ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Seminar for Applied Mathematics SAM

Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion (extended version)

J. A. A. Opschoor and Ch. Schwab and J. Zech

Research Report No. 2020-47

July 2020
Latest revision: November 2022

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

Funding SNF: 18453; 159940

Joost A. A. Opschoor*, Christoph Schwab, and Jakob Zech

Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion (extended version)

Abstract: We establish dimension independent expression rates by deep ReLU networks for certain countably-parametric maps, so-called $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic functions. These are mappings from $[-1,1]^{\mathbb{N}} \to \mathcal{X}$, with $\mathcal{X}$ being a Banach space, that admit analytic extensions to certain polyellipses in each of the input variables. Parametric maps of this type occur in uncertainty quantification for partial differential equations with uncertain inputs from function spaces, upon the introduction of bases. For such maps, we prove (constructive) expression rate bounds by families of deep neural networks, based on multilevel polynomial chaos expansions. We show that $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy implies summability and sparsity of coefficients in generalized polynomial chaos expansions. This, in turn, implies deep neural network expression rate bounds.

We apply the results to Bayesian inverse problems for partial differential equations with distributed, uncertain inputs from Banach spaces. Our results imply the existence of "neural Bayesian posteriors" emulating the posterior densities with expression rate bounds that are free from the curse of dimensionality, and limited only by sparsity of certain gpc expansions. We prove the neural Bayesian posteriors robust in large data or small noise asymptotics (e.g. [44]) which can be emulated in a noise-robust fashion.

Keywords: Bayesian inverse problems, generalized polynomial chaos, deep networks, uncertainty quantification

MSC 2020: 62F15, 65N21, 62M45, 68Q32, 41A25

***Corresponding author: Joost A. A. Opschoor, Christoph Schwab,** ETH Zürich, SAM, ETH Zentrum, HG G57.1, CH8092 Zürich, Switzerland, e-mail: joost.opschoor@sam.math.ethz.ch, christoph.schwab@sam.math.ethz.ch

Jakob Zech, University of Heidelberg, 69120 Heidelberg, Germany, e-mail: jakob.zech@uni-heidelberg.de

1 Introduction

The efficient numerical approximation of solution (manifolds) to parameter dependent partial differential equations (PDEs) has seen significant progress in recent years. We refer for instance to [13, 14]. Similarly, and closely related, the treatment of Bayesian inverse problems for well-posed partial (integro-)differential equations with uncertain input data has drawn considerable attention, see e.g. [21] and the references there. This is, in part, due to the need to efficiently assimilate noisy observation data into predictions subject to constraints given by certain physical laws governing responses of systems of interest. We mention here the surveys [21, 71] and the references there. In the present paper, we study mathematically the ability of deep neural networks to express Bayesian posterior probability measures, subject to given data and to PDE constraints. To this end, we work in an abstract setting accommodating PDE constrained Bayesian inverse problems with function space priors as exposed, e.g., in [21, 41] and in the references there.

Several concrete constructions of function space prior probability measures for Bayesian PDE inversion beyond Gaussian measures on separable Hilbert spaces have been advocated in recent years. We mention in particular so-called *Besov prior measures* [46, 20].

Recently, several proposals have been put forward advocating the use of DNNs for Bayesian PDE inversion from noisy data; we refer to [9, 82, 42]. These references computationally found good numerical efficiency for DNN expression with various architectures of DNNs. Regarding Deep NNs for "learning" solution maps of PDEs, we mention [82, 78]. Expressive power (approximation) rate bounds for solution manifolds of PDEs were obtained in [77]; results in this reference are also key in the present analysis of DNN expression of Bayesian posteriors. Specifically, we quantify uncertainty in PDE inversion conditional on noisy observation data using the Bayesian framework. Particular attention is on general, convex priors on uncertain function space inputs [21, 41].

The Bayesian approach can incorporate most, if not all, uncertainties of engineering interest in PDE inversion and in graph-based data classification in a systematic manner.

Computational UQ for PDEs poses three challenges: large-scale forward problems need to be solved, high dimensional parameter spaces arise in parametrization of distributed uncertain inputs (from Banach spaces), and numerical approximation needs to scale favorably in the presence of "big data", resulting in consistent posteriors in the sense of Diaconis and Freedman [24].

Foundational mathematical developments on the question of universality of NNs are e.g. in [28, 40, 39, 7, 8]. In recent years so-called *deep neural networks*

(DNNs for short) have seen rapid development and successful deployment in a wide range of applications. Evidence for the benefit afforded by depth of NNs on their expressive power has been documented computationally in an increasing number of applications (see, e.g. [49, 50, 82, 70, 87, 42, 65] and the references there). The results reported in these references are mostly computational, and address particular applications. Independent of these numerical experiments exploring the performance of DNN based algorithms, the *approximation theory of DNNs* has also advanced in recent years. Distinct from earlier, universality results e.g. in [40, 39, 7, 8], emphasis in more recent mathematical developments has been on approximation (i.e., "expression") rate bounds for specific function classes and particular DNN architectures. We mention only [10, 63, 67] and the references there. In [77], we proved that ReLU DNNs can express high-dimensional, parametric solution families of elliptic PDEs, at rates which are free from the curse of dimensionality.

Specifically, we adopt the infinite-dimensional formulation of Bayesian inverse problems from [79] and its extensions to general, convex prior measures on input function spaces as presented in [41]. Assuming an affine representation system on the uncertain input data, we adopt uniform prior measures on the parameters in the representation.

We prove that ReLU DNNs allow for expressing the parameter-to-response map and the Bayesian posterior density at rates which are determined only by the size of the domains of holomorphy.

1.1 Recent mathematical results on expressive power of DNNs

Fundamental universality results (amounting to, essentially, statements on density of shallow NN expressions) on DNN expression in the class of continuous functions have been established in the 90ies (see [68] for proof and a review of results), in recent years *expression rate bounds* for approximation by DNNs for specific classes of functions have been in the focus of interest. We mention in particular [31] and [10]. There, it is shown that deep NNs with a particular architecture allow for approximation rate bounds analogous to those of rather general multiresolution systems when measured in terms of the number N of units in the DNN.

In [18], *convolutional DNNs* were shown capable of expressing multivariate functions given in so-called *Hierarchical Tensor formats*, a numerical representation inspired by electron structure calculations in computational quantum chemistry.

In [83, 51], ReLU DNNs were shown to be able to express general uni- and multivariate polynomials on bounded domains with uniform accuracy $\delta > 0$, with complexity (i.e., with the number of NN layers and the number of NN units and

nonzero weights) scaling polylogarithmically with respect to δ. The results in [83, 51] allow transferring approximation results from high order finite and spectral element methods, in particular exponential convergence results, to certain types of DNNs.

In [72], DNN expression rates for multivariate polynomials were investigated, without reference to function spaces. Expression rate bounds explicit in the number of variables and the polynomial degree by deep NNs were obtained. The proofs in [72] depend strongly on a large number of bounded derivatives of the activation function, and do not cover the presently considered case of ReLU DNNs.

In [77] we proved dimension-independent DNN expression rate bounds on functions of countably many variables. In [77] we used, as we do in part of the present paper, approximation rate bounds for N-term truncated, so-called *generalized polynomial chaos expansions* of the parametric function. These have been investigated thoroughly in recent years (e.g. [15, 16, 4, 3] and the references there). For the present analysis, however, we require more specific information of polynomial degree distributions in N-term approximate gpc expansions as the dimension of the space of active parameters increases. This was investigated by some of the authors recently in [86, 85]. In the present article, we shall also draw upon results in these references.

In [56], the authors provided an analysis of expressive power of DNNs for a specific class of multi-parametric maps which have a defined (assumed known) *compositional structure*: they are obtained as (repeated) composition of a possibly large number of simpler functions, depending only on a few variables at a time. It was shown that such functions 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 Contributions

We extend our previous work [77] on ReLU NN expression bounds of countably-parametric solution families and QoI's for PDEs with affine-parametric uncertain input. In a first main result, Theorem 4.9 of Section 4.4, we prove bounds on the expressive power of ReLU DNNs for many-parametric response functions from Bayesian inverse UQ for PDEs and more general operator equations subject to infinitely-parametric, uncertain and "invisible" (i.e. not directly observable) input data. As in [77], we assume that the input-to-solution map has holomorphic dependence on possibly an infinite number of parameters. We have in mind in particular (boundary, eigenvalue, control,...) problems for elliptic or parabolic PDEs

with uncertain coefficients. These may stem from, for example, domains of definition with uncertain geometry (see, e.g., [69, 43, 17, 49]) in diffusion, incompressible flow, or time-harmonic, electromagnetic scattering (see, e.g., [43]). Adopting a countable representation system renders uncertain inputs countably-parametric, and implies likewise countably-parametric output families ("solution-manifolds", "response-surfaces") of the model under consideration.

In [77], expressive power estimates for deep ReLU NNs for countably-parametric solution manifolds were obtained among others for linear, second order elliptic PDEs with uncertain coefficients, in divergence form. Theorems 4.9 and 5.2 extend [77] in two regards. Firstly, we require merely $(\boldsymbol{b}, \varepsilon)$-holomorphy on poly-ellipses, rather than on polydiscs as assumed in [77]. This requires essential modifications of the DNN expression rate analysis in [77], as Legendre polynomial chaos expansions are used rather than Taylor expansions. Secondly, we generalize our result from [77] to parametric PDEs posed on a polytopal physical domain D of dimension $d \geq 2$ (instead of $d = 1$).

In the *Bayesian setting* (see [79, 21, 41] and the references there), it has been shown in [25, 75] that $(\boldsymbol{b}, \varepsilon)$-holomorphy of the QoI is inherited by the Bayesian posterior density, if it exists. In the present paper we analyze expression rates of ReLU DNNs for countably parametric Bayesian posterior densities which arise from PDE inversion subject to noisy data. We show, in particular, extending our analysis [77], that ReLU DNNs afford expression of such densities at dimension-independent rates. The expression rate bounds are, to a large extent, abstracted from particular model PDEs and apply to a wide class of PDEs and inverse problems (e.g., elliptic and parabolic linear PDEs with uncertain coefficients, domains, source terms). We also provide in Section 6.3 novel bounds on the posterior consistency of the DNN emulated Bayesian posterior in the presently considered, general setting.

We refer to [82] for a possible computational approach and detailed numerical experiments, for a 2nd order divergence form PDE with log-Gaussian diffusion coefficient.

1.3 Notation

We adopt standard notation, consistent with our previous works [85, 86]: $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$. We write $\mathbb{R}_+ := \{x \in \mathbb{R} : x \geq 0\}$. The symbol C will stand for a generic, positive constant independent of any asymptotic quantities in an estimate, and may change its value even within the same equation.

