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*
e-mail: asatryan@math.uni-wuppertal.de*;* hanno.gottschalk@uni-wuppertal.de
marieke-lippert@gmx.de*;* rottmann@math.uni-wuppertal.de

[2]*Institute of Mathematical Modelling, Analysis and Computational Mathematics (IMACM)*

[3]*Interdisciplinary Center for Machine Learning and Data Analytics (IZMD)*

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. We assume that the probability density functions of the underlying measure are uniformly bounded, k-times α-Hölder differentiable ($C^{k,\alpha}$) and uniformly bounded away from zero. Under these assumptions, 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 the Jensen-Shannon divergence between the distribution induced by the generator from the adversarial learning procedure and the data generating distribution converges to zero. Under certain regularity assumptions on the density of the data generating process, we also provide rates of convergence based on chaining and concentration.

Keywords and phrases: Generative adversarial learning, inverse Rosenblatt transformation, statistical learning theory, chaining, covering numbers for Hölder spaces.
MSC2020 subject classifications: Primary 62G20; secondary 68T05.

1. Introduction

Generative learning aims at modeling 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 [16] and has resulted in 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), (variational) autoencoders (AE) and generative adversarial networks (GAN), see [12, 1, 15, 18]. 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 [15] 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 [26, 22, 7, 42].

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 modeling 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 for image synthesis [26, 22, 7]. A remarkable improvement was obtained by the so-called CycleGAN framework [42]. 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 parameterized) 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 [21].

In contrast to the rapidly progressing developments on the empirical side, theoretical results on generative adversarial learning[1] remained unexamined

---

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

until the publication [4]. 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 modeled in a rather abstract way, the authors also provide approximation arguments. In 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.

Adversarial learning has also been applied to density estimation in the nonparametric context [33, 36]. Universal approximation theorems for deep neural networks [40] are used to approximate densities. This is however very different from [15] and also from this work, as no generator for sampling from noise input is learned. In [33] generators are considered, however the existence of such generators in certain Besov spaces is assumed (cf. Assumption A4 in Theorem 9), whereas in this article we give proofs and an explicit construction for Hölder spaces.

Further theoretical work on generative adversarial learning that was recently published includes theory on Wasserstein methods [5], as well as theory on domain shifts quantified by means of an adversarial loss that reduces Jensen-Shannon divergence [32].

In this work we extend the approach of [4] 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 [4]. We achieve this proving the Rosenblatt transformation [27] 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 [30].

Statistical learning theory to a large extent depends on compactness properties of the hypothesis space. In infinite dimensional statistics, where the hypotheses are parameterized by functions in bounded regions of some – say Banach – function space [37, 14, 38], the use of the Banach topology is prohibited by Riesz's theorem which characterizes locally compact Banach spaces (see, e.g., [10, 29]). 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 [13]). With the aid of the uniform law of large numbers over compact spaces [11], 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, assuming sufficiently high Hölder regularity of the density $k+\alpha$, this statement can also be made quantitative in order to prove explicit rates of convergence $\sim n^{-\frac{1}{2}}$ up to polylogarithms. Under the same assumptions as ours, [8] derived a rate $n^{-\frac{k+\alpha}{2(k+\alpha)+d}}$ (up to polylogarithms) of convergence for a Wasserstein-like $\mathcal{F}$-divergence based on Hölder functions, which, in the limit $k \to \infty$, matches our rates. However, this $\mathcal{F}$-divergence is weaker than the Jensen-Shannon divergence employed here. On the other hand, [8] invoke neural networks with architectures adaptive to the sample size, whereas here we restrict ourselves to an infinite dimensional statistics setting. In a recent preprint, [3] suggest rates $\sim n^{-1}$ (up to polylogarithms) in the limit of arbitrarily high regularity.

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 [37, 23]). These rates, in contrast to [4], 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 Jensen-Shannon distance, also for this case.

While this article is motivated by generative adversarial learning, our results on the existence of generators might be of independent interest in the fields of copula theory [17], normalizing flows [9, 19] or optimal transport [6, 39].

Outline. This paper is organized as follows: In Section 2 we introduce the framework of adversarial learning [15] and review the connection of its loss function to the Jensen-Shannon divergence. We then make use of the latter to state a decomposition of the estimation error. Thereafter in Section 3, first state our assumptions on the target measure, before we introduce the hypothesis class of k-times α-Hölder 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 the hypothesis space. Furthermore, we introduce the consistent class of discriminators that also contains the optimal one. Furthermore, 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, in Section 4 we give explicit rate estimates based on covering numbers for the embedding of the hypothesis space into $L^{\infty}\left([0,1]^{d}\right)$,

which is used to eliminate certain regularity assumptions. We give a summary and outline future research directions in the final Section 5. Technical results on the analysis of Hölder spaces can be found in Appendix A, whereas Appendix B collects some inequalities from high dimensional probability theory for the convenience of the reader.

2. An outline of generative adversarial learning

In this section, we give an overview into the general framework of generative adversarial learning following [15]. Generative learning aims learning a process that samples from an unknown original measure μ from data Y_i sampled from this measure ($Y_i \sim \mu$). In practice, generative learning is oftentimes utilized for sake of generating realistic images $y \in [0,1]^d$, $d \in \mathbb{N}$. In that context, let μ be a given probability measure on the measurable space $\left([0,1]^d, \mathscr{B}[0,1]^d\right)$ with $\mathscr{B}[0,1]^d$ being the Borel-σ-algebra, and let λ denote the restriction of the d-dimensional Lebesgue measure to $\left([0,1]^d, \mathscr{B}[0,1]^d\right)$.

A strategy that is oftentimes applied in generative learning is to learn a so-called *generator*, namely a measurable function $\varphi : [0,1]^d \to [0,1]^d$ from a *hypothesis space* of candidate functions $\mathscr{H}^G$, such that the distribution $\mu_\varphi = \varphi_* \lambda$ of $\varphi(Z)$ approximates the original measure μ and Z is easy to sample. Here, $\varphi_* \lambda(A) = \lambda(\varphi^{-1}(A))$, $A \in \mathscr{B}[0,1]^d$, stands for the pushforward or image measure of λ under φ. Ideally, Z is a noise random variable $Z \sim \lambda$ with λ being the Lebesgue measure. Simulation of Z allows then for generating data $\varphi(Z)$ from the approximate distribution μ_φ that resembles the original data Y drawn from μ.

Adversarial learning compares original data obtained by Y with generated data obtained by $\varphi(Z)$ and the learning process (performed by stochastic optimization) modifies φ such that examples of $\varphi(Z)$ become indistinguishable from generated examples obtained by $\varphi(Z)$. This is achieved with the help of a so-called *discriminator*, a function $\xi : [0,1]^d \to (0,1)$ which is supposed to estimate the probability of its input stemming from the original measure μ vs. its input stemming from μ_φ. As a suitable discriminator is not known *a priori*, it has to be learned from the data by maximizing the log-likelihood

$$\hat{L}(\varphi, \xi, n) = \frac{1}{2n} \sum_{i=1}^n \log \xi(Y_i) + \frac{1}{2n} \sum_{i=1}^n \log\left[1 - \xi(\varphi(Z_i))\right] \qquad (2.1)$$

over a hypothesis space of discriminators $\mathscr{H}^D$. Here and in the following Y_i are i.i.d. copies of Y and Z_i i.i.d. copies of Z. The generative adversarial learning problem (see also Fig. 1) then amounts to the solution of the minimax problem

$$\min_{\varphi \in \mathscr{H}^G} \max_{\xi \in \mathscr{H}^D} \hat{L}(\varphi, \xi). \qquad (2.2)$$

The advantage of generative adversarial learning lies in the fact that (2.1) avoids the representation of μ_φ by its density, which is hard to compute, unless one employs restrictive neural network architectures [35].
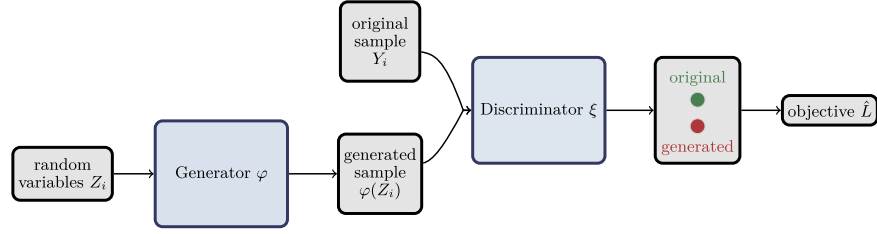
Fig 1. *Illustration of the adversarial learning framework studied in the present manuscript. The discriminator aims at tagging samples from Y_i as original and generated samples from $\varphi(Z_i)$ as generated, based on the estimated probabilities $\xi(\varphi(Z_i))$ and $\xi(Y_i)$. Blue components of the figure are the learnable components subject to optimization.*

Despite generative adversarial learning is intuitive, a deeper theoretical understanding is desirable to explain why it actually works. To this aim, we have to define what it means to have only a small difference between μ_φ and μ. To measure the dissimilarity between two probability distributions, we will use the Kullback-Leibler and the Jensen-Shannon divergences (see [20]). These divergences will be helpful to evaluate a learned measure imitating the given one.

Let ν and μ be two equivalent (i.e., absolutely continuous with respect to one another) probability measures on $[0,1]^d$. Recall that the *Kullback-Leibler divergence* is defined by the formula $d_{KL}(\nu||\mu) := \int \log \frac{d\nu}{d\mu} d\nu$, 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 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 $\nu = \mu$. Here and in the following the integration is over $[0,1]^d$ if not specified.

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 2.1. The *Jensen-Shannon divergence* is defined by the formula

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

In the following we relate (2.1) and the Jensen-Shannon divergence. By the law of large numbers, under suitable conditions (2.1) almost surely converges to

$$L(\varphi,\xi) = \frac{1}{2}\left(\mathbb{E}_{Y\sim\mu}\left[\log(\xi(Y))\right] + \mathbb{E}_{Z\sim\lambda}\left[\log(1 - \xi(\varphi(Z)))\right]\right). \tag{2.3}$$

Thus, the minimax optimization problem (2.2) can be seen as an approximation to the minimax problem for $L(\varphi,\xi)$, namely

$$\min_{\varphi\in\mathscr{H}^G} \max_{\xi\in\mathscr{H}^D} L(\varphi,\xi). \tag{2.4}$$

The following proposition relates the solution of (2.4) to the search of μ_φ which is closest to μ in terms of the Jensen-Shannon divergence (cf. [15]).

Proposition 2.2. *Suppose μ is absolutely continuous with respect to the Lebesgue measure λ and its density f_μ is positive on $[0,1]^d$. Additionally, we assume that μ_φ for $\varphi \in \mathscr{H}^G$ also fulfills the same condition with associated density $f_\varphi > 0$. Finally, we assume that for $\varphi \in \mathscr{H}^G$, $\xi_\varphi = \frac{f_\mu}{f_\mu + f_\varphi} \in \mathscr{H}^D$ holds.*

Then, ξ_φ is the unique solution (up to redefinition on null sets of λ) to

$$\max_{\xi \in \mathscr{H}^D} L(\varphi, \xi) = d_{JS}(\mu, \mu_\varphi) - \log(2).$$

Proof. By setting

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

we rewrite (2.3) in the form

$$L(\varphi, \xi) = \int h(f_\mu, f_\varphi, \xi) \, d\lambda \quad \left(\varphi \in \mathscr{H}^G, \xi \in \mathscr{H}^D \right).$$

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, in view of the assumption $\xi_\varphi \in \mathscr{H}^D$, we have

$$h(f_\mu, f_\varphi, \xi) \leq h(f_\mu, f_\varphi, \xi_\varphi) \quad \left(\varphi \in \mathscr{H}^G, \xi \in \mathscr{H}^D \right).$$

which shows that $\xi_\varphi = \frac{f_\mu}{f_\mu + f_\varphi}$ maximizes $L(\varphi, \cdot)$. Moreover, if $E := \{ x : \xi(x) \neq \xi_\varphi(x) \}$ is a set of positive Lebesgue measure, it follows that

$$h(f_\mu, f_\varphi, \xi)|_E < h(f_\mu, f_\varphi, \xi_\varphi)|_E;$$

and therefore $L(\varphi, \xi) < L(\varphi, \xi_\varphi)$, which shows that ξ is no maximizer.

Finally, writing the Radon-Nikodym derivative as a quotient of densities, we obtain

$$
\begin{aligned}
L(\varphi, \xi_\varphi) &= \frac{1}{2} \left[\int \log \left(\frac{f_\mu}{f_\mu + f_\varphi} \right) f_\mu \, d\lambda + \int \log \left(\frac{f_\varphi}{f_\mu + f_\varphi} \right) f_\varphi \, d\lambda \right] \\
&= \frac{1}{2} \left[\int \log \left(\frac{2 f_\mu}{f_\mu + f_\varphi} \right) f_\mu \, d\lambda + \int \log \left(\frac{2 f_\mu}{f_\mu + f_\varphi} \right) f_\varphi \, d\lambda \right] - \log(2) \\
&= d_{JS}(\mu, \mu_\varphi) - \log(2). \qquad \qquad \square
\end{aligned}
$$

In the next step, we compare the solution of the accessible minimax problem (2.2) to the actual minimization of the Jensen-Shannon divergence by solution of (2.4). This leads to a decomposition of the estimation error made in generative adversarial learning. A similar estimate can be found in [4].

Theorem 2.3. *Suppose that the conditions of Proposition 2.2 hold and that $\left(\hat{\varphi}, \hat{\xi} \right)$ solve (2.2). Then,*

$$d_{JS}(\mu, \mu_{\hat{\varphi}}) \leq \varepsilon_{\text{model}} + 2\varepsilon_{\text{sample}}(n),$$

where $\varepsilon_{\mathrm{model}} := \inf_{\varphi \in \mathscr{H}^G} d_{JS}(\mu, \mu_\varphi)$ *is the model error and*

$$\varepsilon_{\mathrm{sample}}(n) := \sup_{\substack{\varphi \in \mathscr{H}^G \\ \xi \in \mathscr{H}^D}} \left| L(\varphi, \xi) - \hat{L}(\varphi, \xi) \right|$$

is the sampling error at the sample size n.

