ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Deep Learning in High Dimension

Ch. Schwab and J. Zech

Research Report No. 2017-57

December 2017
Latest revision: February 2018

Seminar für Angewandte Mathematik
Eidgenössische Technische Hochschule
CH-8092 Zürich
Switzerland

Funding SNF: 159940

Deep Learning in High Dimension*

Christoph Schwab and Jakob Zech†

February 13, 2018

Abstract

We estimate the expressive power of a class of deep Neural Networks (DNNs for short) on a class of countably-parametric maps $u : U \to \mathbb{R}$ on the parameter domain $U = [-1, 1]^{\mathbb{N}}$. Such maps arise for example as response surfaces of parametric PDEs with distributed uncertain inputs, i.e., input data from function spaces. Equipping these spaces with suitable bases, instances of uncertain inputs become sequences of (coefficient) parameters of representations in these bases.

Dimension-independent approximation rates of generalized polynomial chaos (gpc for short) approximations of countably-parametric maps $u : U \to \mathbb{R}$ depend only on the degree of sparsity of the gpc expansion of u as quantified by the summability exponent of the sequence of their gpc expansion coefficients: for parametric maps which are p-sparse with some $0 < p < 1$, we show that a certain architecture of DNNs afford the same convergence rates in terms of N, the total number of units in the DNN.

So-called $(\boldsymbol{b}, \varepsilon)$-holomorphic maps u with $\boldsymbol{b} \in \ell^p$ for some $p \in (0, 1)$ arise in a number of applications from computational uncertainty quantification. For this class of functions, up to logarithmic factors we prove the dimension independent approximation rate $s = 1/p - 1$, in terms of the total number N of units in the DNN. This shows that the DNN architectures can overcome the curse of dimensionality when expressing possibly infinite-parametric, real-valued maps with a certain sparsity. Examples of such maps comprise response maps of parametric and stochastic PDEs models with distributed uncertain input data.
Key words: generalized polynomial chaos, deep learning, sparsity, Uncertainty Quantification
Subject Classification: 68Q32, 41A25, 41A46

1 Introduction

After foundational developments several decades ago in answering the question of universality of NNs [16, 21, 20, 3, 4], in recent years so-called *deep neural networks* (DNN for short) have undergone rapid development and successful deployment in a wide range of applications. The benefit afforded by depth of NNs on their approximation properties respectively on their expressive power has been documented in an increasing number of applications. In particular, for response surfaces and classification tasks for "complex" systems superiority of deep architectures in a number of applications has been asserted in recent years.

The purpose of the present paper is to establish that certain architectures of DNNs can express functions of a large number of variables as arise as response surfaces of countably-parametric PDE models. Specifically, we show for a broad range of such many-parametric

*Supported in part by SNSF Grant No. 159940
†SAM, ETH Zürich, ETH Zentrum, HG G57.1, CH8092 Zürich, Switzerland

functions which arise as solutions of parametric operator equations with holomorphic parameter dependence, that DNNs afford an ε-error with complexity of $N = O(\varepsilon^{-1/s})$ units where $O(\cdot)$ and the rate $s > 0$ is independent of the number of parameters (thereby overcoming, in particular, the curse of dimensionality), and comparable to convergence rates of the best N-term generalized polynomial chaos (gpc for short) approximation of the parametric solution. As shown here, the rate s depends only on a suitable notion of sparsity in coefficient sequences of gpc expansions of the parametric solution.

1.1 Recent mathematical results on expressive power of DNNs

Being mostly based on empirical observations, the past year has seen significant efforts towards theoretical understanding of the benefits on expressive power of NNs afforded by NN depth. Theoretical results focused on *approximation rate bounds* for particular function classes rather than mere density as in the earlier results.

We mention in particular [17] and [7] where it is shown that deep NNs with a particular architecture allow the same approximation rate bounds as rather general multiresolution systems (such as, for example, wavelet, ridglet and shearlet systems which are widely used in image processing and data compression) when measured in terms of the number N of units in the deep NN.

In [13], *convolutional DNNs* were proved to be able to express multivariate functions given in so-called *Hierarchic Tensor (HT) formats*, a numerical representation which is inspired by electron structure calculations in computational quantum chemistry.

Also, in [28, 23], it has been shown that DNNs can express general uni- and multivariate polynomials on bounded domains with pointwise accuracy $\delta > 0$ with internal complexity (which we assume to comprise the number of NN layers and the number of NN units) which scales polylogarithmically with respect to δ. The results in [28, 23] allow transferring approximation results from high order finite and spectral element approximation results, in particular exponential convergence results, to certain types of DNNs.

Another type of result, closer to the present investigation, is the analysis of NN depth in high-dimensional approximation. In [24] it was shown that multivariate functions which can be written as superpositions (being additive but also compositional) of a possibly large number of "simpler" functions, depending only on a few variables at a time, can be expressed with DNNs at complexity which is bounded by the dimensionality of constituent functions in the composition and the size of the connectivity graph, thereby alleviating the curse of dimensionality for this class.

1.2 Scope of the present results

In the present paper, we investigate the expressive power of DNNs for many-parametric response functions of solutions of many-parametric operator equations, with holomorphic dependence on the parameters. Countably-parametric operator equations with a certain holomorphic dependence of the operator on the parameters arise in a number of applications. We mention only elliptic PDEs with uncertain, spatially heterogeneous, uncertain coefficients (see, e.g., [11, 2] and the references there), and PDEs posed in domains of uncertain geometry (see, e.g., [25, 22, 12]), and time-harmonic, electromagnetic scattering (see, e.g. [22]). Such models are ubiquitous in the area of computational uncertainty quantification in engineering and in the sciences. Holomorphic parametric dependence implies holomorphic dependence of solutions on the parameters, and in particular that response functions (and, in fact, manifolds of parametric solutions) admit *sparse gpc expansions*. This sparsity in turn implies dimension independent approximation rates of various adaptive approximation methods to approximate the parametric PDE solution manifold, and of the response surfaces for so-called *quantities of interest* (QoIs for short), which are real-valued, linear or non-linear solution functionals, i.e. superpositions of the data-to-solution map and of a QoI, being a map from the (Hilbert or Banach) space accomodating the PDE solution into the real numbers. The present DNN approximation results show that response surfaces (or data-to-QoI maps) which depend holomorphically on the high-dimensional parameter vector can be 'expressed' by DNNs at dimension independent rates in terms of N, being the total size of the DNN, which is to say the sum of the number of nodes over all layers of the DNN.

While these remarks pertain to so-called *forward problems* described by parametric PDEs, often also the corresponding *inverse problems* are of interest. The present results are also relevant to these: in the *Bayesian setting* (see [27] and the references there), it has been shown in [15, 26] that parametric holomorphy of the QoI is inherited by the bayesian posterior density, if it exists. The present results therefore imply that DNNs can also express these densities at dimension-independent rates, opening a perspective of "deep bayesian learning" in UQ.

1.3 Notation and preliminaries

As in our previous works [14, 29], we adopt standard notation: $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. The symbol C will stand for a generic, positive constant that is independent of any quantities determining the asymptotic behaviour of an estimate.

Multiindices are denoted by $\boldsymbol{\nu} = (\nu_j)_{j=1}^M \in \mathbb{N}_0^M$ where either $M \in \mathbb{N}$ or $M = \infty$. The *order* of a multiindex $\boldsymbol{\nu}$ is denoted by $|\boldsymbol{\nu}|_1 := \sum_{j=1}^M \nu_j$. The countable set of "finitely supported" multiindices is denoted by

$$\mathcal{F} := \{\boldsymbol{\nu} \in \mathbb{N}_0^{\mathbb{N}} : |\boldsymbol{\nu}|_1 < \infty\}. \tag{1.1}$$

The notation $\operatorname{supp} \boldsymbol{\nu}$ stands for the *support* of the multiindex, i.e. the set $\{j \in \{1, \dots, M\} : \nu_j \neq 0\}$. The size of the support of $\boldsymbol{\nu} \in \mathcal{F}$ is $|\boldsymbol{\nu}|_0 = \#(\operatorname{supp} \boldsymbol{\nu})$. A subset $\Lambda \subseteq \mathcal{F}$ is *downward closed*, if $\boldsymbol{\nu} = (\nu_j)_{j\geq 1} \in \Lambda$ implies $\boldsymbol{\mu} = (\mu)_{j\geq 1} \in \Lambda$ for all $\boldsymbol{\mu} \leq \boldsymbol{\nu}$. Here, the ordering "$\leq$" on $\mathcal{F}$ is defined as $\mu_j \leq \nu_j$, for all $j \geq 1$. For $0 < p < \infty$, denote by $\ell^p(\mathcal{F})$ the space of sequences $\mathbf{t} = (t_{\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}} \subset \mathbb{R}$ satisfying $\|\mathbf{t}\|_{\ell^p(\mathcal{F})} := (\sum_{\boldsymbol{\nu} \in \mathcal{F}} |t_{\boldsymbol{\nu}}|^p)^{1/p} < \infty$. As usual, $\ell^\infty(\mathcal{F})$ equipped with the norm $\|\mathbf{t}\|_{\ell^\infty(\mathcal{F})} := \sup_{\boldsymbol{\nu} \in \mathcal{F}} |t_{\boldsymbol{\nu}}| < \infty$ denotes the space of all uniformly bounded sequences.

