le d),$$

i.e., the Jacobian matrix of ψ is lower triangular. Hence

$$J_\psi = \frac{\partial \psi_1}{\partial y_1} \cdot \frac{\partial \psi_2}{\partial y_2} \cdots \frac{\partial \psi_d}{\partial y_d}.$$

(2.13) and (3.4) give

$$\frac{\partial \psi_j(y_1, \ldots, y_j)}{\partial y_j} = \frac{f_{(Y_1,\ldots,Y_j)}(y_1, \ldots, y_j)}{f_{(Y_1,\ldots,Y_{j-1})}(y_1, \ldots, y_{j-1})} \quad (y_1, \ldots, y_j \in (0,1);\, 2 \le j \le d),$$

which, together with (3.3), yields

$$J_\psi(y_1, \ldots, y_d) = f_{Y_1}(y_1) \cdot \frac{f_{(Y_1,Y_2)}(y_1, y_2)}{f_{Y_1}(y_1)} \cdots \frac{f_{(Y_1,\ldots,Y_d)}(y_1, \ldots, y_d)}{f_{(Y_1,\ldots,Y_{d-1})}(y_1, \ldots, y_{d-1})} = f_{(Y_1,\ldots,Y_d)}(y_1, \ldots, y_d).$$

The proof is complete. □

Corollary 3.7. *Under Assumption 3.4, the inequality*

$$J_\psi(y) \ge \varepsilon \quad (y \in \Omega)$$

holds.

The inverse of the Rosenblatt transformation ψ will be denoted by ϕ. Whenever clarity requires, we will write ϕ_μ instead of ϕ. The second statement of Theorem 3.6 shows that ϕ can be written in the form

$$\phi(x_1, \ldots, x_d) = (\phi_1(x_1), \ldots, \phi_d(x_1, \ldots, x_d)),$$

hence ϕ, like ψ, has a lower triangular Jacobi matrix.

The following simple statement shows that the inverse Rosenblatt transformation ϕ is a generator for the measure μ.

Proposition 3.8. *If μ satisfies Assumption 3.4, then $\phi \in C^{k,\alpha}(\Omega)$ and the equality $\mu = \phi_* \lambda^{(d)}$ holds.*

Proof. Theorems 2.4 and 3.6 show that $\phi = \psi^{-1} \in C^{k,\alpha}(\Omega)$. Next, for any $A \in \mathscr{B}\left([0,1]^d\right)$, the change-of-variables theorem, together with the statement (iii) of Theorem 3.6, gives

$$\mu(A) = \int_A f_\mu d\lambda^{(d)} = \int_A |J_\psi| \, d\lambda^{(d)} = \int_{\psi(A)} d\lambda^{(d)} = \lambda^{(d)}\left(\phi^{-1}(A)\right). \qquad \square$$

Theorem 3.6 and Proposition 3.8 show that the inverse Rosenblatt transformation ϕ is in $\mathscr{H}_{G,K}$ for sufficiently large K and generates the original measure μ.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

3.2 Discriminators

Here we present a set of discriminators that is affiliated to the set of generators in the sense that, if the data generating measure μ is realizable in $\mathscr{H}_{G,K}$, then the optimal discriminator that separates data from $\phi_*\lambda^{(d)}$ and μ is realizable in $\mathscr{H}_{D,K}$, see Theorem 4.2 below.

Every $\varphi \in \mathscr{H}_{G,K}$ generates a distribution F_φ with the density function f_φ. Given a true generator $\varphi \in \mathscr{H}_{G,K}$, the purpose of a discriminator is the estimation of the probability that a given distribution generated by $\varphi' \in \mathscr{H}_{G,K}$ is the original one, i.e., the probability that $\varphi_*\lambda^{(d)} = \varphi'_*\lambda^{(d)}$ almost surely.

Definition 3.9. For $\varphi, \varphi' \in \mathscr{H}_{G,K}$, the discriminator $D_{\varphi,\varphi'}$ is defined by the formula

$$D_{\varphi,\varphi'} := \frac{f_\varphi}{f_\varphi + f_{\varphi'}}.$$

Lemma 2.2 and Proposition 3.2 show that $D_{\varphi,\varphi'} \in C^{k-1,\alpha}(\Omega, \mathbb{R})$ for any $\varphi, \varphi' \in \mathscr{H}_{G,K}$.

Definition 3.10. Let K be a positive constant. The set

$$\mathscr{H}_{D,K} \coloneqq \{D_{\varphi,\varphi'} : \varphi, \varphi' \in \mathscr{H}_{G,K}\}$$

will be called the *hypothesis space of discriminators.*

Proposition 3.11. *$\mathscr{H}_{D,K}$ is compact in $C^{k-1,\alpha'}(\Omega, \mathbb{R})$ for $0 < \alpha' < \alpha$.*

Proof. Take any sequence $\{D_{\varphi_n,\varphi'_n}\}_{n=1}^\infty \subset \mathscr{H}_{D,K}$. Corollary 3.3 and the statement (b) of Theorem 2.4 show that we can find a subsequence $\left\{D_{\varphi_{n_i},\varphi'_{n_i}}\right\}_{i=1}^\infty$ such that $\varphi_{n_i}, \varphi'_{n_i}, \varphi_{n_i}^{-1}, \varphi_{n_i}'^{-1}$ converge in $C^{k,\alpha'}$-norm. It is easy to see that $\left\{D_{\varphi_{n_i},\varphi'_{n_i}}\right\}_{i=1}^\infty$ converges in $C^{k,\alpha'}$-norm and $\lim_{i\to\infty} D_{\varphi_{n_i},\varphi'_{n_i}} \in \mathscr{H}_{D,K}$. $\square$

Lemma 3.12. *There exist constants $B_1, B_2 \in (0,1)$, depending only on K and d, such that*

$$B_1 \le D_{\varphi,\varphi'} \le B_2, \tag{3.5}$$

for all $\varphi, \varphi' \in \mathscr{H}_{G,K}$.

Proof. In view of (3.2) and the equality $f_\varphi = \frac{1}{|J_\varphi \circ \varphi^{-1}|}$, the density f_φ can be estimated as

$$\frac{1}{d!K^d} \le f_\varphi \le K \quad (\varphi \in \mathscr{H}_{G,K}). \tag{3.6}$$

Since $f_{\varphi'}$ admits the same estimate, hence

$$\frac{1}{d!K^{d+1}} \le \frac{f_{\varphi'}}{f_\varphi} \le d!K^{d+1}.$$

Using this for $D_{\varphi,\varphi'} = \left(1 + \frac{f_{\varphi'}}{f_\varphi}\right)^{-1}$, we obtain $B_1 \le D_{\varphi,\varphi'} \le B_2$ with

$$B_1 \coloneqq \frac{1}{1 + d!K^{d+1}} \quad \text{and} \quad B_2 \coloneqq 1 - B_1 = \frac{d!K^{d+1}}{1 + d!K^{d+1}}. \tag{3.7}$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

4 Consistency of Generative Adversarial Learning

In this section we introduce theoretical and empirical loss functions for adversarial learning and study their properties. The adversarial learning problem is typically formulated as a minimax problem, see [12]. Applying the uniform law of large numbers, we prove that the empirical loss almost surely converges to the theoretical one and this convergence is uniform with respect to generators and discriminators. The latter enables us to show that the distributions generated by empirical minimizers converge to the original measure in Jensen-Shannon divergence.