Proof. We first note that

$$L(\varphi^*, \xi_{\varphi^*}) = \min_{\varphi \in \mathscr{H}^G_{K, \hat{K}}} \max_{\xi \in \mathscr{H}^D_{B, C_1, C_2}} L(\varphi, \xi) \leq \max_{\xi \in \mathscr{H}^D_{B, C_1, C_2}} L(\hat{\varphi}_n, \xi) = L(\hat{\varphi}_n, \xi_{\hat{\varphi}_n}),$$

where the last equality follows from Proposition 2.2. Next, in view of the definition of the sampling error, we have

$$\begin{aligned}
L(\hat{\varphi}_n, \xi_{\hat{\varphi}_n}) &\leq \hat{L}(\hat{\varphi}_n, \xi_{\hat{\varphi}_n}, n) + \varepsilon_{\mathrm{sample}}(n) \\
&= \max_{\xi \in \mathscr{H}^D} \hat{L}(\hat{\varphi}_n, \xi, n) + \varepsilon_{\mathrm{sample}}(n) \\
&= \min_{\varphi \in \mathscr{H}^G} \max_{\xi \in \mathscr{H}^D} \hat{L}(\varphi, \xi, n) + \varepsilon_{\mathrm{sample}}(n) \\
&\leq \max_{\xi \in \mathscr{H}^D} \hat{L}(\varphi^*, \xi, n) + \varepsilon_{\mathrm{sample}}(n) \\
&\leq \max_{\xi \in \mathscr{H}^D} \left[L(\varphi^*, \xi) + \varepsilon_{\mathrm{sample}}(n) \right] + \varepsilon_{\mathrm{sample}}(n) \\
&= L(\varphi^*, \xi_{\varphi^*}) + 2\varepsilon_{\mathrm{sample}}(n),
\end{aligned}$$

where $\varphi^* \in \mathscr{H}^G$ is arbitrary. Using the representation of the Jensen-Shannon divergence given in Proposition 2.2, we can subtract $\log(2)$ on both sides of the above inequality and take the infimum over $\varphi^* \in \mathscr{H}^G$ to conclude. $\square$

Let us remark that due to the rather strong assumptions on $\mathscr{H}^D$ in Proposition 2.2, we do not get model errors for discriminators.

Let us also remark that [3] suggests a different estimate of the error, which avoids taking the supremum over $\mathscr{H}^G$ and $\mathscr{H}^D$ in the definition of $\varepsilon_{\mathrm{sample}}(n)$.

In the following section, we give our assumptions on the density f_μ of the measure μ in terms of Hölder regularity and choose $\mathscr{H}^G$ such that the model error $\varepsilon_{\mathrm{model}}$ can be reduced to zero. That is, we show that in the hypothesis space there exist generators φ that fulfill $\mu_\varphi = \mu$ exactly. We furthermore show that one can choose $\mathscr{H}^D$ such that the assumptions of Proposition 2.2 and 2.3 are fulfilled. Then in Section 4, we consider the convergence $\varepsilon_{\mathrm{sample}}$ quantitatively.

3. Generators, discriminators and consistency

In this section, we first state our assumptions on the regularity of the density f_μ in terms of Hölder differentiability and we define suitable hypothesis spaces for generators and discriminators. Several technical results on Hölder functions are collected in Appendix A, where we also prove an inverse function theorem for Hölder spaces as an important auxiliary result.

3.1. Assumptions

For a nonnegative integer k (or $k = \infty$), $C^k\left((0,1)^d, \mathbb{R}^{d_2}\right)$ stands for the set of all $\mathbb{R}^{d_2}$-valued functions with continuous k-th order derivatives and $C^k\left([0,1]^d, \mathbb{R}^{d_2}\right)$ is the set of all $\mathbb{R}^{d_2}$-valued functions whose k-th derivatives have continuous extensions to $[0,1]^d$ (or, equivalently, the k-th derivatives are uniformly continuous on $(0,1)^d$). $C^k\left([0,1]^d, \mathbb{R}^{d_2}\right)$ is a Banach space with respect to the norm

$$\|f\|_{C^k([0,1]^d, \mathbb{R}^{d_2})} := \max_{|\beta| \le k} \sup_{x \in [0,1]^d} |D_\beta f(x)| ;$$

here $\beta = (\beta_1, \ldots, \beta_{d_1}) \in \mathbb{Z}^{d_1}$ is a multi-index, $|\beta| := \sum_{i=1}^{d_1} \beta_i$ is its absolute value and $D_\beta f(x) := \frac{\partial^{|\beta|} f(x)}{\partial x_1^{\beta_1} \cdots \partial x_{d_1}^{\beta_{d_1}}}$. The Fréchet derivative of f at x will be denoted by $Df(x)$. As usual, the space of continuous functions C^0 and its norm will be denoted by C and $\|\cdot\|_\infty$, respectively.

The definition and many of the basic properties of Hölder spaces of real-valued functions, given in [13], extend without difficulty to vector-valued ones:

Definition 3.1. Let $0 < \alpha \le 1$, and let k be a nonnegative integer. The *Hölder space* $C^{k,\alpha}\left([0,1]^d, \mathbb{R}^{d_2}\right)$ consists of all functions $f \in C^k\left([0,1]^d, \mathbb{R}^{d_2}\right)$ for which the following norm is finite:

$$\|f\|_{C^{k,\alpha}([0,1]^d, \mathbb{R}^{d_2})} := \|f\|_{C^k([0,1]^d, \mathbb{R}^{d_2})} + \max_{|\beta|=k} \sup_{\substack{x,y \in (0,1)^d \\ x \ne y}} \frac{|D_\beta f(x) - D_\beta f(y)|}{|x-y|^\alpha}.$$

Our major requirement on the original measure is the following

Assumption 3.2. μ *is absolutely continuous with respect to* λ, *and the Radon-Nikodym derivative* $f_\mu := \frac{d\mu}{d\lambda}$ *is in* $C^{k,\alpha}\left([0,1]^d\right)$ *for some* $\alpha \in (0,1]$ *and* $k \ge 1$. *Furthermore,* f_μ *satisfies the condition*

$$\kappa := \min_{x \in [0,1]^d} f_\mu(x) > 0.$$

Remark 3.3. While the above assumptions seem restrictive, for any given μ one can easily define an approximate problem, such that Assumption 3.2 is fulfilled. To this avail, $[0,1]^d$ is embedded in $\mathbb{R}^d$, convolved with $N(0, \kappa^2 \mathbb{1})$ 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 $N \sim N(0, \kappa^2 \mathbb{1})$ 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.

3.2. The hypothesis space of generators

Our goal is to define suitable hypothesis spaces for generators such that $\mu = f_\mu \lambda$ is realizable under the Assumption 3.2. On the other hand, the hypothesis space should not be too large, in order to keep control on $\varepsilon_{\text{sample}}(n)$, which is the same as avoidance of overfitting.

Definition 3.4. Let K and $\hat{K}$ be positive constants. The set of all $k \geq 1$ times α-Hölder differentiable bijective functions $\varphi : [0,1]^d \to [0,1]^d$, satisfying the conditions $\|\varphi\|_{C^{k,\alpha}} \leq K$ and $\left\|\varphi^{-1}\right\|_{C^{k,\alpha}} \leq \hat{K}$, will be denoted by $\mathscr{H}_{K,\hat{K}}^G$.

The spaces $\mathscr{H}_{K,\hat{K}}^G$ are candidates for generator spaces, provided the constants K and $\hat{K}$ are sufficiently large. Before doing so, we convince ourselves that the requirements of Proposition 2.2 concerning the existence of densities are fulfilled. Also, we have a closer look at the regularity of the densities in terms of their Hölder norms.

Proposition 3.5. $\mathscr{H}_{K,\hat{K}}^G$ *is a closed subset of the space* $C^{k,\alpha}\left([0,1]^d, \mathbb{R}^d\right)$. *For each* $\varphi \in \mathscr{H}_{K,\hat{K}}^G$, *the pushforward measure* $\mu_\varphi = \varphi_* \lambda$ *is absolutely continuous with respect to* λ, *and its density* f_φ *is a* $(k-1)$-*times* α-*Hölder differentiable function on* $[0,1]^d$. *The density* f_φ *is given by the formula*

$$f_\varphi = \left|J_{\varphi^{-1}}\right| \tag{3.1}$$

where $J_{\varphi^{-1}}$ *is the Jacobian determinant of* φ^{-1}. *The following estimate bounds* f_φ *in terms of* K *and* $\hat{K}$ *from above and below*

$$\frac{1}{d!K^d} \leq f_\varphi \leq d!\hat{K}^d. \tag{3.2}$$

Moreover,

$$\|f_\varphi\|_{C^{k-1,\alpha}} \leq 2d!d^{k+1}\hat{K}^d.$$

Proof. Using (A.15), for every $\varphi \in \mathscr{H}_{K,\hat{K}}^G$, we may estimate

$$|J_\varphi| \leq d! \max_{x \in [0,1]^d}\left[\max_{1 \leq i,j \leq d}\left|\frac{\partial \varphi_i(x)}{\partial x_j}\right|\right]^d \leq d!K^d, \tag{3.3}$$

and, similarly,

$$\left|J_{\varphi^{-1}}\right| \leq d!\hat{K}^d. \tag{3.4}$$

The latter, in view of the equality $J_\varphi = \frac{1}{J_{\varphi^{-1} \circ \varphi}}$, gives $|J_\varphi| \geq \frac{1}{d!\hat{K}^d}$. Therefore the statement (b) of Theorem A.11 shows that $\mathscr{H}_{K,\hat{K}}^G$ is closed in $C^{k,\alpha}\left([0,1]^d, \mathbb{R}^d\right)$.

The change-of-variables formula shows that

$$\varphi_* \lambda\left(A\right) = \lambda\left(\varphi^{-1}\left(A\right)\right) = \int_{\varphi^{-1}(A)} d\lambda = \int_A \left|J_{\varphi^{-1}}\right| d\lambda,$$

for any $\varphi \in \mathscr{H}_{K,\hat{K}}^G$ and $A \in \mathscr{B}\left([0,1]^d\right)$; hence each $\varphi \in \mathscr{H}_{K,\hat{K}}^G$ generates a measure $\mu_\varphi = \varphi_* \lambda$ with a density $f_\varphi = \left|J_{\varphi^{-1}}\right|$. Equations (3.3) and (3.4), together with the equality $J_{\varphi^{-1}} = \frac{1}{J_\varphi \circ \varphi^{-1}}$, imply the estimate (3.2).

In view of Corollary A.6,

$$\left\|J_{\varphi^{-1}}\right\|_{C^{k-1,\alpha}} \leq 2d!d^k \max\left\{1, [\mathrm{diam}([0,1]^d)]^{1-\alpha}\right\} \left\|\varphi^{-1}\right\|_{C^{k,\alpha}}^d \leq 2d!d^{k+1}\hat{K}^d,$$

which completes the proof. $\qquad \square$

Compactness is an important property for hypothesis spaces as it plays a central role in the uniform law of large numbers (see e.g. [11]) that is used to control $\varepsilon_{\text{sample}}(n)$ in the limit $n \to \infty$. However, in view of Stone's theorem, we cannot expect compactness for the space $\mathscr{H}_{K,\hat{K}}^G$ in the $C^{k,\alpha}$-topology. Therefore it is crucial that the hypothesis space of generators is compact with respect to a topology that is just a little bit weaker:

Proposition 3.6. $\mathscr{H}_{K,\hat{K}}^G$ *is compact in* $C^{k,\alpha'}\left([0,1]^d, \mathbb{R}^d\right)$ *for* $0 < \alpha' < \alpha$.

Proof. Indeed, $\mathscr{H}_{K,\hat{K}}^G$ is a bounded subset of the space $C^{k,\alpha}\left([0,1]^d, \mathbb{R}^d\right)$. The embedding of $C^{k,\alpha}\left([0,1]^d, \mathbb{R}^d\right)$ into $C^{k,\alpha'}\left([0,1]^d, \mathbb{R}^d\right)$ is compact [13, Lemma 6.33], hence $\mathscr{H}_{K,\hat{K}}^G$ is relatively compact in $C^{k,\alpha'}\left([0,1]^d\right)$. Obviously, the uniform bound K with respect to the $\|\cdot\|_{C^{k,\alpha}}$-norm is stable under $C^{k,\alpha'}$ limits $\varphi_n \to \varphi \in C^{k,\alpha'}([0,1]^d, \mathbb{R}^d)$. Using the inverse function theorem for Hölder spaces (Theorem A.11 (b)), we see that $\|\varphi^{-1}\|_{C^{k,\alpha}} = \lim_n \|\varphi_n^{-1}\|_{C^{k,\alpha}} \leq \hat{K}$. Hence, $\varphi \in \mathscr{H}_{K,\hat{K}}^G$ and thus this space is closed in $C^{k,\alpha'}\left([0,1]^d, \mathbb{R}^d\right)$ and therefore is compact. $\qquad\square$

In the next step, we prove the existence of a generator for μ and compute Hölder norms of the generator and its inverse. Our construction is based on the Rosenblatt transformation [27], which recursively computes a transformation based on conditional cumulative distribution functions and their inverse, i.e. conditional quantile functions.