For $M \in \mathbb{N} \cup \{\infty\}$, we consider $\mathbb{C}^M$ endowed with the product topology. Any subset such as $[-1, 1]^M$ is then equipped with the subspace topology. For a ball of radius $\varepsilon > 0$ in $\mathbb{C}$ we write $B_\varepsilon \subseteq \mathbb{C}$, and $\operatorname{clos}(B_\varepsilon) \subseteq \mathbb{C}$ for its closure. Furthermore $B_\varepsilon^M := \bigtimes_{j=1}^M B_\varepsilon \subseteq \mathbb{C}^M$. Elements of $\mathbb{C}^M$ are denoted by boldface characters such as $\boldsymbol{y} = (y_j)_{j=1}^M \in [-1, 1]^M$. For $\boldsymbol{\nu} \in \mathcal{F}$, standard notations $\boldsymbol{y}^{\boldsymbol{\nu}} := \prod_{j\geq 1} y_j^{\nu_j}$ and $\boldsymbol{\nu}! = \prod_{j\geq 1} \nu_j!$ will be employed (observing that these formally infinite products contain only a finite number of nontrivial factors with the conventions $0! := 1$ and $0^0 := 1$).

1.4 Outline

The structure of this note is as follows: in Section 2, we review gpc approximation rate bounds for so-called $(\boldsymbol{b}, \varepsilon)$-holomorphic functions of an infinite sequence $\boldsymbol{y} = (y_j)_{j\geq 1}$ of arguments. Such functions arise as response surfaces of holomorphic-parametric operator equations with many-parametric inputs, e.g., through a linear functional $G(\cdot) \in X^*$ of the parametric solution $u : U \to X$ taking values in some Banach space X. These functions allow holomorphic extensions w.r. to each argument y_j to the complex domain, with quantitative control of the size of the domain of holomorphy w.r. to the dimension index j.

In Section 3, we present the DNN approximation results. Section 3.1 introduces the architectures which are admitted in our approximation results. Section 3.2 proves a basic result on the expressive power of DNNs for univariate polynomials. This will be used in Section 3.3 to establish the main results of this work, namely the approximation of a $(\boldsymbol{b}, \varepsilon)$-holomorphic parametric response map to pointwise accuracy $\delta > 0$ with $\boldsymbol{b} \in \ell^p(\mathbb{N})$ by a DNN with (essentially, i.e. up to polylogarithmic factors) $O(\log \delta)$ many hidden layers and total number of nodes.

Section 4 presents some conclusions in particular addressing the scope of results.

2 Generalized Polynomial Chaos Approximation

To analyze the expressive power of deep NNs on countably-parametric, real-valued maps, we shall draw upon results from [8, 9, 29] on sparse generalized polynomial chaos approximation of such maps. To state these results, with the parameter domain $U := [-1,1]^{\mathbb{N}}$, we consider maps $u : U \to \mathbb{R}$. We are interested in *holomorphic maps* which admit, with respect to each parameter $y_j \in \boldsymbol{y}$, a holomorphic extension to the complex domain. Crucial in the present paper are approximation results for the parametric maps $u : U \to \mathbb{R}$. We shall use results from [11, 9, 8, 29] and the references there on sparse *Taylor generalized polynomial chaos (gpc) expansions*. These are (formal, at this stage) expressions of the form

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} u_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} , \quad \boldsymbol{y} \in U . \tag{2.1}$$

The summability properties of the Taylor coefficients $(|u_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}$ in (2.1) are crucial in assigning a meaning to gpc series like (2.1). As for every $\boldsymbol{y} \in U$ and for every $\boldsymbol{\nu} \in \mathcal{F}$ it holds that $|\boldsymbol{y}^{\boldsymbol{\nu}}| \le 1$, the summability $(|u_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$ implies the unconditional convergence of (2.1) for every $\boldsymbol{y} \in U$. This summability is, in turn, ensured by a suitable form of analytic continuation of the parametric map $u : U \to \mathbb{R}$; we recapitulate principal definitions and results from [29] and the references there.

2.1 $(\boldsymbol{b}, \varepsilon)$-Holomorphy

To establish convergence rates of partial sums of the Taylor gpc expansion (2.1), p-summability of the sequence of (norms of) the Taylor coefficients $\{u_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu}\in\mathcal{F}} \subset X$ is required for some $0 < p < 1$. A sufficient condition on the parametric map $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in \mathbb{R}$ is the following assumption, which has similarly been stated in [9].

Assumption 2.1 ($(\boldsymbol{b}, \varepsilon)$-Holomorphy)**.** *Assume given a sequence $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}}$ of positive reals b_j such that $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1]$, and such that b_j is monotonically decreasing.*

A poly-radius $\boldsymbol{\rho} \in [1,\infty)^{\mathbb{N}}$ is called $(\boldsymbol{b}, \varepsilon)$-admissible for some $\varepsilon > 0$ if

$$\sum_{j\in\mathbb{N}} b_j(\rho_j - 1) \le \varepsilon . \tag{2.2}$$

The function $u : B_{\boldsymbol{\rho}} \to \mathbb{C}$ is continuous. With

$$O_{\boldsymbol{b}} := \bigcup_{\{\boldsymbol{\rho}\,:\,\boldsymbol{\rho} \text{ is } (\boldsymbol{b},\varepsilon)\text{-admissible}\}} \operatorname{clos}(B_{\boldsymbol{\rho}}) \subseteq \mathbb{C}^{\mathbb{N}}, \tag{2.3}$$

u is separately holomorphic on $O_{\boldsymbol{b}}$ as a function of each z_j. Additionally, there exists a constant $C_u < \infty$ such that $\sup_{\boldsymbol{z}\in O_{\boldsymbol{b}}} \|u(\boldsymbol{z})\|_{X_{\mathbb{C}}} \le C_u$.

In case u satisfies Assumption 2.1, we will also say that u is $(\boldsymbol{b}, \varepsilon)$-holomorphic. Note that by continuity in the above assumption we mean continuity with respect to the subspace topology on $B_{\boldsymbol{\rho}} \subseteq \mathbb{C}^{\mathbb{N}}$, where $\mathbb{C}^{\mathbb{N}}$ is equipped with the product topology.

We recall the well-known fact, that the Taylor expansion in (2.1) converges on finite dimensional polydiscs in $\mathbb{C}^M$, $M \in \mathbb{N}$. In the following, by an absolutely convergent series $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in Y^{\mathcal{S}}$, with Y some Banach space and $\mathcal{S}$ some countable set such as $\mathcal{F}$, we mean a sequence for which there exists a bijection $\pi : \mathbb{N} \to \mathcal{S}$ such that the sum $\sum_{j\in\mathbb{N}} \|t_{\pi(j)}\|_Y$ converges. The sum is meaningful due to the countability of $\mathcal{S}$, and due to the fact that the existence of one such bijection ensures the series to converge for any bijection $\pi : \mathbb{N} \to \mathcal{S}$. In the next proposition, for $\boldsymbol{\rho} = (\rho_j)_{j=1}^M$ we write $\operatorname{clos}(B_{\boldsymbol{\rho}})$ to denote the closed polydisc $\bigtimes_{j\ge1} \operatorname{clos}(B_{\rho_j}) \subseteq \mathbb{C}^{\mathbb{N}}$.

Proposition 2.2. *Let* $\boldsymbol{\rho} = (\boldsymbol{\rho}_j)_{j\geq 1} \in (1,\infty)^{\mathbb{N}}$*. With* $U = [-1,1]^{\mathbb{N}}$*, suppose that* $u : U \to \mathbb{C}$ *is* $(\boldsymbol{b},\varepsilon)$*-holomorphic, i.e. satisfies Assumption 2.1.*

Then, for $\boldsymbol{y} \in U$*,* u *admits the* Taylor gpc expansion

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} u_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \qquad \text{where} \qquad u_{\boldsymbol{\nu}} = \frac{1}{\boldsymbol{\nu}!}(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u)(\boldsymbol{y})\Big|_{\boldsymbol{y}=\boldsymbol{0}}, \tag{2.4}$$

which is unconditionally convergent in $L^\infty(U;\mathbb{R})$*. With* $C_u > 0$ *as in Assumption 2.1, for every* $\boldsymbol{\nu} \in \mathcal{F}$ *and for every* $(\boldsymbol{b},\varepsilon)$*-admissible poly-radius* $\boldsymbol{\rho}$ *(i.e.,* (2.2) *holds) we have*

$$|u_{\boldsymbol{\nu}}| \leq C_u \boldsymbol{\rho}^{-\boldsymbol{\nu}} \,. \tag{2.5}$$

The bound (2.5) is a consequence of the $(\boldsymbol{b},\varepsilon)$-holomorphy on the poly-disc $B_{\boldsymbol{\rho}}$ and of the Cauchy integral theorem [19, Thm. 2.1.2], see the proof of [10, Lemma 2.4]. The unconditional convergence of the series (2.4) on U is for example discussed in [10], also see [19, Sec. 2.1] for convergence in the finite dimensional case.

For future reference we next recall three lemmata required for proving summability of sequences allowing bounds of the type (2.5). Let in the following $\boldsymbol{\alpha} = (\alpha_j)_{j\in\mathbb{N}}$ denote a sequence (not necessarily monotonic) of nonnegative real numbers.

