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: May 2018

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

Funding SNF: 159940

Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ*

Christoph Schwab and Jakob Zech†

May 3, 2018

Abstract

We estimate the expressive power of certain deep Neural Networks (DNNs for short) on a class of countably-parametric, holomorphic maps $u : U \to \mathbb{R}$ on the parameter domain $U = [-1, 1]^{\mathbb{N}}$. Dimension-independent rates of best n-term truncations of generalized polynomial chaos (gpc for short) approximations depend only on the summability exponent of the sequence of their gpc expansion coefficients. So-called $(\boldsymbol{b}, \varepsilon)$-holomorphic maps u, with $\boldsymbol{b} \in \ell^p$ for some $p \in (0, 1)$, are known to allow gpc expansions with coefficient sequences in ℓ^p. Such maps arise for example as response surfaces of parametric PDEs, with applications in PDE uncertainty quantification (UQ) for many mathematical models in engineering and the sciences. Up to logarithmic terms, we establish the dimension independent approximation rate $s = 1/p - 1$ for these functions in terms of the total number N of units and weights in the DNN. It follows that certain DNN architectures can overcome the curse of dimensionality when expressing possibly countably-parametric, real-valued maps with a certain degree of sparsity in the sequences of their gpc expansion coefficients. We also obtain rates of expressive power of DNNs for countably-parametric maps $u : U \to V$, where V is the Banach space $H_0^1([0, 1])$.

Key words: Generalized polynomial chaos, Deep networks, Sparse grids, Uncertainty quantification
Subject Classification: 68Q32, 41A25, 41A46

1 Introduction

After foundational developments several decades ago in answering the question of universality of neural networks (NNs for short) [17, 22, 21, 3, 4], in recent years so-called *deep neural networks* have undergone rapid development and successful deployment in a wide range of applications. Evidence for 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 (see, e.g. [24, 25, 35] and the references there for applications in Finite Element approximation of parametrized problems). 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 DNNs can express certain solution families of parametric PDEs which depend on a large (possibly infinite) number of variables. Specifically, we show that DNNs afford an expression error of size $\delta > 0$ with NN size of

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

$O(\delta^{-1/s})$ units and weights, where $O(\cdot)$ and the rate $s > 0$ are independent of the number of parameters activated in the approximating NN (thereby overcoming, in particular, the curse of dimensionality). The rate s is comparable to convergence rates of the best n-term generalized polynomial chaos approximation of the parametric solution, and depends only on a suitable notion of sparsity in coefficient sequences of gpc expansions of the parametric solution. Our analysis mainly focuses on ReLU NNs, as they are widely used currently. ReLU DNNs also allow the *exact* expression of continuous, piecewise linear spline functions in intervals as they appear in Courant Finite Elements as we show in Lemma 4.5 ahead. In passing, we additionally discuss modifications in the proofs which result from employing smooth activation functions instead of ReLU activations in the DNN expression of gpc expansions. We show that this leads to slightly improved error vs. DNN size bounds on the corresponding expression rates (see Thms. 3.9 and 3.10).

1.1 Recent mathematical results on expressive power of DNNs

The past year has seen significant efforts towards theoretical understanding of the benefits on expressive power of NNs afforded by possibly large NN depth. Whereas density results and approximation rate bounds for shallow NNs have been known for some time (see [29]), recent theoretical results focused on approximation rate bounds of deep NNs for certain function classes. We mention in particular [19] and [7]. There, it is shown that deep NNs with a particular architecture allow the same approximation rate bounds as rather general multiresolution systems when measured in terms of the number N of units in the DNN.

In [15], *convolutional DNNs* were proved to be able to express multivariate functions given in so-called *Hierarchic Tensor (HT) formats*, a numerical representation inspired by electron structure calculations in computational quantum chemistry.

Also, in [36, 26], it has been shown that DNNs can express general uni- and multivariate polynomials on bounded domains with accuracy $\delta > 0$, uniform with respect to the parameters, with complexity (which we assume to comprise the number of NN layers and the number of NN units and weights) which scales polylogarithmically with respect to δ. The results in [36, 26] allow transferring approximation results from high order finite and spectral element methods, 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 [32], DNN expression rates for multivariate polynomials are obtained which are explicit in the number of variables and the polynomial degree. The proofs in [32] depend strongly on possibly a large number of derivatives of the activation function. In [27] 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. Such maps arise in a number of applications. We mention only elliptic PDEs with uncertain, spatially heterogeneous, uncertain coefficients (see, e.g., [13, 2] and the references there), PDEs posed in domains of uncertain geometry (see, e.g., [31, 23, 14, 24]), and time-harmonic, electromagnetic scattering (see, e.g., [23]). Such models are ubiquitous in the area of computational uncertainty quantification in engineering and in the sciences. Holomorphic parametric dependence of uncertain input data implies, for regular parametric operator equations, holomorphic dependence of solutions on the parameters. As a consequence, 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). These 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 (Banach) space accommodating the PDE solution into the real numbers.

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 [34] and the references there), it has been shown in [16, 33] 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. This idea has been employed in the recent paper [35], which reports numerical experiments for a Bayesian inverse problem (for a second order elliptic PDE with uncertain diffusion coefficient) using DNNs.

1.3 Notation

We adopt standard notation, consistent with our previous works [37, 38]: $\mathbb{N} = \{1, 2, \dots\}$ 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. It may change its value even within the same equation.

Multiindices are denoted by $\boldsymbol{\nu} = (\nu_j)_{j\in\mathbb{N}} \in \mathbb{N}_0^{\mathbb{N}}$. The *order* of a multiindex $\boldsymbol{\nu}$ is denoted by $|\boldsymbol{\nu}|_1 := \sum_{j\in\mathbb{N}} \nu_j$. For the countable set of "finitely supported" multiindices we write

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

The notation $\operatorname{supp}\boldsymbol{\nu}$ stands for the *support* of the multiindex, i.e. $\operatorname{supp}\boldsymbol{\nu} = \{j \in \mathbb{N} : \nu_j \neq 0\}$. 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 called *downward closed*, if $\boldsymbol{\nu} = (\nu_j)_{j\in\mathbb{N}} \in \Lambda$ implies $\boldsymbol{\mu} = (\mu)_{j\in\mathbb{N}} \in \Lambda$ for all $\boldsymbol{\mu} \leq \boldsymbol{\nu}$. Here, the ordering "$\leq$" on $\mathcal{F}$ is defined as $\mu_j \leq \nu_j$, for all $j \in \mathbb{N}$. We write $|\Lambda|$ to denote the finite cardinality of a set Λ. 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.

We consider the set $\mathbb{C}^{\mathbb{N}}$ endowed with the product topology. Any subset such as $[-1,1]^{\mathbb{N}}$ is then understood to be equipped with the subspace topology. For $\varepsilon \in (0,\infty)$ we write $B_\varepsilon := \{z \in \mathbb{C} : |z| < \varepsilon\}$. Furthermore $B_\varepsilon^{\mathbb{N}} := \bigtimes_{j\in\mathbb{N}} B_\varepsilon \subset \mathbb{C}^{\mathbb{N}}$. Elements of $\mathbb{C}^{\mathbb{N}}$ will be denoted by boldface characters such as $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in [-1,1]^{\mathbb{N}}$. For $\boldsymbol{\nu} \in \mathcal{F}$, standard notations $\boldsymbol{y}^{\boldsymbol{\nu}} := \prod_{j\in\mathbb{N}} y_j^{\nu_j}$ and $\boldsymbol{\nu}! = \prod_{j\in\mathbb{N}} \nu_j!$ 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 Sec. 2, we introduce so-called $(\boldsymbol{b},\varepsilon)$-holomorphic functions and review approximation rate bounds for their truncated gpc expansion. In Sec. 3, we present the DNN approximation results. Sec. 3.1 introduces the architectures which are admitted in our approximation results. Sec. 3.2 proves a basic result on the expressive power of ReLU DNNs for the multiplication of n numbers. Sec. 3.3 indicates extension to smoother activation functions. In Sec. 3.4 we establish the main results of this work, namely the approximation of a real-valued $(\boldsymbol{b},\varepsilon)$-holomorphic parametric response map u to pointwise accuracy $\delta > 0$, by a DNN with (essentially, i.e. up to polylogarithmic factors) $O(\log\delta)$ many hidden layers. The total number of units in the NN is estimated using the sparsity of gpc expansions of u, which result from the property of $(\boldsymbol{b},\varepsilon)$-holomorphy of the parametric map u. Sec. 4 discusses some consequences of our DNN expression results. We there also give an example of a parametric one dimensional diffusion problem, which demands the approximation of a Banach space valued response map via a NN. In Sec. 5 we give conclusions and possible further directions.

2 Generalized polynomial chaos approximation

To analyze the expressive power of deep NNs on countably-parametric, real-valued (or Banach space-valued) maps, we shall draw upon results from [9, 10, 38] on sparse generalized polynomial chaos approximation of such maps. To state these results in the required generality, with the parameter domain $U := [-1,1]^{\mathbb{N}}$ and a real Banach space V, we consider maps $u : U \to V$. We are interested in *separately holomorphic maps*, i.e. maps which admit a holomorphic extension to the complex domain with respect to each parameter $y_j \in [-1,1]$, where $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in U$, with quantitative control on the size of the domains holomorphy, as formalized in Def. 2.1 below. Under certain assumptions, such maps allow a representation as a sparse *Taylor generalized polynomial chaos expansion*. By this, we mean (formal, at this stage) expressions of the form

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

with Taylor coefficients $t_{\boldsymbol{\nu}} \in V$ for all $\boldsymbol{\nu} \in \mathcal{F}$. The summability properties of the (norms of) Taylor coefficients $(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}}$ 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 $(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$ implies the unconditional convergence in V of (2.1) for every $\boldsymbol{y} \in U$. This summability is, in turn, ensured by a suitable form of holomorphic continuation of the parametric map $u : U \to V$, which takes values in the complexification $V_{\mathbb{C}}$ of V. By "complexification" we mean here the space $V_{\mathbb{C}} = V + \mathrm{i}V$ with the so-called *Taylor norm* $\|v + \mathrm{i}w\|_{V_{\mathbb{C}}} := \sup_{t\in[0,2\pi)} \|\cos(t)v - \sin(t)w\|_V$ for all $v,\ w \in V$, where i denotes the square root of -1 with $\arg(\mathrm{i}) = \pi/2$ (cp. [28]). We recapitulate principal definitions and results from [13, 10, 9, 38] and the references there.

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

The notion of $(\boldsymbol{b},\varepsilon)$-holomorphy, which is defined below, has been found to be a sufficient condition on the parametric map $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in V$, in order that u possesses an expansion of the type (2.1) with coefficients satisfying $(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$ for some $p \in (0,1)$. The fact that the sequence of norms of the Taylor coefficients belongs to $\ell^p(\mathcal{F})$ is the crucial property required to establish convergence rates of certain partial sums of the Taylor gpc expansion. We briefly review this result and its implications in the present and the subsequent subsection.

Definition 2.1 ($(\boldsymbol{b},\varepsilon)$-Holomorphy)**.** *Let V be a Banach space. Assume given a monotonically decreasing 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 $p \in (0,1]$. A poly-radius $\boldsymbol{\rho} \in [1,\infty)^{\mathbb{N}}$ is called $(\boldsymbol{b},\varepsilon)$-admissible for some $\varepsilon > 0$ if*

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

A continuous function $u : U \to V$ is called $(\boldsymbol{b},\varepsilon)$-holomorphic *if there exists a constant $C_u < \infty$ such that the following holds:*

For every $(\boldsymbol{b},\varepsilon)$-admissible $\boldsymbol{\rho}$, there exists an extension $\tilde{u} : B_{\boldsymbol{\rho}} \to V_{\mathbb{C}}$ of u, i.e. $\tilde{u}(\boldsymbol{y}) = u(\boldsymbol{y})$ for all $\boldsymbol{y} \in U$, such that $\boldsymbol{z} \mapsto \tilde{u}(\boldsymbol{z}) : B_{\boldsymbol{\rho}} \to V_{\mathbb{C}}$ is holomorphic as a function of each $z_j \in B_{\rho_j}$, $j \in \mathbb{N}$, and such that $\sup_{\boldsymbol{z}\in B_{\boldsymbol{\rho}}} \|u(\boldsymbol{z})\|_{V_{\mathbb{C}}} \le C_u$.

We remind that continuity in Def. 2.1 means continuity with respect to the subspace topology on $U \subset \mathbb{C}^{\mathbb{N}}$, where $\mathbb{C}^{\mathbb{N}}$ is equipped with the product topology.

Definition 2.1 has been similarly stated in [10]. The sequence $\boldsymbol{b}$ in Definition 2.1 controls the size of the domains of analytic continuation of the parametric map with respect to the parameters $y_j \in \boldsymbol{y}$: the stronger the decrease of $\boldsymbol{b}$, the faster the radii ρ_j of $(\boldsymbol{b},\varepsilon)$-admissible sequences $\boldsymbol{\rho}$ may increase. Whereas the sequence $\boldsymbol{b}$ (or, more precisely, the summability exponent p such that $\boldsymbol{b} \in \ell^p(\mathbb{N})$) will determine the algebraic rate at which the Taylor coefficients tend to 0 (see Thm. 2.7 ahead), the parameter $\varepsilon > 0$ merely influences certain constants and can be considered to be of minor importance. To illustrate the notion of $(\boldsymbol{b},\varepsilon)$-holomorphy, we present as a standard example the solution to an affine-parametric, linear elliptic PDE; we hasten to add that $(\boldsymbol{b},\varepsilon)$-holomorphy applies considerably larger classes of equations [10, 37].