Recall the definition of the conditional density (see [31, Sec. 2.7]). Let $Y_j : [0,1]^d \to \mathbb{R}^{d_j}$ $(j = 1, 2)$ be random variables such that the pair $Y := (Y_1, Y_2)$ has a density $f_Y(y_1, y_2)$. Then Y_2 has a marginal density given by the formula $f_{Y_2}(y_2) = \int_{\mathbb{R}^{d_1}} f_Y(y_1, y_2)\, dy_1$ $\left(y_j \in \mathbb{R}^{d_j}; j = 1, 2\right)$. We define the conditional density $f_{Y_1|Y_2}(\cdot|y_2) = f_{Y_1|Y_2=y_2}(\cdot)$ by the formula

$$f_{Y_1|Y_2}(y_1|y_2) = \begin{cases} \dfrac{f_Y(y_1, y_2)}{f_{Y_2}(y_2)} & \text{if} \quad f_{Y_2}(y_2) > 0, \\ 0 & \text{if} \quad f_{Y_2}(y_2) = 0. \end{cases} \tag{3.5}$$

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

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

After these preparations, we first find the inverse of the generator that generates samples from μ. Proving Hölder differentiability for the inverse along with a lower bound for the Jacobian determinant enables us to apply the inverse function theorem for Hölder functions, which we prove in the appendix (see Theorem

---

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

[A.11]). For now, let $Y = (Y_1, \ldots, Y_d)$ be a random vector with distribution μ satisfying Assumption 3.2. 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 $(y_1, \ldots, y_d) \in [0,1]^d$ (see [27]). The inequalities $0 \le F_{Y_1} \le 1$ and $0 \le F_{Y_j|(Y_1,\ldots,Y_{j-1})} \le 1$ $(2 \le j \le d)$ show that $\psi\left([0,1]^d\right) \subset [0,1]^d$.

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

Proposition 3.7. *Under Assumption 3.2, the Rosenblatt transformation ψ has the following properties:*

(i) ψ is k-times α-Hölder differentiable, i.e., $\psi \in C^{k,\alpha}\left([0,1]^d\right)$.
(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]
$$y_1 = F_{Y_1}^{-1}(x_1),$$
$$y_2 = F_{Y_2|Y_1}^{-1}(x_2|y_1),$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$y_d = F_{Y_d|(Y_1,\ldots,Y_{d-1})}^{-1}(x_d|(y_1,\ldots,y_{d-1})).$$

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

Proof. In view of Assumption 3.2, Proposition A.5 and Lemma A.7, 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 conditional cumulative density 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, $\psi \in C^{k,\alpha}\left([0,1]^d, \mathbb{R}^d\right)$.

In order to prove the bijectivity of ψ, we choose an arbitrary point $x = (x_1, \ldots, x_d) \in [0,1]^d$ and show that the equation $\psi(y) = x$ has a unique solution. Put $\psi_1(y_1) := F_{Y_1}(y_1)$ $(y_1 \in [0,1])$ and

$$\psi_j(y_1, \ldots, y_j) := F_{Y_j|(Y_1,\ldots,Y_{j-1})}(y_j|(y_1,\ldots,y_{j-1})) \quad (y_1, \ldots, y_j \in [0,1]),$$

for $2 \le j \le d$. We rewrite the definition of ψ in the form

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

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

Assumption 3.2 ensures that

$$\frac{\partial \psi_1(y_1)}{\partial y_1} = f_{Y_1}(y_1) > 0 \quad (y_1 \in (0,1)), \tag{3.6}$$

$$\frac{\partial \psi_j(y_1, \ldots, y_j)}{\partial y_j} = f_{Y_j|(Y_1, \ldots, Y_{j-1})}(y_j|(y_1, \ldots, y_{j-1})) > 0, \tag{3.7}$$

$$(y_1, \ldots, y_j \in (0,1); 2 \le j \le d)$$

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 $\frac{\partial \psi_i}{\partial y_j} = 0$ $(1 \le i < j \le d)$, i.e., the Jacobian matrix of ψ is lower triangular. Hence

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

(3.5) and (3.7) 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 (3.6), 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. $\qquad\square$

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 3.7 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 for sufficiently large K and $\hat{K}$ the inverse Rosenblatt transformation ϕ is a generator for the measure μ. Note that the procedure described in Remark 3.3 actually allows to compute suitable K and $\hat{K}$ from the parameters of the regularization.

Theorem 3.8. *If μ satisfies Assumption 3.2, then there exists a generator $\phi \in \mathscr{H}^G_{K,\hat{K}}$ such that $\mu = \phi_* \lambda$ holds, provided K and $\hat{K}$ are sufficiently large. In this case, the model error in Theorem 2.3 vanishes, i.e. $\varepsilon_{\mathrm{model}} = 0$.*

Proof. Theorems 3.7 and A.11 show that $\phi = \psi^{-1} \in C^{k,\alpha}\left([0,1]^d, [0,1]^d\right)$. Next, for any $A \in \mathscr{B}\left([0,1]^d\right)$, the change-of-variables theorem, together with the statement (iii) of Theorem 3.7, gives

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

Remark 3.9. Related results on the existence of generators can also be obtained using optimal transport, see e.g. [39], or rearrangement techniques [6]. It might be of interest that our rather elementary analysis of the Rosenblatt transformation leads to comparable results on the existence and regularity of the generator.

The special, lower triangular shape of the inverse Rosenblatt transformation considerably simplifies the numerical computation of $f_\varphi = |J_{\varphi^{-1}}|$, which would be of interest for a likelihood based (non adversarial) training of the generators, as it is done for normalizing flows [9, 19], in particular in the context of inevitable neural networks [35]. Also, our findings could be of interest in the theory of copulae [17].

3.3. The hypothesis space of discriminators

Here we present the hypothesis space of discriminators $\mathscr{H}^D_{B,C_1,C_2}$ which is affiliated to the set of generators in the sense that if the data generating measure μ is realizable in $\mathscr{H}^G_{K,\hat{K}}$, then the optimal discriminator that separates data from $\phi_* \lambda$ and μ, is realizable in $\mathscr{H}^D_{B,C_1,C_2}$, see Theorem 3.13 below. We therefore show that in our setting the assumptions of Theorem 2.3 hold and we can bound the Jensen Shannon divergence of the true measure with the sampling error.

Definition 3.10. Let $B \in \left(0, \frac{1}{2}\right)$ and $C_i > 0$ $(i = 1, 2)$. The collection of all functions $\xi \in C^{k-1,\alpha}\left([0,1]^d, \mathbb{R}\right)$, satisfying the conditions $B \leq \xi \leq 1 - B$, $\|D\xi\|_\infty \leq C_1$ and $\|\xi\|_{C^{k-1,\alpha}([0,1]^d,\mathbb{R})} \leq C_2$, is called the hypothesis space of discriminators and is denoted by $\mathscr{H}^D_{B,C_1,C_2}$.

Again, to achieve control of the sample error, compactness plays a crucial role. Therefore we establish it in the following proposition:

Proposition 3.11. *$\mathscr{H}^D_{B,C_1,C_2}$ is compact in $C^{k-1,\alpha'}\left([0,1]^d, \mathbb{R}\right)$ for $0 < \alpha' < \alpha$.*

Proof. The argument provided in the proof of Proposition 3.6 carries over to this much simpler case. $\qquad \square$

In the next step, we compute constants B, C_1 and C_2 as functions of K and $\hat{K}$, which guarantees that the assumptions of Proposition 2.2 and hence those of

Theorem 2.3 with respect to the generators are fulfilled. The following lemma, which heavily relies on our analysis of Hölder functions in Appendix A, provides these technical results:

Lemma 3.12. *Let* $\varphi, \varphi' \in \mathscr{H}^G_{K,\hat{K}}$, *and let* $\xi_{\varphi,\varphi'} := \frac{f_\varphi}{f_\varphi + f_{\varphi'}}$. *If K and $\hat{K}$ satisfy the conditions $4d!d^{k+1}\hat{K}^d > 1$ and $d!K^d > 2$, then $\xi_{\varphi,\varphi'} \in \mathscr{H}^D_{B,C_1,C_2}$ with $B = \left[1 + (d!)^2 K^d \hat{K}^d\right]^{-1}$, $C_1 = 2d^3 (d!)^4 K^{2d} \hat{K}^{2d+1}$ and $C_2 = \left[4(d!)^2 d^{k+2} K^d \hat{K}^d\right]^{2^{k+4}}$.*

Proof. Proposition 3.5 and Lemma A.7 show that $\xi_{\varphi,\varphi'} \in C^{k-1,\alpha}\left([0,1]^d, \mathbb{R}\right)$. Since f_φ and $f_{\varphi'}$ admit the same estimate (3.2), hence

$$\frac{1}{(d!)^2 K^d \hat{K}^d} \le \frac{f_{\varphi'}}{f_\varphi} \le (d!)^2 K^d \hat{K}^d.$$

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

$$B := \frac{1}{1 + (d!)^2 K^d \hat{K}^d}.$$

To obtain an upper bound for $\|D\left(\xi_{\varphi,\varphi'}\right)\|_\infty$, we will first find a Lipschitz estimate for the function $\xi_{\varphi,\varphi'}$. We have

$$\xi_{\varphi,\varphi'}(x_1) - \xi_{\varphi,\varphi'}(x_2) = \frac{f_{\varphi'}(x_2)\left[f_\varphi(x_1) - f_\varphi(x_2)\right] - f_\varphi(x_2)\left[f_{\varphi'}(x_1) - f_{\varphi'}(x_2)\right]}{\left[f_\varphi(x_1) + f_{\varphi'}(x_1)\right]\left[f_\varphi(x_2) + f_{\varphi'}(x_2)\right]}.$$

The estimate (3.2) 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 (3.1) and (3.2), yields

$$\begin{aligned}
|\xi_{\varphi,\varphi'}(x_1) - \xi_{\varphi,\varphi'}(x_2)| &\le d!K^d \sum_{h \in \{\varphi,\varphi'\}} f_h(x_1)f_h(x_2)\left|\frac{1}{f_h(x_1)} - \frac{1}{f_h(x_2)}\right| \\
&\le (d!)^3 K^d \hat{K}^{2d} \sum_{h \in \{\varphi,\varphi'\}} \left|J_h\left(h^{-1}(x_1)\right) - J_h\left(h^{-1}(x_2)\right)\right|.
\end{aligned}$$

By induction on d, one can easily prove the inequality

$$|a_1 \cdots a_d - b_1 \cdots b_d| \le d\left[\max_{1 \le j \le d}\{|a_j|, |b_j|\}\right]^{d-1} \max_{1 \le j \le d}|a_j - b_j| \tag{3.8}$$

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

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

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

$$|J_h(y_1) - J_h(y_2)| \leq d \cdot d! K^{d-1} |(Dh)(y_1) - (Dh)(y_2)| \quad (y_1, y_2 \in [0,1]^d).$$

The latter, combined with the estimate $\|Dh\| \leq d \|h\|_{C^1} \ \left(h \in C^1([0,1]^d)\right)$ and with the mean value theorem for vector-valued functions, gives

$$|J_h(y_1) - J_h(y_2)| \leq d \cdot d! K^{d-1} \cdot dK |y_1 - y_2| \quad (y_1, y_2 \in [0,1]^d),$$

therefore

$$|\xi_{\varphi,\varphi'}(x_1) - \xi_{\varphi,\varphi'}(x_2)| \leq d^2 (d!)^4 K^{2d} \hat{K}^{2d} \sum_{h \in \{\varphi, \varphi'\}} \left| h^{-1}(x_1) - h^{-1}(x_2) \right|$$
$$\leq 2d^3 (d!)^4 K^{2d} \hat{K}^{2d+1} |x_1 - x_2| \quad (x_1, x_2 \in [0,1]^d).$$

This inequality implies that $\|D\xi_{\varphi,\varphi'}(x)\| \leq 2d^3 (d!)^4 K^{2d} \hat{K}^{2d+1} \ \left(x \in (0,1)^d\right)$. Indeed, choose an arbitrary $x \in (0,1)^d$, then $x \in (t, 1-t)^d$ for all sufficiently small $t > 0$, and we have

$$\|D(\xi_{\varphi,\varphi'})(x)\| = \sup_{|h|=1} \frac{|D(\xi_{\varphi,\varphi'})(x)ht|}{|ht|} \leq \sup_{|h|=1} \frac{|\xi_{\varphi,\varphi'}(x+ht) - \xi_{\varphi,\varphi'}(x)|}{|ht|}$$
$$+ \sup_{|h|=1} \frac{|D(\xi_{\varphi,\varphi'})(x)ht - [\xi_{\varphi,\varphi'}(x+ht) - \xi_{\varphi,\varphi'}(x)]|}{|ht|}$$
$$\leq \sup_{|h|=1} \frac{|D(\xi_{\varphi,\varphi'})(x)ht - [\xi_{\varphi,\varphi'}(x+ht) - \xi_{\varphi,\varphi'}(x)]|}{|ht|}$$
$$+ 2d^3 (d!)^4 K^{2d} \hat{K}^{2d+1}.$$

Letting $t \to 0$, we obtain the desired inequality.

In view of Proposition 3.5, $\|f_\varphi\|_{C^{k-1,\alpha}} \leq 2d! d^{k+1} \hat{K}^d$, $\|f_\varphi + f_{\varphi'}\|_{C^{k-1,\alpha}} \leq 4d! d^{k+1} \hat{K}^d$ and $f_\varphi + f_{\varphi'} \geq \frac{2}{d! K^d}$. Since $4d! d^{k+1} \hat{K}^d > 1$ and $\frac{d! K^d}{2} > 1$, we may apply Proposition A.9 to obtain

$$\|\xi_{\varphi,\varphi'}\|_{C^{k-1,\alpha}} \leq \left[4(d!)^2 d^{k+1} \hat{K}^d K^d\right]^{2^{k+4}} \max\left\{1, [\text{diam}([0,1]^d)]^{2(1-\alpha)}\right\} \leq C_2.$$

which completes the proof. $\qquad\square$

We can now wrap up our results in the following theorem, which shows that we can drop model errors with regard to generators and discriminators:

Theorem 3.13. *Suppose the Assumptions 3.2 hold and the constants K and $\hat{K}$ are sufficiently large. Furthermore, let B, C_1 and C_2 be given as in Lemma 3.12. Then the assumptions of Theorem 2.3 hold and*

$$d_{JS}(\mu, \mu_{\hat{\varphi}}) \leq 2\varepsilon_{\text{sample}}(n),$$

where $\hat{\varphi}$ solves the minimax problem (2.2).