Lemma 2.3 ([10, Lemma 7.1])**.** *Let* $p \in (0,\infty)$*. The sequence* $(\boldsymbol{\alpha}^{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ *belongs to* $\ell^p(\mathcal{F})$*, iff* $\|\boldsymbol{\alpha}\|_{\ell^p(\mathbb{N})} < \infty$ *and* $\|\boldsymbol{\alpha}\|_{\ell^\infty(\mathbb{N})} < 1$*.*

Lemma 2.4 ([10, Thm. 7.2])**.** *Let* $p \in (0,1]$*. The sequence* $(\boldsymbol{\alpha}^{\boldsymbol{\nu}}|\boldsymbol{\nu}|!/\boldsymbol{\nu}!)_{\boldsymbol{\nu}\in\mathcal{F}}$ *belongs to* $\ell^p(\mathcal{F})$ *iff* $\|\boldsymbol{\alpha}\|_{\ell^p} < \infty$ *and* $\|\boldsymbol{\alpha}\|_{\ell^1} < 1$*.*

Lemma 2.5 ([29, Lemma 2.9])**.** *Let* $\boldsymbol{x} = (x_j)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ *be a monotonically decreasing sequence of nonnegative numbers for some* $p > 0$*. Then* $x_j \leq \|\boldsymbol{x}\|_{\ell^p} j^{-1/p}$ *for all* $j \in \mathbb{N}$*.*

2.2 gpc Approximation

As has been observed in several references (see, e.g., [14, 9]), $(\boldsymbol{b},\varepsilon)$-holomorphy implies dimension-independent N-term gpc approximation rate bounds. Our analysis of the expressive power of DNNs will be based on a version of these approximation rate bounds as stated in the following theorem.

Theorem 2.6. *Let* u *be* $(\boldsymbol{b},\varepsilon)$*-holomorphic for some* $\boldsymbol{b} \in \ell^p$*,* $p \in (0,1)$*, and let* $u_{\boldsymbol{\nu}} \in \mathbb{R}$ *denote the Taylor gpc coefficient as defined in* (2.4)*. Then the sequence* $(u_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ *belongs to* $\ell^p(\mathcal{F})$*. Additionally there exists a finite constant* $C > 0$ *as well as a sequence of nested, finite and downward closed index sets* $\Lambda_n \subset \mathcal{F}$ *such that for all* $M \in \mathbb{N}$ *it holds* $|\Lambda_n| \leq n$ *and*

(i) $\sum_{\boldsymbol{\nu}\notin\Lambda_n} |u_{\boldsymbol{\nu}}| \leq CM^{1-1/p}$*,*

(ii) $\sup_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \leq C(1+\log(M))$*.*

Since it will allow us to discuss results based on related but different presumptions than the ones of Assumption 2.1 (see. Sec. 4 ahead), we provide part of the above theorem as a separate Lemma, before proceeding to the proof of Thm. 2.6.

Lemma 2.7. *Let* $r \in [1,\infty)$ *and* $p \in (0,1)$*. Let* $(u_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \mathbb{R}^{\mathcal{F}}$ *and assume that* $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}} \in \ell^{pr/(r-p)}(\mathbb{N})$ *with* $\beta_j \in (0,1)$ *for all* $j \in \mathbb{N}$*, is such that* $\sum_{\boldsymbol{\nu}\in\mathcal{F}}(\boldsymbol{\beta}^{-\boldsymbol{\nu}}|u_{\boldsymbol{\nu}}|)^r < \infty$*. Then there exists a constant* C *and for every* $n \in \mathbb{N}$ *there exists a downward closed index set* $\Lambda_n \subset \mathcal{F}$ *satisfying* $|\Lambda_n| \leq n$ *as well as (i) and (ii) of Thm. 2.6.*

Proof. Denote in the following by $\mathbf{e}_j \in \mathcal{F}$ the jth unit multiindex, i.e. $(\mathbf{e}_j)_i = 1$ if $i = j$ and $(\mathbf{e}_j)_i = 0$ otherwise. Without loss of generality we will assume that $|u_{\mathbf{e}_j}|$ is monotonically decreasing in j (otherwise we can reorder the dimensions).

First we note that $(u_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$, since with Hölder's inequality it holds

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} |u_{\boldsymbol{\nu}}|^p = \sum_{\boldsymbol{\nu}\in\mathcal{F}} |u_{\boldsymbol{\nu}}|^p \boldsymbol{\beta}^{-p\boldsymbol{\nu}}\boldsymbol{\beta}^{p\boldsymbol{\nu}} \le \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} (|u_{\boldsymbol{\nu}}|\boldsymbol{\beta}^{-\boldsymbol{\nu}})^r\right)^{\frac{p}{r}} \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\beta}^{\boldsymbol{\nu}\frac{pr}{r-p}}\right)^{\frac{r-p}{r}} < \infty, \tag{2.6}$$

where the last sum is finite since $(\boldsymbol{\beta}^{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{pr/(r-p)}$ according to Lemma 2.3 and because $\boldsymbol{\beta} \in \ell^{pr/(r-p)}(\mathbb{N})$ as well as $\|\boldsymbol{\beta}\|_{\ell^\infty} = \max_{j\in\mathbb{N}} \beta_j < 1$ (here we have used $\beta_j \to 0$ which follows by $\boldsymbol{\beta} \in \ell^{pr/(r-p)}$).

Next fix $0 < q < \infty$ such that $pr/(r-p) > q$ and set

$$\alpha_{\boldsymbol{\nu}} := \begin{cases} j^{-1/q} & \text{if } \boldsymbol{\nu} = \mathbf{e}_j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.7}$$

Now define $\zeta_{\boldsymbol{\nu}} := \max\{\boldsymbol{\beta}^{\boldsymbol{\nu}}, \alpha_{\boldsymbol{\nu}}\}$. Then $\zeta_{\boldsymbol{\nu}} \in \ell^{pr/(r-p)}$ since $(\alpha_{\mathbf{e}_j})_{j\in\mathbb{N}} \in \ell^{pr/(r-p)}$ and $(\boldsymbol{\beta}^{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{pr/(r-p)}$. Moreover

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} (\zeta_{\boldsymbol{\nu}}^{-1}|u_{\boldsymbol{\nu}}|)^r \le \sum_{\boldsymbol{\nu}\in\mathcal{F}} (\boldsymbol{\beta}^{-\boldsymbol{\nu}}|u_{\boldsymbol{\nu}}|)^r < \infty. \tag{2.8}$$

Let $\pi : \mathbb{N} \to \mathcal{F}$ be a bijection such that the sequence $(\zeta_{\pi(j)})_{j\ge1}$ is monotonically decreasing in j, and such that $\{\pi(1), \dots, \pi(n)\}$ is downward closed for any $n \in \mathbb{N}$. This is possible, because $\zeta_{\boldsymbol{\nu}}$ is monotonically decreasing in the sense that $\boldsymbol{\nu} \le \boldsymbol{\mu}$ implies $\zeta_{\boldsymbol{\nu}} \ge \zeta_{\boldsymbol{\mu}}$. Define $\Lambda_n := \{\pi(j) : 1 \le j \le n\}$. With $r' \in (1, \infty]$ denoting the Hölder conjugate of r we get

$$\sum_{\boldsymbol{\nu}\in\Lambda_n^c} |u_{\boldsymbol{\nu}}| = \sum_{\boldsymbol{\nu}\in\Lambda_n^c} \zeta_{\boldsymbol{\nu}}\zeta_{\boldsymbol{\nu}}^{-1}|u_{\boldsymbol{\nu}}| \le \|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}} \|(\zeta_{\boldsymbol{\nu}}^{-1}|u_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^r} \le C\|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}}. \tag{2.9}$$

With Lemma 2.5 we conclude that there exists a constant C such that $\zeta_{\pi(j)} \le Cj^{-(r-p)/(pr)}$. Hence, the last quantity in (2.9) can be bounded for $r > 1$, i.e. $r' = r/(r-1) < \infty$, by

$$\|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}} \le \left(C\sum_{j>n} j^{-\frac{r}{r-1}\frac{r-p}{pr}}\right)^{\frac{r-1}{r}} \le C\left(n^{1-\frac{r}{r-1}\frac{r-p}{rp}}\right)^{\frac{r-1}{r}} \le Cn^{\frac{r-1}{r}-\frac{r-p}{rp}} = Cn^{1-1/p}, \tag{2.10}$$

where we have used $(r(r-p))/((r-1)pr) > 1$ which follows by $p \in (0,1)$. For $r = 1$, i.e. $r' = \infty$, we use $\|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}} \le C\sup_{j>n} j^{1-1/p} \le Cn^{1-1/p}$ instead (where we have again employed Lemma 2.5). This shows (i).