In statements about (generalized) polynomial chaos expansions we require multiindices $\boldsymbol{\nu} = (\nu_j)_{j \in \mathbb{N}} \in \mathbb{N}_0^{\mathbb{N}}$. The *total 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

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

Here, $\operatorname{supp}\boldsymbol{\nu} = \{j \in \mathbb{N} : \nu_j \neq 0\}$ denotes the *support* of the multiindex $\boldsymbol{\nu}$. The size of the support of $\boldsymbol{\nu} \in \mathcal{F}$ is $|\boldsymbol{\nu}|_0 = \#(\operatorname{supp}\boldsymbol{\nu})$; it will, subsequently, indicate the number of active coordinates in the multivariate monomial term $\boldsymbol{y}^{\boldsymbol{\nu}} := \prod_{j\in\mathbb{N}} y_j^{\nu_j}$.

A subset $\Lambda \subseteq \mathcal{F}$ is called *downward closed*[1], if $\boldsymbol{\nu} = (\nu_j)_{j\in\mathbb{N}} \in \Lambda$ implies $\boldsymbol{\mu} = (\mu_j)_{j\in\mathbb{N}} \in \Lambda$ 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 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 (throughout, $0! := 1$ and $0^0 := 1$, so that $\boldsymbol{\nu}!$ contains finitely many nontrivial factors). For any index set $\Lambda \subset \mathcal{F}$ we denote $\mathbb{P}_\Lambda := \operatorname{span}\{\boldsymbol{y}^{\boldsymbol{\nu}}\}_{\boldsymbol{\nu}\in\Lambda}$.

For a Banach space X we denote by $\mathcal{P}(X)$ the space of Borel probability measures on X and by $d_H(\cdot,\cdot)$ the Hellinger metric on $\mathcal{P}(X)$.

1.4 Structure of the present paper

The structure of this paper is as follows: in Section 2, we review the mathematical setting of Bayesian inverse problems for PDEs, including results which account for the impact of the PDE discretization error on the Bayesian posterior. In Section 3, we recall the notion of $(\boldsymbol{b},\varepsilon)$-holomorphic functions on polyellipses, taking values in Banach spaces and review approximation rate bounds for their truncated gpc expansion.

Sections 4-5 contain the mathematical core and main technical contributions of this paper: we define the DNN architectures and present, after recapitulating the basic operations of DNN calculus, expression rate bounds for so-called $(\boldsymbol{b},\varepsilon,\mathbb{R})$-holomorphic functions. This function class consists of maps from $[-1,1]^{\mathbb{N}} \to \mathbb{R}$,

1 Index sets with the "downward closed" property are also referred to in the literature [59] as *lower sets*.

which allow holomorphic extensions (in each variable) to certain subsets of $\mathbb{C}^{\mathbb{N}}$. This is subsequently generalized to $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic functions. To keep the network size possibly small, we employ a multilevel strategy by combining approximations to elements in $\mathcal{X}$ at different accuracy levels. Section 5.2 presents an illustrative example of a PDE with uncertain input data which satisfy the preceding, abstract hypotheses. Following this, we apply our result for $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic functions to the Bayesian posterior density in Section 6. We show, in particular, that ReLU DNNs are able to express the posterior density with rates (in terms of the size of the DNN) which are free from the curse of dimensionality. We also show in Section 6.2 that DNNs allow for expression rates which are robust w.r. to certain types of posterior concentration in the small noise respectively the large data limits. Section 6.3 shows that the L^∞-convergence of approximations of the posterior density implies convergence of the approximate posterior measure in the Hellinger and total variation distances. In Section 7 we give conclusions and indicate further directions. In the appendix we provide proofs of several results from the main text, which are not included in the published version [64] of this text.

2 Bayesian inverse UQ

We first present the abstract setting of BIP on function spaces, [79, 25, 75]. We then verify the abstract hypotheses in several examples; in particular, for diffusion equations with uncertain coefficients in polygons.

2.1 Forward model

We consider abstract parametric operator equations, which are possibly nonlinear, whose operators depend on uncertain input data a.

We consider given an uncertain input datum $a \in \tilde{X} \subset X$, where X denotes a Banach space containing the set $\tilde{X}$ of admissible input data of the operator equation. Generally, a is not accessible a priori and, therefore, is considered as uncertain input data. A priori knowledge about the distribution of $a \in X$ for a particular application is encoded through a probability measure μ_0 on X, the Bayesian prior, which is supported on a measurable subset $\tilde{X} \subset X$ of admissible uncertain inputs. This implies, in particular, that $\tilde{X} \in \mathcal{B}(X)$ is μ_0-measurable, and that $\mu_0(\tilde{X}) = 1$; we discuss this in detail in Section 2.2 ahead.

The *abstract forward model* to be considered in the sequel reads: given (a realization of) the uncertain input parameter $a \in \tilde{X}$, and a possibly nonlinear map

$\mathcal{N}(a,\cdot) : \mathcal{X} \to \mathcal{Y}'$,

$$\text{find } u \in \mathcal{X}: \quad \langle \mathcal{N}(a,u), v\rangle = 0 \quad \text{for all} \quad v \in \mathcal{Y}\,. \tag{1}$$

Here, $\langle\cdot,\cdot\rangle$ denotes the $\mathcal{Y}' \times \mathcal{Y}$ duality pairing. Throughout, we admit infinite-dimensional Banach spaces $X, \mathcal{X}, \mathcal{Y}$ (all results apply verbatim for the finite-dimensional settings).

In (1), the nonlinear map $\mathcal{N}(\cdot,\cdot) : X \times \mathcal{X} \to \mathcal{Y}'$ could be thought of as residual map for a PDE with solution space $\mathcal{X}$ and uncertain, distributed input data a from a function space X.

2.2 Bayesian inverse problem

We recapitulate the abstract setting of Bayesian inverse problems (BIPs for short) where the data-to-prediction map is constrained by possibly nonlinear operator equations (1) which are subject to unknown / unobservable input data.

2.2.1 Setup

In the Bayesian inversion of the forward model (1), we in general do not have access to the uncertain input a. Instead, we assume given *noisy observation data* $\delta \in Y$, where Y is a space of observation data. The data $\delta \in Y$ is a response of (1) for some admissible input $a \in \tilde{X}$, which response is corrupted by additive observation noise $\eta \in Y$, i.e.

$$\delta = \mathcal{G}(a) + \eta\,. \tag{2}$$

The *data-to-observation map* $\mathcal{G}(\cdot)$ is composed of the solution operator $G : a \mapsto u$ associated to (1) and a continuous, linear observation map $\mathcal{O} \in \mathcal{L}(\mathcal{X}, Y)$ taking the solution $u(a) \in \mathcal{X}$ with input $a \in X$ to observations $\mathcal{O}(u(a)) \in Y$. Thus $\mathcal{G} : X \to Y : a \mapsto \mathcal{G}(a) := (\mathcal{O} \circ G)(a)$.

We often wish to predict a so-called *quantity of interest (QoI for short)*. In this work, we assume the QoI to be a bounded, linear functional $Q \in \mathcal{L}(\mathcal{X}, Z)$ where Z is a suitable Banach space. In this setup, then, the inverse problem consists in estimating the "most likely" realization of the QoI based on solutions $u = G(a)$ of the forward problem (1), given noisy observation data δ of responses $\mathcal{G}(a)$.

In Bayesian inversion, one assumes given a probability measure μ_0 on the Banach space X of inputs which charges the set $\tilde{X} \subset X$ of admissible inputs and which encodes our prior information about the occurrence of inputs a.

Given a realization of the parameter $a \in \tilde{X}$, and observation data $\delta \in Y$, we denote by $\mu^{\delta|a}$ the probability measure on δ, conditioned on a. Under the

assumption that $\mu^{\delta|a} \ll \mu_{\text{ref}}$ for some reference measure μ_{ref} on Y, and that $\mu^{\delta|a}$ has a density w.r. to μ_{ref} which is positive μ_{ref}-a.e., we may define the *likelihood potential* $\Phi(a;\delta) : X \times Y \to \mathbb{R}$ (the "negative log-likelihood") so that

$$\frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\mu_{\text{ref}}}(\delta) = \exp(-\Phi(a;\delta)), \quad \int_Y \exp(-\Phi(a;\delta))\mathrm{d}\mu_{\text{ref}}(\delta) = 1 \,. \tag{3}$$

Remark 2.1. *Let $Y = \mathbb{R}^K$ and assume that the observation noise $\eta \sim \mathcal{N}(0,\Gamma)$ is additive, centered Gaussian with positive definite covariance matrix $\Gamma \in \mathbb{R}^{K\times K}$.*

Then there exists a measure μ_{ref} on Y, equal to a Γ-dependent constant times the Lebesgue measure on Y, such that

$$\Phi(a;\delta) = \frac{1}{2}\|\Gamma^{-1/2}(\mathcal{G}(a) - \delta)\|_2^2 =: \frac{1}{2}\|\mathcal{G}(a) - \delta\|_\Gamma^2 \,. \tag{4}$$

The potential Φ is an inverse covariance weighted, least squares functional of the response-to-observation misfit for uncertain input parameter $a \in X$ and observation data $\delta \in Y$.

In finite dimensions, Bayes' rule states that the posterior $\mu^{a|\delta}$ (the probability measure of the unknown a conditioned on the data δ) is proportional to the product of the likelihood $\mu^{\delta|a}$ and the prior μ_0. In the present Banach space setting, this *formally* extends to

$$\frac{\mathrm{d}\mu^{a|\delta}}{\mathrm{d}\mu_0}(a) = \frac{1}{Z(\delta)}\exp(-\Phi(a;\delta)), \quad \text{where } Z(\delta) = \int_X \exp(-\Phi(a;\delta))\mathrm{d}\mu_0(a) \,, \tag{5}$$

which can be made rigorous, see the references below. Here $Z(\delta)$ is a normalization constant guaranteeing $\frac{\exp(-\Phi(a;\delta))}{Z(\delta)}$ to be a probability density as a function of $a \in X$ w.r.t. the measure μ_0. In the Bayesian methodology, the posterior probability measure $\mu^{a|\delta}$ is considered an updated version of the prior μ_0 on the uncertain inputs that is informed by the observation data δ. In the following, the posterior probability measure will be denoted by μ^δ.