Proof. Combine the statements of 2.3, Theorem 3.8 and Lemma 3.12. $\qquad\square$

3.4. Consistency of generative adversarial learning

From the construction of discriminators we have seen that the hypothesis spaces $\mathscr{H}_{K,\hat{K}}^G$ and $\mathscr{H}_{B,C_1,C_2}$ are large enough to contain optimal generators and discriminators, respectively, provided the constants $K, \hat{K}, B, C_1$ and C_2 are properly chosen. Therefore we conclude the consistency of generative adversarial learning by a simple application of the uniform law of large numbers:

Theorem 3.14. *Suppose that $k \geq 1$, K nd $\hat{K}$ are sufficiently large and let B, C_1 and C_2 be given as in Lemma 3.12 and let $\mathscr{H}^G = \mathscr{H}_{K,\hat{K}}^G$ and $\mathscr{H}^D = \mathscr{H}_{B,C_1,C_2}^D$. Let $(\hat{\varphi}_n, \hat{\xi}_n)$ solve the minimax Problem 2.2 for sample size n. Then*

$$d_{JS}(\mu, \mu_{\hat{\varphi}_n}) \to 0$$

holds almost surely, i.e. generative adversarial learning is consistent with respect to the Jensen-Shannon divergence.

Proof. By theorem 3.13, it is enough to prove $\varepsilon_{\mathrm{sample}}(n) \to 0$ almost surely. By the Propositions 3.6 and 3.11, for $0 < \alpha' < \alpha$ the hypothesis spaces $\mathscr{H}^G$ and $\mathscr{H}^D$ are compact in the $C^{k,\alpha'}$ and $C^{k-1,\alpha'}$-topology, and thus their direct product $\mathscr{H}^G \times \mathscr{H}^D$ is compact as well.

Furthermore, we note that the single summands $L(\varphi, \xi) - \frac{1}{2}[\log(\xi(Y_i)) + \log(1 - \xi(\phi(Z_i)))$ in the definition of $\varepsilon_{\mathrm{sample}}(n)$ are independent and identically distributed.

In addition, they and uniformly bounded by $2 \max\{-log(B), -\log(1 - B)\}$. Furthermore it is easy to see that these expressions are continuous in (φ, ξ) with respect to the $C^{k,\alpha'} \times C^{k-1,\alpha'}$-topology. Hence, the conditions of the uniform law of large numbers, see e.g. [11], hold and the claim $\varepsilon_{\mathrm{sample}}(n) \to 0$ (a.s.) follows from this theorem. □

4. Quantitative estimates of the sampling error

In the previous Theorem 3.14 we have proven that the sampling error converges to 0 almost surely and that this implies the a.s. convergence of the generates distribution in the Jensen-Shannon distance. This worked for rather weak assumptions on the regularity. In particular, the regularity assumptions $k \geq 1$, $0 < \alpha \leq 1$ were independent of the dimension d. However, we have not obtained any results on the rate of convergence.

The main objective of this section is to estimate the rate of that convergence. To achieve this, we make strong assumptions on the differentiability of generators and discriminators. Especially, if the dimension d is high – in practical applications d can be in the millions – we have to assume that the density f_μ of μ, the generators φ and discriminators ξ are 'almost' C^∞-functions.

To achieve an upper bound for the expectation of the sampling error

$\varepsilon_{\text{sample}}(n)$, i.e., for the quantity

$$\mathbb{E}\left[\sup_{\substack{\varphi\in\mathscr{H}^G_{K,\hat{K}}\\\xi\in\mathscr{H}^D_{B,C_1,C_2}}}\left|\hat{L}(\varphi,\xi,n)-L(\varphi,\xi)\right|\right],$$

we apply the Dudley estimate (B.1) to the random processes $\pm\left[\hat{L}(\varphi,\xi,n)-L(\varphi,\xi)\right]$. The so-called empirical process over $\mathscr{H}^G_{K,\hat{K}}\times\mathscr{H}^D_{B,B_1,C_2}$,

$$\mathscr{H}^G_{K,\hat{K}}\times\mathscr{H}^D_{B,B_1,C_2}\ni(\varphi,\xi)\mapsto\hat{L}(\varphi,\xi,n)-L(\varphi,\xi),$$

clearly is centered. We already know from the previous section that $L(\cdot,\cdot)$ and $\hat{L}(\cdot,\cdot,n)$ are continuous on $\mathscr{H}^G_{K,\hat{K}}\times\mathscr{H}^D_{B,C_1,C_2}$. To apply Dudley's inequality, it remains to check the subgaussian property, see Definition B.2, of the increments of the empirical process $\hat{L}(\varphi,\xi,n)-L(\varphi,\xi)$.

Lemma 4.1. *For each positive integer* n, *the increment* $\hat{L}(\varphi_1,\xi_1,n)-L(\varphi_1,\xi_1)-\hat{L}(\varphi_2,\xi_2,n)+L(\varphi_2,\xi_2)$ *is* $\left[\rho_n\left((\varphi_1,\xi_1),(\varphi_2,\xi_2)\right)\right]^2$-*subgaussian with*

$$\rho_n\left((\varphi_1,\xi_1),(\varphi_2,\xi_2)\right)\coloneqq\frac{2}{B\sqrt{n}}\left(\|\xi_1-\xi_2\|_\infty+C_1\|\varphi_1-\varphi_2\|_\infty\right).$$

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

$$\hat{L}(\varphi_1,\xi_1,n)-\hat{L}(\varphi_2,\xi_2,n)=\frac{1}{2n}\sum_{i=1}^n\left[\log\xi_1(Y_i)-\log\xi_2(Y_i)\right]$$

$$+\frac{1}{2n}\sum_{i=1}^n\left\{\log\left[1-\xi_1(\varphi_1(Z_i))\right]-\log\left[1-\xi_2(\varphi_2(Z_i))\right]\right\},$$

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

Next, we will estimate the deviation $\left\|\hat{L}(\varphi_1,\xi_1,1)-\hat{L}(\varphi_2,\xi_2,1)\right\|_\infty$. We have

$$\left|\hat{L}(\varphi_1,\xi_1,1)-\hat{L}(\varphi_2,\xi_2,1)\right|\leq\frac{1}{2}\left|\log\xi_1(Y_1)-\log\xi_2(Y_1)\right|$$

$$+\frac{1}{2}\left|\log\left(1-\xi_1\left(Y_1^{(\varphi_1)}\right)\right)-\log\left(1-\xi_2\left(Y_1^{(\varphi_1)}\right)\right)\right|$$

$$+\frac{1}{2}\left|\log\left(1-\xi_2\left(Y_1^{(\varphi_1)}\right)\right)-\log\left(1-\xi_2\left(Y_1^{(\varphi_2)}\right)\right)\right|.$$

The inequalities $B\leq\xi_i\leq1-B$ $(i=1,2)$, together with the elementary estimate

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

yield

$$\left| \hat{L}\left(\varphi_1, \xi_1, 1\right) - \hat{L}\left(\varphi_2, \xi_2, 1\right) \right| \leq \frac{1}{2B} \left[\left| \xi_1\left(Y_1\right) - \xi_2\left(Y_1\right) \right| \right.$$
$$\left. + \left| \xi_1\left(Y_1^{(\varphi_1)}\right) - \xi_2\left(Y_1^{(\varphi_1)}\right) \right| + \left| \xi_2\left(Y_1^{(\varphi_1)}\right) - \xi_2\left(Y_1^{(\varphi_2)}\right) \right| \right].$$

The first and the second summands on the right hand side do not exceed $\|\xi_1 - \xi_2\|_\infty$. To find an upper bound for the third summand, we first note that

$$|\xi_2(x_1) - \xi_2(x_2)| \leq \|D\xi_2\|_\infty \cdot |x_1 - x_2| \leq C_1 |x_1 - x_2| \quad \left(x_1, x_2 \in [0,1]^d\right),$$

hence $\|\xi_2 \circ \varphi_1 - \xi_2 \circ \varphi_2\|_\infty \leq C_1 \|\varphi_1 - \varphi_2\|_\infty$. Thus, we arrive at the estimate

$$\begin{aligned}
\left| \hat{L}\left(\varphi_1, \xi_1, 1\right) - \hat{L}\left(\varphi_2, \xi_2, 1\right) \right| &\leq \frac{1}{B} \left[\|\xi_1 - \xi_2\|_\infty + C_1 \|\varphi_1 - \varphi_2\|_\infty \right] \\
&= \frac{1}{2} \rho_1\left(\left(\varphi_1, \xi_1\right), \left(\varphi_2, \xi_2\right)\right).
\end{aligned}$$

From this estimate, the estimate $|L(\varphi_1, \xi_1) - L(\varphi_2, \xi_2)| \leq \frac{1}{2}\rho_1\left(\left(\varphi_1, \xi_1\right), \left(\varphi_2, \xi_2\right)\right)$ immediately follows by taking the expected value on the left hand side. This, together with Hoeffding's Lemma (see [38, Sec. 3.1]), shows that $\hat{L}\left(\varphi_1, \xi_1, 1\right) - \hat{L}\left(\varphi_2, \xi_2, 1\right) - L(\varphi_1, \xi_1) + L(\varphi_2, \xi_2)$ is $\left[\rho_1\left(\left(\varphi_1, \xi_1\right), \left(\varphi_2, \xi_2\right)\right)\right]^2$-subgaussian. $\square$

An essential component in the proof of our main result (see Theorem 4.5) 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 $C\left(\overline{U}\right)$.

For a metric space (T, ρ) and for $\varepsilon > 0$, $N\left(T, \rho, \varepsilon\right)$ will denote the *covering number* of T, i.e., the minimum cardinality of an ε-net for T. The closed ball of radius ε and center $x \in T$ will be denoted by $B_T\left[x, \varepsilon\right]$ (or simply $B\left[x, \varepsilon\right]$). Below we will consider covering numbers only for Cauchy-precompact metric spaces; this, in view of Hausdorff's theorem (see, e.g., [10, Lemma I.6.15]), guarantees the finiteness of all covering numbers.

Theorem 4.2. *If $U \subset \mathbb{R}^{d_1}$ is bounded, open and convex, then*

$$\log N\left(B_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)}[0,1], \|\cdot\|_\infty, \varepsilon\right) \leq \gamma_1 \varepsilon^{-\frac{d_1}{\alpha+k}}, \tag{4.1}$$

where γ_1 is a constant depending only on d_1, d_2 and $\alpha + k$.

Proof. In the case $d_2 = 1$ the estimate (4.1) is proved in [37, Th. 2.7.1]. The general case follows without difficulty; to construct an ε-net with desired properties, we first embed the unit ball $B_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)}[0,1]$ into $\left[B_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)}[0,1]\right]^{d_2}$ and then cover the ball $B_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)}[0,1]$ by the balls $B^\infty_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)}\left[f_j, \frac{\varepsilon}{\sqrt{d_2}}\right]$ $(j \in J)$ in the space $\left(C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right), \|\cdot\|_\infty\right)$ such that the cardinality $|J|$ of the index set J

satisfies the relation

$$\log |J| = \log N \left(B_{C^{k,\alpha}(\overline{U},\mathbb{R})}[0,1], \|\cdot\|_\infty, \tfrac{\varepsilon}{\sqrt{d_2}} \right) \leq \gamma_1^* \left(\tfrac{\varepsilon}{\sqrt{d_2}} \right)^{-\frac{d_1}{\alpha+k}}$$

with a constant γ_1^*, depending only on d_1 and $\alpha + k$. For each $l = (j_1, \ldots, j_{d_2}) \in J^{d_2}$, let $B_{C^{k,\alpha}(\overline{U},\mathbb{R}^{d_2})}^\infty [f^l, \varepsilon]$ denote the ball of radius ε and centered at $f^l := (f_{j_1}, \ldots, f_{j_{d_2}})$ in the space $(C^{k,\alpha}(\overline{U}, \mathbb{R}^{d_2}), \|\cdot\|_\infty)$. Then

$$\left[B_{C^{k,\alpha}(\overline{U},\mathbb{R})}[0,1] \right]^{d_2} \subset \bigcup_{j_1,\ldots,j_{d_2} \in J} \prod_{m=1}^{d_2} B_{C^{k,\alpha}(\overline{U},\mathbb{R})}^\infty \left[f_{j_m}, \tfrac{\varepsilon}{\sqrt{d_2}} \right]$$

$$\subset \bigcup_{l \in J^{d_2}} B_{C^{k,\alpha}(\overline{U},\mathbb{R}^{d_2})}^\infty \left[f^l, \varepsilon \right],$$

$$\log \left| J^{d_2} \right| = d_2 \log |J| \leq \gamma_1^* d_2^{1 + \frac{d_1}{2(\alpha+k)}} \varepsilon^{-\frac{d_1}{\alpha+k}},$$

and we may take $\gamma_1 := \gamma_1^* d_2^{1 + \frac{d_1}{2(\alpha+k)}}$. $\qquad\qquad\qquad\qquad$ □

Dudley's metric entropy estimate (B.1) requires the square root of the right hand side in (4.1) to be integrable at zero. From this, we derive our regularity requirements and obtain a rate estimate for the expected value of the sampling error.