To show (ii) note that by definition of $(\alpha_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and $(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$, it holds $\min\{\zeta_{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \Lambda_n\} \ge n^{-1/q}$. On the other hand, with $c := \sup_{j\in\mathbb{N}} \beta_j < 1$ we have $\sup\{\zeta_{\boldsymbol{\nu}} : |\boldsymbol{\nu}|_1 = d\} \ge \sup\{\boldsymbol{\beta}^{\boldsymbol{\nu}} : |\boldsymbol{\nu}|_1 = d\} \ge c^{|\boldsymbol{\nu}|_1}$. Hence with $f(d) := c^d$

$$\max_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \le f^{-1}(n^{-1/q}) = O(\log(n)), \tag{2.11}$$

which concludes the proof. □

Proof of Thm. 2.6. It is by now well-established, see [10, 11] and in particular the proof of [9, Thm. 2.2], that the Taylor coefficients of $(\boldsymbol{b}, \varepsilon)$-holomorphic maps allow bounds of the following type:

$$|u_{\boldsymbol{\nu}}| \le C\kappa^{|\nu_E|}\frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!}\boldsymbol{\gamma}^{\boldsymbol{\nu}_F}, \tag{2.12}$$

where for some fixed $J \in \mathbb{N}$ we employ the notation $\boldsymbol{\nu}_E := (\nu_1, \dots, \nu_J)$, $\boldsymbol{\nu}_F := (\nu_{J+1}, \dots)$, and $\kappa \in (0,1)$ as well as $\boldsymbol{\gamma} \in \ell^p(\mathbb{N})$ with $\|\boldsymbol{\gamma}\|_{\ell^p} < 1$ are also fixed. This is a consequence of Prop. 2.2, and we refer again to [10, 11, 9] for proofs of such statements, where it is also shown that (2.12) implies $(|u_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$. We now choose a sequence $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}} \in \ell^{p/(1-p)}(\mathbb{N})$ such that $\|\boldsymbol{\beta}\|_{\ell^\infty} < 1$ as follows

$$\beta_j := \begin{cases} \frac{2\kappa}{1+\kappa} & \text{if } j \leq J, \\ \gamma_j^{1-p} & \text{if } j > J. \end{cases} \tag{2.13}$$

Then for $\boldsymbol{\delta}$ with $\delta_j := \gamma_j \beta_{J+j}^{-1}$ for $j > J$ we have

$$\|\boldsymbol{\delta}\|_{\ell^1} = \sum_{j\in\mathbb{N}} \gamma_j^{1-(1-p)} = \|\boldsymbol{\gamma}\|_{\ell^p}^p < 1. \tag{2.14}$$

Now,

$$\begin{aligned} \sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\beta}^{-\boldsymbol{\nu}} |u_{\boldsymbol{\nu}}| &\leq C \sum_{\boldsymbol{\nu}\in\mathcal{F}} \kappa^{|\boldsymbol{\nu}_E|} \left(\frac{1+\kappa}{2\kappa}\right)^{|\boldsymbol{\nu}_E|} \frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!} \boldsymbol{\gamma}^{\boldsymbol{\nu}_F} \boldsymbol{\beta}^{-\boldsymbol{\nu}_F} \\ &= C \sum_{\boldsymbol{\nu}\in\mathbb{N}_0^J} \left(\frac{1+\kappa}{2}\right)^{|\boldsymbol{\nu}|} \sum_{\boldsymbol{\mu}\in\mathcal{F}} \frac{|\boldsymbol{\mu}|!}{\boldsymbol{\mu}!} \boldsymbol{\delta}^{\boldsymbol{\mu}}. \end{aligned} \tag{2.15}$$

By Lemma 2.3 and Lemma 2.4 the last two sums are finite, since $(1+\kappa)/2 < 1$ and because of (2.14). This proves $\sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\beta}^{-\boldsymbol{\nu}} |u_{\boldsymbol{\nu}}| < \infty$. Furthermore, Lemma 2.3 gives $(\boldsymbol{\beta}^{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p/(1-p)}$. We may thus employ Lemma 2.7 with $r = 1$, which concludes the proof. □

3 Deep Neural Network Approximations

3.1 DNN Architecture

We consider so-called *feedforward NNs (FFNNs for short)*. They are composed of layers of computational nodes and define a function $f : \mathbb{R}^n \to \mathbb{R}$. We denote by L the number of hidden layers in the NN, by N_ℓ the number of compute nodes in layer ℓ. The number $N = \sum_{\ell=1}^L N_\ell$ denotes the total number of nodes in the NN and we shall adopt it as a measure of the "number of degrees of freedom" in the NN. In the NN literature, N is also referred to as the *size* of the DNN.

The vector $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ will denote the input of the DNN, z_j^ℓ denotes the output of unit j in layer $\ell+1$, b_j^ℓ denotes the "bias" of unit j in layer ℓ. Outputs between layers of the FFNN are then characterized by the following maps:

$$z_j^{\ell+1} := \sigma\left(\sum_{i=1}^{N_\ell} w_{i,j}^\ell z_i^\ell + b_j^{\ell+1}\right), \quad \ell \in \{1, \dots, L-1\}, \;\; j \in \{1, \dots, N_{\ell+1}\}, \tag{3.1}$$

with the *input layer*

$$z_j^1 := \sigma\left(\sum_{i=1}^{n} w_{i,j}^0 x_i + b_j^1\right), \quad j \in \{1, \dots, N_1\}, \tag{3.2}$$

and the *output layer*

$$f(\boldsymbol{x}) := \sum_{i=1}^{N_L} w_{i,1}^L z_i^L + b_1^{L+1}. \tag{3.3}$$

In (3.1) - (3.2), the function $\sigma(\cdot)$ denotes the *activation function*, which here and in the following is assumed to be the so-called Rectifier Linear Unit (ReLU) $\sigma(x) = \max\{0, x\}$. As is customary in the theory of NNs, the number of layers of a NN is referred to as *depth* and the total number of nodes is referred to as *size* of the NN. The weights are assumed to take values in $\mathbb{R}$, i.e. we do not consider quantization as e.g. in [7].

3.2 Expressive Power of DNNs

To prove complexity bounds on the expressive power of DNNs for high dimensional parametric maps, we exploit the $(\boldsymbol{b}, \varepsilon)$-holomorphy and the resulting sparsity of their Taylor gpc representations (2.4). The point of departure will be the n-term truncated Taylor polynomial of the parametric map $u(\boldsymbol{y}) : U \to \mathbb{R}$, which is obtained via the gpc approximation result Theorem 2.6. In particular, we shall use recent, quantitative bounds on expressing multivariate polynomials by DNNs, from [28]. There it was observed, that deep NNs allow efficient approximation of $x \mapsto x^2$, in the sense that the number of required layers, units and weights only depends logarithmically on the absolute accuracy $\delta > 0$, up to which this function is to be approximated. This yields efficient approximation of multiplication, and ultimately entails corresponding results on the approximation of polynomials. We now recall and present some core statements from [28] in a form required in our subsequent analysis.

As mentioned above, the main task is to approximate $x \mapsto x^2$ for $x \in [0,1]$. This is achieved in [28] through the functions f_m which denote the continuous, piecewise linear spline interpolation of x^2 at the equispaced nodes $j2^{-m}$ for $j = 0, \dots, 2^m$. The pointwise error of this approximation is

$$\sup_{x\in[0,1]} |x^2 - f_m(x)| = 2^{-2m-2}. \tag{3.4}$$

Denote again by $\sigma(x) = \max\{0, x\}$ the ReLU activation function. With $f_0(x) := x = \sigma(x)$ for $x \in [0,1]$, the function f_m can be exactly expressed by a NN via

$$f_m(x) = f_{m-1}(x) - \frac{g_m(x)}{2^{2m}} \qquad \forall\, m \geq 1, \tag{3.5a}$$

where $g_m = g \circ \cdots \circ g$ is the m-fold composition of g (g_m is a “sawtooth function”), and

$$g(x) = 2\sigma(x) - 4\sigma(x - 1/2) + 2\sigma(x-1) = \begin{cases} 2x & \text{if } x < \frac{1}{2}, \\ 2(1-x) & \text{if } x \geq \frac{1}{2}, \end{cases} \tag{3.5b}$$

is the linear combination of 3 ReLUs. This shows that f_m is the output of a DNN with m hidden layers, each exhibiting 4 ReLUs as displayed in Fig. 1 (a). For some fixed $M > 0$ and a, $b \in \mathbb{R}$ with $|a|$, $|b| \leq M$, one can write $ab = M^2((|a+b|/(2M))^2 - (|a|/(2M))^2 - (|b|/(2M))^2)/8$ where $|a+b|/(2M)$, $|a|/(2M)$ and $|b|/(2M)$ are all in the interval $[0,1]$. Replacing the squared terms with the NN yields a NN approximating the multiplication of two numbers in $[-M, M]$. This argument, presented in more detail in [28], gives the following proposition:

Proposition 3.1 ([28, Prop. 3])**.** *Given $M > 0$ and $\delta \in (0,1)$ there exists a ReLU network $\tilde{\times}$ with two input units such that for all $|a|$, $|b| \leq M$, it holds $|\tilde{\times}(a,b) - a \cdot b| \leq \delta$. The depth, number of weights and number of computation units in this network behaves as $O(\log(1/\delta))$ as $\delta \to 0$.*

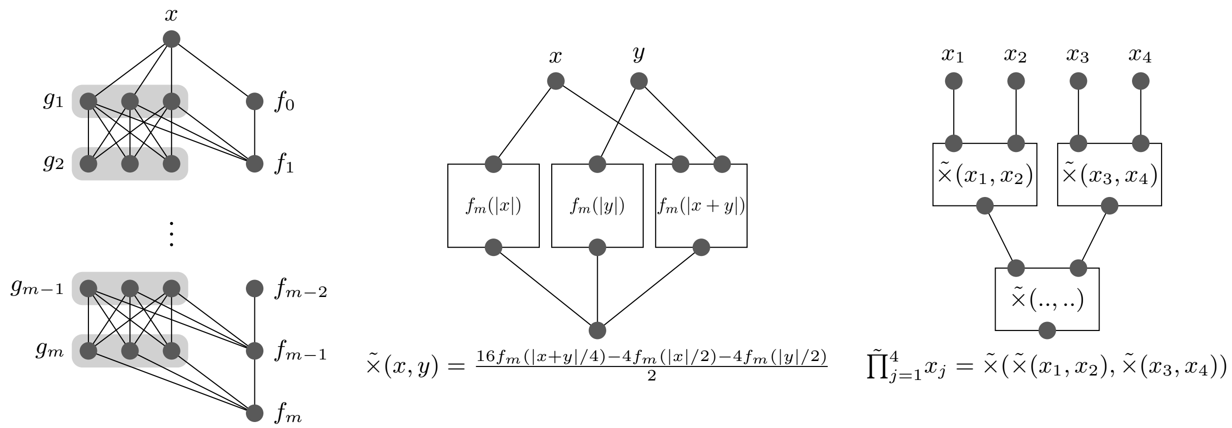
Proposition 3.1 allows to approximate the multiplication $\prod_{j=1}^n x_j$ of n numbers with $n \in \mathbb{N}$ arbitrary. We next provide a proof of this result, which slightly deviates from the constructions employed in [28] (see Rmk. 3.4). The following short Lemma will be required in the proof.