Let F_μ and f_μ denote the cumulative distribution function and the density of the original measure μ, respectively. On the one hand, we want to find a generator φ for which the measure μ_φ is barely distinguishable from the measure μ; on the other hand, we look for a good discriminator capable of distinguishing between μ_φ and μ. Thus, the generator and the discriminator work at cross-purposes, and to formalize this problem, we consider the following (theoretical) loss function introduced in [12]:

$$L(\varphi, D) = \frac{1}{2}\left(\mathbb{E}_{Y\sim F_\mu}\left[\log\left(D\left(Y\right)\right)\right] + \mathbb{E}_{\hat{Y}\sim F_\varphi}\left[\log\left(1 - D\left(\hat{Y}\right)\right)\right]\right) \tag{4.1}$$

with $\varphi \in \mathscr{H}_{G,K}$ and $D \in \mathscr{H}_{D,K}$. Then the best generator φ is a minimizer of the quantity

$$\sup_{D\in\mathscr{H}_{D,K}} L(\varphi, D). \tag{4.2}$$

Proposition 4.1. *L is continuous on $\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}$.*

Proof. Rewrite L in the form

$$L(\varphi, D) = \frac{1}{2}\int_\Omega \left[f_\mu \log D + f_\varphi \log\left(1 - D\right)\right] d\lambda^{(d)} \quad \left(\varphi \in \mathscr{H}_{G,K}, D \in \mathscr{H}_{D,K}\right). \tag{4.3}$$

Estimates (3.5) and (3.6) show that

$$\sup_{x\in\Omega,\varphi\in\mathscr{H}_{G,K},D\in\mathscr{H}_{D,K}} \left|f_\mu(x)\log D(x) + f_\varphi(x)\log\left(1 - D(x)\right)\right| < \infty.$$

Let $(\varphi_n, D_n) \to (\varphi, D)$ in $\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}$, i.e., $\varphi_n \to \varphi$ in $C^{k,\alpha}(\Omega)$-norm and $D_n \to D$ in $C^{k-1,\alpha}(\Omega, \mathbb{R})$-norm. Then $\varphi_n^{-1} \to \varphi^{-1}$ in $C^{k,\alpha}(\Omega)$-norm by the statement (b) of Theorem 2.4, which implies the pointwise convergence $J_{\varphi_n^{-1}}(x) \to J_{\varphi^{-1}}(x) \quad (x \in \Omega)$. Therefore $f_{\varphi_n}(x) \to f_\varphi(x) \quad (x \in \Omega)$, and Lebesgue's dominated convergence theorem shows that $L(\varphi_n, D_n) \to L(\varphi, D)$. □

In view of Corollary 3.3 and Proposition 3.11, $\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}$ is compact in $C^{k,\alpha'} \times C^{k-1,\alpha'}$ for $0 < \alpha' < \alpha$, which, together with the continuity of L, shows that the supremum in (4.2) can be replaced by the maximum. Furthermore, L is uniformly continuous in $C^{k,\alpha'} \times C^{k-1,\alpha'}$-norm for $0 < \alpha' < \alpha$, which implies the continuity of the functional $\sup_{D\in\mathscr{H}_{D,K}} L(\cdot, D)$ in $C^{k,\alpha'}$-norm. Therefore the functional $\sup_{D\in\mathscr{H}_{D,K}} L(\cdot, D)$ attains its minimum on $\mathscr{H}_{G,K}$. Thus, best generators exist, and the quantity $\inf_{\varphi\in\mathscr{H}_{G,K}} \sup_{D\in\mathscr{H}_{D,K}} L(\varphi, D)$ can be rewritten as

$$\min_{\varphi\in\mathscr{H}_{G,K}} \max_{D\in\mathscr{H}_{D,K}} L(\varphi, D). \tag{4.4}$$

Below the positive constant K is assumed to be so large that the inverse Rosenblatt transformation ϕ is in $\mathscr{H}_{G,K}$. By analogy with [3], the maximizers of function $L(\varphi, \cdot)$ admit the following simple characterization:

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Theorem 4.2. *For any $\varphi \in \mathscr{H}_{G,K}$, the function $L(\varphi, \cdot)$ has exactly one maximizer D_φ on $\mathscr{H}_{D,K}$; this maximizer is given by*

$$D_\varphi = \frac{f_\mu}{f_\mu + f_\varphi}.$$

Proof. Putting

$$h(s, r, p) := \frac{1}{2}[s \log p + r \log(1-p)] \quad (s, r > 0; 0 < p < 1),$$

we rewrite (4.3) in the form

$$L(\varphi, D) = \int_\Omega h(f_\mu, f_\varphi, D)\, dx \quad (\varphi \in \mathscr{H}_{G,K}, D \in \mathscr{H}_{D,K}).$$

It is easy to check that for fixed s and r the function $h(s, r, \cdot)$ attains its strict global maximum at $p = \frac{s}{s+r}$. Hence

$$h(f_\mu, f_\varphi, D) \le h(f_\mu, f_\varphi, D_\varphi) \quad (\varphi \in \mathscr{H}_{G,K}, D \in \mathscr{H}_{D,K})$$

which shows that $D_\varphi = \dfrac{f_\mu}{f_\mu + f_\varphi}$ maximizes $L(\varphi, \cdot)$. Moreover, if $D \ne D_\varphi$, then $E := \{x : D(x) \ne D_\varphi(x)\}$ is a set of positive Lebesgue measure and

$$h(f_\mu, f_\varphi, D)|_E < h(f_\mu, f_\varphi, D_\varphi)|_E;$$

therefore $L(\varphi, D) < L(\varphi, D_\varphi)$. □

In the next proposition we reveal the relationship between the loss function L and the Jensen-Shannon divergence and then apply it to describe all minimizers of the quantity (4.2).

Proposition 4.3. *The following statements are true:*

(a) For each $\varphi \in \mathscr{H}_{G,K}$, the Jensen-Shannon divergence between two densities f_μ and f_φ can be expressed in terms of the loss function L by the formula

$$d_{JS}(f_\mu, f_\varphi) = L(\varphi, D_\varphi) + \log(2). \tag{4.5}$$

(b) A generator $\varphi \in \mathscr{H}_{G,K}$ is a minimizer of the quantity (4.2) iff the equality

$$|J_{\varphi^{-1}}| = f_\mu \tag{4.6}$$

holds. In particular, the inverse Rosenblatt transformation ϕ is a minimizer of (4.2).