Proposition 4.3. *If $k > 1 - \alpha + \frac{d}{2}$, then there exists a positive constant γ, depending only on d, α, and k, such that*

$$\mathbb{E}\left[\varepsilon_{\mathrm{sample}}(n) \right] \leq \frac{\gamma}{B} \max \{C_1 K, C_2\} n^{-\frac{1}{2}} \quad (n = 1, 2, \ldots), \qquad (4.2)$$

for each $K > 0, \hat{K} > 0, C_i > 0 \ (i = 1, 2)$ and $B \in \left(0, \frac{1}{2} \right)$.

Proof. Since $\pm \left[\hat{L}(\varphi, \xi, n) - L(\varphi, \xi) \right]$ is a continuous subgaussian processes on the compact space $(\Theta; \rho_n) = \left(\mathcal{H}_{K,\hat{K}}^G \times \mathcal{H}_{B,C_1,C_2}^D; \rho_n \right)$, the entropy bound (B.1) is applicable to $\hat{L}(\varphi, \xi, n) - L(\varphi, \xi)$, which gives

$$\mathbb{E}\left[\varepsilon_{\mathrm{sample}}(n) \right] \leq 12 \int_0^\infty \sqrt{\log N(\Theta, \rho_n, \varepsilon)} d\varepsilon = 12 \int_0^\infty \sqrt{\log N(\Theta, \rho_1, \sqrt{n}\varepsilon)} d\varepsilon$$

$$= \frac{12}{\sqrt{n}} \int_0^\infty \sqrt{\log N(\Theta, \rho_1, \varepsilon)} d\varepsilon. \qquad (4.3)$$

It is easy to see that

$$N(\Theta, \rho_1, \rho_{r_1,r_2}) \leq N \left(\mathcal{H}_{K,\hat{K}}^G, \|\cdot\|_\infty, r_1 \right) N \left(\mathcal{H}_{B,C_1,C_2}^D, \|\cdot\|_\infty, r_2 \right)$$

$$\leq N \left(B_{C^{k,\alpha}}[0,K], \|\cdot\|_\infty, r_1 \right) N \left(B_{C^{k-1,\alpha}}[0,C_2], \|\cdot\|_\infty, r_2 \right),$$

where $B_{C^{l,\alpha}}[0, r]$ denotes the closed ball of radius $r > 0$ centered at 0 in the space $C^{l,\alpha}\left([0, 1]^d\right)$ $(l = k - 1, k)$ and $\rho_{r_1, r_2} := 4B^{-1}\left(r_2 + C_1 r_1\right)$. The equality

$$N\left(B_{C^{l,\alpha}([0,1]^d)}[0, r], \|\cdot\|_\infty, \varepsilon\right) = N\left(B_{C^{l,\alpha}([0,1]^d)}[0, 1], \|\cdot\|_\infty, \varepsilon r^{-1}\right) \quad (l = k - 1, k),$$

together with (4.1), yields

$$\log N\left(B_{C^{l,\alpha}([0,1]^d)}[0, r], \|\cdot\|_\infty, \varepsilon\right) \leq \gamma_1\left(d, \alpha + l\right)\left(r\varepsilon^{-1}\right)^{\frac{d}{\alpha+l}} \quad (r > 0).$$

Using this inequality, we may estimate further

$$\log N\left(\Theta, \rho_1, \rho_{r_1, r_2}\right) \leq \gamma_2\left(d, \alpha, k\right)\left[\left(Kr_1^{-1}\right)^{\frac{d}{\alpha+k}} + \left(C_2 r_2^{-1}\right)^{\frac{d}{\alpha+k-1}}\right]$$

with $\gamma_2\left(d, \alpha, k\right) := \max\left\{\gamma_1\left(d, \alpha + k\right), \gamma_1\left(d, \alpha + k - 1\right)\right\}$. Taking

$$r_1 = \frac{\varepsilon B}{4C_1}, \quad r_2 = \frac{\varepsilon B}{4}$$

and putting

$$\gamma_3 := \left[\gamma_2\left(d, \alpha, k\right)\max\left\{\left(4C_1 K B^{-1}\right)^{\frac{d}{\alpha+k}}, \left(4C_2 B^{-1}\right)^{\frac{d}{\alpha+k-1}}\right\}\right]^{\frac{1}{2}},$$

we obtain

$$\sqrt{\log N\left(\Theta, \rho_1, \varepsilon\right)} \leq \gamma_3\left[\varepsilon^{-\frac{d}{2(\alpha+k)}} + \varepsilon^{-\frac{d}{2(\alpha+k-1)}}\right]. \tag{4.4}$$

We would like to integrate both parts of the last inequality. In view of the condition $k > 1 - \alpha + \frac{d}{2}$, we have $\frac{d}{2(\alpha+k-1)} < 1$ which guaranties the integrability of the right hand side near 0. Put $\delta_1 := \operatorname{diam}\left(\mathscr{H}_{K,\hat{K}}^G \times \mathscr{H}_{B,C_1,C_2}^D; \rho_1\right)$. Since $N\left(\Theta, \rho_1, \varepsilon\right) = 1$ $(\varepsilon > \delta_1)$, hence (4.3) and (4.4) give

$$\mathbb{E}\left[\varepsilon_{\mathrm{sample}}\left(n\right)\right] \leq \frac{12\gamma_3}{\sqrt{n}}\int_0^{\delta_1}\left[\varepsilon^{-\frac{d}{2(\alpha+k)}} + \varepsilon^{-\frac{d}{2(\alpha+k-1)}}\right]d\varepsilon$$

$$= \frac{12\gamma_3}{\sqrt{n}}\left[\delta_1^{1-\frac{d}{2(\alpha+k)}} + \delta_1^{1-\frac{d}{2(\alpha+k-1)}}\right];$$

therefore we may take $\gamma_4 = 12\gamma_3\left[\delta_1^{1-\frac{d}{2(\alpha+k)}} + \delta_1^{1-\frac{d}{2(\alpha+k-1)}}\right]$ to obtain

$$\mathbb{E}\left[\varepsilon_{\mathrm{sample}}\left(n\right)\right] \leq \gamma_4 n^{-\frac{1}{2}} \quad (n = 1, 2, \ldots).$$

Since $\alpha + k - 1 > \frac{d}{2}$, we can estimate

$$\gamma_4 \leq 48B^{-1}\gamma_2^{\frac{1}{2}}\max\left\{C_1 K, C_2\right\}\left[\delta_1^{1-\frac{d}{2(\alpha+k)}} + \delta_1^{1-\frac{d}{2(\alpha+k-1)}}\right],$$

therefore we may choose $\gamma := 48\gamma_2^{\frac{1}{2}}\left[\delta_1^{1-\frac{d}{2(\alpha+k)}} + \delta_1^{1-\frac{d}{2(\alpha+k-1)}}\right]$. $\qquad\square$

We now exploit McDiarmid's concentration inequality (Theorem B.4) to obtain the following result in the spirit of PAC (probably approximately correct)-learning [30]:

Theorem 4.4. *Let $0 < \alpha \leq 1$, and let d and k be positive integers such that $k > 1 - \alpha + \frac{d}{2}$. There exists a positive constant γ, depending only on d, α, and k, such that for every $B \in \left(0, \frac{1}{2}\right), C_i > 0 \ (i = 1, 2)$, for each $K, \hat{K} > 0$ satisfying the condition $\phi \in \mathscr{H}_{K,\hat{K}}^G$, for each positive integer n and for every $t \geq 0$ the following estimate is true:*

$$\mathbb{P}\left(\varepsilon_{\text{sample}}(n) \geq t + \gamma B^{-1} \max\{C_1 K, C_2\} n^{-\frac{1}{2}}\right) \leq \exp\left(-\frac{nt^2}{\log^2 B}\right). \quad (4.5)$$

Proof. To obtain the desired upper bound, we will apply McDiarmid's inequality (B.2). In view of (2.1), the corresponding function f has the form

$$f(y_1, \ldots, y_{2n}) = \sup_{\substack{\varphi \in \mathscr{H}_{K,\hat{K}}^G \\ \xi \in \mathscr{H}_{B,C_1,C_2}^D}} |f_{\varphi,\xi}(y_1, \ldots y_{2n})|$$

with

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

We need to estimate the oscillations of $f(y_1, \ldots y_{2n})$ with respect to each of its arguments. Let us do that for the first argument. Using the triangle inequality and the estimate $B \leq \xi \leq 1 - B$, we have

$$f(y_1'', y_2, \ldots, y_{2n})$$

$$= \sup_{\substack{\varphi \in \mathscr{H}_{K,\hat{K}}^G \\ \xi \in \mathscr{H}_{B,C_1,C_2}^D}} \left| f_{\varphi,\xi}(y_1', y_2, \ldots, y_{2n}) + \frac{1}{2n} \log \xi(y_1'') - \frac{1}{2n} \log \xi(y_1') \right|$$

$$\leq \sup_{\substack{\varphi \in \mathscr{H}_{K,\hat{K}}^G \\ \xi \in \mathscr{H}_{B,C_1,C_2}^D}} |f_{\varphi,\xi}(y_1', y_2, \ldots, y_{2n})| + \frac{1}{2n} \sup_{\xi \in \mathscr{H}_{B,C_1,C_2}^D} [|\log \xi(y_1'')| + |\log \xi(y_1')|]$$

$$\leq f(y_1', y_2, \ldots, y_{2n}) + \frac{1}{2n} \cdot (-2 \log B),$$

hence

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

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

$$\mathbb{P}(\varepsilon_{\text{sample}}(n) - \mathbb{E}[\varepsilon_{\text{sample}}(n)] \geq t) \leq e^{-\frac{nt^2}{\log^2 B}} \quad (t \geq 0).$$

The latter, together with (4.2), implies the desired estimate. $\qquad \square$

In the next theorem, we let K and $\hat{K}$ increase with the sample size n and we thereby adaptively increase the capacity of the hypothesis space of generators. This enables us to fulfill the requirement $\phi \in \mathscr{H}^G_{K,\hat{K}}$ for large n without knowing $\kappa = \inf_{x \in [0,1]^d} f_\mu(x)$ and $\|f_\mu\|_{C^{k,\alpha}}$, the minimum value of the density and its norm, explicitly.

For each $K > 0$, the constants B, C_1, C_2 will be chosen as in Lemma 3.12. Thereby also the space of discriminators is enlarged such that the assumptions of Theorem 2.3 remain valid.

To indicate the dependence of the quantities $\varepsilon_{\text{sample}}(n)$ and $\hat{\varphi}_n$ on K and $\hat{K}$, we use the notation $\varepsilon^{K,\hat{K}}_{\text{sample}}(n)$ and $\hat{\varphi}_{n,K,\hat{K}}$, respectively.

Theorem 4.5. *Let $0 < \alpha \leq 1$, let $\beta > 0$, let d and k be positive integers satisfying the condition $k > 1 - \alpha + \frac{d}{2}$, and let $\left\{K_n\right\}_{n=1}^\infty, \left\{\hat{K}_n\right\}_{n=1}^\infty$ be sequences of positive numbers, tending to ∞, such that $K_n \hat{K}_n = (\log n)^{2^{-k-5}d^{-1}\beta}$ $(n \geq 3)$ with $\beta \geq 1$. Then for almost every ω there exists a positive integer $N = N(\omega)$ such that*

$$\varepsilon^{K_n,\hat{K}_n}_{\text{sample}} \leq \Gamma n^{-\frac{1}{2}} (\log n)^{(2^{-k-5}+2^{-1})\beta} \quad (n > N) \tag{4.6}$$

and

$$d_{JS}\left(f_\mu, f_{\hat{\varphi}_{n,K_n,\hat{K}_n}}\right) \leq 2\Gamma n^{-\frac{1}{2}} (\log n)^{(2^{-k-5}+2^{-1})\beta} \quad (n > N), \tag{4.7}$$

where Γ is a constant, depending only on d, α, and k.

Proof. We may assume, without loss of generality, that $K_n > 1$ and $\hat{K}_n > 1$ for every $n \in \mathbb{N}$. (4.5) gives

$$\mathbb{P}\left(\varepsilon^{K_n,\hat{K}_n}_{\text{sample}} \geq t + \gamma A_n n^{-\frac{1}{2}}\right) \leq \exp\left(-nt^2 \log^{-2} B_n\right) \quad (t \geq 0), \tag{4.8}$$

where $A_n = \left[1 + (d!)^2 K_n^d \hat{K}_n^d\right] \left[4(d!)^2 d^{k+2} K_n^d \hat{K}_n^d\right]^{2^{k+4}}$, $B_n = \left[1+(d!)^2 K_n^d \hat{K}_n^d\right]^{-1}$ (see Lemma 3.12). Clearly, we may assume that $\gamma \geq 1$. Taking $t = \gamma A_n n^{-\frac{1}{2}}$ in (4.8) yields

$$\mathbb{P}\left(\varepsilon^{K_n,\hat{K}_n}_{\text{sample}} \geq 2\gamma A_n n^{-\frac{1}{2}}\right) \leq \exp\left(-\gamma^2 A_n^2 \log^{-2} B_n\right)$$
$$\leq \exp\left(-A_n^2 \log^{-2} B_n\right) \quad (n \in \mathbb{N}).$$

In view of the inequality $\log(1 + x) < x$ $(x > 0)$,

$$A_n^2 \log^{-2} B_n \geq \left[4(d!)^2 d^{k+2} K_n^d \hat{K}_n^d\right]^{2^{k+5}} \geq 2 \log n,$$

which implies the convergence of the series

$$\sum_{n=1}^\infty \mathbb{P}\left(\varepsilon^{K_n,\hat{K}_n}_{\text{sample}} \geq 2\gamma A_n n^{-\frac{1}{2}}\right).$$

The Borel–Cantelli lemma shows that for $\mathbb{P}$-almost every random parameter ω there exists a positive integer $N = N(\omega)$ such that

$$\varepsilon^{K_n,\hat{K}_n}_{\text{sample}} < 2\gamma A_n n^{-\frac{1}{2}} \quad (n > N(\omega)).$$

The latter, together with the estimate

$$
\begin{aligned}
A_n &\leq 2(d!)^2 \big[4(d!)^2 d^{k+2}\big]^{2^{k+4}} \big(K_n \hat{K}_n\big)^{d(1+2^{k+4})} \\
&\leq 2(d!)^2 \big[4(d!)^2 d^{k+2}\big]^{2^{k+4}} (\log n)^{(2^{-k-5}+2^{-1})\beta},
\end{aligned}
$$

implies (4.6).