Lemma 3.2. *Let $a_0 = 1$, $\varepsilon \geq 0$ fixed and $a_{j+1} := a_j^2 + \varepsilon$, $j \in \mathbb{N}_0$. Then $a_j \leq (1+2\varepsilon)^{2^j}$ for all $j \in \mathbb{N}$.*

Proof. We prove by induction that $a_j \leq (1+2\varepsilon)^{2^j} - \varepsilon$. This is true for $j = 1$ since $a_1 = 1 + \varepsilon \leq (1+2\varepsilon)^2 - \varepsilon$. For the induction step we obtain

$$\begin{aligned} a_{j+1} = a_j^2 + \varepsilon &\leq ((1+2\varepsilon)^{2^j} - \varepsilon)^2 + \varepsilon = (1+2\varepsilon)^{2^{j+1}} - 2(1+2\varepsilon)^{2^j}\varepsilon + \varepsilon^2 + \varepsilon \\ &\leq (1+2\varepsilon)^{2^{j+1}} - 2\varepsilon - 4\varepsilon^2 + \varepsilon^2 + \varepsilon \leq (1+2\varepsilon)^{2^{j+1}} - \varepsilon, \end{aligned} \tag{3.6}$$

which shows the claim. ☐



(a) Approximation of x^2 (b) Approximation of xy (c) Approximation of $\prod_{j=1}^n x_j$

Figure 1: Subfigure (a) shows the network approximating $[0,1] \ni x \mapsto x^2$ via $f_j(x) = f_{j-1}(x) - g_j(x)/2^{2j}$, cp. (3.5). Subfigure (b) shows the network approximating $[0,2]^2 \ni (x,y) \mapsto xy$. The boxes contain the network from subfigure (a) e.g. applied to $|x|/2 = \sigma(x/2) + \sigma(-x/2)$. Subfigure (c) shows the network approximating $\prod_{j=1}^n x_j$ for $n = 4$ where $|x_j| \le 1$ for all j. The boxes contain the approximate multiplication $\tilde{\times}$ from (b).

Corollary 3.3. *Let $\delta \in (0,1)$. There exists a NN $\tilde{\prod}$ with n input units such that for $x_1, \dots, x_n$ with $|x_i| \le 1$ for all i, it holds $|\prod_{j=1}^n x_j - \tilde{\prod}(x_1, \dots, x_n)| \le \delta$.*

There exists a constant C such that for every $n \in \mathbb{N}$ and for every accuracy $0 < \delta < 1$, the number of ReLUs and weights in the NN $\tilde{\prod}$ is bounded by $C(1 + n \log(n/\delta))$ and the depth of the network $\tilde{\prod}$ is bounded by $C(1 + \log(n) \log(n/\delta))$.

Proof. Step 1: We construct the network. If $n \notin \{2^j : j \in \mathbb{N}\} =: A$, let $\tilde{n} = \min\{x \ge n : x \in A\}$ and define $x_j := 1$ for all $j \in \{n+1, \dots, \tilde{n}\}$. We note that $\tilde{n} \le 2n$. We let $\tilde{\times}$ as in Prop. 3.1: more precisely, for x, y with $|x|$, $|y| \le 1$ define (see Fig. 1 (b))

$$\tilde{\times}(x,y) := \frac{16 f_m(|x+y|/4) - 4 f_m(|x|/2) - 4 f_m(|y|/2)}{2}, \tag{3.7}$$

where $m = C_1 \log(\tilde{n}/\delta)$ and the constant $C_1 \ge 1/\log(2)$ is to be chosen subsequently independent of $\tilde{n}$ and δ. Then $\tilde{\times}$ is a neural network approximating the multiplication of two numbers in $[-2,2]$ such that

(i) the number of layers, weights and units is bounded by $O(m) = O(C_1 \log(\tilde{n}/\delta))$ (see Fig. 1 (a)),

(ii) for all $|x|$, $|y| \le 2$ it holds

$$\begin{aligned} |\tilde{\times}(x,y) - xy| &= \left| \tilde{\times}(x,y) - \frac{16(|x+y|/4)^2 - 4(x/2)^2 - 4(y/2)^2}{2} \right| \\ &\le \frac{16|f_m(|x+y|/4) - (|x+y|/4)^2| + 4|f_m(|x|/2) - (|x|/2)^2| + 4|f_m(|y|/2) - (|y|/2)^2|}{2} \\ &\le \frac{(16+4+4)2^{-2m-2}}{2} = 3 \cdot 2^{-2m}, \end{aligned} \tag{3.8}$$

where we have used (3.4),

(iii) for all $|x|$, $|y| \leq 2$ it holds $|\tilde{\times}(x,y)| \leq |xy| + 3 \cdot 2^{-2m}$.

For any even positive integer k we define

$$R(y_1,\dots,y_k) := (\tilde{\times}(y_1,y_2),\dots,\tilde{\times}(y_{k-1},y_k)) \in \mathbb{R}^{k/2} \tag{3.9}$$

and (cp. Fig. 1 (c))

$$\tilde{\prod}(x_1,\dots,x_n) := \underbrace{R \circ \cdots \circ R}_{\log_2(\tilde{n})}(x_1,\dots,x_{\tilde{n}}) \in \mathbb{R}. \tag{3.10}$$

In the following, we use the notation $R^{\log_2(\tilde{n})}$ instead which records in the exponent the number of compositions.

Step 2: We now show that, upon choosing C_1 large enough, it holds $R^j(x_1,\dots,x_{\tilde{n}}) \in [-2,2]^{\tilde{n}/2^j}$ for all $j = 1,\dots,\log_2(\tilde{n})$. Define $\varepsilon := 3 \cdot 2^{-2m}$ where as above $m = C_1 \log(\tilde{n}/\delta)$, i.e. $\varepsilon = \varepsilon(\tilde{n})$. By item (iii) and since $|x_j| \leq 1$ for all j, we get $R(x_1,\dots,x_{\tilde{n}}) \in [-1-\varepsilon, 1+\varepsilon]^{\tilde{n}/2}$ and inductively with $a_0 := 1$, $a_{j+1} := a_j^2 + \varepsilon$ it holds $R^j(x_1,\dots,x_{\tilde{n}}) \in [-a_j, a_j]^{\tilde{n}/2^j}$ for $j = 1,\dots,\log_2(\tilde{n})$. Using Lemma 3.2, it suffices to show that $(1+2\varepsilon(\tilde{n}))^{2^{\log_2(\tilde{n})}} = (1+2\varepsilon(\tilde{n}))^{\tilde{n}} \leq 2$ for all $\tilde{n} \in \mathbb{N}$. We have for $0 < \delta \leq 1$

$$\begin{aligned}
(1+2\varepsilon(\tilde{n}))^{\tilde{n}} &= (1 + 2(3 \cdot 2^{-2C_1 \log(\tilde{n}/\delta)}))^{\tilde{n}} \\
&\leq (1 + 6 \cdot 2^{-2C_1 \log(\tilde{n})})^{\tilde{n}} = (1 + 6\tilde{n}^{-2\log(2)C_1})^{\tilde{n}} \\
&= \exp\left(\tilde{n} \log(1 + 6\tilde{n}^{-2\log(2)C_1})\right).
\end{aligned} \tag{3.11}$$

Since $\log(1+x) = x + O(x^2)$ asymptotically as $x \to 0$, the exponent behaves like

$$6\tilde{n}^{1-2\log(2)C_1} \to 0$$

as either ($C_1 > 1/(2\log(2))$ and) $\tilde{n} \to \infty$ or ($\tilde{n} \geq 2$ and) $C_1 \to \infty$. Hence $\sup_{\tilde{n}\in\mathbb{N}}(1+2\varepsilon(\tilde{n}))^{\tilde{n}} \leq 2$ provided that $C_1 > 0$ is large enough.