Example 2.2. *In a bounded Lipschitz domain $D \subset \mathbb{R}^d$, $d \in \mathbb{N}$, we consider the elliptic diffusion equation*

$$-\mathrm{div}(a\nabla u) = f, \tag{2.3}$$

with homogeneous Dirichlet boundary conditions $u|_{\partial D} = 0$. In (2.3), the scalar diffusion coefficient $a \in L^\infty(D)$ is assumed to satisfy $0 < r < a < R < \infty$ almost everywhere in D. Denoting $V = H_0^1(D)$ and $V' = H^{-1}(D)$, for any $f \in V'$, the weak formulation of (2.3): find

$$u \in V \quad \text{such that} \quad \int_D a\nabla u \cdot \nabla v = f(v), \quad \forall\, v \in V, \tag{2.4}$$

admits a unique solution $u \in V$.

We consider affine-parametric diffusion coefficients $a = a(\boldsymbol{y})$, where $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}}$ is a sequence of real-valued parameters ranging in $U = [-1,1]^{\mathbb{N}}$. For $\overline{a} \in L^\infty(D)$ and a sequence of fluctuations $(\psi_j)_{j\in\mathbb{N}} \subset L^\infty(D,\mathbb{R})$ define $b_j := \|\psi_j\|_{L^\infty(D)}$ for all $j \in \mathbb{N}$. Assuming that $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$, the affine parametric diffusion coefficient

$$\forall \boldsymbol{y} \in U: \qquad a(\boldsymbol{y},\cdot) = \bar{a}(\cdot) + \sum_{j\in\mathbb{N}} y_j\psi_j(\cdot) \in L^\infty(D), \tag{2.5}$$

is well-defined. Expansions (2.5) arise for example, from Fourier-, Karhunen-Loève-, spline- or wavelet series representations of a.

Assume that $\overline{a} \equiv 1$ and $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$ and additionally that $\|\boldsymbol{b}\|_{\ell^1(\mathbb{N})} < 1$. Then, we claim that the solution $u(\boldsymbol{y})$ of (2.4) with the $\boldsymbol{y}$-dependent diffusion coefficient $a(\boldsymbol{y},\cdot)$ in (2.5) is then well-defined for all $\boldsymbol{y} \in U$ and $(\boldsymbol{b},\varepsilon)$-holomorphic as long as $\varepsilon > 0$ is small enough: for $\boldsymbol{\rho} \in [1,\infty)^{\mathbb{N}}$ being $(\boldsymbol{b},\varepsilon)$-admissible, and for every $\boldsymbol{z} \in B_{\boldsymbol{\rho}}$,

$$\begin{aligned}\operatorname*{ess\,inf}_{x\in D} \Re(a(\boldsymbol{z},x)) = \operatorname*{ess\,inf}_{x\in D}\left(1 + \Re\sum_{j\in\mathbb{N}} z_j\psi_j(x)\right) &\geq 1 - \sum_{j\in\mathbb{N}}\rho_j b_j \\ &= 1 - \sum_{j\in\mathbb{N}} b_j - \sum_{j\in\mathbb{N}}(\rho_j - 1)b_j \geq 1 - \|\boldsymbol{b}\|_{\ell^1(\mathbb{N})} - \varepsilon > 0\end{aligned}$$

provided that $0 < \varepsilon < 1 - \|\boldsymbol{b}\|_{\ell^1(\mathbb{N})}$. The fact that the real part of the diffusion coefficient $a(\boldsymbol{z},x)$ is strictly positive implies that for all $\boldsymbol{z} \in B_{\boldsymbol{\rho}}$ (in particular for $\boldsymbol{y} \in U \subset B_{\boldsymbol{\rho}}$) there is a unique solution $u(\boldsymbol{z}) \in V_{\mathbb{C}} \simeq H_0^1(D,\mathbb{C})$ of the variational problem (2.4) (this follows by a complex version of the Lax-Milgram Lemma, see for example [8, Lemma 2.6]). Since $\Re(a(\boldsymbol{z},x)) \geq 1 - \|\boldsymbol{b}\|_{\ell^1(\mathbb{N})} - \varepsilon > 0$ independent of of $\boldsymbol{\rho}$, $\boldsymbol{z}$ and x, one can show that there is an upper bound C_u on $\sup_{\boldsymbol{z}\in B_{\boldsymbol{\rho}}} \|u(\boldsymbol{z})\|_{V_{\mathbb{C}}}$ independent of the $(\boldsymbol{b},\varepsilon)$-admissible $\boldsymbol{\rho}$. Continuity of $\boldsymbol{y} \mapsto u(\boldsymbol{y}) : U \to V$ follows from continuity of $\boldsymbol{y} \mapsto a(\boldsymbol{y},\cdot) : U \to L^\infty([0,1])$. This, in turn, follows from $(\|\psi_j(\cdot)\|_{L^\infty(D)})_{j\in\mathbb{N}} = \boldsymbol{b} \in \ell^1(\mathbb{N})$. Finally, holomorphy of $u(\boldsymbol{z}) \in V_{\mathbb{C}}$ as a function of each $z_j \in B_{\rho_j}$ can either be deduced directly considering the difference quotient, or by a holomorphic version of the implicit function theorem. For more details, we refer to [10, 11] and to the references there.

We next recall a proposition, which provides estimates on the norm of the Taylor coefficients

$$t_{\boldsymbol{\nu}} := \frac{1}{\boldsymbol{\nu}!}(\partial_{\boldsymbol{y}}^{\boldsymbol{\nu}} u)(\boldsymbol{y})\Big|_{\boldsymbol{y}=\mathbf{0}} \in V, \qquad \boldsymbol{\nu} \in \mathcal{F}, \tag{2.6}$$

of a $(\boldsymbol{b}, \varepsilon)$-holomorphic function $u : U \to V$, and thus lays the groundwork for showing $(\|t_{\boldsymbol{\nu}}\|_X)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$.

Proposition 2.3. *Suppose that $u : U \to V$ is $(\boldsymbol{b}, \varepsilon)$-holomorphic.*

Then $t_{\boldsymbol{\nu}} \in V$ in (2.6) is well-defined for every $\boldsymbol{\nu} \in \mathcal{F}$ and u admits the Taylor gpc expansion

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}}, \tag{2.7}$$

which is unconditionally convergent for all $\boldsymbol{y} \in U$. With $C_u > 0$ as in Def. 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

$$\|t_{\boldsymbol{\nu}}\|_V \le C_u \boldsymbol{\rho}^{-\boldsymbol{\nu}} \ . \tag{2.8}$$

The bound (2.8) is a consequence of the separate holomorphy of u on the poly-disc $B_{\boldsymbol{\rho}}$ (as stated in Def. 2.1) and of the Cauchy integral theorem [20, Thm. 2.1.2], see the proof of [12, Lemma 2.4]. The unconditional convergence of the series (2.7) on U is for example discussed in [12], also see [20, Sec. 2.1] for convergence in the finite dimensional case.

For future reference we state three lemmata required for proving summability of sequences allowing bounds of the type (2.8). Let in the following $\boldsymbol{\alpha} = (\alpha_j)_{j\in\mathbb{N}}$ denote a sequence (not necessarily monotonic) of nonnegative real numbers.

Lemma 2.4 ([12, 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 2.5 ([12, 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 2.6 ([38, 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 $x_j \le \|\boldsymbol{x}\|_{\ell^p(\mathbb{N})} j^{-1/p}$ for all $j \in \mathbb{N}$.*

2.2 Summability of Taylor coefficients

As has been observed in several references (see, e.g., [37, 10]), $(\boldsymbol{b}, \varepsilon)$-holomorphic functions taking values in a Banach space V with $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $0 < p < 1$ admit sequences $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ of Taylor coefficients whose $\|\circ\|_V$-norms belong to $\ell^p(\mathcal{F})$. This "p-summability" implies dimension-independent n-term gpc approximation rate bounds. Our analysis of the expressive power of DNNs on such parametric solutions families will be based on a version of these results as stated in the next theorem. In the following, we denote 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.

Theorem 2.7. *Let u be $(\boldsymbol{b}, \varepsilon)$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0, 1)$ and, for $\boldsymbol{\nu} \in \mathcal{F}$ let $t_{\boldsymbol{\nu}} \in V$ denote the Taylor coefficient defined in (2.6). Then $(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$. Furthermore, 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 $n \in \mathbb{N}$ it holds $|\Lambda_n| = n$ and*

(i) $\sum_{\boldsymbol{\nu}\notin\Lambda_n} \|t_{\boldsymbol{\nu}}\|_V \le C n^{-1/p+1}$,

(ii) $\sup_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \le C(1 + \log(n))$.

Moreover, $\mathbf{e}_j \in \Lambda_n$ implies $\mathbf{e}_i \in \Lambda_n$ for all $i \le j$.

Since it will allow us to discuss results based on related but different hypotheses than the ones of Def. 2.1 (see. Sec. 4 ahead), we provide part of the above theorem as a separate Lemma, before proceeding to the proof of Thm. 2.7.

Lemma 2.8. *Let $r \in [1,\infty)$ and $p \in (0,1)$. Let $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in (0,\infty)^{\mathcal{F}}$ and assume that $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}} \in \ell^{pr/(r-p)}(\mathbb{N})$ is monotonically decreasing with $\beta_j \in (0,1)$ for all $j \in \mathbb{N}$, and such that $\sum_{\boldsymbol{\nu}\in\mathcal{F}}(\boldsymbol{\beta}^{-\boldsymbol{\nu}} t_{\boldsymbol{\nu}})^r < \infty$.*

Then $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$. Moreover, there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$ there exists a downward closed index set $\Lambda_n \subset \mathcal{F}$ with $|\Lambda_n| = n$ and such that (i) $\sum_{\boldsymbol{\nu}\notin\Lambda_n} t_{\boldsymbol{\nu}} \leq Cn^{-1/p+1}$ and (ii) $\sup_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \leq C(1+\log(n))$. Moreover, $\mathbf{e}_j \in \Lambda_n$ implies $\mathbf{e}_i \in \Lambda_n$ for all $i \leq j$.

Proof. First we show $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$. This follows with Hölder's inequality, since it holds

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}} {t_{\boldsymbol{\nu}}}^p = \sum_{\boldsymbol{\nu}\in\mathcal{F}} {t_{\boldsymbol{\nu}}}^p \boldsymbol{\beta}^{-p\boldsymbol{\nu}} \boldsymbol{\beta}^{p\boldsymbol{\nu}} \leq \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} (t_{\boldsymbol{\nu}}\boldsymbol{\beta}^{-\boldsymbol{\nu}})^r\right)^{\frac{p}{r}} \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\beta}^{\boldsymbol{\nu}\frac{pr}{r-p}}\right)^{\frac{r-p}{r}} < \infty, \tag{2.9}$$

where the last sum is finite since $(\boldsymbol{\beta}^{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{pr/(r-p)}(\mathcal{F})$ according to Lemma 2.4 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 $\boldsymbol{\beta} \in \ell^{pr/(r-p)}(\mathbb{N})$).

Next fix $0 < q < \infty$ such that $pr/(r-p) > q$ and set for $\boldsymbol{\nu} \in \mathcal{F}$

$$\alpha_{\boldsymbol{\nu}} := \begin{cases} j^{-1/q} & \text{if } \boldsymbol{\nu} = \mathbf{e}_j, \\ 0 & \text{otherwise.} \end{cases} \tag{2.10}$$

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

$$C_0 := \sum_{\boldsymbol{\nu}\in\mathcal{F}} (\zeta_{\boldsymbol{\nu}}^{-1} t_{\boldsymbol{\nu}})^r \leq \sum_{\boldsymbol{\nu}\in\mathcal{F}} (\boldsymbol{\beta}^{-\boldsymbol{\nu}} t_{\boldsymbol{\nu}})^r < \infty,$$

by assumption.

Let $\pi : \mathbb{N} \to \mathcal{F}$ be a bijection such that the sequence $(\zeta_{\pi(j)})_{j\in\mathbb{N}}$ is monotonically decreasing in j, and such that $\{\pi(1),\dots,\pi(n)\} \subset \mathcal{F}$ is downward closed for any $n \in \mathbb{N}$. This is possible, because $\zeta_{\boldsymbol{\nu}}$ is monotonically decreasing in the sense that $\boldsymbol{\nu} \leq \boldsymbol{\mu}$ implies $\zeta_{\boldsymbol{\nu}} \geq \zeta_{\boldsymbol{\mu}}$. Define $\Lambda_n := \{\pi(j) : 1 \leq j \leq n\}$ and $\Lambda_n^c := \mathcal{F}\backslash\Lambda_n$. Since $(\beta_j)_{j\in\mathbb{N}}$ is monotonically decreasing, both $(\alpha_{\mathbf{e}_j})_{j\in\mathbb{N}}$ and $(\boldsymbol{\beta}^{\mathbf{e}_j})_{j\in\mathbb{N}}$ are monotonically decreasing in j. Thus the same is true for $(\zeta_{\mathbf{e}_j})_{j\in\mathbb{N}}$. Consequently, if $\mathbf{e}_j \in \Lambda_n$ and $i \leq j$, then $\zeta_{\mathbf{e}_i} \geq \zeta_{\mathbf{e}_j}$ and we can choose π such that $\mathbf{e}_j \in \Lambda_n$ necessarily implies $\mathbf{e}_i \in \Lambda_n$ for all $i \leq j$.