In view of (2.3), the terms $L(\hat{\varphi}_{n,K_n,\hat{K}_n}, \xi_{\hat{\varphi}_{n,K_n,\hat{K}_n}}), L(\varphi^*, \xi_{\varphi^*})$ of the error decomposition given in Theorem 2.3 can be replaced by the Jensen-Shannon divergences $d_{JS}\left(f_\mu, f_{\hat{\varphi}_{n,K_n,\hat{K}_n}}\right), d_{JS}(f_\mu, f_{\varphi^*})$. Since Theorem 3.13 states that $f_\mu = f_{\varphi^*}$ and we know that φ^* is a member of $\mathscr{H}^G_{K_n,\hat{K}_n}$ given by the inverse Rosenblatt transformation ϕ (see Proposition 3.8), Theorem 2.3 gives

$$
d_{JS}\left(f_\mu, f_{\hat{\varphi}_{n,K_n,\hat{K}_n}}\right) \leq 2\varepsilon^{K_n,\hat{K}_n}_{\text{sample}}(n),
$$

which, together with (4.6), implies (4.7). $\square$

Note that, under the additional condition $k > 1 - \alpha + \frac{d}{2}$, Theorem 3.14 follows from Theorem 4.5.

Remark 4.6. Let $0 < \alpha \leq 1$, and let d and k be positive integers satisfying the condition $k > \frac{d}{2} - \alpha$. Using approximation of densities by ReLU networks in Besov spaces, A. Uppal, S. Singh, and B. Póczos [36] assume the existence of a sequence of generators $\{\varphi_n\}$ with certain properties such that almost surely

$$
d_{JS}(f_\mu, f_{\varphi_n}) \leq C(d, \alpha, k) n^{-\frac{1}{2}} \sqrt{\log n}. \tag{4.9}
$$

In practical applications we usually deal with large data sets which correspond to a large dimension d. If the regularity k is increased by one, i.e., if k satisfies the condition $k > 1 + \frac{d}{2} - \alpha$, then, in view of Theorem 4.5, the estimate (4.7) holds. Note that the estimates (4.7) and (4.9) are essentially of the same nature. The advantage of Theorem 4.5 is that it describes a way to construct a sequence $\{\varphi_n\}_{n=1}^\infty$ with desired properties so that here we do not rely on assumptions. Furthermore, the rates for Hölder generators given here are much faster than the rate $\sim n^{-d^{-1}}$ obtained in [5] for the case of Wasserstein GAN, where the authors however do not require Hölder regularity of the density and in addition provide approximations with neural networks.

Recently, better rates of convergence up to $\sim n^{-1}$ were suggested in [3]. In fact, from the viewpoint of parametric statistics in finite dimensional spaces, where $\mu_{\hat{\theta}_n}$ estimates $\hat{\mu}_{\theta_0}$, a convergence $d_{JS}(\mu, \hat{\mu}) \sim n^{-\frac{1}{2}}$ is not optimal, as from maximum likelihood theory $\hat{\theta}_n - \theta_0 \sim n^{-\frac{1}{2}}$ (see e.g. [11]) and thus $d_{JS}(\mu_{\theta_0}, \mu_{\hat{\theta}_n}) \sim n^{-1}$ is possible in this setting, as $\theta' \mapsto d_{JS}(\mu_{\theta_0}, \mu_{theta'})$ will

approximately behave quadratic around the minimum θ_0. We therefore suggest that the rates obtained in this work can well be improved.

5. 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 of 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 precompact 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 research lines 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 [40, 24], including deep convolutional neural networks [41, 25], 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 [40] 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 CycleGAN [42] (in moderate dimension).

Appendix A: Analysis of Hölder functions

In this section we collect some analytical results that are not directly connected to generative adversarial learning. In particular, we derive certain technical bounds for products and quotients of Hölder differentiable functions, prove a version of the inverse function theorem for Hölder differentiable maps, which might be of independent interest.

A.1. General properties of Hölder functions

It is easy to see that $C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)$ is a Banach space (see [13, Sec. 4.1]); its members 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

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

moreover, this embedding is compact, i.e., each 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., [13, 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., [28, 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

$$C^k\left(\overline{U}, \mathbb{R}^{d_2}\right) \subset C^{k-1,1}\left(\overline{U}, \mathbb{R}^{d_2}\right) \quad (k \geq 1). \qquad (A.2)$$

Moreover, the embedding (A.2) is continuous; in fact,

$$\|f\|_{C^{k-1,1}\left(\overline{U}, \mathbb{R}^{d_2}\right)} \leq 2 \, \|f\|_{C^k\left(\overline{U}, \mathbb{R}^{d_2}\right)} \quad \left(f \in C^k\left(\overline{U}, \mathbb{R}^{d_2}\right)\right).$$

A.2. Bounds for products and quotients of differentiable functions

In this subsection we first compute C^k-bounds for products and quotients of differentiable functions in order to prepare the ground for the corresponding results for Hölder functions.

Proposition A.1. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a nonnegative integer, and let $f_i \in C^k\left(\overline{U}, \mathbb{R}\right) \; (1 \leq i \leq m)$. Then*

$$\|f_1 f_2 \cdots f_m\|_{C^k(\overline{U}, \mathbb{R})} \leq m^k \, \|f_1\|_{C^k(\overline{U}, \mathbb{R})} \, \|f_2\|_{C^k(\overline{U}, \mathbb{R})} \cdots \|f_m\|_{C^k(\overline{U}, \mathbb{R})}. \qquad (A.3)$$

Proof. For simplicity of notation, let us write C^k instead of $C^k\left(\overline{U}, \mathbb{R}\right)$. We proceed by induction on k. In the case $k = 0$ the statement is obvious. Now assume that the estimate is true for the norm $\|\cdot\|_{C^{k-1}}$ with $k \geq 1$. To estimate the quantity $\|f_1 f_2 \cdots f_m\|_{C^k}$, note that

$$\|h\|_{C^k} = \max\left\{ \|h\|_{C^{k-1}}, \left\|h'_{x_1}\right\|_{C^{k-1}}, \ldots, \left\|h'_{x_{d_1}}\right\|_{C^{k-1}} \right\} \quad \left(h \in C^k\right). \qquad (A.4)$$

For every $j = 1, 2, \ldots, d_1$ we have

$$\left\|(f_1 f_2 \cdots f_m)'_{x_j}\right\|_{C^{k-1}} = \left\|\sum_{i=1}^{m}\left(\frac{\partial f_i}{\partial x_j}\prod_{\substack{1 \le l \le m \\ l \ne i}} f_l\right)\right\|_{C^{k-1}} \le \sum_{i=1}^{m}\left\|\frac{\partial f_i}{\partial x_j}\prod_{\substack{1 \le l \le m \\ l \ne i}} f_l\right\|_{C^{k-1}}$$

$$\le m^{k-1}\sum_{i=1}^{m}\left\|\frac{\partial f_i}{\partial x_j}\right\|_{C^{k-1}}\prod_{\substack{1 \le l \le m \\ l \ne i}}\|f_l\|_{C^{k-1}}$$

$$\le m \cdot m^{k-1}\prod_{l=1}^{m}\|f_l\|_{C^k},$$

which, together with (A.4), yields the desired estimate for $\|f_1 f_2 \cdots f_m\|_{C^k}$. $\quad\square$

Lemma A.2. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a nonnegative integer, and let $v \in C^k\left(\overline{U}, \mathbb{R}\right)$ with $\inf\limits_{x \in U}|v(x)| > 0$. Then*

$$\left\|\tfrac{1}{v}\right\|_{C^k} \le 4^{2^k - k - 1}\|v\|_{C^k}^{2^k - 1}\left[\inf_{x \in U}|v(x)|\right]^{-2^k}.$$

Proof. The case $k = 0$ is trivial. Let $k \ge 1$. Applying (A.4) to $h = \frac{1}{v}$ gives

$$\left\|\tfrac{1}{v}\right\|_{C^m} = \max\left\{\left\|\tfrac{1}{v}\right\|_{C^{m-1}}, \left\|v'_{x_1}\cdot\tfrac{1}{v^2}\right\|_{C^{m-1}}, \ldots, \left\|v'_{x_{d_1}}\cdot\tfrac{1}{v^2}\right\|_{C^{m-1}}\right\} \quad (1 \le m \le k).$$

In view of Proposition A.1,

$$\left\|v'_{x_j}\cdot\tfrac{1}{v^2}\right\|_{C^{m-1}} \le 2^{m-1}\left\|v'_{x_j}\right\|_{C^{m-1}}\left\|\tfrac{1}{v^2}\right\|_{C^{m-1}} \le 4^{m-1}\|v\|_{C^m}\left\|\tfrac{1}{v}\right\|_{C^{m-1}}^2 \quad (j \le d_1),$$

which, together with the previous equality, yields

$$\left\|\tfrac{1}{v}\right\|_{C^m} \le \left\|\tfrac{1}{v}\right\|_{C^{m-1}}\max\left\{1, 4^{m-1}\|v\|_{C^m}\left\|\tfrac{1}{v}\right\|_{C^{m-1}}\right\}.$$

Proposition A.1 shows that

$$1 = \left\|v\cdot\tfrac{1}{v}\right\|_{C^{m-1}} \le 2^{m-1}\|v\|_{C^{m-1}}\left\|\tfrac{1}{v}\right\|_{C^{m-1}} \le 4^{m-1}\|v\|_{C^m}\left\|\tfrac{1}{v}\right\|_{C^{m-1}},$$

hence the estimate for $\left\|\tfrac{1}{v}\right\|_{C^m}$ reduces to

$$\left\|\tfrac{1}{v}\right\|_{C^m} \le 4^{m-1}\|v\|_{C^m}\left\|\tfrac{1}{v}\right\|_{C^{m-1}}^2 \quad (1 \le m \le k).$$

Using this inequality and applying induction on m, we can easily prove that

$$\left\|\tfrac{1}{v}\right\|_{C^m} \le \left[\inf_{x \in U}|v(x)|\right]^{-2^m}\prod_{j=1}^{m}\left(4^{j-1}\|v\|_{C^j}\right)^{2^{m-j}} \quad (1 \le m \le k). \tag{A.5}$$

(A.5) implies

$$\left\|\tfrac{1}{v}\right\|_{C^k} \le 4^{k\sum\limits_{i=0}^{k-1}2^i - \sum\limits_{i=0}^{k-1}(i+1)2^i}\|v\|_{C^k}^{\sum\limits_{i=0}^{k-1}2^i}\left[\inf_{x \in U}|v(x)|\right]^{-2^k}$$

which, together with the equalities $\sum_{i=0}^{k-1}2^i = 2^k - 1$ and $\sum_{i=0}^{k-1}(i+1)2^i = (k-1)2^k + 1$, yield the desired estimate. $\quad\square$

The previous two results now easily combine to an estimate of the C^k-norm for quotients of C^k functions.

Corollary A.3. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a non negative integer, and let $f, g \in C^k\left(\overline{U}, \mathbb{R}\right)$ with $\inf_{x \in U} |g(x)| > 0$. Then*

$$\left\| \frac{f}{g} \right\|_{C^k} \leq 2^{2^{k+1}-k-2} \left[\inf_{x \in U} |g(x)| \right]^{-2^k} \|f\|_{C^k} \|g\|_{C^k}^{2^k-1}.$$

Proof. The proof follows from Proposition A.1 and Lemma A.2. □

A.3. Hölder norm bounds for products and quotients of Hölder differentiable functions

Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let $0 < \alpha \leq 1$, and let k be a nonnegative integer. For $f \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)$ put

$$[f]_{k,\alpha} := \|f\|_{C^{k,\alpha}} - \|f\|_{C^k} = \max_{|\beta|=k} \sup_{\substack{x,y \in U \\ x \neq y}} \frac{|D_\beta f(x) - D_\beta f(y)|}{|x-y|^\alpha}.$$

Lemma A.4. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a nonnegative integer, let $0 < \alpha \leq 1$, and let $f_i \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)$ $(1 \leq i \leq m)$. Then $f_1 f_2 \cdots f_m \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)$ and*

$$[f_1 f_2 \cdots f_m]_{k,\alpha} \leq m^{k+1} \max_{1 \leq i \leq m} \|f_i\|_{C^k}^{m-1} \cdot \max_{\substack{1 \leq i \leq m \\ 0 \leq j \leq k}} [f_i]_{j,\alpha}.$$

Proof. In the case $k = 0$ we use (3.8) to obtain

$$[f_1 f_2 \cdots f_m]_{0,\alpha} \leq m \max_{1 \leq i \leq m} \|f_i\|_C^{m-1} \cdot \max_{1 \leq i \leq m} \sup_{\substack{x,y \in U \\ x \neq y}} \frac{|f_i(x) - f_i(y)|}{|x-y|^\alpha}$$

$$= m \max_{1 \leq i \leq m} \|f_i\|_C^{m-1} \cdot \max_{1 \leq i \leq m} [f_i]_{0,\alpha}.$$

Now suppose the statement is true for $C^{k-1,\alpha}\left(\overline{U}, \mathbb{R}\right)$. It is clear that

$$[f_1 \cdots f_m]_{k,\alpha}$$

$$= \max_{1 \leq j \leq d_1} \max_{|\beta|=k-1} \sup_{\substack{x,y \in U \\ x \neq y}} \frac{\left| \left(D_\beta \frac{\partial}{\partial x_j}\right)(f_1 \cdots f_m)(x) - \left(D_\beta \frac{\partial}{\partial x_j}\right)(f_1 \cdots f_m)(y) \right|}{|x-y|^\alpha}.$$