Proof. (a) Straightforward computation following [12] shows that

$$\begin{aligned} d_{JS}(f, f_\varphi) &= \frac{1}{2}\left[\int_\Omega f_\mu \log\left(\frac{2 f_\mu}{f_\mu + f_\varphi}\right) d\lambda^{(d)} + \int_\Omega f_\varphi \log\left(\frac{2 f_\varphi}{f_\mu + f_\varphi}\right) d\lambda^{(d)}\right] \\ &= \frac{1}{2}\int_\Omega [f_\mu \log(D_\varphi) + f_\varphi \log(1 - D_\varphi)]\, d\lambda + \log(2) = L(\varphi, D_\varphi) + \log(2). \end{aligned}$$

(b) Rewrite (4.5) in the form

$$\max_{D \in \mathscr{H}_{D,K}} L(\varphi, D) = d_{JS}(f_\mu, f_\varphi) - \log(2).$$

Since $d_{JS}(f_\mu, f_\varphi) \ge 0$ and $d_{JS}(f_\mu, f_\varphi) = 0$ iff $f_\mu = f_\varphi$, the generator φ is a minimizer of (4.2) iff $f_\mu = f_\varphi$, which, in view of (3.1), is the desired conclusion. □

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

We approximate the theoretical loss function L by the empirical loss function

$$\hat{L}(\varphi, D, n) = \frac{1}{2n}\sum_{i=1}^{n} \log D(Y_i) + \frac{1}{2n}\sum_{i=1}^{n} \log\left[1 - D(\varphi(Z_i))\right], \tag{4.7}$$

where Y_i and Z_i $(i = 1, 2, \ldots, n)$ are samples with densities f_μ and 1, respectively; and n is the size of the training data set. For simplicity of notation, we will suppress the dependence of $\hat{L}$ on the samples Y_i and Z_i. Note that the theoretical loss L is the expectation of the empirical loss $\hat{L}$. Since $\log \in C^{k,\alpha}([a,b], \mathbb{R})$ for any $[a,b] \subset (0,1)$, the statement (b) of Lemma 2.3 and Lemma 3.12 easily imply the continuity of the function $\hat{L}$ on $\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}$; it is easy to see that the continuity holds even in $C(\Omega, \mathbb{R}^d) \times C(\Omega, \mathbb{R})$-norm.

Denote by

$$\varepsilon_{sampling}(n) := \sup_{\varphi \in \mathscr{H}_{G,K}, D \in \mathscr{H}_{D,K}} \left|\hat{L}(\varphi, D, n) - L(\varphi, D)\right| \tag{4.8}$$

the sampling error. In order to prove that $\varepsilon_{sampling}(n) \to 0$, we need the following version of the uniform law of large numbers (see [8, Sec. 16]):

Theorem 4.4 (Uniform law of large numbers)**.** *Let (Θ, ρ) be a compact metric space, and let the function $U : \mathbb{R} \times \Theta \to \mathbb{R}$ be measurable with respect to x for any fixed $\theta \in \Theta$ and continuous with respect to θ for (almost) all x. Further, let $X_1, X_2, \ldots$ be independent and identically distributed random variables. If there exists a Borel function $\xi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[\xi(X_1)] < \infty$ and $|U(x, \theta)| \le \xi(x)$ $(x \in \mathbb{R}, \theta \in \Theta)$, then*

$$\lim_{n\to\infty} \sup_{\theta \in \Theta} \left| \frac{1}{n}\sum_{j=1}^{n} U(X_j, \theta) - \mu(\theta)\right| = 0$$

almost surely with $\mu(\theta) = \mathbb{E}[U(X_1, \theta)]$.

Now we are in a position to state and prove our first main result:

Theorem 4.5. *The sampling error $\varepsilon_{sampling}(n)$ converges to 0 almost surely, as $n \to \infty$.*

Proof. Put $Y_i^{(\varphi)} := \varphi(Z_i)$ $(i = 1, 2, \ldots, n)$ for brevity. We have

$$\begin{aligned} 2\varepsilon_{sampling}(n) \le &\sup_{\varphi \in \mathscr{H}_{G,K}, D \in \mathscr{H}_{D,K}} \left| \mathbb{E}_{Y \sim F}\left[\log D(Y)\right] - \frac{1}{n}\sum_{i=1}^{n} \log D(Y_i)\right| \\ &+ \sup_{\varphi \in \mathscr{H}_{G,K}, D \in \mathscr{H}_{D,K}} \left| \mathbb{E}_{Y \sim F_\varphi}\left[\log(1 - D(Y))\right] - \frac{1}{n}\sum_{i=1}^{n} \log\left(1 - D\left(Y_i^{(\varphi)}\right)\right)\right|, \end{aligned}$$

and in order to finish the proof, it is enough to see that the uniform law of large numbers is applicable to the functions

$$U_1(x, D) := \log D(x) \quad \text{and} \quad U_2(x, \varphi, D) := \log(1 - D(\varphi(x))).$$

Indeed, the parameter space $\Theta = \mathscr{H}_{G,K} \times \mathscr{H}_{D,K}$ is compact in $C^{k,\alpha'} \times C^{k-1,\alpha'}$ for $0 < \alpha' < \alpha$. Since $\log \in C^{k,\alpha}([a,b], \mathbb{R})$ for any $[a,b] \subset (0,1)$, the statement (b) of Lemma 2.3 and Lemma 3.12 easily imply the uniform boundedness and continuity of the functions U_1 and U_2. □

By analogy with (4.4), the continuity of the empirical loss $\hat{L}$ implies the existence of

$$\min_{\varphi \in \mathscr{H}_{G,K}} \max_{D \in \mathscr{H}_{D,K}} \hat{L}(\varphi, D, n) \tag{4.9}$$

for every $n \in \mathbb{N}$. The following theorem reveals the relationship between (4.4) and (4.9). Due to our construction that guarantees *realizability* both of generators and discriminators, we simplify the proof of a related result from [3]:

Theorem 4.6. *If φ^* is a minimizer of the quantity (4.2) on $\mathscr{H}_{G,K}$ and if $\hat{\varphi}_n$ is a minimizer of the quantity*

$$\max_{D \in \mathscr{H}_{D,K}} \hat{L}(\varphi, D, n)$$

on $\mathscr{H}_{G,K}$ for each $n \in \mathbb{N}$, then $f_{\varphi^*} = f_\mu$ and $d_{JS}(f_\mu, f_{\hat{\varphi}_n}) \to 0$, as $n \to \infty$.