With $r' \in (1,\infty]$ denoting the Hölder conjugate of r we get

$$\sum_{\boldsymbol{\nu}\in\Lambda_n^c} t_{\boldsymbol{\nu}} = \sum_{\boldsymbol{\nu}\in\Lambda_n^c} \zeta_{\boldsymbol{\nu}}\zeta_{\boldsymbol{\nu}}^{-1} t_{\boldsymbol{\nu}} \leq \|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}(\Lambda_n^c)} \|(\zeta_{\boldsymbol{\nu}}^{-1} t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^r(\Lambda_n^c)} \leq C_0^{1/r} \|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}(\Lambda_n^c)}. \tag{2.11}$$

With Lemma 2.6 we conclude that there exists a constant C such that $\zeta_{\pi(j)} \leq Cj^{-(r-p)/(pr)}$ for all $j \in \mathbb{N}$. Hence, the last quantity in (2.11) can be bounded for $r > 1$, i.e. $r' = r/(r-1) < \infty$, by

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

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 $\|(\zeta_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n^c}\|_{\ell^{r'}(\Lambda_n^c)} \le C\sup_{j>n} j^{1-1/p} \le Cn^{-1/p+1}$ instead (where we have again employed Lemma 2.6). This shows (i).

To show (ii) note that by definition of $(\alpha_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ (cp. (2.10)) 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 $\sup\{\zeta_{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \mathcal{F},\, |\boldsymbol{\nu}|_1 = d\} \ge \sup\{\boldsymbol{\beta}^{\boldsymbol{\nu}} : \boldsymbol{\nu} \in \mathcal{F},\, |\boldsymbol{\nu}|_1 = d\} \ge c^d$. Now, if $n^{-1/q} > c^{d_0}$ for some $d_0 \in \mathbb{N}$, then it must hold $\max_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 < d_0$. Hence with $f(d) := c^d$ and $f^{-1}(x) = \log(x)/\log(c)$,

$$\max_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \le f^{-1}(n^{-1/q}) = O(\log(n)) \qquad \text{as } n \to \infty,$$

which concludes the proof. □

Proof of Thm. 2.7. The Taylor coefficients of $(\boldsymbol{b},\varepsilon)$-holomorphic maps admit bounds of the following type (see [12, 13] and in particular the proof of [10, Thm. 2.2])

$$\|t_{\boldsymbol{\nu}}\|_V \le C\kappa^{|\nu_E|} \frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!} \boldsymbol{\gamma}^{\boldsymbol{\nu}_F}, \quad \boldsymbol{\nu} \in \mathcal{F}\,. \tag{2.12}$$

Here, $J \in \mathbb{N}$, $\kappa \in (0,1)$ as well as $\boldsymbol{\gamma} \in \ell^p(\mathbb{N})$ (monotonically decreasing) with $\|\boldsymbol{\gamma}\|_{\ell^p(\mathbb{N})} < 1$ are fixed, and $\boldsymbol{\nu}_E := (\nu_1,\dots,\nu_J) \in \mathbb{N}_0^J$ as well as $\boldsymbol{\nu}_F := (\nu_{J+1},\nu_{J+2},\dots) \in \mathbb{N}_0^{\mathbb{N}}$. This is a consequence of Prop. 2.3, and we refer again to [12, 13, 10] for proofs of such statements. There, it is also shown that (2.12) implies $(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$. Moreover, since $\sup_{j\in\mathbb{N}} \gamma_j \le \|\boldsymbol{\gamma}\|_{\ell^1(\mathbb{N})} < 1$, by increasing κ if necessary, there is no loss of generality in assuming that $\kappa \in (0,1)$ and

$$\frac{2\kappa}{1+\kappa} > \sup_{j\in\mathbb{N}} \gamma_j^{1-p}. \tag{2.13}$$

We now choose a particular sequence $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}} \in \ell^{p/(1-p)}(\mathbb{N})$ with $\|\boldsymbol{\beta}\|_{\ell^\infty(\mathbb{N})} < 1$. It is defined as follows

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

Evidently, $(\beta_j)_{j\in\mathbb{N}}$ is monotonically decreasing, because of (2.13) and because $\boldsymbol{\gamma}$ is monotonically decreasing. Then for $\boldsymbol{\delta} = (\delta_j)_{j\in\mathbb{N}}$ with $\delta_j := \gamma_j \beta_{J+j}^{-1} = \gamma_j^{1-(1-p)}$ we have

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

Now, with (2.12),

$$\begin{aligned}
\sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\beta}^{-\boldsymbol{\nu}} \|t_{\boldsymbol{\nu}}\|_V &\le C \sum_{\boldsymbol{\nu}\in\mathcal{F}} \kappa^{|\boldsymbol{\nu}_E|} \frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!} \boldsymbol{\gamma}^{\boldsymbol{\nu}_F} \left(\frac{1+\kappa}{2\kappa}\right)^{|\boldsymbol{\nu}_E|} \prod_{j>J} \beta_j^{-\nu_j} \\
&= C \sum_{\boldsymbol{\nu}\in\mathcal{F}} \kappa^{|\nu_E|} \left(\frac{1+\kappa}{2\kappa}\right)^{|\boldsymbol{\nu}_E|} \frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!} \prod_{j\in\mathbb{N}} \gamma_j^{\nu_{J+j}} \beta_{J+j}^{-\nu_{J+j}} \\
&= C \sum_{\boldsymbol{\nu}\in\mathbb{N}_0^J} \left(\frac{1+\kappa}{2}\right)^{|\boldsymbol{\nu}|} \sum_{\boldsymbol{\mu}\in\mathcal{F}} \frac{|\boldsymbol{\mu}|!}{\boldsymbol{\mu}!} \boldsymbol{\delta}^{\boldsymbol{\mu}}.
\end{aligned}$$

By Lemma 2.4 and Lemma 2.5 the last two sums are finite, since $(1+\kappa)/2 < 1$ and because of (2.14). This proves $\sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\beta}^{-\boldsymbol{\nu}} \|t_{\boldsymbol{\nu}}\|_V < \infty$. Furthermore, Lemma 2.4 gives $(\boldsymbol{\beta}^{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p/(1-p)}(\mathcal{F})$. We may thus employ Lemma 2.8 with $r = 1$, and use that $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}}$ is a monotonically decreasing sequence by construction 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 for $\ell \in \{1, \dots, L\}$.

The vector $\boldsymbol{x} = (x_1, ..., x_n) \in \mathbb{R}^n$ will denote the input of the DNN, and we set $N_0 := n$. Next, z_j^ℓ denotes the output of unit j in layer ℓ, b_j^ℓ denotes the *bias* of unit j in layer ℓ, and $w_{i,j}^\ell$ is the *weight* connecting the ith unit in layer $\ell - 1$ with the jth unit in layer ℓ. For an *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$, outputs between layers of the FFNN are then characterized by the following maps: the first hidden layer

$$z_j^1 := \sigma\left(\sum_{i=1}^{n} w_{i,j}^1 x_i + b_j^1\right), \quad j \in \{1, \dots, N_1\}, \tag{3.1}$$

the $L - 1$ remaining hidden layers

$$z_j^{\ell+1} := \sigma\left(\sum_{i=1}^{N_\ell} w_{i,j}^{\ell+1} z_i^\ell + b_j^{\ell+1}\right), \quad \ell \in \{1, \dots, L-1\}, \ j \in \{1, \dots, N_{\ell+1}\}, \tag{3.2}$$

and the *output layer*

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

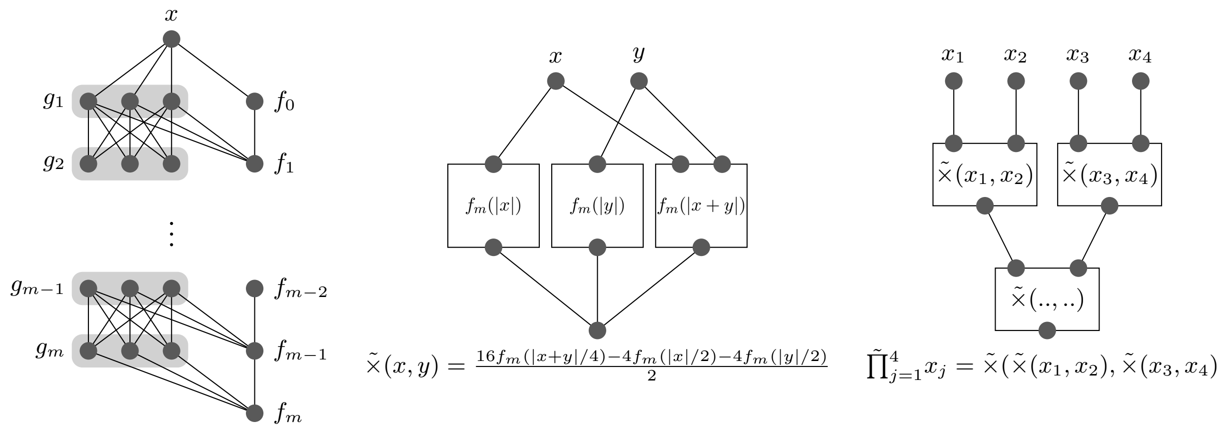
As an activation function $\sigma(\cdot)$ in (3.1) - (3.2) we either consider the so-called *rectified linear unit* (ReLU) given by $\sigma(x) = \max\{0, x\}$ for $x \in \mathbb{R}$ or general, smooth, nonlinear activation functions. We refer to Prop. 3.7 ahead for the precise statement of the term "nonlinear" activation function. As is customary in the theory of NNs, the number of hidden layers L of a NN is referred to as *depth* and the total number of nodes and nonzero weights is referred to as *size* of the NN. Similarly, by the *number of weights of a network*, we always mean the number of nonzero weights. With a DNN f as in (3.1)-(3.3), we define

$$\mathrm{size}(f) := |\{(i,j,\ell) : w_{i,j}^\ell \neq 0\}| + \sum_{\ell=0}^{L} N_\ell \qquad \text{and} \qquad \mathrm{depth}(f) := L.$$

The weights $w_{i,j}^\ell$ for $\ell \in \{1, \dots, L+1\}$, $i \in \{1, \dots, N_{\ell-1}\}$ and $j \in \{1, \dots, N_\ell\}$, 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 ReLU 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 (2.7). Our point of departure will be the n-term truncation of the Taylor polynomial chaos expansion of the parametric map $u(\boldsymbol{y}) : U \to \mathbb{R}$, which is obtained via the gpc approximation result Theorem 2.7. In particular, we shall use recent, quantitative bounds on expressing multivariate polynomials by DNNs, from [36]. 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 [36] in a form required in our subsequent analysis.



(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 $|x_j| \leq 1$ for all j. The boxes contain the approximate multiplication $\tilde{\times}$ from (b).

As mentioned above, the main task is to approximate $x \mapsto x^2$ for $x \in [0,1]$. This is achieved in [36] 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 $ab = 2M^2((|a+b|/(2M))^2 - (|a|/(2M))^2 - (|b|/(2M))^2)$ 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 [36, Prop. 3], allows to approximate the multiplication of two numbers in $[-M, M]$ with accuracy $\delta > 0$ by a network $\tilde{\times} : M \times M \to \mathbb{R}$ of size and depth $O(\log(1/\delta))$. We next generalize this result, by additionally taking into account the approximation of the gradient of the multiplication $(a,b) \mapsto ab$.

Proposition 3.1. *Let $M > 0$ and $\delta \in (0,1)$. There exists a ReLU NN $\tilde{\times}$ with two input units such that*

$$\sup_{|a|,|b|\le M} |ab - \tilde{\times}(a,b)| \le \delta \qquad \text{and} \qquad \operatorname*{ess\,sup}_{|a|,|b|\le M} \max\left\{ \left| b - \frac{d}{da}\tilde{\times}(a,b)\right|, \left| a - \frac{d}{db}\tilde{\times}(a,b)\right| \right\} \le \delta, \tag{3.6}$$

where $\frac{d}{da}\tilde{\times}(a,b)$ and $\frac{d}{db}\tilde{\times}(a,b)$ denote weak derivatives. It holds $\mathrm{size}(\tilde{\times}) = O(\log(1/\delta))$ *and* $\mathrm{depth}(\tilde{\times}) = O(\log(1/\delta))$ *as $\delta \downarrow 0$. Moreover, for every $a \in M$, there exists a finite set $\mathcal{N}_a \subseteq M$ such that $b \mapsto \tilde{\times}(a,b)$ is strongly differentiable at all $b \in (-M,M)\backslash\mathcal{N}_a$.*

Proof. Set

$$\tilde{\times}(a,b) = 2M^2\left(f_m\Big(\frac{|a+b|}{2M}\Big) - f_m\Big(\frac{|a|}{2M}\Big) - f_m\Big(\frac{|b|}{2M}\Big)\right), \tag{3.7}$$

where f_m is the piecewise linear interpolant of x^2 at the nodes $x_j = 2^{-m}j$, $j = 0,\dots,2^m$. Then

$$\begin{aligned}\operatorname*{ess\,sup}_{x\in[0,1]} |2x - f_m'(x)| &= \sup_{j=0,\dots,2^m-1}\ \sup_{x\in[x_j,x_{j+1}]} |2x - f_m'(x)| = \sup_{j=0,\dots,2^m-1}\ \sup_{x\in[x_j,x_{j+1}]} \left|2x - \frac{x_{j+1}^2 - x_j^2}{x_{j+1}-x_j}\right| \\ &= \sup_{j=0,\dots,2^m-1}\ \sup_{x\in[x_j,x_{j+1}]} |2x - (x_{j+1}+x_j)| = \sup_{j=0,\dots,2^m-1} |x_{j+1}-x_j| = 2^{-m}. \end{aligned}\tag{3.8}$$

For every a, $b \in [0,M]\backslash\{2Mx_j : j = 0,\dots,2^m\}$ such that $a+b \notin \{2Mx_j : j = 0,\dots,2^m\}$,

$$\left| b - \frac{d}{da}\tilde{\times}(a,b)\right| = M\left| \frac{2(a+b)}{2M} - \frac{2a}{2M} - f_m'\Big(\frac{a+b}{2M}\Big) + f_m'\Big(\frac{a}{2M}\Big)\right| \le 2M2^{-m}.$$

Choosing $m = \lceil -\log_2(\delta/(2M))\rceil = O(\log(1/\delta))$ gives $\operatorname{ess\,sup}_{0\le a,b\le M} |b - \frac{d}{da}\tilde{\times}(a,b)| \le \delta$. Due to the symmetry in a and b, and by distinguishing between all cases where a, b and $a+b$ are either negative or positive, we obtain the second inequality in (3.6). The proof of the first inequality is similar by using $\sup_{x\in[0,1]} |x^2 - f_m(x)| \le 2^{-2m}$ instead of (3.8). Finally, as depicted in Fig. 1, the network f_m has m hidden layers and a total size of $O(m) = O(\log(1/\delta))$, which gives the statement about the size and depth of $\tilde{\times}$ in (3.7).