Remark 2.2. *Note that* (5) *is independent of the choice of reference measure μ_{ref} in* (3)*: Let $\tilde{\mu}_{\text{ref}}$ be another (equivalent) reference measure such that $\frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\tilde{\mu}_{\text{ref}}}(\delta) = \exp(-\tilde{\Phi}(a;\delta))$. Then*

$$\exp(-\Phi(a;\delta)) = \frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\mu_{\text{ref}}}(\delta) = \frac{\mathrm{d}\mu^{\delta|a}}{\mathrm{d}\tilde{\mu}_{\text{ref}}}(\delta)\frac{\mathrm{d}\tilde{\mu}_{\text{ref}}}{\mathrm{d}\mu_{\text{ref}}}(\delta) = \exp(-\tilde{\Phi}(a;\delta))\frac{\mathrm{d}\tilde{\mu}_{\text{ref}}}{\mathrm{d}\mu_{\text{ref}}}(\delta).$$

Hence $\tilde{\Phi}(a;\delta) = \Phi(a;\delta) + \log(c)$ with the a-independent constant $c = \frac{\mathrm{d}\tilde{\mu}_{\text{ref}}}{\mathrm{d}\mu_{\text{ref}}}(\delta)$. The constant c will merely influence the normalization $Z(\delta)$ in (5)*, but either choice of μ_{ref} or $\tilde{\mu}_{\text{ref}}$ leads to the same formula for $a \mapsto \frac{\mathrm{d}\mu^{a|\delta}}{\mathrm{d}\mu_0}(a)$.*

We refer to [79, 21, 41] for a detailed discussion and further references, in particular [21, Section 3.4.2], which is an application of the more general discussion in [21, Section 3.2]. Our definition of the likelihood potential Φ in (4) is consistent with [21, Equation (10.39)]. It is shifted with respect to the definition in [21, Equation (10.30)] by adding to Φ a function depending on δ, but not on a, see Remark 2.2 and also [21, Remark 5].

2.2.2 Assumptions

Based on [79, 20, 21, 41], we now formalize the preceding concepts. To this end, we introduce a set of assumptions on the prior and on the forward map which ensure well-posedness and continuous dependence of the BIP.

Assumption 2.3 ([41, Assumption 2.1])**.** *In the Banach space X of uncertain parameters and the Banach space Y of observation data, the potential $\Phi : X \times Y \to \mathbb{R}$ satisfies:*

(i) (bounded below) There is some $\alpha_1 \geq 0$ such that for every $r > 0$ exists a constant $M(\alpha_1, r) \in \mathbb{R}$ such that for every $u \in X$ and for every data $\delta \in Y$ with $\|\delta\|_Y < r$ holds

$$\Phi(u;\delta) \geq M - \alpha_1 \|u\|_X \ .$$

(ii) (boundedness above) For every $r > 0$ exists $K(r) > 0$ such that for every $u \in X$ and for every $\delta \in Y$ with $\max\{\|u\|_X, \|\delta\|_Y\} < r$ holds

$$\Phi(u;\delta) \leq K \ .$$

(iii) (Lipschitz continuous dependence on u) For every $r > 0$ exists a constant $L(r) > 0$ such that for every $u_1, u_2 \in X$ and for every $\delta \in Y$ with $\max\{\|u_1\|_X, \|u_2\|_X, \|\delta\|_Y\} < r$ holds

$$|\Phi(u_1;\delta) - \Phi(u_2;\delta)| \leq L\|u_1 - u_2\|_X \ .$$

(iv) (Lipschitz continuity w.r. to observation data $\delta \in Y$) For some $\alpha_2 \geq 0$ and for every $r > 0$ exists $C(\alpha_2, r) \in \mathbb{R}$ such that for every $\delta_1, \delta_2 \in Y$ with $\max\{\|\delta_1\|_Y, \|\delta_2\|_Y\} < r$ and for every $u \in X$ holds

$$|\Phi(u;\delta_1) - \Phi(u;\delta_2)| \leq \exp\left(\alpha_2\|u\|_X + C\right) \|\delta_1 - \delta_2\|_Y \ .$$

(v) (Radon prior measure) The prior measure μ_0 is a Radon probability measure charging a measurable subset $\tilde{X} \subseteq X$ with $\tilde{X} \in \mathcal{B}(X)$ of admissible uncertain parameters, i.e. $\mu_0(\tilde{X}) = 1$.

(vi) (exponential tails) The prior measure μ_0 on the Banach space X has exponential tails:

$$\exists \kappa > 0 : \quad \int_X \exp(\kappa \|u\|_X) \mathrm{d}\mu_0(u) < \infty \,. \tag{6}$$

Remark 2.4. *Assumption (v) on the prior μ_0 being a Radon probability measure is always satisfied when X is separable.*

2.2.3 Well-posedness

We shall consider well-posedness of the BIP in the following sense.

Definition 2.5 (Well-posedness of the BIP, [41, Definition 1.4])**.** *For Banach spaces X, Y, with $d_H(\cdot,\cdot)$ denoting the Hellinger metric on the space $\mathcal{P}(X)$ of Borel probability measures on X, for a prior $\mu_0 \in \mathcal{P}(X)$ and for the likelihood potential Φ, the BIP (5) is well-posed if the following holds:*

(i) *(existence and uniqueness) For every data $\delta \in Y$ exists a unique posterior measure $\mu^\delta \in \mathcal{P}(X)$ which is absolutely continuous w.r. to the prior μ_0 and which satisfies (5),*

(ii) *(stability) for every $\varepsilon > 0$ and $r > 0$ there exists a constant $C_\varepsilon(r) > 0$ such that for every $\delta, \delta' \in Y$ with $\max\{\|\delta\|_Y, \|\delta'\|_Y\} < r$ and $\|\delta - \delta'\|_Y \leq C_\varepsilon$, there holds*

$$d_H(\mu^\delta, \mu^{\delta'}) < \varepsilon \,.$$

2.2.4 Existence and continuous dependence

We are now in position to state sufficient conditions for well-posedness of the BIP and for existence and uniqueness of the posterior μ^δ. We work in the abstract setting Assumption 2.3, deferring the verification of the items in Assumption 2.3 to the ensuing discussion of concrete model problems.

Theorem 2.6 ([41, Theorems 2.4 and 2.6])**.** *Given Banach spaces X and Y and a likelihood function $\Phi : X \times Y \to \mathbb{R}$ satisfying Assumption 2.3, items (i), (ii), (iii) with some $\alpha_1 > 0$. Moreover, the prior measure $\mu_0 \in \mathcal{P}(X)$ satisfies Assumption 2.3, items (v) and (vi) with some constant $\kappa > 0$.*

Then it holds:

(i) *If $\kappa \geq \alpha_1$, for every $\delta \in Y$, the posterior measure μ^δ defined in Equation (5) is well-defined, and a Radon probability measure on X.*

(ii) *(Lipschitz continuity of the posterior w.r. to the data) If Φ satisfies in addition Assumption 2.3, item (iv) with some constant $\alpha_2 \geq 0$, and if the constant κ from Assumption 2.3, item (vi), satisfies $\kappa \geq \alpha_1 + 2\alpha_2$, then for every $r > 0$ exists a constant $C(r) > 0$ such that, for every $\delta, \delta' \in Y$ with $\max\{\|\delta\|_Y, \|\delta'\|_Y\} < r$, the posteriors $\mu^\delta, \mu^{\delta'} \in \mathcal{P}(X)$ satisfy*

$$d_H(\mu^\delta, \mu^{\delta'}) \leq C(r)\|\delta - \delta'\|_Y \,. \tag{7}$$

A proof of this result is, for example, in [41, Theorems 2.4 and 2.6].

2.2.5 Consistent approximation

In the numerical approximation of posteriors μ^δ where the input-to-observation map $\mathcal{G} = \mathcal{O} \circ G : X \to Y$ involves a well-posed, parametric forward operator equation (1), we will in general have to resort to approximate, numerical solutions of (1). Generically, we tag such approximate solution maps by a subscript $N \in \mathbb{N}$ which should be understood as the "number of degrees of freedom" involved in the discretization of the parametric equation (1). In this way, we denote the data-to-solution map of the nonlinear equation (1) by $G_N : X \to \mathcal{X}$, the corresponding data-to-observation map by $\mathcal{G}_N = \mathcal{O} \circ G_N$, and the likelihood potential by Φ_N.

Approximation of the forward model (1), e.g. by consistent discretization, leads to an *approximate Bayesian inverse problem*, which is of the form

$$\frac{d\mu_N^\delta}{d\mu_0}(a) = \frac{1}{Z_N(\delta)} \exp(-\Phi_N(a;\delta)), \quad \text{where} \quad Z_N(\delta) := \int_X \exp(-\Phi_N(a;\delta))\mathrm{d}\mu_0(a) \,. \tag{8}$$

Assuming exact observations $\mathcal{O}(\cdot)$ at hand, the approximate potential Φ_N in (8) is

$$\Phi_N(a;\delta) = \frac{1}{2}\|\Gamma^{-1/2}\big((\mathcal{O} \circ G_N)(a) - \delta\big)\|_2^2 \,, \quad a \in \tilde{X}, \delta \in Y \,.$$

The posterior μ^δ would, consequently, also be approximated by the corresponding numerical posterior, which we denote by μ_N^δ.

It is of interest to identify sufficient conditions so that, as $N \to \infty$, the approximate posteriors $\{\mu_N^\delta\}_{N \geq 1}$ tend to the posterior μ^δ in $\mathcal{P}(X)$.

Definition 2.7 (consistent posterior approximation, [41, Definition 1.5])**.** *The approximate Bayesian inverse problem* (8) *is said to be a* consistent approximation of (5) for a prior $\mu_0 \in \mathcal{P}(X)$ and a potential Φ *if the approximate potential Φ_N is such that for every data $\delta \in Y$, as $N \to \infty$, there holds*

$$|\Phi(a;\delta) - \Phi_N(a;\delta)| \to 0 \quad \textit{implies} \quad d_H(\mu^\delta, \mu_N^\delta) \to 0 \,.$$

Apart from consistency in the sense of Definition 2.7, in the numerical approximation of BIPs we are also interested in *convergence rates*: if the numerical approximation G_N of the forward solution map converges with a certain rate, say $\psi(N)$, with ψ a nonnegative function such that $\psi(N) \downarrow 0$ as $N \to \infty$, then the corresponding posteriors μ_N^δ should converge with a rate related to $\psi(N)$. The following theorem, which is proved in [21, Theorem 18], gives sufficient conditions for posterior convergence.

Theorem 2.8 ([21, Theorem 18])**.** *Let Banach spaces X and Y of uncertain parameters a and observation data δ, resp., be given.*

Let $\mu_0 \in \mathcal{P}(X)$ be a Borel probability measure on X which satisfies Assumption 2.3, items (i) - (vi), so that for observation data $\delta \in Y$ the BIPs (5), (8) *for $\mu^\delta, \mu_N^\delta \in \mathcal{P}(X)$ are well-defined.*

Assume also that the likelihood potentials Φ and Φ_N satisfy Assumption 2.3, items (i), (ii) with constant $\alpha_1 \geq 0$ which is uniform w.r. to N, and that for some $\alpha_3 \geq 0$ exists $C(\alpha_3) > 0$ independent of N such that for every $a \in \tilde{X}$ holds

$$|\Phi(a;\delta) - \Phi_N(a;\delta)| \leq C \exp(\alpha_3 \|a\|_X)\psi(N) \tag{9}$$

with $\psi(N) \downarrow 0$ as $N \to \infty$.