Step 3: We estimate the error. By item (ii) it holds $|\tilde{\times}(x,y) - xy| \leq \varepsilon = 3 \cdot 2^{-2m}$ for all $|x|$, $|y| \leq 2$. We claim that for all $r \in \mathbb{N}$ and all $b_1,\dots,b_{2^r}$ such that $R^j(b_1,\dots,b_{2^r}) \in [-2,2]^{\tilde{n}/2^j}$ for all $j = 0,\dots,r$, it holds that $|R^r(b_1,\dots,b_{2^r}) - \prod_{j=1}^{2^r} b_j| \leq (4^r - 1)\varepsilon$. Note that for $C_1 \geq 1/\log(2)$ our global choice $m = C_1 \log(\tilde{n}/\delta)$ ensures $\varepsilon = 3 \cdot 2^{-2m} \leq 3(\delta/\tilde{n})^2 \leq 3\delta/\tilde{n}^2$. With $r = \log_2(\tilde{n})$ and the statement from Step 2, this will prove the desired bound

$$\left|\prod_{j=1}^{n} x_j - \tilde{\prod}(x_1,\dots,x_n)\right| = \left|\prod_{j=1}^{\tilde{n}} x_j - \tilde{\prod}(x_1,\dots,x_{\tilde{n}})\right| \leq (\tilde{n}^2-1)\varepsilon \leq (\tilde{n}^2-1)3\delta/\tilde{n}^2 \leq 3\delta\ . \tag{3.12}$$

We proceed by induction over r to verify the above claim. The case $r = 1$ is trivial since by assumption $|R(b_1,b_2) - b_1 b_2| = |\tilde{\times}(b_1,b_2) - b_1 b_2| \leq \varepsilon \leq (4^r - 1)\varepsilon$. For the induction step, note that $R^r(b_1,\dots,b_{2^r}) = \tilde{\times}(R^{r-1}(b_1,\dots,b_{2^{r-1}}), R^{r-1}(b_{2^{r-1}+1},\dots,b_{2^r}))$. We get

$$\begin{aligned}
\left|\prod_{j=1}^{2^r} b_j - R^r(b_1,\dots,b_{2^r})\right| &= \left|\prod_{j=1}^{2^{r-1}} b_j \prod_{i=2^{r-1}+1}^{2^r} b_i - \tilde{\times}(R^{r-1}(b_1,\dots,b_{2^{r-1}}), R^{r-1}(b_{2^{r-1}+1},\dots,b_{2^r}))\right| \\
&\leq \left|\tilde{\times}(R^{r-1}(b_1,\dots,b_{2^{r-1}}), R^{r-1}(b_{2^{r-1}+1},\dots,b_{2^r})) - R^{r-1}(b_1,\dots,b_{2^{r-1}}) \cdot R^{r-1}(b_{2^{r-1}+1},\dots,b_{2^r})\right| \\
&\quad + \left|\prod_{j=1}^{2^{r-1}} b_j\right| \left|\prod_{i=2^{r-1}+1}^{2^r} b_i - R^{r-1}(b_{2^{r-1}+1},\dots,b_{2^r})\right| \\
&\quad + \left|R^{r-1}(b_{2^{r-1}+1},\dots,b_{2^r})\right| \left|\prod_{j=1}^{2^{r-1}} b_j - R^{r-1}(b_1,\dots,b_{2^{r-1}})\right|.
\end{aligned} \tag{3.13}$$

We denote the last three terms by $T_1 + T_2 + T_3$. To bound the first term we use $s_1 := R^{r-1}(b_1, \ldots, b_{2^r-1})$, $s_2 := R^{r-1}(b_{2^{r-1}+1}, \ldots, b_{2^r}) \in [-2, 2]$ by assumption, which gives $T_1 = |\tilde{\times}(s_1, s_2) - s_1 \cdot s_2| \le \varepsilon$ by (ii). For T_2 we use $|\prod_{j=1}^{2^{r-1}} b_j| \le 1$ (since $|b_i| \le 1$ for all i) and the induction hypothesis which gives $T_2 \le (4^{r-1}-1)\varepsilon$. In the same way we obtain $T_3 \le 2(4^{r-1}-1)\varepsilon$, where we employed $|R^{r-1}(b_{2^{r-1}+1}, \ldots, b_{2^r})| \le 2$. Overall $T_1+T_2+T_3 \le (1+4^{r-1}-1+2(4^{r-1}-1))\varepsilon \le (4^r - 1)\varepsilon$, which proves the claim.

Step 4: Finally we sum up all layers, weights and units. The operator $\tilde{\prod}$ in (3.10) describes a NN with $O(\log_2(\tilde{n}) \log(\tilde{n}/\delta))$ layers and with $O(\tilde{n} \log(\tilde{n}/\delta))$ weights and ReLUs: first note that we may use one layer to create the values $1 = x_{n+1} = \cdots = x_{\tilde{n}}$ as $1 = x_j = \sigma(1 + 0 \cdot x_1)$ for $j = n+1, \ldots, \tilde{n}$ and write $x_j = \sigma(x_j) - \sigma(-x_j)$ for $j = 1, \ldots, n$ to copy the n input values $x_1, \ldots, x_n$ to the second layer (which is the first hidden layer). Next, the first application of R employs $\tilde{n}/2$ times the neural network $\tilde{\times}$ with the inputs from the first hidden layer. Hence, by (i) this adds $O(\log(\tilde{n}/\delta))$ layers and in total $O(\log(\tilde{n}/\delta)\tilde{n}/2)$ weights and units. For the second application of R we employ the neural network $\tilde{\times}$ exactly $\tilde{n}/4$ times, which adds another $O(\log(\tilde{n}/\delta))$ layers and $O(\log(\tilde{n}/\delta)\tilde{n}/4)$ weights and units. After $\log_2(\tilde{n})$ steps we end up with $O(\log_2(\tilde{n}) \log(\tilde{n}/\delta))$ layers and $O(\log(\tilde{n}/\delta)\tilde{n})$ weights and units. Since $\tilde{n} \le 2n$, this shows that the network uses the stated number of units, weights and layers. □

Remark 3.4. *The proof in [28] uses $\tilde{\times}(a_1, \tilde{\times}(a_2, \ldots, \tilde{\times}(a_{n-1}, a_n)))$ to approximate $\prod_{j=1}^n a_j$. This would give $O(n \log(n/\delta))$ layers in Cor. 3.3. On the other hand, this construction has the advantage of giving all products $\prod_{j=l}^n a_j$ for $l = 1, \ldots, n$ in between, which is convenient when approximating a polynomial $\sum_{j=1}^n c_j x^j$ where all values $x, \ldots, x^n$ are needed.*

Remark 3.5. *Similar results in terms of the depth and number of units as we have cited here were obtained in [23] using a NN composed of ReL and BiS ("binary step") units.*

3.3 DNN Approximation of $(\boldsymbol{b}, \varepsilon)$-holomorphic maps

We now give a result on the expressive power of neural networks concerning $(\boldsymbol{b}, \varepsilon)$-holomorphic functions. It states that, up to logarithmic terms, ReLU based DNNs are capable of approximating $(\boldsymbol{b}, \varepsilon)$-holomorphic maps at rates equivalent to those achieved by best n-term approximation. Here, the notion "rate" is understood in terms of the network size, i.e., in terms of the used total number of units, weights and layers. [Notation: $N \to n$]

Theorem 3.6. *Let $u : U \to \mathbb{R}$ be $(\boldsymbol{b}, \varepsilon)$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, and with some $p \in (0, 1)$.*

Then, there exists a constant C and for every $n \in \mathbb{N}$ there exists a ReLU network $\tilde{u}(y_1, \ldots, y_n)$ with n input units whose size, i.e., the total number of units and weights, is bounded by $C(1 + n \log(n) \log\log(n))$, and whose depth is bounded by $C(1 + \log(n) \log\log(n))$ which satisfies the uniform error bounds

$$\sup_{\boldsymbol{y} \in U} |u(\boldsymbol{y}) - \tilde{u}(y_1, \ldots, y_n)| \le C n^{1-1/p} .$$

Proof. According to Prop. 2.2 and Thm. 2.6, there exist downward closed index sets Λ_n with $|\Lambda_n| \le n$ such that

$$\sup_{\boldsymbol{y} \in U} \left| u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_n} u_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \right| \le \sum_{\boldsymbol{\nu} \in \Lambda_n^c} |u_{\boldsymbol{\nu}}| \le C n^{-1/p+1}. \tag{3.14}$$

It thus suffices to approximate $u_n := \sum_{\boldsymbol{\nu} \in \Lambda_n} u_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}}$ with a DNN. Fix n and let $\pi : \{1, \ldots, n\} \to \Lambda_n$ be such that $|u_{\pi(j)}|$ is monotonically decreasing in j. By Thm. 2.6, $(u_{\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}} \in \ell^p(\mathcal{F})$ so that by Lemma 2.5 we obtain $|u_{\pi(j)}| \leq Cj^{-1/p}$. Note that C does not depend on n here. Define $\varepsilon_j := (j/n)^{1/p}$. Below we shall construct a NN which yields, for every $\boldsymbol{y} \in U$, the output

$$\sum_{\boldsymbol{\nu} \in \Lambda_n} u_{\boldsymbol{\nu}} f_{\boldsymbol{\nu}}(\boldsymbol{y}) \tag{3.15}$$

where $\sup_{\boldsymbol{y} \in U} |\boldsymbol{y}^{\pi(j)} - f_{\pi(j)}(\boldsymbol{y})| \leq \varepsilon_j$ for all $j \in \{1, \ldots, n\}$. Then by (3.14) the total error is bounded by

$$\begin{aligned} \sup_{\boldsymbol{y} \in U} \left| u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_n} u_{\boldsymbol{\nu}} f_{\boldsymbol{\nu}}(\boldsymbol{y}) \right| &\leq Cn^{-1/p+1} + \sum_{j=1}^{n} |u_{\pi(j)}| \sup_{\boldsymbol{y} \in U} |\boldsymbol{y}^{\pi(j)} - f_{\pi(j)}(\boldsymbol{y})| \\ &\leq Cn^{-1/p+1} + \sum_{j=1}^{n} Cj^{-1/p} \left(\frac{j}{n} \right)^{1/p} \leq Cn^{-1/p+1}. \end{aligned} \tag{3.16}$$