The existence of $\mathcal{N}_a$ follows by the fact that f_m is a piecewise linear interpolant of x^2 in $2^m + 1$ nodes and by the definition of $\tilde{\times}$ in (3.7). □

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 [36] (see Rmk. 3.4). The following short Lemma will be required in the proof.

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

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

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

which shows the claim. □

Proposition 3.3. *Let $\delta \in (0,1)$. There exists a ReLU 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 $0 < \delta < 1$ it holds $\mathrm{size}(\tilde{\prod}) \le C(1 + n\log(n/\delta))$ *and* $\mathrm{depth}(\tilde{\prod}) \le C(1 + \log(n)\log(n/\delta))$.

Proof. Step 1: We construct the network. With $A := \{2^j : j \in \mathbb{N}\}$ let $\tilde{n} = \min\{x \geq n : x \in A\}$ and define $x_j := 1$ for every $j \in \{n+1, \dots, \tilde{n}\}$. We note that $\tilde{n} \leq 2n$. We introduce $\tilde{\times}$ similar as in Prop. 3.1:

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

where $m = C_1 \log(\tilde{n}/\delta)$ and the constant $C_1 \geq 1/\log(2)$ is to be chosen subsequently independent of $\tilde{n}$ and δ. Then $\tilde{\times}$ is a NN 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 $|x|$, $|y| \leq 2$ it holds

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

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

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

and (cp. Fig. 1 (c))

$$\tilde{\prod}(x_1, \dots, x_n) := \underbrace{R \circ \cdots \circ R}_{\log_2(\tilde{n})}(x_1, \dots, x_{\tilde{n}}) \in \mathbb{R}. \tag{3.9}$$

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 $0 < \delta \leq 1$

$$\begin{aligned} (1 + 2\varepsilon(\tilde{n}))^{\tilde{n}} &= (1 + 2(3 \cdot 2^{-2C_1 \log(\tilde{n}/\delta)}))^{\tilde{n}} \\ &\leq (1 + 6 \cdot 2^{-2C_1 \log(\tilde{n})})^{\tilde{n}} = (1 + 6\tilde{n}^{-2\log(2)C_1})^{\tilde{n}} \\ &= \exp\left(\tilde{n} \log(1 + 6\tilde{n}^{-2\log(2)C_1})\right). \end{aligned} \tag{3.10}$$

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 $j = 0, \ldots, r$, it holds that $|R^r(b_1, \ldots, b_{2^r}) - \prod_{j=1}^{2^r} b_j| \leq (4^r - 1)\varepsilon$. 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

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

To verify the above claim, we proceed by induction over r. 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 $R^r(b_1, \ldots, b_{2^r}) = \tilde{\times}(R^{r-1}(b_1, \ldots, b_{2^{r-1}}), R^{r-1}(b_{2^{r-1}+1}, \ldots, b_{2^r}))$. We get

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

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

Step 4: Finally we sum up all layers, weights and units. The operator $\tilde{\prod}$ in (3.9) 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 first hidden layer. Next, the first application of R employs $\tilde{n}/2$ times the NN $\tilde{\times}$ 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 NN $\tilde{\times}$ 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 $\log_2(\tilde{n})$ applications of R 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 $\tilde{n} \leq 2n$, this shows that the network uses the stated number of units, weights and layers. □

Remark 3.4. *The proof in [36] 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 Prop. 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 [26] using a NN composed of ReL and BiS ("binary step") units.*

Remark 3.6. *Feedforward ReLU networks cannot exactly represent the multiplication of two (or* $n \geq 3$*) real numbers, since the output is necessarily a piecewise linear function. Other architectures, such as the recently proposed "sum-product-networks" [30], make stronger assumptions:*

they stipulate availability of so-called "multiplication units", i.e. of units returning the exact *product of two inputs. In deep sum-product-networks, multiplying n numbers becomes trivial. Results analogous to the present ones will also hold for such sum-product-networks, with better complexity bounds. As ReLU NNs are widely used currently, we do not elaborate.*

3.3 Smoother activation functions

In [32], the authors show that for smooth (nonlinear) activation functions σ_s, the function $x \mapsto x^2$ can in fact be approximated to arbitrary accuracy with a *fixed number of units*. The idea is to assume that there exists $x_0 \in \mathbb{R}$ such that $\sigma_s(x_0+h) = c_0 + c_1 h + c_2 h^2 + O(x^3)$ as $h \to 0$, for $c_0 = \sigma_s(x_0)$, $c_1 = \sigma_s'(x_0)$ and $c_2 = \sigma_s''(x_0) \neq 0$ (since σ_s is nonlinear, there must be at least one point in $\mathbb{R}$ at which the second derivative is nonzero). Then

$$f(h) := \frac{\sigma_s(x_0+h) + \sigma_s(x_0-h) - 2c_0}{2c_2} = h^2 + O(h^3) \qquad \text{as } h \to 0.$$

Introducing the scaling factor $\lambda > 0$, it holds

$$|\lambda^2 f(x/\lambda) - x^2| = |\lambda^2 O(x^3/\lambda^3)| \qquad \text{as } x/\lambda \to 0.$$

This shows that for arbitrary $M > 0$, for sufficiently large $\lambda > 0$ the NN $\lambda^2 f(x/\lambda)$ (which is a one-layer network comprising two units) approximates $x \mapsto x^2$ for all $x \in [-M, M]$ at any prescribed accuracy. With similar constructions as displayed in Fig. 1, one then obtains the following result, which is a particular case of Thm. II.1 in [32].

Proposition 3.7. *Let $\sigma_s : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable and asssume that there exists $x_0 \in \mathbb{R}$ with $\sigma_s''(x_0) \neq 0$. Let $n \in \mathbb{N}$. Then, for every $\delta > 0$ and every $M > 0$, there exists a FFNN $\tilde{\prod}$ employing the activation function σ_s, such that $\tilde{\prod}_s$ has n inputs and for all $x_1, \ldots, x_n$ with $|x_i| \leq 1$ it holds $|\prod_{j=1}^n x_j - \tilde{\prod}_s(x_1, \ldots, x_n)| \leq \delta$.*

There exists a constant C (independent of δ) such that for all $n \in \mathbb{N}$ it holds $\mathrm{size}(\tilde{\prod}_s) \leq Cn$ *and* $\mathrm{depth}(\tilde{\prod}_s) \leq C(1 + \log(n))$.

3.4 DNN approximation of $(\boldsymbol{b}, \varepsilon)$-holomorphic maps

We now give a result on the expressive power of NNs concerning $(\boldsymbol{b}, \varepsilon)$-holomorphic functions. It states that, up to logarithmic terms, DNNs are capable of approximating real-valued $(\boldsymbol{b}, \varepsilon)$-holomorphic maps at rates equivalent to those achieved by best n-term gpc approximation. Here, *the notion "rate" is understood in terms of the NN size*, i.e., in terms of the total number of units and weights in the DNN.

Lemma 3.8. *Let $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in (0,\infty)^{\mathcal{F}}$, $p \in (0,1)$ and $\Lambda_n \subset \mathcal{F}$ for every $n \in \mathbb{N}$ be as in Lemma 2.8. Then there exists a sequence of ReLU NNs $(f_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n}$ such that $f_{\boldsymbol{\nu}}$ has $|\mathrm{supp}\,\boldsymbol{\nu}|$ many input variables $(y_j)_{j\in\mathrm{supp}\,\boldsymbol{\nu}}$, and there exists a constant C depending on $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ and p such that for every $n \in \mathbb{N}$*

$$\sup_{\boldsymbol{y}\in U} \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} |\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\mathrm{supp}\,\boldsymbol{\nu}})| \leq Cn^{-1/p+1} \tag{3.11}$$

and

$$\begin{aligned} \sum_{\boldsymbol{\nu}\in\Lambda_n} \mathrm{size}(f_{\boldsymbol{\nu}}) &\leq C(1 + n\log(n)\log\log(n)), \\ \max_{\boldsymbol{\nu}\in\Lambda_n} \mathrm{depth}(f_{\boldsymbol{\nu}}) &\leq C(1 + \log(n)\log\log(n)). \end{aligned} \tag{3.12}$$

Moreover, for every $n \in \mathbb{N}$ holds $\sup_{\boldsymbol{\nu}\in\Lambda_n} \sup_{\boldsymbol{y}\in U} |f_{\boldsymbol{\nu}}((y_j)_{j\in\mathrm{supp}\,\boldsymbol{\nu}})| \leq 2$.

Proof. Let $\sigma(\xi) = \max\{0, \xi\}$, $\xi \in \mathbb{R}$. Fix $n \in \mathbb{N}$ and $\boldsymbol{\nu} \in \Lambda_n$. For every $j \in \operatorname{supp} \boldsymbol{\nu}$, we may create ν_j copies of y_j as an output of the first hidden layer in the network by setting $y_j = \sigma(y_j) - \sigma(-y_j)$. This will require $O(|\boldsymbol{\nu}|_1)$ units and weights. We use these $|\boldsymbol{\nu}|_1$ values (which contain y_j exactly ν_j times) as an input to the approximate multiplication of Prop. 3.3, with accuracy $\delta_{\boldsymbol{\nu}} := \min\{1, 1/(t_{\boldsymbol{\nu}} n^{1/p})\}$. This results in a NN $f_{\boldsymbol{\nu}}$ such that both depth and size of $f_{\boldsymbol{\nu}}$ are bounded by

$$C(1 + |\boldsymbol{\nu}|_1 \log(|\boldsymbol{\nu}|_1/\delta_{\boldsymbol{\nu}})) = C(1 + |\boldsymbol{\nu}|_1(\log(|\boldsymbol{\nu}|_1) + \log(\max\{1, t_{\boldsymbol{\nu}} n^{1/p}\})), \tag{3.13}$$

where $C > 0$ is independent of $\boldsymbol{\nu}$ and of n. Since

$$\sup_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \le C(1 + \log(n)) \tag{3.14}$$

by (ii) of Lemma 2.8, and because $\sup_{\boldsymbol{\nu}\in\Lambda_n} \log(\max\{1, t_{\boldsymbol{\nu}} n^{1/p}\}) \le C\log(n)$ (here we used $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F}) \hookrightarrow \ell^\infty(\mathcal{F})$) we obtain the estimate on the depth in (3.12). For the NN size estimate in (3.12), we get with (3.13), (3.14) and, because $|\Lambda_n| = n$,

$$\begin{aligned}\sum_{\boldsymbol{\nu}\in\Lambda_n} \operatorname{size}(f_{\boldsymbol{\nu}}) &\le C \sum_{\boldsymbol{\nu}\in\Lambda_n} (1 + \log(n)\log\log(n)) + C \sum_{\boldsymbol{\nu}\in\Lambda_n} (1 + \log(n)\log(\max\{1, t_{\boldsymbol{\nu}} n^{1/p}\})) \\ &\le C(1 + n\log(n)\log\log(n)) + C(1 + \log(n)) \sum_{\boldsymbol{\nu}\in\Lambda_n} \log(\max\{1, t_{\boldsymbol{\nu}} n^{1/p}\}).\end{aligned}$$

For the sum in the upper bound, we obtain with $\log(x) \le x$ for $x \ge 1$ the bound

$$\sum_{\boldsymbol{\nu}\in\Lambda_n} \log(\max\{1, t_{\boldsymbol{\nu}} n^{1/p}\}) = \frac{1}{p} \sum_{\boldsymbol{\nu}\in\Lambda_n} \log(\max\{1, t_{\boldsymbol{\nu}}^p n\}) \le \frac{n}{p} \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}}^p \le Cn,$$

with $C = \|(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^p(\mathcal{F})}^p / p < \infty$ independent of n. In all this proves (3.12).

Estimate (3.11) is obtained by

$$\sup_{\boldsymbol{y}\in U} |\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| = \sup_{\boldsymbol{y}\in U} \left| \prod_{j\in\operatorname{supp}_{\boldsymbol{\nu}}} y_j^{\nu_j} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}) \right| \le \delta_{\boldsymbol{\nu}} = \min\{1, t_{\boldsymbol{\nu}}^{-1} n^{-1/p}\} . \tag{3.15}$$

This implies

$$\sup_{\boldsymbol{y}\in U} \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} |\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| \le \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} \min\{1, t_{\boldsymbol{\nu}}^{-1} n^{-1/p}\} \le |\Lambda_n| n^{-1/p} = n^{-1/p+1} .$$

The uniform boundedness of the NNs $\boldsymbol{y} \mapsto f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})$ with respect to $\boldsymbol{y} \in U$ and with respect to $\boldsymbol{\nu} \in \Lambda_n$ follows immediately by (3.15). □

Theorem 3.9. *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 $\tilde{u}_n(y_1, \ldots, y_n)$ with n input units such that

$$\operatorname{size}(\tilde{u}_n) \le C(1 + n\log(n)\log\log(n)), \qquad \operatorname{depth}(\tilde{u}_n) \le C(1 + \log(n)\log\log(n))$$

and such that $\tilde{u}_n$ satisfies the uniform error bound

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

Proof. According to Prop. 2.3 there exist coefficients $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \mathbb{R}^{\mathcal{F}}$ such that $u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}}$ in the sense of unconditional convergence for every $\boldsymbol{y} \in U$. By Thm. 2.7, for every $n \in \mathbb{N}$ there exists a downward closed index set $\Lambda_n \subset \mathcal{F}$ with $|\Lambda_n| = n$ such that

$$\sup_{\boldsymbol{y}\in U} \left| u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \right| \leq \sum_{\boldsymbol{\nu}\in\Lambda_n^c} |t_{\boldsymbol{\nu}}| \leq C n^{-1/p+1}. \tag{3.16}$$

It thus suffices to approximate $\sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}}$ with a DNN.