If furthermore Assumption 2.3, item (vi) holds with $\kappa \geq \alpha_1 + 2\alpha_3$, then for every $r > 0$ exists a constant $D(r) > 0$ such that for every $\delta \in Y$ with $\|\delta\|_Y < r$ holds

$$\forall N \in \mathbb{N}: \quad d_H(\mu^\delta, \mu_N^\delta) \leq D\psi(N) \,.$$

Here, the constant $D(r)$ generally depends on the covariance Γ of the centered Gaussian observation noise η in (2).

2.3 Prior modeling

The modeling of prior probability measures on function spaces of distributed, uncertain PDE input data a in the model (1) has been developed in several references in recent years. The 'usual construction' is based on (a) coordinate representations of (realizations of) instances of a in terms of a suitable basis $\{\psi_j\}_{j\geq 1}$ (thereby implying a will take values in a separable subset $\tilde{X}$ of X) and on (b) construction of the prior as countable product probability measure of probability measures on the co-ordinate spaces.

This approach, which is inspired by N. Wiener's construction of the Wiener process by placing Gaussian measures on coefficient realizations of Fourier series, has been realized for example in [46, 20, 35] for Besov spaces, and in [41, 80] and the references there for more general priors.

2.4 Examples

The foregoing, abstract setting (1) accommodates a wide range of PDE boundary value, eigenvalue, control, and shape optimization problems with uncertain function space input $a \in X$. We illustrate the scope by listing several examples which are covered by the ensuing, abstract DNN expression rate bounds. In all examples, $\mathrm{D} \subset \mathbb{R}^d$ shall denote an open, bounded and connected, polytopal domain in physical Euclidean space of dimension $d \geq 2$. In dimension $d = 1$, D shall denote an open, bounded interval of positive length.

2.4.1 Diffusion equation

We consider the linear, 2nd order, diffusion equation with uncertain coefficients in $\mathrm{D} \subset \mathbb{R}^2$. Holomorphic dependence of solutions on coefficient data was shown in [5] and the numerical analysis, including Finite-Element discretization in D on corner-refined families of triangulations, with approximation rate estimates for both, the parametric solution and the Karhunen-Loeve expansion terms, was provided in [36]. Given a source term $f \in H^{-1}(\mathrm{D}) = (H^1_0(\mathrm{D}))^*$, and an isotropic diffusion coefficient $a \in \tilde{X} \subset \{a \in L^\infty(\mathrm{D}) : \operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} a(\boldsymbol{x}) > 0\}$ the diffusion problem reads: find $u \in H^1_0(\mathrm{D})$ such that

$$\mathcal{N}(a,u)(\boldsymbol{x}) := f(\boldsymbol{x}) + \nabla \cdot (a(\boldsymbol{x})\nabla u(\boldsymbol{x})) = 0 \ \text{ in } \mathrm{D}, \quad u|_{\partial\mathrm{D}} = 0\,. \tag{10}$$

It falls into the variational setting (1) with $\mathcal{X} = \mathcal{Y} = H^1_0(\mathrm{D})$, $X = L^\infty(\mathrm{D})$. In [5, 36], also anisotropic diffusion coefficients a and advection and reaction terms were admitted.

For $a \in \tilde{X}$, the weak formulation (1) of (10) is uniquely solvable and the data-to-solution map $G : \tilde{X} \to \mathcal{X} : a \mapsto u(a)$ is continuous. Equipping $\mathcal{X}$ with the norm $\|v\|_{\mathcal{X}} = \|\nabla v\|_{L^2(\mathrm{D})}$, there holds

$$\|u\|_{\mathcal{X}} \leq \frac{\|f\|_{H^{-1}(\mathrm{D})}}{\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} a(\boldsymbol{x})}\,.$$

Assuming *affine-parametric* uncertain input [76, 15, 16], i.e., given $a_0 \in X$ with

$$a_- := \operatorname*{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} a_0(\boldsymbol{x}) > 0,$$

for $\{\psi_j\}_{j\geq 1} \subset X$ with $\sum_{j\geq 1} \|\psi_j\|_X < a_-$, we choose the prior such that its support is contained in the set

$$\tilde{X} := \{a \in X : a = a(\boldsymbol{y}) := a_0 + \sum_{j\geq 1} y_j\psi_j,\ \boldsymbol{y} = (y_j)_{j\geq 1} \in [-1,1]^{\mathbb{N}}\}. \tag{11}$$

For every $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$ and $a(\boldsymbol{y}) \in \tilde{X}$, problem (10) admits a unique parametric solution $u(\boldsymbol{y}) \in \mathcal{X}$ such that $\mathcal{N}(a(\boldsymbol{y}), u(\boldsymbol{y})) = 0$ in $H^{-1}(\mathrm{D})$.

2.4.2 Elliptic eigenvalue problem with uncertain coefficient

For $a \in \tilde{X}$ as defined in (11), for every $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$ we seek solutions $(\lambda(\boldsymbol{y}), w(\boldsymbol{y})) \in \mathbb{R} \times H^1_0(\mathrm{D}) \backslash \{0\}$ of the eigenvalue problem

$$\mathcal{N}(a(\boldsymbol{y}), (\lambda(\boldsymbol{y}), w(\boldsymbol{y}))) = 0 \ \text{ in } \ H^{-1}(\mathrm{D}) \,, \tag{12}$$

where, for every $a \in \tilde{X}$, $\mathcal{N}(a, (\lambda, w)) : \mathbb{R} \times H^1_0(\mathrm{D}) \to H^{-1}(\mathrm{D}) : (\lambda, w) \mapsto \lambda w + \nabla \cdot (a \nabla w)$. For every $\boldsymbol{y}$, the EVP (12) admits a sequence $\{(\lambda_k(\boldsymbol{y}), w_k(\boldsymbol{y})) : k = 1, 2, ...\}$ of real eigenvalues $\lambda_k(\boldsymbol{y})$ (which we assume enumerated according to their size, with multiplicity counted) with associated eigenfunctions $w_k(\boldsymbol{y}) \in H^1_0(\mathrm{D})$ (which form a dense set in $H^1_0(\mathrm{D})$). It is known (e.g. [29, Proposition 2.4]) that the first eigenpair $\{(\lambda_1(\boldsymbol{y}), w_1(\boldsymbol{y})) : \boldsymbol{y} \in [-1,1]^{\mathbb{N}}\}$ is isolated, admits a uniform (w.r. to $\boldsymbol{y} \in [-1,1]^{\mathbb{N}}$) spectral gap.

3 Generalized polynomial chaos surrogates

3.1 Uncertainty parametrization

Let $\mathcal{Z}$ and $\mathcal{X}$ be two complex Banach spaces and let $(\psi_j)_{j \in \mathbb{N}}$ be a sequence in $\mathcal{Z}$. Additionally suppose that $O \subseteq \mathcal{Z}$ is open and let $\mathfrak{u} : O \to \mathcal{X}$ be complex differentiable. With the parameter domain $U := [-1,1]^{\mathbb{N}}$ we consider the infinite parametric map

$$u(\boldsymbol{y}) := \mathfrak{u}\left(\sum_{j \in \mathbb{N}} y_j \psi_j\right) \qquad \forall \boldsymbol{y} = (y_j)_{j \in \mathbb{N}} \in U, \tag{13}$$

which is well-defined for instance if $(\|\psi_j\|_{\mathcal{Z}})_{j \in \mathbb{N}} \in \ell^1(\mathbb{N})$. Here the map $U \to O : \boldsymbol{y} \mapsto \sum_{j \in \mathbb{N}} y_j \psi_j$ is understood as an (affine) parametrization of the uncertain input a and $\mathfrak{u}$ denotes the map which relates the input to the solution of the model under consideration.

Under certain assumptions, such maps allow a representation as a sparse *Taylor generalized polynomial chaos expansion* [15, 16], i.e. for $\boldsymbol{y} \in U$

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu} \in \mathcal{F}} t_{\boldsymbol{\nu}} \boldsymbol{y}^{\boldsymbol{\nu}}, \qquad t_{\boldsymbol{\nu}} = \frac{1}{\boldsymbol{\nu}!} \partial^{\boldsymbol{\nu}}_{\boldsymbol{y}} u(\boldsymbol{y}) \mid_{\boldsymbol{y} = \mathbf{0}} \in \mathcal{X}, \tag{14}$$

or as a sparse *Legendre generalized polynomial chaos expansion* [13], i.e.

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} l_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}}(\boldsymbol{y}), \qquad l_{\boldsymbol{\nu}} = \int_U L_{\boldsymbol{\nu}}(\boldsymbol{y}) u(\boldsymbol{y}) \mathrm{d}\mu_U(\boldsymbol{y}) \in \mathcal{X}, \tag{15}$$

where $L_{\boldsymbol{\nu}}(\boldsymbol{y}) = \prod_{j\in\mathbb{N}} L_{\nu_j}(y_j)$ and $L_n : [-1,1] \to \mathbb{R}$ denotes the n-th Legendre polynomial normalized in $L^2([-1,1], \lambda/2)$, where λ denotes the Lebesgue measure on $[-1,1]$, i.e. $\lambda/2$ is a uniform probability measure on $[-1,1]$. Also, $\mu_U := \otimes_{j\in\mathbb{N}} \frac{\lambda}{2}$ denotes the uniform probability measure on $U = [-1,1]^{\mathbb{N}}$ equipped with the product σ-algebra. Then by [61, §18.3]

$$\|L_n\|_{L^\infty([-1,1])} \leq (1+2n)^{\frac{1}{2}} \qquad \forall n \in \mathbb{N}_0. \tag{16}$$

The summability properties of the ($\mathcal{X}$-norms of) Taylor or Legendre gpc coefficients $(\|t_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}$, $(\|l_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}$ are key for assigning a meaning to such formal gpc expansions like (14) and (15). For example, as for every $\boldsymbol{y} \in U$ and for every $\boldsymbol{\nu} \in \mathcal{F}$ it holds that $|\boldsymbol{y}^{\boldsymbol{\nu}}| \leq 1$, the summability $(\|t_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$ guarantees unconditional convergence in $\mathcal{X}$ of the series in (14) for every $\boldsymbol{y} \in U$. As we shall recall in Section 3.3, this summability is in turn ensured by a suitable form of holomorphic continuation of the parameter-to-response map $u : U \to \mathcal{X}$.