It remains to construct the NN as stated above. First observe that u_n depends on at most n variables, since $|\Lambda_n| \leq n$ and Λ_n is downward closed: for each $j \in \operatorname{supp} \boldsymbol{\nu}$ for some $\boldsymbol{\nu} \in \Lambda_n$, it must hold $\mathbf{e}_j \in \Lambda_n$ which implies

$$|\{j \,:\, j \in \operatorname{supp} \boldsymbol{\nu} \wedge \boldsymbol{\nu} \in \Lambda_n\}| \leq |\Lambda_n| \leq n. \tag{3.17}$$

We use $y_1, \ldots, y_n$ as the input of our NN. Next, writing $y_j = \max\{0, y_j\} - \max\{0, -y_j\}$, we observe that it is possible to obtain $\sum_{\boldsymbol{\nu} \in \Lambda} \nu_j$ copies of y_j for each $j = 1, \ldots, n$, by using one layer with $2 \sum_{\boldsymbol{\nu} \in \Lambda_n} |\boldsymbol{\nu}|_1$ ReLUs and weights. By item (ii) of Thm. 2.6, it holds $2 \sum_{\boldsymbol{\nu} \in \Lambda_n} |\boldsymbol{\nu}|_1 \leq C(1 + n \log(n))$. Next, we fix $\boldsymbol{\nu} = \pi(j) \in \Lambda_n$ and invoke Cor. 3.3 to approximate $\boldsymbol{y}^{\boldsymbol{\nu}}$: Wlog let $\boldsymbol{y}^{\boldsymbol{\nu}} = \prod_{i=1}^{n} y_i$ (in the general case some y_i may occur multiple times and $\operatorname{supp} \boldsymbol{\nu}$ can be any subset of $\{1, \ldots, n\}$). Define $f_{\boldsymbol{\nu}}(\boldsymbol{y}) := \tilde{\prod}(y_1, \ldots, y_n)$ as in Cor. 3.3 with $\delta = \varepsilon_j$. Then Cor. 3.3 gives $\sup_{\boldsymbol{y} \in U} |\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}(\boldsymbol{y})| \leq \varepsilon_j$. By this corollary, the number of layers required by the NN $f_{\boldsymbol{\nu}}$ is bounded by $C(1 + \log(|\boldsymbol{\nu}|_1) \log(|\boldsymbol{\nu}|_1/\varepsilon_j))$ which is bounded by $C(1 + \log(n) \log\log(n))$ since $|\boldsymbol{\nu}|_1 \leq C(1 + \log(n))$ according to Thm. 2.6, and $n^{-1/p} \leq \varepsilon_j \leq 1$ for $\boldsymbol{\nu} \in \Lambda_n$ by definition. Since $\boldsymbol{\nu} \in \Lambda_n$ was arbitrary, this gives the upper bound $O(\log(n) \log\log(n))$ on the depth of the NN (3.15). Finally, employing again Cor. 3.3 and $|\pi(j)|_1 \leq C(1 + \log(n))$ for $j = 1, \ldots, n$, we sum up the number of weights and units as

$$\begin{aligned} \sum_{j=1}^{n} C \left(1 + |\pi(j)|_1 \left| \log \left(C + C \frac{|\pi(j)|_1}{\varepsilon_j} \right) \right| \right) &\leq Cn + C \sum_{j=1}^{n} \log(n) \log \left(C \log(n) \left(\frac{n}{j} \right)^{1/p} \right) \\ &\leq Cn + Cn \log(n) \log\log(n) + C \log(n) \sum_{j=1}^{n} \log \left(\frac{n}{j} \right). \end{aligned} \tag{3.18}$$

The last sum can be estimated by

$$\begin{aligned} \sum_{j=1}^{n} \log \left(\frac{n}{j} \right) &\leq \log(n) + \int_{1}^{n} \log(n/x) \mathrm{d}x = \log(n) - n \int_{1/n}^{1} \log(y) \mathrm{d}y \\ &= \log(n) - n \left(-1 - \frac{1}{n} \log \left(\frac{1}{n} \right) + \frac{1}{n} \right) = O(n), \end{aligned} \tag{3.19}$$

which concludes the proof. □

We remark that the network size $N(n)$ is bounded by $Cn^{1+\varepsilon}$ for arbitrary small $\varepsilon > 0$, in terms of the number n of input parameters. Thus Thm. 3.6 implies the convergence $N^{-1/p+1+\varepsilon}$ with $\varepsilon > 0$ arbitrary in terms of the network size.

4 Examples, Generalizations and Conclusions

We present a brief summary of our main results, illustrate their relevance to parametric PDEs with a simple example (a linear elliptic PDE with uncertain, parametric coefficient) and indicate several generalizations which can be obtained by combining known gpc approximation rate results with the presently proposed NN approximations of polynomials.

We have proved convergence rate bounds for a class of DNNs for a class of real-valued functions f which depend holomorphically on a sequence $\boldsymbol{y} = (y_j)_{j\geq 1}$ of (possibly infinitely many) parameters, and which allow for a sparse gpc expansion with respect to these parameters. Their relevance stems from the fact that functions of this type are obtained as response surfaces of operator equations with distributed, uncertain input data in function spaces. Specifically, we considered functions of countably many parameters y_j which admit Taylor gpc expansions (2.4) that are sparse in the sense that the sequence $(u_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ of Taylor gpc coefficients is p-summable for some $0 < p < 1$. Such functions arise as response surfaces of broad classes of countably-parametric PDEs which model systems with distributed uncertain inputs in engineering and in the sciences. Our main result, Theorem 3.6, implies that such response surfaces can be learned with accuracy $\delta > 0$ (uniform w.r. to the parameter vector $\boldsymbol{y}$) by DDNs of size bounded (up to logarithmic factors) by $C\delta^{-1/s}$ where $s = 1/p-1$ and with a constant $C > 0$ that is independent of the dimension of the input data (3.14). For $\boldsymbol{b} \in \ell^p(\mathbb{N})$ with $0 < p < 1$, the DNN approximation essentially reproduces the gpc n-term approximation rates.

The argument in the proof of Theorem 3.6 consisted in reapproximating n-term truncated gpc expansions of the response surface by DNNs, and in showing that this reapproximation could be achieved by DNNs with $O(\log\varepsilon)$ many nodes for each monomial term. The curse of dimensionality was overcome by exploiting the separability of the gpc bases (here $\boldsymbol{y}^{\boldsymbol{\nu}}$).

Analogous results will hold also for other types of gpc expansions, where $\boldsymbol{y}^{\boldsymbol{\nu}}$ is replaced by tensor products of other systems of polynomials such as, for example, Tschebyscheff or Jacobi polynomials [1, 2]. Also, when higher convergence rates of n-term gpc approximations are available, these can be expected to translate into improved rates of expressive power of DNNs with N units.

Finally, let us present one example of a parametric partial differential equation where the foregoing theory can be applied. It also serves to indicate that the preceding statements admit certain refinements, and analogous proofs in the nonanalytic case.

In a bounded domain $D \subset \mathbb{R}^d$, consider the elliptic diffusion equation

$$-\mathrm{div}(a\nabla u) = f, \tag{4.1}$$

for a given right-hand side $f \in L^2(D)$, with homogeneous Dirichlet boundary conditions $u|_{\partial D} = 0$. The scalar diffusion coefficient a is assumed to be spatially variable. Using the notation $V = H_0^1(D)$ and $V' = H^{-1}(D)$, for any $f \in V'$, the weak formulation of (4.1) in V,

$$\int_D a\nabla u \cdot \nabla v = \langle f, v\rangle_{V',V}, \quad v \in H_0^1(D), \tag{4.2}$$

admits a unique solution $u \in V$ provided the diffusion coefficient a satisfies $0 < r < a < R < \infty$.

We consider diffusion coefficients having a parametrized form $a = a(\boldsymbol{y})$, where the parameter sequence $\boldsymbol{y} = (y_j)_{j\geq 1}$ is a sequence of real-valued parameters ranging in $U = [-1, 1]^{\mathbb{N}}$. The resulting solution map

$$\boldsymbol{y} \mapsto u(\boldsymbol{y}), \tag{4.3}$$

acts from the parameter domain U to the solution space V. *Affine-parametric input* representations arise, for example, from Fourier-, Karhunen-Loève-, spline- or wavelet series representations of a:

$$a = a(\boldsymbol{y}) = \bar{a} + \sum_{j\geq 1} y_j\psi_j. \tag{4.4}$$

Here, $\bar{a}$ and $(\psi_j)_{j\geq 1}$ are known functions in $L^\infty(D)$. Under the so-called *uniform ellipticity assumption*

$$\sum_{j\geq 1} |\psi_j| \leq \bar{a} - r, \quad \text{a.e. on } D, \tag{4.5}$$

for some $r > 0$, the parametric solution $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in V$ is well-defined.