In the proof of Thm. 2.7 we constructed a monotonically decreasing sequence $\boldsymbol{\beta} \in \ell^{p/(1-p)}(\mathbb{N})$ of positive real numbers with $\|\boldsymbol{\beta}\|_{\ell^\infty(\mathbb{N})} < 1$, such that $(\|t_{\boldsymbol{\nu}}\|_V \boldsymbol{\beta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$. The set $\Lambda_n \subset \mathcal{F}$ was then chosen as in Lemma 2.8. Hence, by Lemma 3.8 there exists a sequence of DNNs $(f_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n}$ satisfying (3.12) and

$$\sup_{\boldsymbol{y}\in U} \sum_{\boldsymbol{\nu}\in\Lambda_n} |t_{\boldsymbol{\nu}}| |\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| \leq C n^{-1/p+1}. \tag{3.17}$$

Next, we claim that

$$S := \{j \in \mathbb{N} : \exists \boldsymbol{\nu} \in \Lambda_n \text{ s.t. } j \in \operatorname{supp}\boldsymbol{\nu}\} \subseteq \{1, \ldots, n\}. \tag{3.18}$$

To show it, we note that due to the downward closedness of Λ_n, for each $j \in \operatorname{supp}\boldsymbol{\nu}$ for some $\boldsymbol{\nu} \in \Lambda_n$, it must hold $\mathbf{e}_j \in \Lambda_n$. Therefore $|S| \leq |\Lambda_n| = n$. Now suppose that $r \in S$ for some $r > n$. By Thm. 2.7 $\mathbf{e}_r \in \Lambda_n$ implies $\mathbf{e}_j \in \Lambda_n$ for every $j \leq r$. Thus $|S| > n$, which is a contradiction and proves (3.18). Define

$$\tilde{u}_n((y_j)_{j\in S}) := \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}). \tag{3.19}$$

By (3.18) $\tilde{u}_n$ does not depend on y_j for $j > n$, so that we can take $y_1, \ldots, y_n$ as the input of the NN $\tilde{u}_n$. Using (3.16) and (3.17), the NN $\tilde{u}_n$ satisfies the error bound

$$\begin{aligned}\sup_{\boldsymbol{y}\in U} |u(\boldsymbol{y}) - \tilde{u}_n(y_1, \ldots, y_n)| &\leq C n^{-1/p+1} + \sup_{\boldsymbol{y}\in U} \left| \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} - \sum_{\boldsymbol{\nu}\in\Lambda_n} t_{\boldsymbol{\nu}} f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}) \right| \\ &\leq C n^{-1/p+1} + \sup_{\boldsymbol{y}\in U} \sum_{\boldsymbol{\nu}\in\Lambda_n} |t_{\boldsymbol{\nu}}| |\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| \leq C n^{-1/p+1}.\end{aligned}$$

It remains to estimate the size and depth of $\tilde{u}_n$ in (3.19). With the input layer consisting of the values $y_1, \ldots, y_n$, starting from the second layer we can compute all the values $f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})$ with the networks $f_{\boldsymbol{\nu}}$ in parallel. To this end each y_j needs to connect to all networks $f_{\boldsymbol{\nu}}$ for which $j \in \operatorname{supp}\boldsymbol{\nu}$. The number of nonzero weights used to create these connections can simply be bounded by $\sum_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \leq C(1 + n\log(n))$, since $\sup_{\boldsymbol{\nu}\in\Lambda_n} |\boldsymbol{\nu}|_1 \leq C(1 + \log(n))$ by (ii) of Thm. 2.7. In the output layer, the sum in (3.19) is computed. Thus, up to a constant, the total depth of the NN $\tilde{u}_n$ is bounded by $\max_{\boldsymbol{\nu}\in\Lambda_n} \operatorname{depth}(f_{\boldsymbol{\nu}}) \leq C(1 + \log(n)\log\log(n))$ by Lemma 3.8. The size can be bounded by

$$C(1 + n\log(n)) + \sum_{\boldsymbol{\nu}\in\Lambda_n} \operatorname{size}(f_{\boldsymbol{\nu}}) \leq C(1 + n\log(n)\log\log(n)), \tag{3.20}$$

where we again used Lemma 3.8. This concludes the proof. □

Based on Prop. 3.7, we obtain the following improved result for NNs employing smooth activation functions.

Theorem 3.10. *Let $u : U \to \mathbb{R}$ be $(\boldsymbol{b}, \varepsilon)$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, and with some $p \in (0,1)$. Let further the activation $\sigma_{\mathrm{s}} : \mathbb{R} \to \mathbb{R}$ be three times continuously differentiable and nonlinear.*

Then, there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$ there exists a FFNN $\tilde{u}_n$ using σ_{s} as activation function, where $\tilde{u}_n$ has n input units, $\mathrm{size}(\tilde{u}_n) \le C(1 + n\log(n))$ and $\mathrm{depth}(\tilde{u}_n) \le C(1 + \log\log(n))$ so that the expression error is bounded as

$$\sup_{\boldsymbol{y} \in U} |u(\boldsymbol{y}) - \tilde{u}_n(y_1, \dots, y_n)| \le C n^{1-1/p} \,. \tag{3.21}$$

Proof. The proof is analogous to the one of Lemma 3.8 and Thm. 3.9, employing Prop. 3.7 instead of Prop. 3.3. Therefore we only sketch the argument. Define $\tilde{u}_n$ as in (3.19), but replace $f_{\boldsymbol{\nu}}((y_j)_{j \in \mathrm{supp}\,\boldsymbol{\nu}})$ with $f_{\mathrm{s};\boldsymbol{\nu}}((y_j)_{j \in \mathrm{supp}\,\boldsymbol{\nu}}) = \tilde{\prod}_{\mathrm{s}}(y_{i_1}, \dots, y_{i_{|\boldsymbol{\nu}|_1}})$, where $\tilde{\prod}_{\mathrm{s}}$ denotes the approximate multiplication of Prop. 3.7 and $(i_j)_{j=1}^{|\boldsymbol{\nu}|_1} \in \mathbb{N}^{|\boldsymbol{\nu}|_1}$ is such that each $j \in \mathrm{supp}\,\boldsymbol{\nu}$ occurs exactly ν_j times. By Prop. 3.7, the NN $f_{\mathrm{s};\boldsymbol{\nu}}(\boldsymbol{y})$ can approximate $\boldsymbol{y}^{\boldsymbol{\nu}}$ to any precision uniformly with respect to $\boldsymbol{y} \in U$ using $O(\log(|\boldsymbol{\nu}|_1))$ layers and $O(|\boldsymbol{\nu}|_1)$ weights and units. In particular, we may assume that (3.17) still holds if $f_{\boldsymbol{\nu}}$ is replaced by $f_{\mathrm{s};\boldsymbol{\nu}}$. Consequently, the same calculations as in the proof of Thm. 3.9 imply the bound (3.21). The depth of $\tilde{u}_n$ is then, up to a constant, the maximum depth of one of the subnetworks $f_{\mathrm{s};\boldsymbol{\nu}}$, $\boldsymbol{\nu} \in \Lambda_n$, and thus bounded by $\max_{\boldsymbol{\nu} \in \Lambda_n} C(1 + \log(|\boldsymbol{\nu}|_1))$. According to item (ii) of Thm. 2.7, this can be further estimated by $C(1 + \log\log(n))$. The total number of weights and units can be bounded by n, which is the number of inputs $(y_1, \dots, y_n)$, and by $\sum_{\boldsymbol{\nu} \in \Lambda_n} \mathrm{size}(f_{\mathrm{s};\boldsymbol{\nu}}) \le C \sum_{\boldsymbol{\nu} \in \Lambda_n} |\boldsymbol{\nu}|_1$. With item (ii) of Thm. 2.7, this gives

$$\mathrm{size}(\tilde{u}_n) \le n + \sum_{\boldsymbol{\nu} \in \Lambda_n} C(1 + |\boldsymbol{\nu}|_1) \le C(1 + n\log(n))$$

as an upper bound on the total number of units and weights. □

We remark that for every $n \in \mathbb{N}$, the network size $N(n)$ in Thms. 3.9 and 3.10 is bounded by $C_\gamma n^{1+\gamma}$ for arbitrarily small $\gamma > 0$, in terms of the number n of input parameters. Thus Thms. 3.9 and 3.10 imply the convergence rate $-1/p + 1 + \gamma$ with $\gamma > 0$ arbitrary in terms of the NN size N.

4 Examples and generalizations

4.1 Response surfaces of parametric PDEs

In Ex. 2.2 we considered the parametric weak variational formulation: find $u(\boldsymbol{y}) \in V = H_0^1(D)$ such that

$$\int_D a(\boldsymbol{y}, x) \nabla u(\boldsymbol{y}, x) \cdot \nabla v(x) \,\mathrm{d}x = {}_{V'}\langle f, v \rangle_V, \quad \forall\, v \in H_0^1(D), \tag{4.1}$$

with the parametric diffusion coefficient

$$a(\boldsymbol{y}, x) = \overline{a}(x) + \sum_{j \in \mathbb{N}} y_j \psi_j(x), \qquad x \in D, \tag{4.2}$$

where $D \subseteq \mathbb{R}^d$, $d \ge 1$, is a bounded Lipschitz domain, $f \in V'$ and $(\|\psi_j\|_{L^\infty(D)}) \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$. In the following, we occasionally omit the x argument, and write e.g. $u(\boldsymbol{y}) \in V$ instead of $u(\boldsymbol{y}, \cdot) \in V$. In Ex. 2.2, we argued that the solution $u(\boldsymbol{y}) \in V$ is $(\boldsymbol{b}, \varepsilon)$-holomorphic under certain assumptions (which we recall in the theorem below). In order to apply Thms. 3.9 and 3.10 (which were formulated for real-valued functions), let $G \in V'$ be a bounded linear functional. It is straightforward to verify that $G \circ u : U \to \mathbb{R}$ is then also $(\boldsymbol{b}, \varepsilon)$-holomorphic.

Furthermore, with the unconditionally convergent expansion $u(\boldsymbol{y}) = \sum_{\boldsymbol{y}\in U} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \in V$ as in (2.7), it holds $G(u(\boldsymbol{y})) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} G(t_{\boldsymbol{\nu}})\boldsymbol{y}^{\boldsymbol{\nu}} \in \mathbb{R}$ for $\boldsymbol{y} \in U$. Since $|G(t_{\boldsymbol{\nu}})| \le \|G\|_{V'}\|t_{\boldsymbol{\nu}}\|_V$ for all $\boldsymbol{\nu} \in \mathcal{F}$, we immediately get

$$\left\|(|G(t_{\boldsymbol{\nu}})|)_{\boldsymbol{\nu}\in\mathcal{F}}\right\|_{\ell^p(\mathcal{F})} \le \|G\|_{V'}\left\|(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}}\right\|_{\ell^p(\mathcal{F})}.$$

We arrive at the following result.

Corollary 4.1. *Let $p \in (0,1)$, let $\bar{a} = 1$ and assume that $(\psi_j)_{j\in\mathbb{N}} \subset L^\infty(D)$ is such that with $b_j := \|\psi_j\|_{L^\infty(D)}$ the sequence $\boldsymbol{b} = (b_j)_{j\in\mathbb{N}}$ is monotonically decreasing and it holds $\|\boldsymbol{b}\|_{\ell^1(\mathbb{N})} < 1$ and $\|\boldsymbol{b}\|_{\ell^p(\mathbb{N})} < \infty$. Then the weak solution $u(\boldsymbol{y}) \in V = H_0^1(D)$ of (4.1)-(4.2) is $(\boldsymbol{b},\varepsilon)$-holomorphic. Moreover, there exists a constant C such that for every $G \in V'$ and for every $n \in \mathbb{N}$ there exists a ReLU network $\tilde{g}_n(y_1,\ldots,y_n)$ with n input units such that* $\mathrm{size}(\tilde{g}_n) \le C(1 + n\log(n)\log\log(n))$, $\mathrm{depth}(\tilde{g}_n) \le C(1 + \log(n)\log\log(n))$ *and*

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

In case the expansion functions $\psi_j \in L^\infty(D)$ in (4.2) have local supports, the convergence rate can be slightly improved. We discuss our result in this setting, based on [1].

Theorem 4.2 ([1, Thm. 1.1])**.** *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 monotonically decreasing 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,*

$$\theta := \left\| \frac{\sum_{j\in\mathbb{N}} \beta_j^{-1}|\psi_j(\cdot)|}{\bar{a}(\cdot)} \right\|_{L^\infty(D)} < 1 \ . \tag{4.3}$$

Then with $t_{\boldsymbol{\nu}}$ as in (2.6) it holds $\sum_{\boldsymbol{\nu}\in\mathcal{F}} (\boldsymbol{\beta}^{-\boldsymbol{\nu}}\|t_{\boldsymbol{\nu}}\|_V)^2 < \infty$, and in particular $(\|t_{\boldsymbol{\nu}}\|_V)_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$.