Assume that the first maximum in the right hand side is attained at $j = j_0$. Since

$$\left| \left(D_\beta \frac{\partial}{\partial x_{j_0}}\right)(f_1 \cdots f_m)(x) - \left(D_\beta \frac{\partial}{\partial x_{j_0}}\right)(f_1 \cdots f_m)(y) \right|$$

$$\leq \sum_{i=1}^{m} \left| D_\beta \left(\frac{\partial f_i}{\partial x_{j_0}} \prod_{\substack{1 \leq l \leq m \\ l \neq i}} f_l \right)(x) - D_\beta \left(\frac{\partial f_i}{\partial x_{j_0}} \prod_{\substack{1 \leq l \leq m \\ l \neq i}} f_l \right)(y) \right|,$$

we obtain

$$[f_1 f_2 \cdots f_m]_{k,\alpha} \leq \sum_{i=1}^{m} \left[\frac{\partial f_i}{\partial x_{j_0}} \prod_{\substack{1 \leq l \leq m \\ l \neq i}} f_l \right]_{k-1,\alpha}.$$

The induction hypothesis shows that

$$\left[\frac{\partial f_i}{\partial x_{j_0}} \prod_{\substack{1 \leq l \leq m \\ l \neq i}} f_l \right]_{k-1,\alpha} \leq m^k \max \left\{ \left\| \frac{\partial f_i}{\partial x_{j_0}} \right\|_{C^{k-1}}^{m-1}, \max_{\substack{0 \leq l \leq m \\ l \neq i}} \|f_l\|_{C^{k-1}}^{m-1} \right\}$$

$$\cdot \max_{0 \leq j \leq k-1} \left\{ \left[\frac{\partial f_i}{\partial x_{j_0}} \right]_{j,\alpha}, \max_{\substack{0 \leq l \leq m \\ l \neq i}} [f_l]_{j,\alpha} \right\}$$

$$\leq m^k \max_{0 \leq l \leq m} \|f_l\|_{C^k}^{m-1} \max_{\substack{0 \leq l \leq m \\ 0 \leq j \leq k}} [f_i]_{j,\alpha},$$

for every $j = 1, 2, \ldots, m$; which immediately implies the desired estimate. $\square$

Proposition A.5. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a nonnegative integer, let $0 < \alpha \leq 1$, and let $f_i \in C^{k,\alpha}\left(\overline{U}, \mathbb{R} \right)$ $(1 \leq i \leq m)$. Then $f_1 f_2 \cdots f_m \in C^{k,\alpha}\left(\overline{U}, \mathbb{R} \right)$ and*

$$\|f_1 f_2 \cdots f_m\|_{C^{k,\alpha}}$$
$$\leq 2m^{k+1} \max\left\{ 1, [\operatorname{diam}(U)]^{1-\alpha} \right\} \max_{1 \leq i \leq m} \|f_i\|_{C^k}^{m-1} \max_{1 \leq i \leq m} \|f_i\|_{C^{k,\alpha}}.$$

Proof. The mean value theorem for vector-valued functions shows that

$$[f]_{j,\alpha} \leq [\operatorname{diam}(U)]^{1-\alpha} \|f\|_{C^{j+1}} \quad \left(f \in C^j\left(\overline{U}, \mathbb{R} \right), 0 \leq j \leq k-1 \right), \quad (A.6)$$

hence

$$[f]_{j,\alpha} \leq \max\left\{ 1, [\operatorname{diam}(U)]^{1-\alpha} \right\} \|f\|_{C^{k,\alpha}} \quad \left(f \in C^{k,\alpha}\left(\overline{U}, \mathbb{R} \right), 0 \leq j \leq k \right), \quad (A.7)$$

which, together with Lemma A.4, implies

$$[f_1 f_2 \cdots f_m]_{k,\alpha} \leq m^{k+1} \max\left\{ 1, [\operatorname{diam}(U)]^{1-\alpha} \right\} \max_{1 \leq i \leq m} \|f_i\|_{C^k}^{m-1} \max_{1 \leq i \leq m} \|f_i\|_{C^{k,\alpha}}.$$

The latter, together with Proposition A.1 and with the equality

$$\|f\|_{C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2} \right)} := \|f\|_{C^k\left(\overline{U}, \mathbb{R}^{d_2} \right)} + [f]_{k,\alpha} \quad \left(f \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2} \right) \right), \quad (A.8)$$

implies the required estimate. $\square$

Using Proposition A.5 and the definition of the determinant, we obtain the following:

Corollary A.6. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a nonnegative integer, let $0 < \alpha \leq 1$, and let $A = (a_{ij})_{d_2 \times d_2} \in C^{k,\alpha}\left(\overline{U}, M_{d_2 d_2}(\mathbb{R})\right)$. Then $\det(A) \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)$ and*

$$\|\det(A)\|_{C^{k,\alpha}} \leq 2d_2! d_2^{k+1} \max\{1, [\operatorname{diam}(U)]^{1-\alpha}\}$$
$$\cdot \max_{1 \leq i,j \leq d_2} \|a_{ij}\|_{C^k}^{d_2-1} \max_{1 \leq i,j \leq d_2} \|a_{ij}\|_{C^{k,\alpha}}.$$

Lemma A.7. *Let $U \subset \mathbb{R}^{d_1}$ be a bounded open set, let k be a nonnegative integer, let $0 < \alpha \leq 1$, and let $v \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)$ with $\inf_{x \in U} |v(x)| > 0$. Then $\frac{1}{v} \in C^{k,\alpha}\left(\overline{U}, \mathbb{R}\right)$ and*

$$\left\|\tfrac{1}{v}\right\|_{C^{k,\alpha}} \leq 2^{k+1} \max\{1, [\operatorname{diam}(U)]^{1-\alpha}\} \max\{\|v\|_{C^k}, 2^k \left\|\tfrac{1}{v}\right\|_{C^k}^2\}$$
$$\cdot \max\{\|v\|_{C^{k,\alpha}}; 2^k \left\|\tfrac{1}{v}\right\|_{C^k}^2\}.$$

Proof. We first prove that

$$\left[\tfrac{1}{v}\right]_{k,\alpha} \leq 2^k \max\{1, [\operatorname{diam}(U)]^{1-\alpha}\} \max\{\|v\|_{C^k}, \left\|\tfrac{1}{v^2}\right\|_{C^k}\}$$
$$\cdot \max\{\|v\|_{C^{k,\alpha}}; \left\|\tfrac{1}{v^2}\right\|_{C^k}\}. \tag{A.9}$$

In the case $k = 0$, (A.9) follows from the inequality $\left[\tfrac{1}{v}\right]_{0,\alpha} \leq \left\|\tfrac{1}{v^2}\right\|_C \cdot [v]_{0,\alpha}$. Assuming $k \geq 1$, the reasoning used in the proof of Lemma A.4 can be applied:

$$\left[\tfrac{1}{v}\right]_{k,\alpha} = \max_{1 \leq j \leq d_1} \max_{|\beta|=k-1} \sup_{\substack{x,y \in U \\ x \neq y}} \frac{\left|\left(D_\beta \frac{\partial}{\partial x_j}\right)\left(\frac{1}{v}\right)(x) - \left(D_\beta \frac{\partial}{\partial x_j}\right)\left(\frac{1}{v}\right)(y)\right|}{|x-y|^\alpha}$$

$$= \max_{1 \leq j \leq d_1} \max_{|\beta|=k-1} \sup_{\substack{x,y \in U \\ x \neq y}} \frac{\left|D_\beta\left(v'_{x_j} \cdot \frac{1}{v^2}\right)(x) - D_\beta\left(v'_{x_j} \cdot \frac{1}{v^2}\right)(y)\right|}{|x-y|^\alpha}$$

$$= \max_{1 \leq j \leq d_1} \left[v'_{x_j} \cdot \tfrac{1}{v^2}\right]_{k-1,\alpha}.$$

Applying Lemma A.4 to $\left[v'_{x_j} \cdot \tfrac{1}{v^2}\right]_{k-1,\alpha}$ yields

$$\left[\tfrac{1}{v}\right]_{k,\alpha} \leq 2^k \max_{1 \leq j \leq d_1} \left[\max\{\left\|v'_{x_j}\right\|_{C^{k-1}}, \left\|\tfrac{1}{v^2}\right\|_{C^{k-1}}\}\right.$$

$$\left. \cdot \max\{[v'_{x_j}]_{0,\alpha}; \ldots; [v'_{x_j}]_{k-1,\alpha}; [\tfrac{1}{v^2}]_{0,\alpha}; \ldots; [\tfrac{1}{v^2}]_{k-1,\alpha}\}\right]$$

$$\leq 2^k \max\{\|v\|_{C^k}, \left\|\tfrac{1}{v^2}\right\|_{C^{k-1}}\}$$

$$\cdot \max\{[v]_{1,\alpha}; \ldots; [v]_{k,\alpha}; [\tfrac{1}{v^2}]_{0,\alpha}; \ldots; [\tfrac{1}{v^2}]_{k-1,\alpha}\}.$$

The latter, combined with (A.6) and (A.7), implies (A.9).

In view of (A.3), $\left\|\tfrac{1}{v^2}\right\|_{C^k} \leq 2^k \left\|\tfrac{1}{v}\right\|_{C^k}^2$, which, together with (A.9), gives

$$\left[\tfrac{1}{v}\right]_{k,\alpha} \leq 2^k \max\{1, [\operatorname{diam}(U)]^{1-\alpha}\} \max\{\|v\|_{C^k}, 2^k \left\|\tfrac{1}{v}\right\|_{C^k}^2\}$$

$$\cdot \max\big\{\|v\|_{C^{k,\alpha}}\, ; 2^k \big\|\tfrac{1}{v}\big\|_{C^k}^2\big\}. \tag{A.10}$$

Next, by Proposition A.1,

$$\big\|\tfrac{1}{v}\big\|_{C^k} = \big\|v \cdot \tfrac{1}{v^2}\big\|_{C^k} \leq 2^k \|v\|_{C^k} \cdot \big\|\tfrac{1}{v^2}\big\|_{C^k} \leq 2^k \|v\|_{C^k} \cdot 2^k \big\|\tfrac{1}{v}\big\|_{C^k}^2,$$

hence $\big\|\tfrac{1}{v}\big\|_{C^k}$ does not exceed the right hand side of (A.10). This fact, together with (A.8) and (A.10), implies the desired estimate for $\big\|\tfrac{1}{v}\big\|_{C^{k,\alpha}}$. □

Proposition A.5 and Lemma A.7 immediately imply the following

Corollary A.8. *Let* $u_n \to u$ *and* $v_n \to v$ *in* $C^{k,\alpha}\big(\overline{U}, \mathbb{R}\big)$*-norm, and let* $\inf\limits_{x \in U, n \in \mathbb{N}} |v_n(x)| > 0$*. Then* $\tfrac{u_n}{v_n} \to \tfrac{u}{v}$ *in* $C^{k,\alpha}\big(\overline{U}, \mathbb{R}\big)$*-norm.*

Lemmas A.2, A.7 and Proposition A.5 yield the following estimate for the $C^{k,\alpha}$-norm of the quotient of two $C^{k,\alpha}$-functions:

Proposition A.9. *Let* $U \subset \mathbb{R}^{d_1}$ *be a bounded open set, let* k *be a nonnegative integer, let* $0 < \alpha \leq 1$*, and let functions* $f, g \in C^{k,\alpha}\big(\overline{U}, \mathbb{R}\big)$ *satisfy the conditions* $\|f\|_{C^{k,\alpha}} \leq L, \|g\|_{C^{k,\alpha}} \leq L$ *and* $\inf\limits_{x \in U} |g(x)| \geq l^{-1}$ *with real numbers* $L \geq 1$ *and* $l \geq 1$*. Then*

$$\Big\|\tfrac{f}{g}\Big\|_{C^{k,\alpha}} \leq (2Ll)^{2^{k+5}} \max\big\{1, [\mathrm{diam}(U)]^{2(1-\alpha)}\big\}.$$

Proof. In view of Lemma A.2,

$$\Big\|\tfrac{1}{g}\Big\|_{C^k} \leq 2^{2^{k+1}-2k-2}\Big[\inf\limits_{x \in U}|g(x)|\Big]^{-2^k}\|g\|_{C^{k,\alpha}}^{2^k-1} \leq 2^{2^{k+1}-2k}(Ll)^{2^k},$$

hence

$$\max\Big\{\|f\|_{C^k}, \Big\|\tfrac{1}{g}\Big\|_{C^k}\Big\} \leq 2^{2^{k+1}}(Ll)^{2^k} \quad \text{and} \quad 2^k\Big\|\tfrac{1}{g}\Big\|_{C^k}^2 \leq 2^{2^{k+2}}(Ll)^{2^{k+1}}.$$

Therefore Lemma A.7 gives

$$\Big\|\tfrac{1}{g}\Big\|_{C^{k,\alpha}} \leq \max\big\{1, [\mathrm{diam}(U)]^{1-\alpha}\big\} \cdot 2^{2^{k+3}+k+1}(Ll)^{2^{k+2}}.$$

The latter, together with Proposition A.5, yields

$$\begin{aligned}
\Big\|\tfrac{f}{g}\Big\|_{C^{k,\alpha}} &\leq 2^{k+2}\max\big\{1, [\mathrm{diam}(U)]^{2(1-\alpha)}\big\} \cdot 2^{2^{k+1}}(Ll)^{2^k} \cdot 2^{2^{k+3}+k+1}(Ll)^{2^{k+2}} \\
&\leq (2Ll)^{2^{k+5}}\max\big\{1, [\mathrm{diam}(U)]^{2(1-\alpha)}\big\}. \qquad \qquad \square
\end{aligned}$$

A.4. Composition of differentiable functions

Let $M_{d_2 d_3}(\mathbb{R})$ stand for the set of all $d_2 \times d_3$ matrices with real entries. The collection of all matrix-valued functions $(a_{ij})_{d_2 \times d_3} : \overline{U} \to M_{d_2 d_3}(\mathbb{R})$, whose entries are k-times Hölder differentiable on $\overline{U}$, will be denoted by $C^{k,\alpha}\left(\overline{U}, M_{d_2 d_3}(\mathbb{R})\right)$. 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., [13, 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 (see Proposition A.5). Obviously, these assertions remain true for matrix-valued functions. Moreover, the quotient of k-times Hölder differentiable real-valued functions is also k-times Hölder differentiable provided that the absolute value of the denominator has a positive lower bound (see Lemma A.7).