Proof. We first note that

$$L(\varphi^*, D_{\varphi^*}) = \min_{\varphi \in \mathscr{H}_{G,K}} \max_{D \in \mathscr{H}_{D,K}} L(\varphi, D) \leq \max_{D \in \mathscr{H}_{D,K}} L(\hat{\varphi}_n, D) = L(\hat{\varphi}_n, D_{\hat{\varphi}_n}).$$

Next, in view of (4.8), we have

$$\begin{aligned} L(\hat{\varphi}_n, D_{\hat{\varphi}_n}) &\leq \hat{L}(\hat{\varphi}_n, D_{\hat{\varphi}_n}, n) + \varepsilon_{sampling}(n) \leq \max_{D \in \mathscr{H}_{D,K}} \hat{L}(\hat{\varphi}_n, D, n) + \varepsilon_{sampling}(n) \\ &= \min_{\varphi \in \mathscr{H}_{G,K}} \max_{D \in \mathscr{H}_{D,K}} \hat{L}(\varphi, D, n) + \varepsilon_{sampling}(n) \leq \max_{D \in \mathscr{H}_{D,K}} \hat{L}(\varphi^*, D, n) + \varepsilon_{sampling}(n) \\ &\leq \max_{D \in \mathscr{H}_{D,K}} [L(\varphi^*, D) + \varepsilon_{sampling}(n)] + \varepsilon_{sampling}(n) = L(\varphi^*, D_{\varphi^*}) + 2\varepsilon_{sampling}(n). \end{aligned}$$

Thus,

$$0 \leq L(\hat{\varphi}_n, D_{\hat{\varphi}_n}) - L(\varphi^*, D_{\varphi^*}) \leq 2\varepsilon_{sampling}(n), \tag{4.10}$$

and Theorem 4.5 implies that $L(\hat{\varphi}_n, D_{\hat{\varphi}_n}) \to L(\varphi^*, D_{\varphi^*})$. The latter, together with (4.5), gives $d_{JS}(f_\mu, f_{\hat{\varphi}_n}) \to d_{JS}(f_\mu, f_{\varphi^*})$, as $n \to \infty$. (3.1) and (4.6), together with the statement (b) of Proposition 4.3, show that $f_{\varphi^*} = f_\mu$, and the proof is complete. □

5 Quantitative Estimates of the Sampling Error

In the previous section we have proven that the sampling error converges to 0 almost surely. The main objective of this section is to estimate the rate of that convergence. Assuming that the order of smoothness of generators and discriminators is large enough, we first give an upper bound for the expectation of the sampling error. Combining the obtained result with McDiarmid's inequality gives the desired estimate.

To achieve an upper bound for the expectation of the sampling error $\varepsilon_{sampling}(n)$, i.e., for the quantity

$$\mathbb{E}\left[\sup_{\substack{\varphi \in \mathscr{H}_{G,K} \\ D \in \mathscr{H}_{D,K}}} \left|\hat{L}(\varphi, D, n) - L(\varphi, D)\right|\right],$$

we will apply the Dudley estimate (2.14) to the random processes $\pm\left[\hat{L}(\varphi, D, n) - L(\varphi, D)\right]$. The expectation of the difference $\hat{L}(\varphi, D, n) - L(\varphi, D)$ is clearly 0, and we already know from the previous section that $L(\cdot,\cdot)$ and $\hat{L}(\cdot,\cdot,n)$ are continuous on $\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}$; hence it remains to state the subgaussian property of the increments $\hat{L}(\varphi_1, D_1, n) - \hat{L}(\varphi_2, D_2, n)$.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Lemma 5.1. *The increment* $\hat{L}(\varphi_1, D_1, n) - \hat{L}(\varphi_2, D_2, n)$ *is* $[\rho_n((\varphi_1, D_1), (\varphi_2, D_2))]^2$*-subgaussian with*

$$\rho_n((\varphi_1, D_1), (\varphi_2, D_2)) := \frac{1 + d!K^{d+1}}{\sqrt{n}} \left[\|D_1 - D_2\|_\infty + d^2 (d!)^3 K^{3d+2} \|\varphi_1 - \varphi_2\|_\infty\right].$$

Proof. It is easy to check that the arithmetic mean of n independent and identically distributed σ^2-subgaussian random variables is $\frac{\sigma^2}{n}$-subgaussian. This fact, together with the representation

$$\begin{aligned}\hat{L}(\varphi_1, D_1, n) - \hat{L}(\varphi_2, D_2, n) = &\frac{1}{2n} \sum_{i=1}^{n} [\log D_1(Y_i) - \log D_2(Y_i)] \\ &+ \frac{1}{2n} \sum_{i=1}^{n} \{\log[1 - D_1(\varphi_1(Z_i))] - \log[1 - D_2(\varphi_2(Z_i))]\},\end{aligned}$$

shows that it suffices to prove the subgaussian property in the case $n = 1$.

Next, we will estimate the deviation

$$\left\|\hat{L}(\varphi_1, D_1, 1) - \hat{L}(\varphi_2, D_2, 1)\right\|_\infty.$$

We have

$$\begin{aligned}\left|\hat{L}(\varphi_1, D_1, 1) - \hat{L}(\varphi_2, D_2, 1)\right| \le &\frac{1}{2} |\log D_1(Y_i) - \log D_2(Y_i)| \\ &+ \frac{1}{2} \left|\log\left(1 - D_1\left(Y_i^{(\varphi_1)}\right)\right) - \log\left(1 - D_2\left(Y_i^{(\varphi_1)}\right)\right)\right| \\ &+ \frac{1}{2} \left|\log\left(1 - D_2\left(Y_i^{(\varphi_1)}\right)\right) - \log\left(1 - D_2\left(Y_i^{(\varphi_2)}\right)\right)\right|.\end{aligned}$$

In view of Lemma 3.12, there exists a segment $[B_1, B_2] \subset (0, 1)$ such that $D_{\varphi,\varphi'} \in [B_1, B_2]$. The elementary estimate

$$|\log x_1 - \log x_2| = \left|\int_{x_1}^{x_2} \frac{dt}{t}\right| \le \frac{|x_1 - x_2|}{\min\{x_1, x_2\}} \quad (x_1, x_2 > 0)$$

gives

$$|\log x - \log y| \le \frac{1}{B_1} |x - y| \quad (x, y \in [B_1, B_2]),$$

$$|\log(1 - x) - \log(1 - y)| \le \frac{1}{1 - B_2} |x - y| = \frac{1}{B_1} |x - y| \quad (x, y \in [B_1, B_2]).$$

The last two inequalities imply

$$\begin{aligned}\left|\hat{L}(\varphi_1, D_1, 1) - \hat{L}(\varphi_2, D_2, 1)\right| \le &\frac{1}{2B_1} \Big[|D_1(Y_i) - D_2(Y_i)| \\ &+ \left|D_1\left(Y_i^{(\varphi_1)}\right) - D_2\left(Y_i^{(\varphi_1)}\right)\right| \\ &+ \left|D_2\left(Y_i^{(\varphi_1)}\right) - D_2\left(Y_i^{(\varphi_2)}\right)\right|\Big].\end{aligned}$$

The first and the second summands on the right-hand side do not exceed $\|D_1 - D_2\|_\infty$. To find an upper bound for the third summand, we first note that

$$\begin{aligned} D_{\varphi,\varphi'}(x_1) - D_{\varphi,\varphi'}(x_2) &= \frac{f_\varphi(x_1)}{f_\varphi(x_1) + f_{\varphi'}(x_1)} - \frac{f_\varphi(x_2)}{f_\varphi(x_2) + f_{\varphi'}(x_2)} \\ &= \frac{f_{\varphi'}(x_2)\left[f_\varphi(x_1) - f_\varphi(x_2)\right] - f_\varphi(x_2)\left[f_{\varphi'}(x_1) - f_{\varphi'}(x_2)\right]}{\left[f_\varphi(x_1) + f_{\varphi'}(x_1)\right]\left[f_\varphi(x_2) + f_{\varphi'}(x_2)\right]}. \end{aligned}$$