We remark that the proof of Theorem 4.2 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 $t_{\boldsymbol{\nu}}$.

As in the case of $(\boldsymbol{b},\varepsilon)$-holomorphy, if $\boldsymbol{b} \in \ell^p(\mathbb{N})$ for some $p \in (0,1)$, one can show that the Taylor gpc expansion

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}} \ , \tag{4.4}$$

converges unconditionally for all $\boldsymbol{y} \in U$ for the parametric solution $u(\boldsymbol{y})$ in Thm. 4.2. Using Lemma 2.8 with $r = 2$, a proof completely analogous to the one of Thm. 3.9 then yields:

Corollary 4.3. *Let $p \in (0,1)$ and $q = 2p/(2-p)$, i.e. $1/p = 1/q + 1/2$. Then, under the assumptions of Thm. 4.2, there exists a constant C such that for every $G \in V'$ and for every $n \in \mathbb{N}$ there exists a ReLU network $\tilde{g}_n(y_1,\ldots,y_n)$ with n input units such that* $\mathrm{size}(\tilde{g}_n) \le C(1 + n\log(n)\log\log(n))$, $\mathrm{depth}(\tilde{g}_n) \le C(1 + \log(n)\log\log(n))$ *and there hold the uniform error bounds*

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

We conclude that Thm. 3.9 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. Using Thm. 3.10 instead of Thm. 3.9 one can infer variants of Corollaries 4.1 and 4.3 for NNs based on a smooth activation function σ_{s} as in Thm. 3.10.

4.2 Solution manifolds of parametric PDEs

For the results of Sec. 4.1, we considered only real-valued countably parametric maps. Often, however, rather than linear functionals $G \in V'$ of the parametric solution, also the approximation of the parametric solution manifold $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in V$ itself is of interest. Here, also a *"spacial approximation"* of $u(\boldsymbol{y}) \in V$ is sought, where V is a Banach space. The results proved imply in particular that DNNs can express such solution manifolds. Rather than developing this in the most general setting and in order to keep technicalities to a minimum, we illustrate this for the physical domain $D = (0,1)$ in space dimension $d = 1$. To present the setting, we assume further in the parametric weak formulation (4.1) that $f \in L^2(D)$, and that $U \ni \boldsymbol{y} \mapsto a(\boldsymbol{y}) \in W^{1,\infty}(D)$. Standard regularity results then imply that for every $\boldsymbol{y} \in U$, $u(\boldsymbol{y}) \in X := (H^2 \cap H^1_0)(D)$. Furthermore, the sequence of $H^2(D)$-norms of the Taylor gpc coefficients $t_{\boldsymbol{\nu}}$ in (4.4) of $u(\boldsymbol{y})$ is sparse as expressed in the following gpc summability result from [1, Thm. 2.1].

Theorem 4.4. *Let $0 < q < \infty$ and $0 < p < 2$ be such that $\frac{1}{p} = \frac{1}{q} + \frac{1}{2}$. Assume that $\bar{a} \in L^\infty(D)$ is such that $\operatorname{ess\,inf} \bar{a} > 0$, and that there exists a sequence $\boldsymbol{\beta} = (\beta_j)_{j\in\mathbb{N}} \in (0,1)^{\mathbb{N}}$ such that $\boldsymbol{\beta} \in \ell^q(\mathbb{N})$ and such that*

$$\theta := \left\| \frac{\sum_{j\in\mathbb{N}} \beta_j^{-1} |\psi_j(\cdot)|}{\bar{a}(\cdot)} \right\|_{L^\infty(D)} < 1 \ . \tag{4.5a}$$

Assume in addition that $f \in L^2(D)$ and that $\bar{a}$ and all functions ψ_j belong to $W^{1,\infty}(D)$ and that

$$\left\| \sum_{j\in\mathbb{N}} \beta_j^{-1} |\psi_j'(\cdot)| \right\|_{L^\infty(D)} < \infty \ . \tag{4.5b}$$

Denote by $u(\boldsymbol{y})$ the solution of (4.1)-(4.2). Then, with $t_{\boldsymbol{\nu}}$ as in (2.6), $\sum_{\boldsymbol{\nu}\in\mathcal{F}} (\boldsymbol{\beta}^{-\boldsymbol{\nu}} \|t_{\boldsymbol{\nu}}\|_{H^2(D)})^2 < \infty$ and in particular $(\|t_{\boldsymbol{\nu}}\|_{H^2(D)})_{\nu\in\mathcal{F}} \in \ell^p(\mathcal{F})$.

The p-summability of the $H^2(D)$-norms of the Taylor gpc coefficients $t_{\boldsymbol{\nu}}$ is key for the analysis of FE approximations of $t_{\boldsymbol{\nu}}$ in D. Given $\boldsymbol{y} \in U$, we consider an approximation of $u(\boldsymbol{y}) \in X$ by continuous, piecewise linear finite elements in D: for a (uniform) meshwidth $h = 1/m_h$ with $1 < m_h \in \mathbb{N}$, we denote by $\mathcal{T}_h$ the mesh of m_h subintervals $K^h_j = (x^h_{j-1}, x^h_j) \subset D$, with the (equispaced) nodes $x^h_j = jh$, $j = 0, 1, ..., m_h$. We denote by

$$\mathbb{R} \ni \xi \mapsto \hat{\varphi}(\xi) := \begin{cases} 1 - |\xi| & \text{if} \quad |\xi| < 1 \ , \\ 0 & \text{else} \ , \end{cases} \tag{4.6}$$

the standard "hat" continuous, piecewise linear function. With g as in (3.5b), observe that $\hat{\varphi}$ can be exactly represented by a one-layer ReLU NN via

$$\hat{\varphi}(\xi) = g((\xi+1)/2) = 2\sigma(\xi/2 + 1/2) - 4\sigma(\xi/2) + 2\sigma(\xi/2 - 1/2) \ , \qquad \xi \in \mathbb{R} \ . \tag{4.7}$$

Alternatively, the following two-layer network, which also uses 3 ReLUs, achieves the same

$$\hat{\varphi}(\xi) = \sigma(1 - [\sigma(\xi) + \sigma(-\xi)]) \ , \qquad \xi \in \mathbb{R} \ . \tag{4.8}$$

The Courant FE basis functions φ^h_j in V are then defined by translating and scaling $\hat{\varphi}$:

$$\varphi^h_j(x) := \hat{\varphi}(h^{-1}(x - x^h_j)) = \sigma(1 - h^{-1}[\sigma(x - x^h_j) + \sigma(-x + x^h_j)]) \ , j = 1, ..., m_h - 1 \ . \tag{4.9}$$

The FE space $V_h \subset V$ is their span, i.e.,

$$V_h := \operatorname{span}\{\varphi^h_j : j = 1, ..., m_h - 1\} \ . \tag{4.10}$$

Evidently, V_h is a subspace of $V = H_0^1(D)$ of finite dimension $\dim(V_h) = m_h - 1$. For $w \in X = (H^2 \cap H_0^1)(D) \subset C^1(D)$, we denote by $I_h w \in V_h$ the continuous, piecewise linear function on $\mathcal{T}_h$ which interpolates w in the nodes x_j^h, $j = 0, 1, ..., m_h$. There holds the approximation error bound

$$\|w - I_h w\|_{L^2(D)} + h\|w' - (I_h w)'\|_{L^2(D)} \leq C_X h^2 \|w''\|_{L^2(D)} \,. \tag{4.11}$$

Using the approximate multiplication from Prop. 3.1 and $\hat{\varphi}$ as in (4.7) or (4.8), up to a certain error one can express tensorized translated and scaled hat functions. Together with well-established FEM theory, this yields results about the approximation of Sobolev functions in higher dimensional spaces via NNs (also sparse grids, and as we saw in Sec. 3.2, higher order polynomials could be used). As mentioned above, for ease of exposition we prove this in the one dimensional case and for continuous, piecewise linear Finite Elements only.

Lemma 4.5. *For every $2 \leq m_h \in \mathbb{N}$, every continuous, piecewise linear function $w \in V_h$ on the uniform partition $\mathcal{T}_h$ of $D = (0,1)$ into m_h intervals of length $h = 1/m_h$ can be* exactly expressed *by a one layer network using m_h ReLU units. More precisely*

$$w(x) = \sum_{j=0}^{m_h - 1} c_j \sigma(x - x_j^h), \qquad x \in [0,1], \tag{4.12}$$

where

$$c_0 := \frac{w(x_1^h)}{h} \qquad \text{and} \qquad c_j := \frac{w(x_{j+1}^h) - w(x_j^h)}{h} - \sum_{i=0}^{j-1} c_i \quad \forall\, j = 1, \ldots, m_h - 1.$$

Proof. Denote in the following the right-hand side of (4.12) by $\tilde{w}$. We show by induction over j that $\tilde{w}|_{[0,x_j^h]} = w|_{[0,x_j^h]}$ for all $j = 1, \ldots, m_h$. First let $j = 1$ and note that $x_0^h = 0$ and $x_1^h = h$. Since for all $i = 0, \ldots, m_h - 1$

$$\operatorname{supp} \sigma(\cdot - x_i^h) = [x_i^h, \infty), \tag{4.13}$$

for $x \in [0, x_1^h]$ it holds $\tilde{w}(x) = (w(x_1^h)/h)\sigma(x - x_0^h) = (w(x_1^h)/h)x$ so $\tilde{w}|_{[0,x_1^h]}$ is a linear function with $\tilde{w}(0) = 0$ and $\tilde{w}(x_1^h) = w(x_1^h)$, and thus $\tilde{w}|_{[0,x_1^h]} = w|_{[0,x_1^h]}$. For the induction step, assume that $\tilde{w}|_{[0,x_j^h]} = w|_{[0,x_j^h]}$ for some $j < n$, so that in particular $\tilde{w}(x_j^h) = w(x_j^h)$. It then suffices to show that $\tilde{w}|_{[x_j^h, x_{j+1}^h]}$ is affine with derivative $(w(x_{j+1}^h) - w(x_j^h))/h$, since this implies $\tilde{w}(x_{j+1}^h) = w(x_j^h) + \int_{x_j^h}^{x_{j+1}^h} \tilde{w}'(x) \mathrm{d}s = w(x_{j+1}^h)$. Because of (4.13), for $x \in [x_j^h, x_{j+1}^h]$

$$\begin{aligned} \tilde{w}(x) &= \sum_{i=0}^{j-1} c_i \sigma(x - x_i^h) + \left(\frac{w(x_{j+1}^h) - w(x_j^h)}{h} - \sum_{i=0}^{j-1} c_i \right) \sigma(x - x_j^h) \\ &= \sum_{i=0}^{j-1} c_i (x - x_i^h) + \left(\frac{w(x_{j+1}^h) - w(x_j^h)}{h} - \sum_{i=0}^{j-1} c_i \right) (x - x_j^h). \end{aligned}$$

This shows that $\tilde{w}|_{[x_j^h, x_{j+1}^h]}$ is affine with derivative $(w(x_{j+1}^h) - w(x_j^h))/h$, which concludes the proof. ☐

Remark 4.6. *By definition of c_{j-1} it holds $\sum_{i=0}^{j-1} c_i = c_{j-1} + \sum_{i=0}^{j-2} c_i = (w(x_j^h) - w(x_{j-1}^h))/h$ for all $j \geq 1$ and thus*

$$c_j = \frac{w(x_{j+1}^h) - 2w(x_j^h) + w(x_{j-1}^h)}{h}$$

for all $j = 1, \ldots, m_h - 1$ in Lemma 4.5.

A dimension-independent convergence rate for the DNN expression rate of the parameter-to-solution map

$$U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in V$$

can now be obtained from the following result, which is a particular case of [1, Thm. 3.1], specialized to the present assumptions. We state it in slightly greater generality than required subsequently, with the parameter $\gamma > 0$ denoting the rate of convergence in V of the FE discretization in the physical domain D; based on (4.11), we will choose $\gamma = 1$.

Lemma 4.7. *Let $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$, $(b_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in (0,\infty)^{\mathcal{F}}$ and $\boldsymbol{\beta} \in (0,1)^{\mathbb{N}}$ be such that $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p_a}(\mathcal{F})$, $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^2(\mathcal{F})$ and $\boldsymbol{\beta} \in \ell^{2p_b/(2-p_b)}(\mathbb{N})$ for some $0 < p_b \le p_a < 1$. Assume additionally that $\boldsymbol{\beta}$ is monotonically decreasing. For every $n \in \mathbb{N}$, let $\Lambda_n \subset \mathcal{F}$ be as in Lemma 2.8 for the sequence $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^2(\mathcal{F})$ and let $\gamma > 0$ be given.*

Then there exists a constant $C > 0$ and, for every $n \in \mathbb{N}$, there exists a sequence $(m_{n;\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n} \in \mathbb{N}^{\Lambda_n}$ such that with $\mathcal{N}_n := \sum_{\boldsymbol{\nu}\in\Lambda_n} m_{n;\boldsymbol{\nu}} \ge n$ it holds

$$n^{-1/p_b+1} + \sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}} m_{n;\boldsymbol{\nu}}^{-\gamma} + \sum_{\boldsymbol{\nu}\in\Lambda_n^c} b_{\boldsymbol{\nu}} \le C\mathcal{N}_n^{-r}$$

where the constant $C > 0$ is independent of n and where

$$r = \gamma \min\left\{1, \frac{1/p_b - 1}{\gamma + 1/p_b - 1/p_a}\right\}. \tag{4.14}$$

Proof. We proceed similarly as in [1, Sec. 3], see also [18, Sec. 2].