Remark 3.1. *We assume here $\mathcal{X}$ to be a complex space. If $\mathcal{X}$ is a Banach space over $\mathbb{R}$, one can consider* $\mathfrak{u}$ *as a map to the complexification $\mathcal{X}_{\mathbb{C}} = \mathcal{X} + \mathrm{i}\mathcal{X}$ of $\mathcal{X}$ equipped with the so-called* Taylor norm $\|v + \mathrm{i}w\|_{\mathcal{X}_{\mathbb{C}}} := \sup_{t\in[0,2\pi)} \|\cos(t)v - \sin(t)w\|_{\mathcal{X}}$ *for all v, $w \in \mathcal{X}$ (cp. [58]). Here,* $\mathrm{i} = \sqrt{-1}$ *with* $\arg(\mathrm{i}) = \pi/2$.

3.2 $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy

To prove expressive power estimates for DNNs, we use parametric holomorphic maps from a compact parameter domain U into a Banach space $\mathcal{X}$ with quantified sizes of domains of holomorphy. To introduce such maps, we recapitulate principal definitions and results from [16, 13, 12, 86] and the references there. The notion of $(\boldsymbol{b}, \varepsilon)$-holomorphy (given in Definition 3.3 ahead), which stipulates holomorphic parameter dependence of a function $u : U \to \mathcal{X}$ in each variable on certain product domains $\mathcal{O} = \bigtimes_{j\in\mathbb{N}} O_j \subseteq \mathbb{C}^{\mathbb{N}}$, has been found to be a sufficient condition on a parametric map $U \ni \boldsymbol{y} \mapsto u(\boldsymbol{y}) \in \mathcal{X}$, in order that u admits *gpc* expansions with p-summable coefficients for some $p \in (0,1)$, see, e.g., [13, 77] and also Section 3.3 ahead. In the following, we extend the results from [77] in the sense that we admit smaller domains of holomorphy: each $O_j = \mathcal{E}_{\rho_j}$ is a *Bernstein-ellipse* defined by

$$\mathcal{E}_\rho := \left\{ \frac{z + z^{-1}}{2} \,:\, z \in \mathbb{C},\ 1 \leq |z| < \rho \right\} \subseteq \mathbb{C},$$

rather than a complex disc $O_j = B_{\rho_j}$ as in [77].

Remark 3.2. *Let $\mathcal{J} \subseteq \mathbb{N}$. Throughout,* continuity of a function defined on a cylindrical set $\bigtimes_{j\in\mathcal{J}} O_j$ *with $O_j \subseteq \mathbb{C}$ for all $j \in \mathcal{J}$ will be understood as continuity with respect to the subspace topology on $\bigtimes_{j\in\mathcal{J}} O_j \subset \bigtimes_{j\in\mathcal{J}} \mathbb{C}$, where $\bigtimes_{j\in\mathcal{J}} \mathbb{C}$ is assumed to be equipped with the product topology by our convention (see Section 1.3). In this topology, the parameter domain $U = [-1,1]^{\mathbb{N}}$ is compact by Tychonoff's theorem [57, Theorem 37.3].*

In the following, if $\boldsymbol{\rho} = (\rho_j)_{j=1}^N \subseteq (1,\infty)$ for some $N \in \mathbb{N}$, we define the poly-ellipse $\mathcal{E}_{\boldsymbol{\rho}} := \bigtimes_{j=1}^N \mathcal{E}_{\rho_j} \subseteq \mathbb{C}^N$, and similarly in case $\boldsymbol{\rho} = (\rho_j)_{j\in\mathbb{N}} \subseteq (1,\infty)$

$$\mathcal{E}_{\boldsymbol{\rho}} := \bigtimes_{j\geq 1} \mathcal{E}_{\rho_j} \subseteq \mathbb{C}^{\mathbb{N}} .$$

Definition 3.3 ($(\boldsymbol{b},\varepsilon,\mathcal{X})$-Holomorphy)**.** *Let $\mathcal{X}$ be a complex 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 $p \in (0,1]$.*

We say that a map $u : U \to \mathcal{X}$ is $(\boldsymbol{b},\varepsilon,\mathcal{X})$-holomorphic if there exists a constant $M < \infty$ such that

(i) *$u : U \to \mathcal{X}$ is continuous,*

(ii) *for every sequence $\boldsymbol{\rho} = (\rho_j)_{j\in\mathbb{N}} \subset (1,\infty)^{\mathbb{N}}$ which is* $(\boldsymbol{b},\varepsilon)$-admissible*, i.e. which satisfies*

$$\sum_{j\in\mathbb{N}} b_j(\rho_j - 1) \leq \varepsilon , \tag{17}$$

u admits a separately holomorphic extension (again denoted by u) onto the poly-ellipse $\mathcal{E}_{\boldsymbol{\rho}}$,

(iii) *for each $(\boldsymbol{b},\varepsilon)$-admissible $\boldsymbol{\rho}$ holds*

$$\sup_{\boldsymbol{z}\in\mathcal{E}_{\boldsymbol{\rho}}} \|u(\boldsymbol{z})\|_{\mathcal{X}} \leq M . \tag{18}$$

If it is clear from the context that $\mathcal{X} = \mathbb{C}$, then we will omit $\mathcal{X}$ in notation.

Remark 3.4. *We note that for $\boldsymbol{b} \in \ell^1(\mathbb{N})$ as in Definition 3.3, $b_j \to 0$ as $j \to \infty$. By (17), $(\boldsymbol{b},\varepsilon)$-admissible polyradii $\boldsymbol{\rho}$ can satisfy $\rho_j \to \infty$, implying that the component sets $\mathcal{E}_{\rho_j}$ will grow as $j \to \infty$. We also observe the following, elementary geometric fact:*

$$\forall \rho > 1 : \quad \mathcal{E}_\rho \supset B_{(\rho - 1/\rho)/2} . \tag{19}$$

In particular, $\mathcal{E}_\rho \supset \overline{B_1} \supset [-1,1]$ for all $\rho > 1 + \sqrt{2}$. Bernstein ellipses $\mathcal{E}_\rho$ are moreover useful if the domain of holomorphy of u does not contain B_1. Moreover,

if $\rho_j \to \infty$, after all but a (possibly small) finite number of parameters, the domains of holomorphy $\mathcal{E}_{\rho_j}$ contain a polydisc with radius $(\rho_j - 1/\rho_j)/2 > 1$. We shall see in Section 4 below that multivariate monomials can be expressed by smaller DNNs than, e.g., multivariate Legendre, or Jacobi polynomials. In particular, for the emulation of tensor products of Taylor monomials the product network is of smaller size than that for the emulation of tensor product Legendre polynomials. The reason is that the L^∞-norm of Taylor monomials equals 1*, whereas for $\boldsymbol{\nu} \in \mathcal{F}$ it holds that $\|L_{\boldsymbol{\nu}}\|_{L^\infty(U)} \leq \prod_{j \in \operatorname{supp} \boldsymbol{\nu}} \sqrt{1+2\nu_j}$ (cf.* (16)*). Due to the growth of this bound, to achieve the same absolute accuracy a larger relative accuracy is required, and therefore a larger product network size (see Proposition 4.3). We therefore use in our expression rate bounds "Taylor DNN emulations" as in [77] for all but a fixed, finite number of dimensions. There, we use an* exponential expression rate bound *from [62] for the ReLU DNN approximation of tensor product Legendre polynomials (Proposition 4.6).*

Definition 3.3 has been similarly stated in [13]. The sequence $\boldsymbol{b}$ in Definition 3.3 quantifies 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. 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 gpc coefficients tend to 0 (see Theorem 3.7 ahead). The notion of $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy applies to large classes of parametric operator equations, notably including functions of the type (13). This statement is given in the next lemma which is proven in [84, Lemma 2.2.7], see also [86, Lemma 3.3] (for a version based on holomorphy on polydiscs rather than on polyellipses).

Lemma 3.5. *Let $\mathfrak{u} : O \to \mathcal{X}$ be holomorphic where $O \subseteq \mathcal{Z}$ is open. Assume that $(\psi_j)_{j\in\mathbb{N}} \subseteq \mathcal{Z}$, $\psi_j \neq 0$ for all j, with $(\|\psi_j\|_{\mathcal{Z}})_{j\in\mathbb{N}} \in \ell^1(\mathbb{N})$ and $\{\sum_{j\in\mathbb{N}} y_j\psi_j : \boldsymbol{y} \in U\} \subseteq O$. Then there exists $\varepsilon > 0$ such that $u(\boldsymbol{y}) = \mathfrak{u}(\sum_{j\in\mathbb{N}} y_j\psi_j)$, $\boldsymbol{y} \in U$ defines a $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic function with $b_j := \|\psi_j\|_{\mathcal{Z}}$.*

3.3 Summability of gpc coefficients

As mentioned above, the relevance of $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy lies in that it guarantees such functions to possess gpc expansions with coefficients whose norms are p-summable for some $p \in (0, 1)$. This p-summability is the crucial property required to establish convergence rates of certain partial sums. Our analysis of the expressive power of DNNs of such parametric solution families will be based on a version of these results as stated in the next theorem. To reduce the asymptotic size of

the networks, we consider gpc expansions combining both multivariate monomials and multivariate Legendre polynomials, as motivated in Remark 3.4. While p-summability of the norms of both the Taylor and the Legendre coefficients of such functions is well-known (under suitable assumptions), Theorem 3.7 below is not available in the literature. For this reason we provide a proof but stress that the general line of arguments closely follows earlier works such as [15, 16, 13, 85].

In the next theorem we distinguish between low- and high-dimensional co-ordinates: We shall use in "low dimensions" indexed by $j \in \{1, ..., J\}$ Legendre expansions, whereas in the co-ordinates indexed by $j > J$ we resort to Taylor gpc expansions. For $1 \leq j \leq J$, we thus exploit holomorphy on poly-ellipses $\mathcal{E}_{\rho_j}$ and Legendre gpc expansions. For $j > J$, we emulate by ReLU DNNs the corresponding Taylor gpc expansions in these co-ordinates using [77] and the fact that sufficiently large Bernstein ellipses with foci ± 1 contain discs with radius > 1 centered at the origin (as pointed out in Remark 3.4).