The estimate (3.6) gives

$$\frac{f_\varphi(x_2)}{\left[f_\varphi(x_1)+f_{\varphi'}(x_1)\right]\left[f_\varphi(x_2)+f_{\varphi'}(x_2)\right]} \le d!K^d, \quad \frac{f_{\varphi'}(x_2)}{\left[f_\varphi(x_1)+f_{\varphi'}(x_1)\right]\left[f_\varphi(x_2)+f_{\varphi'}(x_2)\right]} \le d!K^d,$$

which, together with (3.6), yields

$$\begin{aligned} \left|D_{\varphi,\varphi'}(x_1) - D_{\varphi,\varphi'}(x_2)\right| &\le d!K^d \sum_{h\in\{\varphi,\varphi'\}} \left|f_h(x_1) - f_h(x_2)\right| \\ &= d!K^d \sum_{h\in\{\varphi,\varphi'\}} f_h(x_1) f_h(x_2) \left|\frac{1}{f_h(x_1)} - \frac{1}{f_h(x_2)}\right| \\ &\le d!K^{d+2} \sum_{h\in\{\varphi,\varphi'\}} \left|J_h\left(h^{-1}(x_1)\right) - J_h\left(h^{-1}(x_2)\right)\right|. \end{aligned}$$

By induction on d, one can easily prove the inequality

$$|a_1\cdots a_d - b_1\cdots b_d| \le d\left[\max_{1\le j\le d}\{|a_j|,|b_j|\}\right]^{d-1} \max_{1\le j\le d}|a_j - b_j|$$

where $a_j, b_j \ (1\le j\le d)$ are arbitrary real numbers. Using this inequality and the definition of a determinant, we can easily obtain the estimate

$$\left|\det(a_{ij})_{d\times d} - \det(b_{ij})_{d\times d}\right| \le d\cdot d!\left[\max_{1\le i,j\le d}\{|a_{ij}|,|b_{ij}|\}\right]^{d-1} \max_{1\le i,j\le d}|a_{ij} - b_{ij}|,$$

for any matrices $(a_{ij})_{d\times d}, (b_{ij})_{d\times d} \in M_{d,d}(\mathbb{R})$. Applying this estimate to the Jacobian matrices $J_h(y_1)$ and $J_h(y_2)$ yields

$$|J_h(y_1) - J_h(y_2)| \le d\cdot d!K^{d-1}\left|(Dh)(y_1) - (Dh)(y_2)\right| \quad (y_1, y_2 \in \Omega).$$

The latter, combined with the mean value theorem for vector-valued functions, gives

$$|J_h(y_1) - J_h(y_2)| \le d\cdot d!K^{d-1}\cdot dK\,|y_1 - y_2| \quad (y_1, y_2 \in \Omega),$$

therefore

$$\left|D_{\varphi,\varphi'}(x_1) - D_{\varphi,\varphi'}(x_2)\right| \le d^2 (d!)^2 K^{2d+2} \sum_{h\in\{\varphi,\varphi'\}} \left|h^{-1}(x_1) - h^{-1}(x_2)\right|.$$

Using (2.7), we estimate further

$$\left|D_{\varphi,\varphi'}(x_1) - D_{\varphi,\varphi'}(x_2)\right| \le 2d^2 (d!)^3 K^{3d+2}|x_1 - x_2|,$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

hence

$$\|D_{\varphi,\varphi'} \circ \varphi_1 - D_{\varphi,\varphi'} \circ \varphi_2\|_\infty \le 2d^2 (d!)^3 K^{3d+2} \|\varphi_1 - \varphi_2\|_\infty .$$

Thus, we arrive at the estimate

$$\left|\hat{L}(\varphi_1, D_1, 1) - \hat{L}(\varphi_2, D_2, 1)\right| \le \frac{1}{B_1} \left[\|D_1 - D_2\|_\infty + d^2 (d!)^3 K^{3d+2} \|\varphi_1 - \varphi_2\|_\infty\right].$$

Using (3.7), we can rewrite this estimate in the form

$$\left|\hat{L}(\varphi_1, D_1, 1) - \hat{L}(\varphi_2, D_2, 1)\right| \le \rho_1 ((\varphi_1, D_1), (\varphi_2, D_2)),$$

which, together with Hoeffding's Lemma, shows that $\hat{L}(\varphi_1, D_1, 1) - \hat{L}(\varphi_2, D_2, 1)$ is $[\rho_1 ((\varphi_1, D_1), (\varphi_2, D_2))]^2$-subgaussian. □

Proposition 5.2. *Let $k > 1 - \alpha + \frac{d^2}{2}$.*

(i) There exists a positive constant C, depending only on d, α, k and K, such that

$$\mathbb{E}[\varepsilon_{sampling}(n)] \le Cn^{-\frac{1}{2}} \quad (n = 1, 2, \dots).$$

(ii) There exists a positive constant γ, depending only on d, α, and k, such that

$$\mathbb{E}[\varepsilon_{sampling}(n)] \le \gamma K^{4(d+1)} n^{-\frac{1}{2}}, \tag{5.1}$$

for each $n = 1, 2, \dots$ and $K > 1$.

Proof. Since $\pm\left[\hat{L}(\varphi, D, n) - L(\varphi, D)\right]$ are continuous subgaussian processes on the compact space $(\Theta; \rho_n) = (\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}; \rho_n)$, the entropy bound (2.14) is applicable to $\hat{L}(\varphi, D, n) - L(\varphi, D)$, which gives

$$\begin{aligned}\mathbb{E}[\varepsilon_{sampling}(n)] &\le 12 \int_0^\infty \sqrt{\log N(\Theta, \rho_n, \varepsilon)}\, d\varepsilon = 12 \int_0^\infty \sqrt{\log N\left(\Theta, \rho_1, \sqrt{n}\varepsilon\right)}\, d\varepsilon \\ &= \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\Theta, \rho_1, \varepsilon)}\, d\varepsilon. \end{aligned} \tag{5.2}$$

It is easy to see that

$$\begin{aligned} N(\Theta, \rho_1, \rho_{r_1,r_2}) &\le N(\mathscr{H}_{G,K}, \|\cdot\|_\infty, r_1)\, N(\mathscr{H}_{D,K}, \|\cdot\|_\infty, r_2) \\ &\le N(B_{C^{k,\alpha}}[0,K], \|\cdot\|_\infty, r_1)\, N(B_{C^{k-1,\alpha}}[0,K], \|\cdot\|_\infty, r_2), \end{aligned}$$

where $B_{C^{l,\alpha}}[0,K]$ denotes the closed ball of radius K centered at 0 in the space $C^{l,\alpha}(\Omega)$ $(l = k-1, k)$ and

$$\rho_{r_1,r_2} := \left(1 + d! K^{d+1}\right)\left[r_2 + d^2 (d!)^3 K^{3d+2} r_1\right].$$