Fix $n \in \mathbb{N}$ and choose $\Lambda_n \subset \mathcal{F}$, $|\Lambda_n| = n$, as in Lemma 2.8 with the sequence $(\boldsymbol{\beta}^{-\boldsymbol{\nu}}b_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^2(\mathcal{F})$ (i.e. $r = 2$ in Lemma 2.8). Then Λ_n satisfies all properties stated in Lemma 2.8 and in particular $\sum_{\boldsymbol{\nu}\in\Lambda_n^c} b_{\boldsymbol{\nu}} \le Cn^{-1/p_b+1}$. To choose $m_{n;\boldsymbol{\nu}}$, we minimize $\sum_{\boldsymbol{\nu}\in\Lambda_n} m_{n;\boldsymbol{\nu}}$ under the constraint $\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}} m_{n;\boldsymbol{\nu}}^{-\gamma} \le n^{-1/p_b+1}$. Allowing for now $m_{n;\boldsymbol{\nu}}$ to take positive real values this can be solved using a Lagrange multiplier λ. To this end, we define $F((\tilde{m}_{n;\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n}, \lambda) := \sum_{\boldsymbol{\nu}\in\Lambda_n} \tilde{m}_{n;\boldsymbol{\nu}} + \lambda(\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}}\tilde{m}_{n;\boldsymbol{\nu}}^{-\gamma} - n^{-1/p_b+1})$. The first order necessary condition $\nabla F = 0$ results in

$$\tilde{m}_{n;\boldsymbol{\nu}} = n^{(1/p_b-1)/\gamma} a_{\boldsymbol{\nu}}^{1/(1+\gamma)} \left(\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}}^{1/(1+\gamma)}\right)^{1/\gamma}.$$

Define

$$m_{n;\boldsymbol{\nu}} := \lceil \tilde{m}_{n;\boldsymbol{\nu}} \rceil \in \mathbb{N}, \qquad \forall\, \boldsymbol{\nu} \in \Lambda_n,\ n \in \mathbb{N}.$$

By construction $\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}}\tilde{m}_{n;\boldsymbol{\nu}}^{-\gamma} = n^{-1/p_b+1}$ and thus $\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}} m_{n;\boldsymbol{\nu}}^{-\gamma} \le n^{-1/p_b+1}$. This implies that there exists $C > 0$ such that for every $n \in \mathbb{N}$

$$n^{-1/p_b+1} + \sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}} m_{n;\boldsymbol{\nu}}^{-\gamma} + \sum_{\boldsymbol{\nu}\in\Lambda_n^c} b_{\boldsymbol{\nu}} \le Cn^{-1/p_b+1}. \tag{4.15}$$

To complete the proof, we show that there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$ holds $n^{-1/p_b+1} \le C\mathcal{N}_n^{-r}$. We observe that $m_{n;\nu} \ge 1$ by construction. Hence, it holds $\mathcal{N}_n \ge \sum_{\boldsymbol{\nu}\in\Lambda_n} 1 = n$. To give an upper bound for $\mathcal{N}_n$ we first note

$$\mathcal{N}_n = \sum_{\boldsymbol{\nu}\in\Lambda_n} m_{n;\boldsymbol{\nu}} \le \sum_{\boldsymbol{\nu}\in\Lambda_n} (1 + \tilde{m}_{n;\boldsymbol{\nu}}) = n + n^{(1/p_b-1)/\gamma} \left(\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}}^{1/(1+\gamma)}\right)^{(1+\gamma)/\gamma}. \tag{4.16}$$

In the following, we distinguish between the two cases

$$p_a \le \frac{1}{1+\gamma} \qquad \text{and} \qquad p_a > \frac{1}{1+\gamma}. \tag{4.17}$$

In the first case in (4.17), by (4.16)

$$\mathcal{N}_n \le n + n^{(1/p_b-1)/\gamma}\|(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^{1/(1+\gamma)}(\mathcal{F})}^{1/\gamma} \le n + n^{(1/p_b-1)/\gamma}\|(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^{p_a}(\mathcal{F})}^{1/\gamma}.$$

Thus

$$n^{-1/p_b+1} \le C\left(\mathcal{N}_n^{-1/p_b+1} + \mathcal{N}_n^{-\gamma}\right) \le C\mathcal{N}_n^{-r},$$

since $r \le 1/p_b - 1$ due to $1/p_b - 1/p_a \ge 0$.

In the second case, by Hölder's inequality and using $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p_a}(\mathcal{F})$, there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$ holds

$$\mathcal{N}_n \le n + n^{(1/p_b-1)/\gamma}\left(\left(\sum_{\boldsymbol{\nu}\in\Lambda_n} a_{\boldsymbol{\nu}}^{p_a}\right)^{1/(p_a(1+\gamma))} n^{1-1/(p_a(1+\gamma))}\right)^{(1+\gamma)/\gamma} \le n + Cn^{(1/p_b+\gamma-1/p_a)/\gamma}.$$

Therefore

$$n^{-1/p_b+1} \le C\left(\mathcal{N}_n^{-1/p_b+1} + \mathcal{N}_n^{\frac{-1/p_b+1}{(1/p_b+\gamma-1/p_a)/\gamma}}\right). \tag{4.18}$$

This is again bounded by $C\mathcal{N}_n^{-r}$, since as before $1/p_b - 1 \le r$, and the second exponent on the right-hand side of (4.18) is (up to its sign) exactly the second term in the minimum in (4.14). ☐

Theorem 4.8. *Let $0 < q_V \le q_X < 2$ and denote $p_V := (1/q_V + 1/2)^{-1} \in (0,1)$ and $p_X := (1/q_X + 1/2)^{-1} \in (0,1)$. Let $\boldsymbol{\beta}_V = (\beta_{V;j})_{j\in\mathbb{N}} \in (0,1)^{\mathbb{N}}$ and $\boldsymbol{\beta}_X = (\beta_{X;j})_{j\in\mathbb{N}} \in (0,1)^{\mathbb{N}}$ be two monotonically decreasing sequences such that $\boldsymbol{\beta}_V \in \ell^{q_V}(\mathbb{N})$ and $\boldsymbol{\beta}_X \in \ell^{q_X}(\mathbb{N})$, and such that the parametric diffusion coefficient satisfies (4.3) with $\boldsymbol{\beta} = \boldsymbol{\beta}_V$ and (4.5) with $\boldsymbol{\beta} = \boldsymbol{\beta}_X$. Assume that $f \in L^2(D)$ in (4.1). Denote for every $\boldsymbol{y} \in U$ by $u(\boldsymbol{y}, \cdot) \in V$ the solution of (4.1) for the affine-parametric diffusion coefficient in (4.2). Then, there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$ there exists a ReLU network $\tilde{u}_n(y_1, \ldots, y_n, x)$ with $n+1$ input units and for a number $\mathcal{N}_n \ge n$ with $r = \min\{1, (1+p_V^{-1})/(1+p_V^{-1}-p_X^{-1})\}$ there holds the bound*

$$\sup_{\boldsymbol{y}\in U}\|u(\boldsymbol{y},\cdot) - \tilde{u}_n(y_1,\ldots,y_n,\cdot)\|_V \le C\mathcal{N}_n^{-r}. \tag{4.19}$$

Moreover, for every $n \in \mathbb{N}$,

$$\text{size}(\tilde{u}_n) \le C(1 + \mathcal{N}_n\log(\mathcal{N}_n)\log\log(\mathcal{N}_n)), \quad \text{depth}(\tilde{u}_n) \le C(1 + \log(\mathcal{N}_n)\log\log(\mathcal{N}_n)).$$

Proof. Step 1: For every $n \in \mathbb{N}$, define $\Lambda_n \subset \mathcal{F}$ and $(m_{n;\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n} \in \mathbb{N}^{\Lambda_n}$.

For $\boldsymbol{\nu} \in \mathcal{F}$ denote by $t_{\boldsymbol{\nu}} \in V = H_0^1(D)$ the Taylor coefficient of $u(\boldsymbol{y})$ defined in (2.6). Additionally, we define $a_{\boldsymbol{\nu}} := \|t_{\boldsymbol{\nu}}\|_{H^2(D)}$ and $b_{\boldsymbol{\nu}} := \|t_{\boldsymbol{\nu}}\|_V$ (in fact $t_{\boldsymbol{\nu}} \in H^2(D)$ by Thm. 4.4). By Thm. 4.2, we have $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}_V^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^2(\mathcal{F})$ and by Thm. 4.4 it holds $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p_X}(\mathcal{F})$.

For every $n \in \mathbb{N}$, let Λ_n and $(m_{n;\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n}$ be as in Lemma 4.7. Then, by Lemma 4.7, there exists a constant $C > 0$ such that for every $n \in \mathbb{N}$ holds

$$n^{-1/p_V+1} + \sum_{\boldsymbol{\nu}\in\Lambda_n}\|t_{\boldsymbol{\nu}}\|_X m_{n;\boldsymbol{\nu}}^{-1} + \sum_{\boldsymbol{\nu}\in\Lambda_n^c}\|t_{\boldsymbol{\nu}}\|_V \le C\mathcal{N}_n^{-r}, \tag{4.20}$$

with $r = \min\{1, (1+p_V^{-1})/(1+p_V^{-1}-p_X^{-1})\}$. Hence, by (4.11), there exists $C > 0$ such that for all $n \in \mathbb{N}$ holds, with $h_{n;\boldsymbol{\nu}} := 1/m_{n;\boldsymbol{\nu}}$, the bound

$$\sum_{\boldsymbol{\nu}\in\Lambda_n}\|t_{\boldsymbol{\nu}} - I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}}\|_V \le C\sum_{\boldsymbol{\nu}\in\Lambda_n} h_{n;\boldsymbol{\nu}}\|t_{\boldsymbol{\nu}}\|_X \le C\mathcal{N}_n^{-r}. \tag{4.21}$$

Step 2: We construct the network.

By Lemma 4.7 and Lemma 2.8 the sets Λ_n satisfy $\sup_{\boldsymbol{\nu}\in\Lambda_n}|\boldsymbol{\nu}|_1 \le C(1+\log(n))$. Applying Lemma 3.8 again with the sequence $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}_V^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^2(\mathcal{F})$, there exists a sequence of ReLU NNs $(f_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n}$, such that $f_{\boldsymbol{\nu}}$ only depends on $(y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}$, and such that

$$\sup_{\boldsymbol{y}\in U} |f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| \le 2. \tag{4.22}$$

Moreover, by Lemma 3.8, for some constant $C>0$ that is independent of n there holds

$$\sup_{\boldsymbol{y}\in U}\sum_{\boldsymbol{\nu}\in\Lambda_n}\|t_{\boldsymbol{\nu}}\|_V|\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| = \sup_{\boldsymbol{y}\in U}\sum_{\boldsymbol{\nu}\in\Lambda_n} b_{\boldsymbol{\nu}}|\boldsymbol{y}^{\boldsymbol{\nu}} - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| \le Cn^{-1/p_V+1}. \tag{4.23}$$

The size and depth of the NNs $(f_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ are bounded according to (3.12). Due to the continuous embedding $V=H_0^1(D)\hookrightarrow L^\infty(D)$, we have

$$\begin{aligned} M := \max\left\{2, \sup_{\boldsymbol{\nu}\in\mathcal{F}}\|t_{\boldsymbol{\nu}}\|_{L^\infty(D)}\right\} &\le C\max\left\{2,\sup_{\boldsymbol{\nu}\in\mathcal{F}}\|t_{\boldsymbol{\nu}}\|_V\right\} \\ &\le C\max\left\{2, \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}}(\boldsymbol{\beta}_V^{-\boldsymbol{\nu}}\|t_{\boldsymbol{\nu}}\|_V)^2\right)^{1/2}\right\} < \infty, \end{aligned} \tag{4.24}$$

where we used $\boldsymbol{\beta}^{-\boldsymbol{\nu}} > 1$ for all $\boldsymbol{\nu}\in\mathcal{F}$ since $\boldsymbol{\beta}\in(0,1)^{\mathbb{N}}$. Here the factor 2 in the definition of M ensures by (4.22) that $|f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})| \le M$ for all $\boldsymbol{\nu}\in\mathcal{F}$ and all $\boldsymbol{y}\in U$. Denote by $(a,b)\mapsto\tilde{\times}(a,b)$ the approximate multiplication from Prop. 3.1, with M in (4.24) and with accuracy

$$\delta_n := \mathcal{N}_n^{-r-1}. \tag{4.25}$$

Define $\tilde{u}_n$ as

$$\tilde{u}_n(y_1,\dots,y_n,x) := \sum_{\boldsymbol{\nu}\in\Lambda_n}\tilde{\times}\big(f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}),(I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}})(x)\big), \quad x\in D,\ \boldsymbol{y}\in U\ . \tag{4.26}$$

By Lemma 4.5, each $I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}}$ is a NN with one hidden layer and of size $O(m_{n;\boldsymbol{\nu}})$. Hence, $\tilde{u}_n$ in (4.26) is a NN with $n+1$ input units.