Accordingly, we introduce the following notation: for some fixed $J \in \mathbb{N}$ (defined in the following) and $\boldsymbol{\nu} \in \mathcal{F}$ set

$$\boldsymbol{\nu}_E := (\nu_1, \dots, \nu_J), \qquad \boldsymbol{\nu}_F := (\nu_{J+1}, \nu_{J+2}, \dots)$$

and $\mathcal{F}_E := \mathbb{N}_0^J$, and we will write $\boldsymbol{\nu} = (\boldsymbol{\nu}_E, \boldsymbol{\nu}_F)$. Moreover $U_E := [-1, 1]^J$ and $U_F := \bigtimes_{j>J}[-1, 1]$, and for $\boldsymbol{y} = (y_j)_{j\in\mathbb{N}} \in U$ define $\boldsymbol{y}_E := (y_j)_{j=1}^J \in U_E$ and $\boldsymbol{y}_F := (y_j)_{j>J} \in U_F$. In particular we will employ the notation $\boldsymbol{y}_F^{\boldsymbol{\nu}_F} = \prod_{j>J} y_j^{\nu_j}$. Additionally, for a function $u : U \to \mathcal{X}$, by $u(\boldsymbol{y}_E, \mathbf{0})$ we mean u evaluated at $(y_1, \dots, y_J, 0, 0, \dots) \in U$. In terms of the Lebesgue measure λ on $[-1, 1]$ define $\mu_E := \otimes_{j=1}^J \frac{\lambda}{2}$ on U_E and $\mu_F := \otimes_{j>J} \frac{\lambda}{2}$ on U_F.

Lemma 3.6. *Let $C_0 := 4/9$. Then $B^{\mathbb{C}}_{C_0\rho} \subseteq \mathcal{E}_\rho$ for all $\rho \geq 3$.*

Proof. By Remark 3.4 it holds $B_{(\rho-\rho^{-1})/2} \subseteq \mathcal{E}_\rho$, so it suffices to check $(\rho-\rho^{-1})/2 \geq C_0\rho$ for all $\rho \geq 3$. For $\rho = 3$ this follows by elementary calculations, and for $\rho > 3$ it follows by the fact that $\rho \mapsto (\rho - \rho^{-1})/(2\rho) = (1 - \rho^{-2})/2$ is monotonically increasing for $\rho \geq 3$. □

Theorem 3.7. *Let u be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0, 1)$ and $\varepsilon > 0$. Then there exists $J \in \mathbb{N}$ such that*

(i) *for each $\boldsymbol{\nu} \in \mathcal{F}$*

$$c_{\boldsymbol{\nu}} := \int_{U_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E, \mathbf{0})}{\boldsymbol{\nu}_F!} \mathrm{d}\mu_E(\boldsymbol{y}_E) \in \mathcal{X} \tag{20}$$

is well-defined and it holds

$$(\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F}),$$

(ii) *it holds*

$$u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} \in \mathcal{X},$$

with absolute and uniform convergence for all $\boldsymbol{y} \in U$,

(iii) *there exist constants* $C_1,\ C_2 > 0$ *and a monotonically increasing sequence* $\boldsymbol{\delta} = (\delta_j)_{j\in\mathbb{N}} \subseteq (1,\infty)$ *such that* $(\delta_j^{-1})_{j\in\mathbb{N}} \in \ell^{p/(1-p)}(\mathbb{N})$, $\delta_j \leq C_1 j^{2/p}$ *for all* $j \in \mathbb{N}$ *and*

$$(\boldsymbol{\delta}^{\boldsymbol{\nu}} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F}). \tag{21}$$

Furthermore with

$$\Lambda_\tau := \{\boldsymbol{\nu} \in \mathcal{F} : \boldsymbol{\delta}^{-\boldsymbol{\nu}} \geq \tau\}$$

it holds for all $\tau \in (0,1)$ *that* $|\Lambda_\tau| > 0$ *and*

$$\sup_{\boldsymbol{y}\in U} \left\| u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} \right\|_{\mathcal{X}} \leq C_2 |\Lambda_\tau|^{-\frac{1}{p}+1}.$$

The proof is given in Appendix A.1. We next give more details on the structure of the sets $(\Lambda_\tau)_{\tau\in(0,1)}$ that will be required in establishing the ensuing DNN expression rate bounds. To this end let us introduce the quantities

$$m(\Lambda) := \sup_{\boldsymbol{\nu}\in\Lambda} |\boldsymbol{\nu}|_1 \qquad \text{and} \qquad d(\Lambda) := \sup_{\boldsymbol{\nu}\in\Lambda} |\operatorname{supp}\boldsymbol{\nu}|. \tag{22}$$

Proposition 3.8. *Let the assumptions of Theorem 3.7 be satisfied, and let* $J \in \mathbb{N}$ *and* $(\Lambda_\tau)_\tau \in (0,1)$ *be as in the statement of Theorem 3.7. Then*

(i) Λ_τ *is finite and downward closed for all* $\tau \in (0,1)$,

(ii) $m(\Lambda_\tau) = O(\log(|\Lambda_\tau|))$ *and* $d(\Lambda_\tau) = o(\log(|\Lambda_\tau|))$ *as* $\tau \to 0$,

(iii) $|\{\boldsymbol{\nu}_E : \boldsymbol{\nu} \in \Lambda_\tau\}| = O(\log(|\Lambda_\tau|)^J)$ *as* $\tau \to 0$,

(iv) *for all* $\tau \in (0,1)$, *if* $\mathbf{e}_j \in \Lambda_\tau$ *for some* $j \in \mathbb{N}$ *then for all* $i < j$ *it holds that* $\mathbf{e}_i \in \Lambda_\tau$.

Proof. To show (i), for downward closedness, let $\boldsymbol{\nu} \leq \boldsymbol{\mu}$ and $\boldsymbol{\mu} \in \Lambda_\tau$ be given. Then $\tau \leq \boldsymbol{\rho}^{-\boldsymbol{\mu}} \leq \boldsymbol{\rho}^{-\boldsymbol{\nu}}$ and thus $\boldsymbol{\nu} \in \Lambda_\tau$. Item (ii) was shown in [84, Lemma 1.4.15] and [84, Example 1.4.23].

Item (iii) is a consequence of $m(\Lambda_\tau) = O(\log(|\Lambda_\tau|))$, which holds by (ii). Finally, (iv) is a direct consequence of the monotonicity of $(\delta_j)_{j\in\mathbb{N}}$, which holds by Theorem 3.7. □

Remark 3.9. *We note that in the proof of Theorem 3.7, in particular Equation* (53)*, the sequence* $\boldsymbol{\delta}$ *is defined in terms of only* $\mathbf{b}$, p *and* ε.[2] *The index sets*

2 The sequence $\boldsymbol{\delta}$ depends on ε through $\gamma_2 \in (1,\kappa)$, for κ satisfying Equation (45).

$(\Lambda_\tau)_{\tau\in(0,1)}$ depend solely on $\boldsymbol{\delta}$ and τ. Thus, in principle, ε and the sequence $\boldsymbol{b}$ are sufficient to determine these index sets. For example, in the situation of Lemma 3.5, it holds $b_j = \|\psi_j\|_{\mathcal{Z}}$, $j \in \mathbb{N}$, which is known (or can be estimated) for many function systems $\{\psi_j\}_{j\geq 1}$.

4 DNN surrogates of real-valued functions

We now turn to the statement and proofs of the main results of this work. We first recapitulate in Section 4.1 the DNNs which we consider for the approximation, then present in Section 4.2 mathematical operations on DNNs. In Section 4.3, we recapitulate quantitative approximation rate bounds for polynomials by ReLU NNs, from [51, 77, 62, 47] which we use subsequently to reapproximate N-term gpc approximations of $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic functions.

As in [77], we develop the DNN expression rate bounds (which are free from the curse of dimensionality of the parametric maps) in Sections 4.4 and 4.5 in an abstract setting, for countably-parametric, scalar-valued maps with quantified control on the size of holomorphy domains.

4.1 Network architecture

We will use the same DNN architecture as in previous works (e.g. [62]). In Sections 4.1–4.3 we now restate results from [62, Section 2].

We consider *deep neural networks (DNNs for short)* of feed-forward type. Such a NN f can mathematically be described as a repeated composition of linear transformations with a nonlinear activation function. More precisely: For an *activation function* $\sigma : \mathbb{R} \to \mathbb{R}$, a fixed *number of hidden layers* $L \in \mathbb{N}_0$, numbers $N_\ell \in \mathbb{N}$ of *computation nodes in layer* $\ell \in \{1, \dots, L+1\}$, $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ is realized by a feedforward neural network, if for certain *weights* $w_{i,j}^\ell \in \mathbb{R}$, and *biases* $b_j^\ell \in \mathbb{R}$ it holds for all $\boldsymbol{x} = (x_i)_{i=1}^{N_0}$

$$z_j^1 = \sigma\left(\sum_{i=1}^{N_0} w_{i,j}^1 x_i + b_j^1\right), \quad j \in \{1, \dots, N_1\}, \tag{23a}$$

and

$$z_j^{\ell+1} = \sigma\left(\sum_{i=1}^{N_\ell} w_{i,j}^{\ell+1} z_i^\ell + b_j^{\ell+1}\right), \quad \ell \in \{1, \dots, L-1\}, \ \ j \in \{1, \dots, N_{\ell+1}\}, \tag{23b}$$

and finally

$$f(\boldsymbol{x}) = (z_j^{L+1})_{j=1}^{N_{L+1}} = \left(\sum_{i=1}^{N_L} w_{i,j}^{L+1} z_i^L + b_j^{L+1} \right)_{j=1}^{N_{L+1}} . \tag{23c}$$

In this case N_0 is the dimension of the input and N_{L+1} is the dimension of the output. Furthermore z_j^ℓ denotes the output of unit j in layer ℓ. The weight $w_{i,j}^\ell$ has the interpretation of connecting the ith unit in layer $\ell - 1$ with the jth unit in layer ℓ. If $L = 0$, then (23c) holds with $z_i^0 := x_i$ for $i = 1, \dots, N_0$.

Except when explicitly stated, we will not distinguish between the network (which is defined through σ, the $w_{i,j}^\ell$ and b_j^ℓ) and the function $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_{L+1}}$ it realizes. We note in passing that this relation is typically not one-to-one, i.e. different NNs may realize the same function as their output. Let us also emphasize that we allow the weights $w_{i,j}^\ell$ and biases b_j^ℓ for $\ell \in \{1, \dots, L+1\}$, $i \in \{1, \dots, N_{\ell-1}\}$ and $j \in \{1, \dots, N_\ell\}$ to take any value in $\mathbb{R}$, i.e. we do not consider quantization as e.g. in [10, 67].

As is customary in the theory of NNs, the number of hidden layers L of a NN is referred to as *depth*[3] and the total number of nonzero weights and biases as the *size* of the NN. Hence, for a DNN f as in (23), we define

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

In addition, $\mathrm{size}_{\mathrm{in}}(f) := |\{(i,j) : w_{i,j}^1 \neq 0\}| + |\{j \,:\, b_j^1 \neq 0\}|$ and $\mathrm{size}_{\mathrm{out}}(f) := |\{(i,j) : w_{i,j}^{L+1} \neq 0\}| + |\{j \,:\, b_j^{L+1} \neq 0\}|$, which are the number of nonzero weights and biases in the input layer of f and in the output layer, respectively.