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

Lemma A.10. *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.

The proof of Lemma A.10 is quite simple, hence we will only outline it. First note 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, we proceed inductively, using the equality $D^m(g \circ f) = D^{m-1}\left[(Dg \circ f)Df\right]$ $(1 \le m \le k)$.

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 $C^{k,\alpha}\left(\overline{U}, \mathbb{R}^{d_2}\right)$.

A.5. The inverse function theorem for Hölder differentiable maps

We are now in the position to prove the main theorem of this appendix, namely the inverse function theorem for Hölder differentiable functions on $\overline{U} \subset \mathbb{R}^d$. Note that this is different from the the inverse function theorem for Fréchet differentiable functionals *on* the Banach space of Hölder functions, which is a consequence of the Banach space version of the standard inverse function theorem.

Theorem A.11. *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

$$\inf_{x \in U} |J_\varphi(x)| > 0, \tag{A.11}$$

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

$$\left\| D\left(\varphi^{-1}\right) \right\|_\infty \leq \frac{d! \|D\varphi\|_\infty^{d-1}}{\inf_{x \in U} |J_\varphi(x)|}. \tag{A.12}$$

(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-valued function $\operatorname{adj}(D\varphi)$ (see Proposition A.5). In view of (A.11) and Lemma A.7, 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., [10, Lemma I.5.8]). The inverse function theorem for C^1-maps (see, e.g., [28, Th. 9.24] or [34, 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

$$D\varphi^{-1} = (D\varphi)^{-1} \circ \varphi^{-1}. \tag{A.13}$$

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 (A.13) 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 (A.1) and (A.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 A.10; therefore (A.13) implies that $\varphi^{-1} \in C^{k,\alpha}\left(\overline{U}\right)$.

To prove (A.12), we need an estimate for the operator norm of an invertible matrix. If $A = (a_{ij})_{d \times d} \in M_{dd}(\mathbb{R})$ is invertible, then $A^{-1} = \frac{1}{\det(A)} \operatorname{adj}(A)$. The operator norm and the determinant of any matrix $B = (b_{ij})_{d \times d} \in M_{dd}(\mathbb{R})$ can be easily estimated as

$$|B| \leq d \max_{1 \leq i,j \leq d} |b_{ij}| \tag{A.14}$$

and

$$|\det(B)| \leq d! \left[\max_{1 \leq i,j \leq d} |b_{ij}|\right]^d, \tag{A.15}$$

respectively. Therefore the absolute value of each entry of the matrix $\mathrm{adj}\,(A)$ does not exceed $(d-1)! \left[\max_{1 \leq i,j \leq d} |a_{ij}|\right]^{d-1}$, which, together with (A.14), yields

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

Applying this estimate to $A = (D\varphi)(x)$ and using (A.13), we obtain (A.12).

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

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

(A.12), combined with the mean value theorem for vector-valued functions, gives

$$\left|\varphi_n^{-1}(y_1) - \varphi_n^{-1}(y_2)\right| \leq C |y_1 - y_2| \quad \left(y_1, y_2 \in \overline{U}; n \in \mathbb{N}\right), \tag{A.16}$$

therefore

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

Letting $n \to \infty$, we obtain the inequality $|x_1 - x_2| \leq C |\varphi(x_1) - \varphi(x_2)|$ $\left(x_1, x_2 \in \overline{U}\right)$, 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 $x := \lim_{i \to \infty} x_{n_i}$. 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,

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

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 $\left\{\varphi_n^{-1}\right\}_{n=1}^\infty$ is relatively compact in $C\left(\overline{U}\right)$ and that φ^{-1} is the only possible accumulation point of $\left\{\varphi_n^{-1}\right\}_{n=1}^\infty$. Indeed, $\left\{\varphi_n^{-1}\right\}_{n=1}^\infty$ is uniformly bounded since $\varphi_n^{-1}\left(\overline{U}\right) \subset \overline{U}$ $(n \in \mathbb{N})$. Furthermore, the estimate (A.16) shows that $\left\{\varphi_n^{-1}\right\}_{n=1}^\infty$ is equicontinuous on $\overline{U}$. Hence the Arzela-Ascoli theorem implies the relative compactness of $\left\{\varphi_n^{-1}\right\}_{n=1}^\infty$. If ψ is an accumulation point for $\left\{\varphi_n^{-1}\right\}_{n=1}^\infty$, 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 $\varphi_{n_i}^{-1} \circ \varphi_{n_i} = I$ and $\varphi_{n_i} \circ \varphi_{n_i}^{-1} = I$ letting $i \to \infty$, we easily conclude that $\psi \circ \varphi = I$ and $\varphi \circ \psi = I$, therefore $\psi = \varphi^{-1}$.

Using (A.13), Corollary A.8 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 (A.13) and Corollary A.8, 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 (A.2), implies that $\varphi_n^{-1} \to \varphi^{-1}$ in $C^{k-1,1}$-norm. Using this convergence, (A.13) and the statement (b) of Lemma A.10, we finally conclude that $\varphi_n^{-1} \to \varphi^{-1}$ in $C^{k,\alpha}$-norm. $\qquad\square$

Appendix B: Suprema of subgaussian random processes

In this section, we collect some known results from probability for the convenience of the reader.

To estimate the sampling error in generative adversarial learning, we used some results on suprema of subgaussian random processes. We give the necessary definitions.

Definition B.1. A random variable X is called σ^2-subgaussian if $\mathbb{E}\left(|X|\right) < \infty$ and
$$\mathbb{E}\left[e^{\tau[X-\mathbb{E}(x)]}\right] \le e^{\frac{\tau^2\sigma^2}{2}} \quad (\tau \ge 0).$$

The following can be seen as a stochastic variant of the Lipschitz property for stochastic processes:

Definition B.2. 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.

Provided stochastic processes are subgaussian, we have the following metric entropy estimate for the expected value of the supremum. Here $N(T, \rho, \varepsilon)$ stands for the covering number of the set T, i.e. the smallest number of ρ-balls of radius ε covering T.

Theorem B.3 (Dudley)**.** *If* $\{\pm X_t\}_{t\in T}$ *are continuous subgaussian processes on the compact metric space* (T, ρ)*, then the estimate (the so-called entropy bound)*

$$\mathbb{E}\left[\sup_{t\in T}|X_t|\right] \le 12 \int\limits_0^\infty \sqrt{\log N\left(T, \rho, \varepsilon\right)}d\varepsilon \tag{B.1}$$

holds.

The proof of (B.1) can be found in [38, Sec. 5.3].

The following concentration inequality due to C. McDiarmid (see [23]) in addition allows us to control the deviation from the expected value. This enables us in Sec. 4 to pass from estimates for the expectation values of the maxima of random processes to estimates for the maxima of those processes themselves.

Theorem B.4 (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* $f : \prod_{k=1}^{n} A_k \to \mathbb{R}$ *is a function with the* $(c_1, \ldots, c_n)$*-bounded differences property: for each* $k = 1, 2, \ldots, n$ *and for any vectors* $x, y \in \prod_{k=1}^{n} A_k$ *that differ only in the* k*-th coordinate, the inequality* $f(x) - f(y) \leq c_k$ *holds. Then*

$$\mathbb{P}\left(f(x) - \mathbb{E}[f(x)] \geq t\right) \leq e^{-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}} \quad (t \geq 0). \tag{B.2}$$

Acknowledgments

References

[1] ACKLEY, D. H., HINTON, G. E. and SEJNOWSKI, T. J. (1988). *A Learning Algorithm for Boltzmann Machines* In *Connectionist Models and Their Implications: Readings from Cognitive Science* 285–307. Ablex Publishing Corp., USA.

[2] ARJOVSKY, M., CHINTALA, S. and BOTTOU, L. (2017). Wasserstein Generative Adversarial Networks. (D. PRECUP and Y. W. TEH, eds.). *Proceedings of Machine Learning Research* **70** 214–223. PMLR, International Convkingention Centre, Sydney, Australia.

[3] BELOMESTNY, D., MOULINES, E., NAUMOV, A., PUCHKIN, N. and SAMSONOV, S. (2021). Rates of convergence for density estimation with GANs.

[4] BIAU, G., CADRE, B., SANGNIER, M. and TANIELIAN, U. (2020). Some theoretical properties of GANs. *Ann. Statist.* **48** 1539–1566. MR4124334

[5] BIAU, G., SANGNIER, M. and TANIELIAN, U. (2020). Some Theoretical Insights into Wasserstein GANs.

[6] BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Communications on pure and applied mathematics* **44** 375–417.

[7] BROCK, A., DONAHUE, J. and SIMONYAN, K. (2018). Large Scale GAN Training for High Fidelity Natural Image Synthesis. *CoRR* **abs/1809.11096**.

[8] CHEN, M., LIAO, W., ZHA, H. and ZHAO, T. (2020). Distribution Approximation and Statistical Estimation Guarantees of Generative Adversarial Networks. *arXiv preprint arXiv:2002.03938*.

[9] DINH, L., KRUEGER, D. and BENGIO, Y. (2014). Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*.

[10] DUNFORD, N. and SCHWARTZ, J. T. (1988). *Linear operators. Part I.* John Wiley & Sons, New York. MR1009162

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

[12] GHAHRAMANI, Z. (2004). *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* 72–112. Springer, Berlin-Heidelberg.

[13] GILBARG, D. and TRUDINGER, N. S. (2001). *Elliptic partial differential equations of second order.* Springer, Berlin. MR1814364

[14] GINÉ, E. and NICKL, R. (2016). *Mathematical foundations of infinite-dimensional statistical models* **40**. Cambridge University Press, USA.

[15] GOODFELLOW, I., POUGET-ABADIE, J., MIRZA, M., XU, B., WARDE-FARLEY, D., OZAIR, S., COURVILLE, A. and BENGIO, Y. (2014). Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27* 2672–2680. Curran Associates, Montreal.

[16] HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

[17] JOE, H. (2014). *Dependence modeling with copulas.* CRC press.

[18] KINGMA, D. P., WELLING, M. et al. (2019). An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning* **12** 307–392.

[19] KOBYZEV, I., PRINCE, S. J. and BRUBAKER, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE transactions on pattern analysis and machine intelligence* **43** 3964–3979.

[20] KULLBACK, S. (1997). *Information theory and statistics.* Dover Publications, Mineola. MR1461541

[21] LALA, S., SHADY, M., BELYAEVA, A. and LIU, M. (2018). Evaluation of mode collapse in generative adversarial networks. *High Performance Extreme Computing.*

[22] LEDIG, C., THEIS, L., HUSZÁR, F., CABALLERO, J., CUNNINGHAM, A., ACOSTA, A., AITKEN, A., TEJANI, A., TOTZ, J., WANG, Z. et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition* 4681–4690.

[23] MCDIARMID, C. (1989). *On the method of bounded differences.* In *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference. London Mathematical Society Lecture Note Series* 148–188. Cambridge University Press.

[24] PETERSEN, P. and VOIGTLAENDER, F. (2018). Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Networks* **108** 296–330.

[25] PETERSEN, P. and VOIGTLAENDER, F. (2020). Equivalence of approximation by convolutional neural networks and fully-connected networks. *Proceedings of the American Mathematical Society* **148** 1567–1581.

[26] RADFORD, A., METZ, L. and CHINTALA, S. (2016). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Net-

works.

[27] Rosenblatt, M. (1952). Remarks on a multivariate transformation. *Ann. Math. Statistics* **23** 470–472. MR49525

[28] Rudin, W. (1976). *Principles of Mathematical Analysis.* McGraw-Hill, New York. MR0385023

[29] Rudin, W. (1991). *Functional analysis.* McGraw-Hill, New York.

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

[31] Shiryaev, A. N. (2016). *Probability-1.* Springer, New York. MR3467826

[32] Shui, C., Chen, Q., Wen, J., Zhou, F., Gagné, C. and Wang, B. (2020). Beyond $\mathcal{H}$-Divergence: Domain Adaptation Theory With Jensen-Shannon Divergence.

[33] Singh, S., Uppal, A., Li, B., Li, C.-L., Zaheer, M. and Póczos, B. (2018). Nonparametric Density Estimation under Adversarial Losses. In *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) **31** 10225–10236. Curran Associates, Inc.

[34] Taylor, M. E. (2011). *Partial differential equations I. Basic theory.* Springer, New York. MR2744150

[35] Teshima, T., Ishikawa, I., Tojo, K., Oono, K., Ikeda, M. and Sugiyama, M. (2020). Coupling-based invertible neural networks are universal diffeomorphism approximators. *Advances in Neural Information Processing Systems* **33** 3362–3373.

[36] Uppal, A., Singh, S. and Póczos, B. (2020). Nonparametric Density Estimation & Convergence Rates for GANs under Besov IPM Losses.

[37] van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer, New York. MR1385671

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

[39] Villani, C. (2009). *Optimal transport: old and new* **338**. Springer.

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

[41] Yarotsky, D. (2018). Universal approximations of invariant maps by neural networks. *CoRR* **abs/1804.10306**.

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