In view of (2.12), the estimate

$$\log N\left(B_{C^{k,\alpha}(\Omega)}[0,1], \|\cdot\|_\infty, \varepsilon\right) \le \frac{C_1 \left[\lambda^{(d)}(\Omega^1)\right]^d}{\varepsilon^{\frac{d^2}{\alpha+k}}}$$

holds, where

$$\Omega^1 := \{x : \operatorname{dist}(x, \Omega) < 1\}$$

and C_1 is a constant depending only on d and $\alpha + k$. Since $\Omega^1 \subset [-1,2]^d$, hence

$$\lambda^{(d)}\left(\Omega^1\right) \leq \lambda^{(d)}\left([-1,2]^d\right) = 3^d.$$

Furthermore,

$$N\left(B_{C^{k,\alpha}(\Omega)}[0,K], \|\cdot\|_\infty, \varepsilon\right) = N\left(B_{C^{k,\alpha}(\Omega)}[0,1], \|\cdot\|_\infty, \frac{\varepsilon}{K}\right).$$

Thus, we obtain

$$\log N\left(B_{C^{k,\alpha}(\Omega)}[0,K], \|\cdot\|_\infty, \varepsilon\right) \leq C_1 \cdot 3^{d^2} \cdot \left(\frac{K}{\varepsilon}\right)^{\frac{d^2}{\alpha+k}}.$$

Using this inequality, we may estimate

$$\log N\left(\Theta, \rho_1, \rho_{r_1,r_2}\right) \leq C_2(d,\alpha,k)\left[\left(\frac{K}{r_1}\right)^{\frac{d^2}{\alpha+k}} + \left(\frac{K}{r_2}\right)^{\frac{d^2}{\alpha+k-1}}\right]$$

with

$$C_2(d,\alpha,k) := 3^{d^2}\max\left\{C_1(d,\alpha+k), C_1(d,\alpha+k-1)\right\}.$$

Taking

$$r_1 = \frac{\varepsilon}{2d^2(d!)^3 K^{3d+2}\left(1+d!K^{d+1}\right)}, \quad r_2 = \frac{\varepsilon}{2\left(1+d!K^{d+1}\right)}$$

and putting

$$C_3 := \left[C_2(d,\alpha,k)\max\left\{\left(2d^2(d!)^3K^{3d+3}\left(1+d!K^{d+1}\right)\right)^{\frac{d^2}{\alpha+k}}, \left(2K\left(1+d!K^{d+1}\right)\right)^{\frac{d^2}{\alpha+k-1}}\right\}\right]^{\frac{1}{2}},$$

we obtain

$$\sqrt{\log N(\Theta,\rho_1,\varepsilon)} \leq C_3\left[\varepsilon^{-\frac{d^2}{2(\alpha+k)}} + \varepsilon^{-\frac{d^2}{2(\alpha+k-1)}}\right]. \tag{5.3}$$

We would like to integrate both parts of the last inequality. In view of the condition $k > 1 - \alpha + \frac{d^2}{2}$, we have $\frac{d^2}{2(\alpha+k-1)} < 1$ which guaranties the integrability of the right-hand side near 0. Put $\delta_1 := \operatorname{diam}\left(\mathscr{H}_{G,K} \times \mathscr{H}_{D,K}; \rho_1\right)$. Since $N(\Theta,\rho_1,\varepsilon) = 1 \quad (\varepsilon > \delta_1)$, hence (5.2) and (5.3) give

$$\mathbb{E}\left[\varepsilon_{sampling}(n)\right] \leq \frac{12C_3}{\sqrt{n}}\int_0^{\delta_1}\left[\varepsilon^{-\frac{d^2}{2(\alpha+k)}} + \varepsilon^{-\frac{d^2}{2(\alpha+k-1)}}\right]d\varepsilon = \frac{12C_3}{\sqrt{n}}\left[\delta_1^{1-\frac{d^2}{2(\alpha+k)}} + \delta_1^{1-\frac{d^2}{2(\alpha+k-1)}}\right];$$

therefore we may take $C = 12C_3\left[\delta_1^{1-\frac{d^2}{2(\alpha+k)}} + \delta_1^{1-\frac{d^2}{2(\alpha+k-1)}}\right]$ for (i).

To prove (ii), assume $K > 1$. Then

$$C_3 \leq \left[C_2(d,\alpha,k)\right]^{\frac{1}{2}}\left\{\left[2d^2(d!)^3K^{3d+3}\left(1+d!K^{d+1}\right)\right]^{\frac{d^2}{\alpha+k-1}}\right\}^{\frac{1}{2}}$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

$$\leq \left[C_2\left(d,\alpha,k\right)\right]^{\frac{1}{2}}\left[2d\left(d!\right)^2\right]^{\frac{d^2}{\alpha+k-1}} K^{\frac{2d^2(d+1)}{\alpha+k-1}} \leq 4d^2\left(d!\right)^4\left[C_2\left(d,\alpha,k\right)\right]^{\frac{1}{2}} K^{4(d+1)},$$

hence

$$\begin{aligned}C &= 12C_3\left[\delta_1^{1-\frac{d^2}{2(\alpha+k)}}+\delta_1^{1-\frac{d^2}{2(\alpha+k-1)}}\right]\\ &\leq 48d^2\left(d!\right)^4\left[C_2\left(d,\alpha,k\right)\right]^{\frac{1}{2}} K^{4(d+1)}\left[\delta_1^{1-\frac{d^2}{2(\alpha+k)}}+\delta_1^{1-\frac{d^2}{2(\alpha+k-1)}}\right],\end{aligned}$$

and we may choose

$$\gamma := 48d^2\left(d!\right)^4\left[C_2\left(d,\alpha,k\right)\right]^{\frac{1}{2}}\left[\delta_1^{1-\frac{d^2}{2(\alpha+k)}}+\delta_1^{1-\frac{d^2}{2(\alpha+k-1)}}\right]. \qquad \square$$

Theorem 5.3. *Let $0<\alpha\leq 1$, and let d and k be positive integers such that $k>1-\alpha+\frac{d^2}{2}$. There exists a positive constant γ, depending only on $d,\alpha,$ and k, such that for every $\delta>0$, for each positive integer n and for each $K>1$ satisfying the condition $\phi\in\mathscr{H}_{G,K}$, the following estimate is true:*

$$\mathbb{P}\left(\varepsilon_{sampling}\left(n\right)\geq 2\gamma K^{4(d+1)}n^{\delta-\frac{1}{2}}\right)\leq \exp\left\{-\frac{\gamma^2K^{8(d+1)}n^{2\delta}}{\log^2\left(1+d!K^{d+1}\right)}\right\}. \tag{5.4}$$