Step 3: We now estimate the approximation error of the network. Since $I_{h_{n;\boldsymbol{\nu}}}$ is a nodal interpolant, it holds (with the constant M as in (4.24)) $\|I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}}\|_{L^\infty(D)} \le \|t_{\boldsymbol{\nu}}\|_{L^\infty(D)} \le M$. Therefore, with $\tilde{\times}$ as in Prop. 3.1, the choice of M (4.24) implies that the error bound (3.6) holds, i.e. with δ_n in (4.25)

$$|\tilde{\times}(f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}),(I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}})(x)) - f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})(I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}})(x)| \le \delta_n, \quad x\in D,\ \boldsymbol{y}\in U\ . \tag{4.27}$$

Fix $\boldsymbol{y}\in U$ and $\boldsymbol{\nu}\in\mathcal{F}$. To bound the error in the norm of V it suffices to bound the $H^1(D)$-seminorm of it. To this end, we claim that for almost every $x\in D$ there holds

$$\left|\frac{d}{dx}\big(\tilde{\times}(f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}),(I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}})(x))\big) - \frac{d}{dx}f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})(I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}})(x)\right| \le \delta_n\left|\frac{d}{dx}(I_{h_{n;\boldsymbol{\nu}}}t_{\boldsymbol{\nu}})(x)\right|\ . \tag{4.28}$$

To prove this, we observe that by Prop. 3.1, there exists a finite set $\mathcal{N}_{\boldsymbol{y},\boldsymbol{\nu}}\subset[-M,M]$ such that for all $b\in[-M,M]\backslash\mathcal{N}_{\boldsymbol{y},\boldsymbol{\nu}}$ there exists the strong derivative

$$\frac{d}{db}\tilde{\times}\Big(f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}),b\Big)$$

of $(a,b) \mapsto \tilde{\times}(a,b)$ w.r.t. the second argument. Thus, by Prop. 3.1, for every $x \in D$ such that $I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}(x) \notin \mathcal{N}_{\boldsymbol{y},\boldsymbol{\nu}}$, the left-hand side of (4.28) is bounded by

$$\left| \tfrac{d}{db} \tilde{\times}\Big(f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}), (I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) \Big) \tfrac{d}{dx}(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) - f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}) \tfrac{d}{dx}(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) \right|$$
$$\leq \delta_n \left| \tfrac{d}{dx}(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) \right| .$$

Here, the derivative of the continuous, piecewise linear function $x \mapsto I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}(x)$ is understood in the weak sense. On the other hand define $\mathcal{S}_{\boldsymbol{y},\boldsymbol{\nu}} := \{x \in D : I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}(x) \in \mathcal{N}_{\boldsymbol{y},\boldsymbol{\nu}}\}$. Since $I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}$ is piecewise linear and because $\mathcal{N}_{\boldsymbol{y},\boldsymbol{\nu}}$ is finite, $\mathcal{S}_{\boldsymbol{y},\boldsymbol{\nu}}$ is the union of a finite number of distinct points in D and possibly a finite number of intervals $[x_j, x_{j+1}]$ such that $I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}|_{[x_j,x_{j+1}]}$ takes a constant value in $\mathcal{N}_{\boldsymbol{y},\boldsymbol{\nu}}$. If such an interval exists, for all $\boldsymbol{y} \in U$ and for all $x \in (x_j, x_{j+1})$

$$\frac{d}{dx} \tilde{\times}\Big(f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}), (I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) \Big) = \frac{d}{dx} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}})(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) = 0$$

so that for every $x \in (x_j, x_{j+1})$

$$\left| \frac{d}{dx}\Big(\tilde{\times}(f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}), (I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x)) \Big) - \frac{d}{dx} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}})(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) \right| \leq \delta_n \left| \frac{d}{dx}(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(x) \right|.$$

In all, for every $\boldsymbol{y} \in U$ and for every $\boldsymbol{\nu} \in \mathcal{F}$ there exists a finite set of points $\mathcal{P}_{\boldsymbol{y},\boldsymbol{\nu}}$, such that (4.28) holds in the classical sense for all $x \notin \mathcal{P}_{\boldsymbol{y},\boldsymbol{\nu}}$. Since $\mathcal{F}$ is countable, the set $\mathcal{N}_{\boldsymbol{y}} = \bigcup_{\boldsymbol{\nu}\in\mathcal{F}} \mathcal{P}_{\boldsymbol{y},\boldsymbol{\nu}} \subset D$ has Lebesgue measure zero. Hence, for every fixed $\boldsymbol{y} \in U$, (4.28) holds for all $\boldsymbol{\nu} \in \mathcal{F}$ almost everywhere in D. The NN's expression error is then bounded by

$$\begin{aligned}
\sup_{\boldsymbol{y}\in U} \|u(\boldsymbol{y},\cdot) - \tilde{u}_n(y_1,\dots,y_n,\cdot)\|_V &\leq \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{y}^{\boldsymbol{\nu}} t_{\boldsymbol{\nu}}(\cdot) - \sum_{\boldsymbol{\nu}\in\Lambda_n} \boldsymbol{y}^{\boldsymbol{\nu}} t_{\boldsymbol{\nu}}(\cdot) \right\|_V \\
&+ \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\Lambda_n} \boldsymbol{y}^{\boldsymbol{\nu}} t_{\boldsymbol{\nu}}(\cdot) - \sum_{\boldsymbol{\nu}\in\Lambda_n} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}) t_{\boldsymbol{\nu}}(\cdot) \right\|_V \\
&+ \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\Lambda_n} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}) t_{\boldsymbol{\nu}}(\cdot) - \sum_{\boldsymbol{\nu}\in\Lambda_n} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}})(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(\cdot) \right\|_V \\
&+ \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\Lambda_n} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}})(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(\cdot) - \sum_{\boldsymbol{\nu}\in\Lambda_n} \tilde{\times}\big(f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}), (I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})(\cdot)\big) \right\|_V \\
&\leq C \Big(\mathcal{N}_n^{-r} + n^{-1/p_V+1} + \mathcal{N}_n^{-r} + |\Lambda_n|\delta_n \Big).
\end{aligned} \tag{4.29}$$

Here the first term was estimated by (4.20), the second by (4.23), the third by (4.21) and (4.22) and finally the fourth via (4.27) by

$$\begin{aligned}
&\sup_{\boldsymbol{y}\in U} \sum_{\boldsymbol{\nu}\in\Lambda_n} \Bigg(\|\tilde{\times}(f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}), I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}) - f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}) I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}\|_{L^2(D)} \\
&\qquad + \left\| \frac{d}{dx} \tilde{\times}(f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}), I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}) - \frac{d}{dx} f_{\boldsymbol{\nu}}((y_j)_{j\in \operatorname{supp}\boldsymbol{\nu}}) I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}} \right\|_{L^2(D)} \Bigg) \\
&\quad \leq (1 + \sup_{\boldsymbol{\nu}\in\mathcal{F}} \|I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}\|_V) |\Lambda_n| \delta_n \leq C |\Lambda_n| \delta_n.
\end{aligned}$$

Using Lemma 4.7 we have $|\Lambda_n| = n \leq \mathcal{N}_n$, and with (4.20) as well as $\delta_n = \mathcal{N}_n^{-r-1}$ we may further estimate (4.29) by $C(\mathcal{N}_n^{-r} + \mathcal{N}_n \mathcal{N}_n^{-r-1}) \leq C\mathcal{N}_n^{-r}$.

Step 4: It remains to bound the size and depth of the NN (4.26). First, by Prop. 3.1 and because $\delta_n = \mathcal{N}_n^{-r-1}$, each of the $|\Lambda_n| = n$ applications of $\tilde{\times}$ in (4.26) requires $C(1+\log(1/\delta_n)) \leq C(1+\log(\mathcal{N}_n))$ ReLUs and a depth of $C(1+\log(\mathcal{N}_n))$. By (3.12), we find the bounds

$$\max_{\boldsymbol{\nu}\in\Lambda_n} \mathrm{depth} f_{\boldsymbol{\nu}} \leq C(1+\log(n)\log\log(n)) \quad \text{and} \quad \sum_{\boldsymbol{\nu}\in\Lambda_n} \mathrm{size}(f_{\boldsymbol{\nu}}) \leq C(1+n\log(n)\log\log(n)).$$

Finally, expressing (exactly, with ReLU activations) the continuous, piecewise linear interpolant $I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}}$ requires a NN of depth one with $m_{n;\boldsymbol{\nu}}$ ReLUs by Lemma 4.5. Except for summing over all $\boldsymbol{\nu} \in \Lambda_n$ in (4.26) in the output layer, there are no connections between the subnetworks $\tilde{\times}(f_{\boldsymbol{\nu}}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}}), I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})$. By definition of $\mathcal{N}_n$ in Lemma 4.7, $\mathcal{N}_n = \sum_{\boldsymbol{\nu}\in\Lambda_n} m_{n;\boldsymbol{\nu}}$. Hence, $\tilde{u}_n$ is a DNN of total size

$$C\sum_{\boldsymbol{\nu}\in\Lambda_n}\Big(1+\underbrace{\mathrm{size}(I_{h_{n;\boldsymbol{\nu}}} t_{\boldsymbol{\nu}})}_{\leq C m_{n;\boldsymbol{\nu}}}+\mathrm{size}(f_{\boldsymbol{\nu}})+\underbrace{\mathrm{size}(\tilde{\times})}_{\leq C(1+\log(\mathcal{N}_n))}\Big) \leq C\Big(n+\mathcal{N}_n+n\log(n)\log\log(n)+n\log(\mathcal{N}_n)\Big)$$

and the NN $\tilde{u}_n$ has total depth $C(1+\log(\mathcal{N}_n)+\log(n)\log\log(n))$. We conclude the proof by recalling that $n \leq \mathcal{N}_n$ according to Lemma 4.7. □

5 Conclusions and further directions

We have established bounds on the rate of expression by a class of certain DNNs for many-variate, real-valued functions f which depend holomorphically on a sequence $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}}$ of (possibly infinitely many) parameters. Specifically, we considered functions of countably many parameters y_j which are $(\boldsymbol{b},\varepsilon)$-holomorphic for $\boldsymbol{b} \in \ell^p$ and some $p \in (0,1)$. This implies that they admit Taylor gpc expansions (2.7) that are sparse in the sense that the sequence $(t_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ of Taylor gpc coefficients is p-summable. The relevance of such functions stems from the fact that they arise as response surfaces of operator equations with distributed, uncertain input data in function spaces (see [13, 9] and the references there). Our main results, Theorems 3.9 and 3.10, imply that such real-valued response surfaces can be expressed with arbitrary prescribed accuracy $\delta > 0$ (uniform w.r. to the parameter vector $\boldsymbol{y}$) by DNNs 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 (3.16). We thus prove expression rates for deep ReLU NNs which are essentially equal to the gpc n-term approximation rates obtained in [13, 9]. In the case of one spacial dimension, we have shown for a model problem that the parametric solution, taking values in the Banach space $H_0^1([0,1])$, can be expressed by deep ReLU NNs. The expressive power bounds in terms of the NN size are essentially the same as known convergence rates for approximation by multilevel stochastic collocation. Since the error bound (4.11) is also valid in $L^t([0,1])$ for $t \in (1,\infty)\backslash\{2\}$, results similar to those of Thm. 4.8 hold also in the Banach spaces $V = W_0^{1,t}([0,1])$ and $X = V \cap W^{2,t}([0,1])$. 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.

We address further possible directions and applications of the present results. In the *Bayesian Inversion* of many-parametric PDE models in the presence of noisy data (see, e.g., [34] 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 [34] are satisfied. We refer to [33] 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 as has recently been proposed in a number of applications; we refer to [35] and the references there. Let us also 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]. 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 although the present results quantifying the "expressive power" of DNNs are approximation results, the proof of Theorem 3.9 is constructive. In principle, when combined with results from [38] on the localization of the sets Λ_n of active Taylor gpc coefficients of $(\boldsymbol{b}, \varepsilon)$-holomorphic maps, 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 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 mathematically developed elsewhere.

References

[1] M. Bachmayr, A. Cohen, D. Dũng, and C. Schwab. Fully discrete approximation of parametric and stochastic elliptic PDEs. *SIAM J. Numer. Anal.*, 55(5):2151–2186, 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] C. Băcuţă, H. Li, and V. Nistor. Differential operators on domains with conical points: precise uniform regularity estimates. *Rev. Roumaine Math. Pures Appl.*, 62(3):383–411, 2017.

[9] 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.

[10] 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.

[11] A. Cohen and R. DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015.

[12] 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.

[13] 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.

[14] 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)).

[15] 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.

[16] 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.

[17] 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.

[18] C. J. Gittelson. Convergence rates of multilevel and sparse tensor approximations for a random elliptic PDE. *SIAM J. Numer. Anal.*, 51(4):2426–2447, 2013.

[19] 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.

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

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

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

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

[24] L. Liang, F. Kong, C. Martin, T. Pham, Q. Wang, J. Duncan, and W. Sun. Machine learning-based 3-D geometry reconstruction and modeling of aortic valve deformation using 3-D computed tomography images. *Int. J. Numer. Methods Biomed. Eng.*, 33(5):e2827, 13, 2017.

[25] L. Liang, M. Liu, C. Martin, and W. Sun. A deep learning approach to estimate stress distribution: a fast and accurate surrogate of finite-element analysis. *J. R. Soc. Interface*, 15:20170844, 2018.

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

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

[28] G. A. Muñoz, Y. Sarantopoulos, and A. Tonge. Complexifications of real Banach spaces, polynomials and multilinear maps. *Studia Math.*, 134(1):1–33, 1999.

[29] A. Pinkus. Approximation theory of the MLP model in neural networks. In *Acta numerica, 1999*, volume 8 of *Acta Numer.*, pages 143–195. Cambridge Univ. Press, Cambridge, 1999.

[30] H. Poon and P. Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pages 689–690, Nov 2011.

[31] 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.

[32] D. Rolnik and M. Tegmark. The power of deeper networks for expressing natural functions. Technical Report 1705.05502v1, ArXiv, 2017.

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

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

[35] R. Tripathy and I. Bilionis. Deep UQ: Learning deep neural network surrogate models for high dimensional uncertainty quantification. Technical Report 1802.00850, ArXiv, 2018.

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

[37] J. Zech, D. Dung, and C. Schwab. Multilevel approximation of parametric and stochastic pdes. Technical Report 2018-05, Seminar for Applied Mathematics, ETH Zürich, 2018. (in review).

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