The proofs of our main results are constructive, in the sense that we explicitly provide NN architectures and constructions of instances of DNNs with these architectures which are sufficient (but possibly larger than necessary) for achieving the claimed expression rates. We construct these NNs by assembling smaller networks, using the operations of concatenation and parallelization, as well as so-called "identity-networks" which realize the identity mapping. Below, we recall the definitions.

3 In other recent references (e.g. [63]), slightly different terminology for the number L of layers in the DNN differing from the convention in the present paper by a constant factor, is used. This difference will be inconsequential for all results that follow.

4.2 Basic operations

Throughout, as activation function σ we consider either the ReLU activation function

$$\sigma_1(x) := \max\{0, x\} \qquad x \in \mathbb{R} \tag{24}$$

or, as suggested in [54, 55, 47], for $r \in \mathbb{N}$, $r \geq 2$, the RePU activation function

$$\sigma_r(x) := \max\{0, x\}^r = \sigma_1(x)^r \qquad x \in \mathbb{R}. \tag{25}$$

See [62, Remark 2.1] for a historical note on rectified power units. If a NN uses σ_r as activation function, we refer to it as σ_r-NN. ReLU NNs are referred to as σ_1-NNs. It is assumed throughout that all activations in a DNN are of equal type.

We now recall the parallelization and concatenation of networks, as well networks realizing the identity. The constructions are mostly straightforward. For details and proofs we refer to [67, 63, 27, 62].

4.2.1 Parallelization

Let f, g be two NNs with the same depth $L \in \mathbb{N}_0$, input dimensions n_f, n_g and output dimensions m_f, m_g respectively. There exists a NN $(f,g)_{\mathrm{d}}$ such that

$$(f,g)_{\mathrm{d}} : \mathbb{R}^{n_f} \times \mathbb{R}^{n_g} \to \mathbb{R}^{m_f} \times \mathbb{R}^{m_g} : (\boldsymbol{x}, \tilde{\boldsymbol{x}}) \mapsto (f(\boldsymbol{x}), g(\tilde{\boldsymbol{x}})).$$

It holds $\operatorname{depth}((f,g)_{\mathrm{d}}) = L$, $\operatorname{size}((f,g)_{\mathrm{d}}) = \operatorname{size}(f) + \operatorname{size}(g)$, $\operatorname{size}_{\mathrm{in}}((f,g)_{\mathrm{d}}) = \operatorname{size}_{\mathrm{in}}(f) + \operatorname{size}_{\mathrm{in}}(g)$ and $\operatorname{size}_{\mathrm{out}}((f,g)_{\mathrm{d}}) = \operatorname{size}_{\mathrm{out}}(f) + \operatorname{size}_{\mathrm{out}}(g)$, see [67, 27].

In case $n_f = n_g = n$, there exists a NN (f,g) with the same depth and size as $(f,g)_d$, such that

$$(f,g) : \mathbb{R}^n \to \mathbb{R}^{m_f} \times \mathbb{R}^{m_g} : \boldsymbol{x} \mapsto (f(\boldsymbol{x}), g(\boldsymbol{x})).$$

4.2.2 Identity

By [67, Lemma 2.3], for all $n \in \mathbb{N}$, $L \in \mathbb{N}_0$ there exists a σ_1-*identity network* $\mathrm{Id}_{\mathbb{R}^n}$ of depth L such that $\mathrm{Id}_{\mathbb{R}^n}(\boldsymbol{x}) = \boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n$. It holds that

$$\operatorname{size}(\mathrm{Id}_{\mathbb{R}^n}) \leq 2n(L+1), \qquad \operatorname{size}_{\mathrm{in}}(\mathrm{Id}_{\mathbb{R}^n}) \leq 2n, \qquad \operatorname{size}_{\mathrm{out}}(\mathrm{Id}_{\mathbb{R}^n}) \leq 2n.$$

Analogously, by [62, Proposition 2.3], for all r, $n \in \mathbb{N}$, $r \geq 2$ and $L \in \mathbb{N}_0$ there exists a σ_r-*identity network* $\mathrm{Id}_{\mathbb{R}^n}$ of depth L such that $\mathrm{Id}_{\mathbb{R}^n}(\boldsymbol{x}) = \boldsymbol{x}$. It holds that

$$\operatorname{size}(\mathrm{Id}_{\mathbb{R}^n}) \leq nL(4r^2 + 2r), \qquad \operatorname{size}_{\mathrm{in}}(\mathrm{Id}_{\mathbb{R}^n}) \leq 4nr, \qquad \operatorname{size}_{\mathrm{out}}(\mathrm{Id}_{\mathbb{R}^n}) \leq n(2r+1).$$

4.2.3 Sparse concatenation

Let f and g be σ_1-NNs, such that the output dimension of g equals the input dimension of f. Let n_g be the input dimension of g and m_f the output dimension of f. Then, the *sparse concatenation of the NNs f and g* realizes the function

$$f \circ g : \mathbb{R}^{n_g} \to \mathbb{R}^{m_f} : \boldsymbol{x} \mapsto f(g(\boldsymbol{x})). \tag{26}$$

In the following, by abuse of notation, "$\circ$" can either stand for the composition of functions or the sparse concatenation of networks. The meaning will be clear from the context. By [67, Remark 2.6], $\operatorname{depth}(f \circ g) = \operatorname{depth}(f) + 1 + \operatorname{depth}(g)$,

$$\operatorname{size}(f \circ g) \leq \operatorname{size}(f) + \operatorname{size}_{\text{in}}(f) + \operatorname{size}_{\text{out}}(g) + \operatorname{size}(g) \leq 2\operatorname{size}(f) + 2\operatorname{size}(g) \tag{27}$$

and

$$\operatorname{size}_{\text{in}}(f \circ g) \leq \begin{cases} \operatorname{size}_{\text{in}}(g) & \operatorname{depth}(g) \geq 1, \\ 2\operatorname{size}_{\text{in}}(g) & \operatorname{depth}(g) = 0, \end{cases}$$

$$\operatorname{size}_{\text{out}}(f \circ g) \leq \begin{cases} \operatorname{size}_{\text{out}}(f) & \operatorname{depth}(f) \geq 1, \\ 2\operatorname{size}_{\text{out}}(f) & \operatorname{depth}(f) = 0. \end{cases}$$

Similarly, for $r \geq 2$ there exists a sparse concatenation of σ_r-NNs (we denote the concatenation operator again by $\circ$) satisfying the following size and depth bounds from [62, Proposition 2.4]: Let f, g be two σ_r-NNs such that the output dimension k of g equals the input dimension of f, and suppose that $\operatorname{size}_{\text{in}}(f)$, $\operatorname{size}_{\text{out}}(g) \geq k$. Then $\operatorname{depth}(f \circ g) = \operatorname{depth}(f) + 1 + \operatorname{depth}(g)$,

$$\begin{aligned} \operatorname{size}(f \circ g) &\leq \operatorname{size}(f) + (2r-1)\operatorname{size}_{\text{in}}(f) + (2r+1)k + (2r-1)\operatorname{size}_{\text{out}}(g) + \operatorname{size}(g) \\ &\leq \operatorname{size}(f) + 2r\operatorname{size}_{\text{in}}(f) + (4r-1)\operatorname{size}_{\text{out}}(g) + \operatorname{size}(g) \\ &\leq (2r+1)\operatorname{size}(f) + 4r\operatorname{size}(g), \end{aligned} \tag{28}$$

and

$$\operatorname{size}_{\text{in}}(f \circ g) \leq \begin{cases} \operatorname{size}_{\text{in}}(g) & \operatorname{depth}(g) \geq 1, \\ 2r\operatorname{size}_{\text{in}}(g) + 2rk \leq 4r\operatorname{size}_{\text{in}}(g) & \operatorname{depth}(g) = 0, \end{cases}$$

$$\operatorname{size}_{\text{out}}(f \circ g) \leq \begin{cases} \operatorname{size}_{\text{out}}(f) & \operatorname{depth}(f) \geq 1, \\ 2r\operatorname{size}_{\text{out}}(f) + k \leq (2r+1)\operatorname{size}_{\text{out}}(f) & \operatorname{depth}(f) = 0. \end{cases}$$

Combining identity networks with the sparse concatenation, we can parallelize networks of different depth. The next lemma shows this for ReLU-NNs (a proof is given in Appendix A.2).

Lemma 4.1. *For all k, $n \in \mathbb{N}$ and σ_1-NNs $f_1, \ldots, f_k$ with the same input dimension n and output dimensions $m_1, \ldots, m_k \in \mathbb{N}$, there exists a σ_1-NN $(f_1, \ldots, f_k)_{\mathrm{s}}$ called the* parallelization of $f_1, \ldots, f_k$ with shared identity network. *It has input dimension n, output dimension $m := \sum_{t=1}^{k} m_t$, it realizes $\mathbb{R}^n \to \mathbb{R}^m : \boldsymbol{x} \mapsto (f_1(\boldsymbol{x}), \ldots, f_k(\boldsymbol{x}))$, has depth $L := \max_{t=1,\ldots,k} \operatorname{depth}(f_t)$ and its size is bounded as follows:*

$$\operatorname{size}((f_1, \ldots, f_k)_{\mathrm{s}}) \leq \sum_{t=1}^{k} \operatorname{size}(f_t) + \sum_{t=1}^{k} \operatorname{size}_{\mathrm{in}}(f_t) + 2nL \leq 2 \sum_{t=1}^{k} \operatorname{size}(f_t) + 2nL,$$

$$\operatorname{size}_{\mathrm{in}}((f_1, \ldots, f_k)_{\mathrm{s}}) \leq \sum_{t=1}^{k} \operatorname{size}_{\mathrm{in}}(f_t) + 2n,$$

$$\operatorname{size}_{\mathrm{out}}((f_1, \ldots, f_k)_{\mathrm{s}}) \leq \sum_{t=1}^{k} 2 \operatorname{size}_{\mathrm{out}}(f_t).$$

Remark 4.2. *The term $2nL$ in the size bound corresponds to the nonzero weights (and biases) of the identity network used to construct the parallelization. We point out that this number is independent of the number k of networks $(f_t)_{t=1}^{k}$, since our construction allows the k networks to share one identity network.*

4.3 Approximation of polynomials

As in other recent works (e.g. [77, 62, 23, 63]), the ensuing DNN expression rate analysis of possibly countably-parametric posterior densities will rely on *DNN reapproximation of sparse generalized polynomial chaos approximations* of these densities. It has been observed in [83, 51] that ReLU DNNs can represent high order polynomials on bounded intervals rather efficiently. We recapitulate several results of this type, from [62, Section 2], and from [77] which we will require in the following.

4.3.1 Approximate multiplication

Contrary to [83], the next result bounds the DNN expression error in $W^{1,\infty}([-M, M]^2)$ (instead of the $L^\infty([-M, M]^2)$-norm).