Proof. To obtain the desired upper bound, we will apply the McDiarmid's inequality (2.15). In view of (4.7), the corresponding function f has the form

$$f\left(y_1,\ldots,y_{2n}\right)=\sup_{\substack{\varphi\in\mathscr{H}_{G,K}\\ D\in\mathscr{H}_{D,K}}}\left|f_{\varphi,D}\left(y_1,\ldots y_{2n}\right)\right|$$

with

$$f_{\varphi,D}\left(y_1,\ldots,y_{2n}\right)=\frac{1}{2n}\sum_{i=1}^{n}\log D\left(y_i\right)+\frac{1}{2n}\sum_{i=1}^{n}\log\left[1-D\left(\varphi\left(y_{n+i}\right)\right)\right]-L\left(\varphi,D\right).$$

We need to estimate the oscillations of $f\left(y_1,\ldots y_{2n}\right)$ with respect to its each argument. Let us do that for the first argument. Using (3.5), (3.6), (3.7) and the triangle inequality, we have

$$\begin{aligned}f\left(y_1'',y_2,\ldots,y_{2n}\right) &= \sup_{\substack{\varphi\in\mathscr{H}_{G,K}\\ D\in\mathscr{H}_{D,K}}}\left|f_{\varphi,D}\left(y_1',y_2,\ldots,y_{2n}\right)+\frac{1}{2n}\log D\left(y_1''\right)-\frac{1}{2n}\log D\left(y_1'\right)\right|\\ &\leq \sup_{\substack{\varphi\in\mathscr{H}_{G,K}\\ D\in\mathscr{H}_{D,K}}}\left|f_{\varphi,D}\left(y_1',y_2,\ldots,y_{2n}\right)\right|+\frac{1}{2n}\sup_{D\in\mathscr{H}_{D,K}}\left[\left|\log D\left(y_1''\right)\right|+\left|\log D\left(y_1'\right)\right|\right]\\ &\leq f\left(y_1',y_2,\ldots,y_{2n}\right)+\frac{1}{2n}\cdot\left(-2\log B_1\right),\end{aligned}$$

hence

$$f\left(y_1'',y_2,\ldots,y_{2n}\right)-f\left(y_1',y_2,\ldots,y_{2n}\right)\leq-\frac{\log B_1}{n}.$$

Thus, the oscillation of $f\left(y_1,\ldots y_{2n}\right)$ with respect to its first argument does not exceed $-\frac{\log B_1}{n}$; and the same is true for all other arguments of f. Hence (2.15) gives

$$\mathbb{P}\left(\varepsilon_{sampling}\left(n\right)-\mathbb{E}\left[\varepsilon_{sampling}\left(n\right)\right]\geq t\right)\leq e^{-\frac{2n^2t^2}{2n\log^2 B_1}}=e^{-\frac{nt^2}{\log^2\left(1+d!K^{d+1}\right)}} \quad \left(t\geq 0\right).$$

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

The latter, together with (5.1), implies

$$\mathbb{P}\left(\varepsilon_{sampling}\left(n\right) \geq t + \gamma K^{4(d+1)} n^{\delta-\frac{1}{2}}\right) \leq e^{-\frac{nt^2}{\log^2(1+d!K^{d+1})}} \quad \left(t, \delta > 0; n = 1, 2, \dots\right),$$

and taking $t = \gamma K^{4(d+1)} n^{\delta-\frac{1}{2}}$, we obtain the desired estimate. $\square$

In the next theorem we allow K to increase with the sample size n. This enables us to fulfill the requirement $\phi \in \mathscr{H}_{G,K}$ for large n without knowing K explicitly. To indicate the dependence of the quantities $\varepsilon_{sampling}\left(n\right)$ and $\hat{\varphi}_n$ on K, we shall use for them the notations $\varepsilon^{K}_{sampling}\left(n\right)$ and $\hat{\varphi}_{n,K}$, respectively.

Theorem 5.4. *Let $0 < \alpha \leq 1$, and let d and k be positive integers such that $k > 1 - \alpha + \frac{d^2}{2}$. Then $\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right)$ and $d_{JS}\left(f_\mu, f_{\hat{\varphi}_{n,n^{\frac{1}{32(d+1)}}}}\right)$ converge to 0 almost surely, as $n \to \infty$.*

Proof. In (5.4) putting $K = n^{\frac{1}{32(d+1)}}$ and $\delta = \frac{1}{4}$, we obtain

$$\mathbb{P}\left(\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right) \geq 2\gamma n^{-\frac{1}{8}}\right) \leq \exp\left\{-\frac{\gamma^2 n^{\frac{3}{4}}}{\log^2\left(1 + d! n^{\frac{1}{32}}\right)}\right\}.$$

This estimate clearly implies the convergence of the series $\sum_{n=1}^{\infty} \mathbb{P}\left(\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right) \geq 2\gamma n^{-\frac{1}{8}}\right)$. In particular, we have

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right) \geq \frac{1}{m}\right) < \infty,$$

for evere positive integer m. The Borel–Cantelli lemma shows that

$$\mathbb{P}\left(\limsup_{n\to\infty}\left\{\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right) \geq \frac{1}{m}\right\}\right) = 0 \quad \left(m = 1, 2, \dots\right),$$

hence

$$P\left(\left\{\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right) \to 0\right\}\right) = 1.$$

In view of (4.5), the terms $L(\hat{\varphi}_n, D_{\hat{\varphi}_n})$, $L(\varphi^*, D_{\varphi^*})$ of the error decomposition given in (4.10) can be replaced by the Jensen-Shannon divergences $d_{JS}\left(f_\mu, f_{\hat{\varphi}_n}\right), d_{JS}\left(f_\mu, f_{\varphi^*}\right)$. Since Theorem 4.6 states that $f_\mu = f_{\varphi^*}$ and we know that φ^* is a member of $\mathscr{H}_{G,K}$ given by the inverse Rosenblatt transformation ϕ (see Proposition 3.8), (4.10) gives

$$d_{JS}\left(f_\mu, f_{\hat{\varphi}_{n,n^{\frac{1}{32(d+1)}}}}\right) \leq 2\varepsilon^{n^{\frac{1}{32(d+1)}}}_{sampling}\left(n\right),$$

therefore $d_{JS}\left(f_\mu, f_{\hat{\varphi}_{n,n^{\frac{1}{32(d+1)}}}}\right) \to 0$ almost surely, as $n \to \infty$. $\square$

Note that, under the additional condition $k > 1 - \alpha + \frac{d^2}{2}$, Theorems 4.5 and 4.6 follow from Theorem 5.4.

6 Conclusion and Outlook

In this work we have proven that generative adversarial learning can be successfully pursued in the large data limit for generators in $C^{k,\alpha}$-Hölder spaces if fulfilling a uniform bound. The crucial observation was the realizability of arbitrary distributions with $C^{k,\alpha}$-Hölder density, a suitable bound on the norm within the so-defined hypothesis space and a consistent formulation of the hypothesis space of discriminators. The key technical ingredients were a thorough investigation into the analytical properties of the Rosenblatt transformation based on an inverse function theorem for Hölder spaces, which seems not to be known in the literature. At the same time, Hölder spaces provide us with very flexible possibilities for pre-compact embeddings, both qualitatively in $C^{k,\alpha'}$-spaces and quantitatively in L^∞ where explicit estimates on covering numbers are known. Both kinds of embeddings were exploited in the proofs of learnability with varying regularity requirements, without and with explicit convergence rates.