The parametric solution $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in V$, admits the *Taylor gpc expansion*

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} u_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \ , \qquad \boldsymbol{y} \in U \ , \tag{4.6}$$

with unconditional convergence in the norm of V at every $\boldsymbol{y} \in U$. Here, the affine-parametric nature of the parametric coefficient $a(\boldsymbol{y})$ allows for the following improved summability result for the V-norms of the Taylor gpc coefficients $u_{\boldsymbol{\nu}}$.

Theorem 4.1 ([1, Thm.1])**.** *Let $0 < q < \infty$ and $0 < p < 2$ be such that $1/p = 1/q + 1/2$. Assume that $\bar{a} \in L^\infty(D)$ is such that $\operatorname{ess\,inf} \bar{a} > 0$, and that there exists a sequence $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}}$ of positive numbers strictly smaller than 1 such that $\boldsymbol{\beta} \in \ell^q(\mathbb{N})$ and such that*

$$\theta := \left\| \frac{\sum_{j\geq 1} \beta_j^{-1} |\psi_j|}{\bar{a}} \right\|_{L^\infty} < 1 \ . \tag{4.7}$$

Then $\sum_{\boldsymbol{\nu}\in\mathcal{F}} (\boldsymbol{\beta}^{-\boldsymbol{\nu}} \|u_{\boldsymbol{\nu}}\|_V)^2 < \infty$, and in particular $(\|u_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$.

We remark that the proof of Theorem 4.1 is *not based on holomorphy*, but rather on real variable arguments combined with induction w.r. to the total differentiation order $|\boldsymbol{\nu}|_1$ of the Taylor coefficient $u_{\boldsymbol{\nu}}$. Using Lemma 2.7 with $r = 2$, a proof completely analogous to the one of Thm. 3.6 yields:

Corollary 4.2. *Let $p \in (0, 1)$ and $q = 2p/(2 - p)$, i.e. $1/p = 1/q + 1/2$. Then, under the assumptions of Thm. 4.1, there exists a constant C such that for every $G(\cdot) \in V'$ and for every $n \in \mathbb{N}$ there exists a ReLU network $\tilde{g}(y_1, \ldots, y_n)$ with n input units whose number of units and weights is bounded by $C(1 + n \log(n) \log\log(n))$, its depth is bounded by $C(1 + \log(n) \log\log(n))$ and there hold the uniform error bounds*

$$\sup_{\boldsymbol{y}\in U} |G(u(\boldsymbol{y})) - \tilde{g}(y_1, \ldots, y_n)| \leq C\|G\|_{V'} n^{1-1/p} \ .$$

The assertion of this corollary follows from Theorem 3.6 with the function $\boldsymbol{y} \mapsto g(\boldsymbol{y}) := G(u(\boldsymbol{y}))$, where $\boldsymbol{y} \mapsto u(\boldsymbol{y}) \in V$ denotes the solution of the PDE (4.1) with parametric coefficient (4.4) and with the DNN $\tilde{u}$ from Theorem 3.6 denoted by $\tilde{g}$.

We conclude that Theorem 3.6 shows that response functions of many-parametric operator equations can, in principle, be expressed by deep ReLU NNs with error vs. network size N at an approximation rate which is free from the curse of dimensionality. Moreover, this approximation rate is only limited by the sparsity of the parametric solutions' gpc expansion.

We also mention that in the *Bayesian Inversion* of many-parametric PDE models in the presence of noisy data (see, e.g., [27] for the mathematical formulation), the expectations of quantities of interest conditional on the observation data, can be expressed as high dimensional integrals w.r. to a posterior Bayesian density which is $(\boldsymbol{b}, \varepsilon)$-holomorphic when the assumptions of the abstract theory in [27] is satisfied. We refer to [26] for a verification in the above, affine-parametric setting. The present results therefore open the perspective of deep learning of Bayesian posteriors for PDEs.

This opens a new direction in the approximation of responses of complex PDE models in the sciences and in engineering, by machine learning methodologies combined with suitable DNN architectures. Let us mention that DNN approximations with unsupervised training by stochastic gradient descent have recently been reported to be effective in the valuation of financial derivatives on large baskets of risky assets [5, 18]. The structure of the value function of such contracts does not readily fit into the class of $(\boldsymbol{b}, \varepsilon)$-holomorphic, many-parametric functions, so that the present results do not imply corresponding approximation results.

We emphasize in closing that the present results quantifying the expressive power of DNNs are approximation results. However, we remark that the proof of Theorem 3.6 is constructive. In principle, when combined with results from [29] on the localization of the sets Λ_n of active gpc coefficients, this information could be used in so-called "supervised learning" approaches for training the corresponding DNNs.

In practice, however, "non-supervised" training methodologies are often preferred. The present results, together with the (empirically) observed good performance of widely used training algorithms for DNNs such as stochastic gradient descent (see, e.g. [6] and the references there) imply also new perspectives on the numerical solution of forward and inverse problems of parametric and stochastic PDEs. This aspect will be developed elsewhere.

References

[1] M. Bachmayr, A. Cohen, D. Düng, and C. Schwab. Fully discrete approximation of parametric and stochastic elliptic PDEs. Technical Report 2017-08, Seminar for Applied Mathematics, ETH Zürich, 2017.

[2] M. Bachmayr, A. Cohen, and G. Migliorati. Sparse polynomial approximation of parametric elliptic PDEs. Part I: Affine coefficients. *ESAIM Math. Model. Numer. Anal.*, 51(1):321–339, 2017.

[3] A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics (Spetses, 1990)*, volume 335 of *NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci.*, pages 561–576. Kluwer Acad. Publ., Dordrecht, 1991.

[4] A. R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945, 1993.

[5] C. Beck, W. E, and A. Jentzen. Machine learning approximation algorithms for high-dimensional fully nonlinear partial differential equations and second-order backward stochastic differential equations. Technical Report 2017-49, Seminar for Applied Mathematics, ETH Zürich, 2017.

[6] Y. Bengio. Practical recommendations for gradient-based training of deep architectures, 2012. https://arxiv.org/pdf/1206.5533.

[7] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks, 2017. arXiv:1705.01714.

[8] A. Chkifa, A. Cohen, and C. Schwab. High-dimensional adaptive sparse polynomial interpolation and applications to parametric pdes. *Journ. Found. Comp. Math.*, 14(4):601–633, 2013.

[9] A. Cohen, A. Chkifa, and C. Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric pdes. *Journ. Math. Pures et Appliquees*, 103(2):400–428, 2015.

[10] A. Cohen, R. DeVore, and C. Schwab. Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. *Found. Comput. Math.*, 10(6):615–646, 2010.

[11] A. Cohen, R. Devore, and C. Schwab. Analytic regularity and polynomial approximation of parametric and stochastic elliptic PDE's. *Anal. Appl. (Singap.)*, 9(1):11–47, 2011.

[12] A. Cohen, C. Schwab, and J. Zech. Shape Holomorphy of the stationary Navier-Stokes Equations. Technical Report 2016-45, Seminar for Applied Mathematics, ETH Zürich, 2016. (to appear in SIAM Journ. Math. Analysis (2018)).

[13] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Proc. of 29th Ann. Conf. Learning Theory*, pages 698–728, 2016. arXiv:1509.05009v3.

[14] D. Dũng, C. Schwab, and J. Zech. Sparse approximation of solutions of holomorphic parametric operator equations. 2017. in preparation.

[15] J. Dick, R. N. Gantner, Q. T. Le Gia, and C. Schwab. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. *Math. Models Methods Appl. Sci.*, 27(5):953–995, 2017.

[16] K.-I. Funahashi. Approximate realization of identity mappings by three-layer neural networks. *Electron. Comm. Japan Part III Fund. Electron. Sci.*, 73(11):61–68 (1991), 1990.

[17] P. Grohs, T. Wiatowski, and H. H. Boelcskei. Deep convolutional neural networks on cartoon functions. Technical Report 2016-25, Seminar for Applied Mathematics, ETH Zürich, 2016.

[18] J. Han, A. Jentzen, and W. E. Overcoming the curse of dimensionality: Solving high-dimensional partial differential equations using deep learning. Technical Report 2017-44, Seminar for Applied Mathematics, ETH Zürich, 2017.

[19] M. Hervé. *Analyticity in infinite-dimensional spaces*, volume 10 of *de Gruyter Studies in Mathematics*. Walter de Gruyter & Co., Berlin, 1989.

[20] K. Hornik. Approximation capabilities of multilayer feedforward networks. *Neural networks*, 4(2):251–257, 1991.

[21] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[22] C. Jerez-Hanckes, C. Schwab, and J. Zech. Electromagnetic wave scattering by random surfaces: Shape holomorphy. *Math. Mod. Meth. Appl. Sci.*, 27, 2017.

[23] S. Liang and R. Srikant. Why deep neural networks for function approximation? In *Proc. of ICLR 2017*, pages 1 – 17, 2017. arXiv:1610.04161.

[24] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: an approximation theory perspective. *Anal. Appl. (Singap.)*, 14(6):829–848, 2016.

[25] A. Quarteroni and G. Rozza. Numerical solution of parametrized Navier–Stokes equations by reduced basis methods. *Numerical Methods for Partial Differential Equations*, 23(4):923–948, 2007.

[26] C. Schwab and A. M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Problems*, 28(4):045003, 32, 2012.

[27] A. M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010.

[28] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 94:103–114, 2017.

[29] J. Zech and C. Schwab. Convergence rates of high dimensional Smolyak quadrature. Technical Report 2017-27, Seminar for Applied Mathematics, ETH Zürich, 2017.