From the insight generated through this work, some new of lines of research seem to be promising. While this paper focuses on the theory of infinite dimensional generative learning, the understanding of generative adversarial networks can profit from such an analysis. In fact, deep (and shallow) neural networks (D(S)NN) on the one hand have the universal approximation property which in recent times has also been studied quantitatively [29, 19], including deep convolutional neural networks [30, 18], too. Thus, the approximation of the Rosenblatt transformation by DNN seems to be feasible. DNN with smooth sigmoid activation functions could provide a 'conformal' approximation within $C^{k,\alpha}$. Note that the rates of [29] are yet to be established for this class of networks. ReLU Networks would correspond to an exterior, 'non-conformal' approximation, where the estimate of the sampling error would require a revision in order to obtain convergence rates that are independent of the network's weight count. While both strategies seem feasible, a certain level of technicalities is to be expected to achieve their implementation. Obviously, the 'non-conformal' strategy would be very valuable for the understanding of the success of the contemporary GAN technology.

Secondly, the 'triangular' structure of the Rosenblatt transformation implies that the learning problem could be split into a hierarchy of consecutive learning problems starting from one input and output dimension and proceeding to j-dimensional input in the j-th learning step, whereas always only one additional dimension of output has to be learned at a time. Note that this does not necessarily have to happen in a lexicographical order of input channels, but one could easily combine this with linear transformations that consecutively train multiscale hierarchies like wavelet coefficients of images. Also this triangular structure is more easy to invert numerically, which might be of interest in the construction of a cycle-GAN (in moderate dimension).

Acknowledgment. H.A. acknowledges financial support from bergisch.smart.mobility funded by the Ministry of Economic Affairs, Innovation, Digitalization and Energy of the State of North Rhine-Westphalia.

References

[1] D. H. Ackley, G. E. Hinton and T. J. Sejnowski. “A Learning Algorithm for Boltzmann Machines”. In: *Connectionist Models and Their Implications: Readings from Cognitive Science*. USA: Ablex Publishing Corp., 1988, pp. 285–307.

Preprint – Preprint – Preprint – Preprint – Preprint

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

[2] M. Arjovsky, S. Chintala and L. Bottou. "Wasserstein Generative Adversarial Networks". In: ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 214–223.

[3] G. Biau, B. Cadre, M. Sangnier and U. Tanielian. "Some theoretical properties of GANs". In: *Ann. Statist.* 48.3 (2020), pp. 1539–1566. DOI: 10.1214/19-AOS1858.

[4] G. Biau, M. Sangnier and U. Tanielian. *Some Theoretical Insights into Wasserstein GANs.* 2020. arXiv: 2006.02682 [cs.LG].

[5] V. I. Bogachev. *Measure theory. Vol. I, II.* Springer, Berlin, 2007. DOI: 10.1007/978-3-540-34514-5.

[6] A. Brock, J. Donahue and K. Simonyan. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *CoRR* abs/1809.11096 (2018). arXiv: 1809.11096.

[7] N. Dunford and J. T. Schwartz. *Linear operators. Part I.* John Wiley & Sons, New York, 1988.

[8] T. S. Ferguson. *A Course in Large Sample Theory.* Boca Raton (Fl): Chapman & Hall, 1996.

[9] Z. Ghahramani. "Unsupervised Learning". In: *Advanced Lectures on Machine Learning: ML Summer Schools 2003, Canberra, Australia, February 2 - 14, 2003, Tübingen, Germany, August 4 - 16, 2003, Revised Lectures.* Ed. by O. Bousquet, U. von Luxburg and G. Rätsch. Springer, Berlin-Heidelberg, 2004, pp. 72–112. DOI: 10.1007/978-3-540-28650-9_5.

[10] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order.* Springer, Berlin, 2001.

[11] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models.* Vol. 40. USA: Cambridge University Press, 2016.

[12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27.* Curran Associates, Montreal, 2014, pp. 2672–2680.

[13] R. van Handel. *Probability in High Dimension.* APC 550 Lecture Notes. Princeton University, 2016.

[14] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pp. 97–109. DOI: 10.1093/biomet/57.1.97. eprint: http://biomet.oxfordjournals.org/cgi/reprint/57/1/97.pdf.

[15] S. Kullback. *Information theory and statistics.* Dover Publications, Mineola, 1997.

[16] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al. "Photo-realistic single image super-resolution using a generative adversarial network". In: *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2017, pp. 4681–4690.

[17] C. McDiarmid. "On the method of bounded differences". In: *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference.* Ed. by J. Siemons. London Mathematical Society Lecture Note Series. Cambridge University Press, 1989, pp. 148–188. DOI: 10.1017/CBO9781107359949.008.

Preprint – Preprint – Preprint – Preprint – Preprint – Preprint

[18] P. Petersen and F. Voigtlaender. "Equivalence of approximation by convolutional neural networks and fully-connected networks". In: *Proceedings of the American Mathematical Society* 148.4 (2020), pp. 1567–1581.

[19] P. Petersen and F. Voigtlaender. "Optimal approximation of piecewise smooth functions using deep ReLU neural networks". In: *Neural Networks* 108 (2018), pp. 296–330.

[20] A. Radford, L. Metz and S. Chintala. *Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks*. 2016. arXiv: 1511.06434 [cs.LG].

[21] M. Rosenblatt. "Remarks on a multivariate transformation". In: *Ann. Math. Statistics* 23 (1952), pp. 470–472. DOI: 10.1214/aoms/1177729394.

[22] W. Rudin. *Functional analysis*. McGraw-Hill, New York, 1991.

[23] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, New York, 1976.

[24] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. USA: Cambridge University Press, 2014.

[25] A. N. Shiryaev. *Probability-1*. Springer, New York, 2016.

[26] C. Shui, Q. Chen, J. Wen, F. Zhou, C. Gagné and B. Wang. *Beyond $\mathcal{H}$-Divergence: Domain Adaptation Theory With Jensen-Shannon Divergence*. 2020. arXiv: 2007.15567 [cs.LG].

[27] M. E. Taylor. *Partial differential equations I. Basic theory*. Springer, New York, 2011. DOI: 10.1007/978-1-4419-7055-8.

[28] A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, New York, 1996. DOI: 10.1007/978-1-4757-2545-2.

[29] D. Yarotsky. "Error bounds for approximations with deep ReLU networks". In: *Neural Networks* 94 (2017), pp. 103–114.

[30] D. Yarotsky. "Universal approximations of invariant maps by neural networks". In: *CoRR* abs/1804.10306 (2018). arXiv: 1804.10306.

[31] J. Zhu, T. Park, P. Isola and A. A. Efros. "Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks". In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 2242–2251. DOI: 10.1109/ICCV.2017.244.