Proposition 4.3 ([77, Proposition 3.1]). *For any $\delta \in (0,1)$ and $M \geq 1$ there exists a σ_1-NN $\tilde{\times}_{\delta,M} : [-M,M]^2 \to \mathbb{R}$ such that*

$$\begin{aligned} \sup_{|a|,|b|\leq M} |ab - \tilde{\times}_{\delta,M}(a,b)| \leq \delta, \\ \operatorname*{ess\,sup}_{|a|,|b|\leq M} \max\left\{ \left| b - \frac{\partial}{\partial a}\tilde{\times}_{\delta,M}(a,b) \right|, \left| a - \frac{\partial}{\partial b}\tilde{\times}_{\delta,M}(a,b) \right| \right\} \leq \delta, \end{aligned} \tag{29}$$

where $\frac{\partial}{\partial a}\tilde{\times}_{\delta,M}(a,b)$ and $\frac{\partial}{\partial b}\tilde{\times}_{\delta,M}(a,b)$ denote weak derivatives. There exists a constant $C > 0$ independent of $\delta \in (0,1)$ and $M \geq 1$ such that $\mathrm{size}_{\mathrm{in}}(\tilde{\times}_{\delta,M}) \leq C$, $\mathrm{size}_{\mathrm{out}}(\tilde{\times}_{\delta,M}) \leq C$,

$$\mathrm{depth}(\tilde{\times}_{\delta,M}) \leq C(1+\log_2(M/\delta)), \qquad \mathrm{size}(\tilde{\times}_{\delta,M}) \leq C(1+\log_2(M/\delta)).$$

Moreover, for every $a \in [-M,M]$, there exists a finite set $\mathcal{N}_a \subseteq [-M,M]$ such that $b \mapsto \tilde{\times}_{\delta,M}(a,b)$ is strongly differentiable at all $b \in (-M,M)\backslash\mathcal{N}_a$.

Proposition 4.3 implies the existence of networks approximating the multiplication of n different numbers.

Proposition 4.4 ([77, Proposition 3.3]). *For any $\delta \in (0,1)$, $n \in \mathbb{N}$ and $M \geq 1$ there exists a σ_1-NN $\tilde{\prod}_{\delta,M} : [-M,M]^n \to \mathbb{R}$ such that*

$$\sup_{(x_i)_{i=1}^n \in [-M,M]^n} \left| \prod_{j=1}^n x_j - \tilde{\prod}_{\delta,M}(x_1,\ldots,x_n) \right| \leq \delta. \tag{30}$$

There exists a constant C independent of $\delta \in (0,1)$, $n \in \mathbb{N}$ and $M \geq 1$ such that

$$\mathrm{size}(\tilde{\prod}_{\delta,M}) \leq C(1+n\log(nM^n/\delta)), \quad \mathrm{depth}(\tilde{\prod}_{\delta,M}) \leq C(1+\log(n)\log(nM^n/\delta)). \tag{31}$$

Remark 4.5. *In [77], Propositions 4.3 and 4.4 are shown for $M = 1$. The result for $M > 1$ is obtained by a simple scaling argument. See [62, Proposition 2.6] for more details.*

4.3.2 ReLU DNN approximation of tensor product Legendre polynomials

Based on the ReLU DNN emulation of products in Proposition 4.3, we constructed ReLU DNN approximations of multivariate Legendre polynomials in [62]. For the statement recall $m(\Lambda)$ in (22).

Proposition 4.6 ([62, Proposition 2.13]). *For every finite $\Lambda \subset \mathbb{N}_0^d$ and every $\delta \in (0,1)$, there exists a σ_1-NN $f_{\Lambda,\delta} = (\tilde{L}_{\boldsymbol{\nu},\delta})_{\boldsymbol{\nu}\in\Lambda}$ with input dimension d and output dimension $|\Lambda|$ such that the outputs $\{\tilde{L}_{\boldsymbol{\nu},\delta}\}_{\boldsymbol{\nu}\in\Lambda}$ of $f_{\Lambda,\delta}$ satisfy for every $\boldsymbol{\nu} \in \Lambda$*

$$\|L_{\boldsymbol{\nu}} - \tilde{L}_{\boldsymbol{\nu},\delta}\|_{W^{1,\infty}([-1,1]^d)} \leq \delta, \quad \sup_{\boldsymbol{y}\in[-1,1]^d} \left|\tilde{L}_{\boldsymbol{\nu},\delta}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})\right| \leq (2m(\Lambda)+2)^d.$$

Furthermore, there exists $C > 0$ such that for every d, Λ and δ

$$\begin{aligned}\operatorname{depth}(f_{\Lambda,\delta}) &\leq C(1+d\log d)(1+\log_2 m(\Lambda))\big(m(\Lambda)+\log_2(1/\delta)\big),\\ \operatorname{size}(f_{\Lambda,\delta}) &\leq C\left[d^2 m(\Lambda)^2 + dm(\Lambda)\log_2(1/\delta) + d^2|\Lambda|\big(1+\log_2 m(\Lambda)+\log_2(1/\delta)\big)\right].\end{aligned}$$

4.3.3 RePU DNN emulation of polynomials

The approximation of polynomials by neural networks can be significantly simplified if instead of the ReLU activation σ_1 we consider as activation function the so-called *rectified power unit* ("RePU" for short) $\sigma_r(x) = \max\{0,x\}^r$ for $r \geq 2$. In contrast to σ_1-NNs, as shown in [47], for every $r \in \mathbb{N}$, $r \geq 2$ there exist RePU networks of depth 1 realizing the multiplication of two real numbers *without error*. This yields the following result, slightly improving [47, Theorem 9], in that the constant C is independent of d. This is relevant, as in Section 4.5 ahead the number of active parameters $d(\Lambda_\tau)$ increases with decreasing accuracy τ.

Proposition 4.7 ([62, Proposition 2.14]). *Fix $d \in \mathbb{N}$ and $r \in \mathbb{N}$, $r \geq 2$. Then there exists a constant $C > 0$ depending on r but independent of d such that for any finite downward closed $\Lambda \subseteq \mathbb{N}_0^d$ and any $p \in \mathbb{P}_\Lambda$ there is a σ_r-network $\tilde{p} : \mathbb{R}^d \to \mathbb{R}$ which realizes p exactly and such that* $\operatorname{size}(\tilde{p}) \leq C|\Lambda|$ *and* $\operatorname{depth}(\tilde{p}) \leq C\log_2(|\Lambda|)$.

Remark 4.8. *Similar results hold for other, widely used activation functions ψ. As discussed in [62, Remark 2.15], if the product of two numbers can be approximated by ψ-NNs up to* arbitrary accuracy *and with NN size and depth* independent of the accuracy*, then polynomials can be approximated with size and depth bounded as* $\operatorname{size}(\tilde{p}) \leq C|\Lambda|$ *and* $\operatorname{depth}(\tilde{p}) \leq C\log_2(|\Lambda|)$*, for C independent of the arbitrarily small accuracy.*

Activation functions for which this holds include (i) $\psi \in C^2$ for which there exists $x \in \mathbb{R}$ where $\psi''(x) \neq 0$, (ii) ψ which are continuous and sigmoidal of order *$k \geq 2$ (see also [62, Remark 2.1]), and (iii) NNs with rational activations. We refer to [62, Remark 2.15] for a more detailed discussion.*

4.4 ReLU DNN approximation of $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic maps

We now present a result about the expressive power for $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic functions, in the sense of Remark 3.1. Theorem 4.9 generalizes [77, Theorem 3.9], as it shows that less regular functions[4] can be emulated with the same convergence rate (see Remark 3.4). In particular, we obtain that up to logarithmic terms, ReLU DNNs are capable of approximating $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic maps at rates equivalent to those achieved by best n-term gpc approximations. Here, *"rate" is understood in terms of the NN size*, i.e., in terms of the total number of nonzero weights in the DNN.

In the following, for $\Lambda_\tau \subset \mathcal{F}$ as in Theorem 3.7, we define its support

$$S_{\Lambda_\tau} := \cup_{\boldsymbol{\nu} \in \Lambda_\tau} \operatorname{supp} \boldsymbol{\nu} \subset \mathbb{N}. \tag{32}$$

Theorem 4.9. *Let $u : U \to \mathbb{R}$ be $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$ and $\varepsilon > 0$. For $\tau \in (0,1)$ let $\Lambda_\tau \subset \mathcal{F}$ be as in Theorem 3.7.*

Then there exists $C > 0$ depending on $\boldsymbol{b}$, ε and u, such that for all $\tau \in (0,1)$ there exists a σ_1-NN $\tilde{u}_\tau$ with input variables $(y_j)_{j \in S_{\Lambda_\tau}}$ such that

$$\begin{aligned} \operatorname{size}(\tilde{u}_\tau) &\leq C(1 + |\Lambda_\tau| \cdot \log |\Lambda_\tau| \cdot \log\log |\Lambda_\tau|), \\ \operatorname{depth}(\tilde{u}_\tau) &\leq C(1 + \log |\Lambda_\tau| \cdot \log\log |\Lambda_\tau|). \end{aligned}$$

Furthermore, $\tilde{u}_\tau$ satisfies the uniform error bound

$$\sup_{\boldsymbol{y} \in U} \left| u(\boldsymbol{y}) - \tilde{u}_\tau((y_j)_{j \in S_{\Lambda_\tau}}) \right| \leq C |\Lambda_\tau|^{-1/p+1}. \tag{33}$$

In case $|\Lambda_\tau| = 1$, the statement holds with $\log\log |\Lambda_\tau|$ replaced by 0.

The proof is given in Appendix A.3.

Remark 4.10. *Let $K \in \mathbb{N}$ and let $v : U \to \mathbb{R}^K$ be $(\boldsymbol{b}, \varepsilon, \mathbb{R}^K)$-holomorphic. Then Theorem 4.9 can be applied to each component of v. This at most increases the bound on the network size by a factor K, but it does not affect the depth and the convergence rate. In fact, only the dimension of the output layer has to be increased, the hidden layers of the DNN can be the same as for $K = 1$. This corresponds to reusing the same polynomial basis for the approximation of all components of v.*

4 Theorem 4.9 only assumes quantified holomorphy in *polyellipses* in a suitable, finite number of the parameters y_j, whereas [77, Theorem 3.9] required holomorphy in poly*discs*. The presently obtained expression rates are identical to those in [77, Theorem 3.9], but are shown to hold for maps with smaller domains of holomorphy.

4.5 RePU DNN approximation of $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic maps

We next provide an analogue of Theorem 4.9 (which used σ_1-NNs) for σ_r-NNs, $r \geq 2$. The smaller multiplication networks of Proposition 4.7 allow to prove the same approximation error for slightly smaller networks in this case.

Theorem 4.11. *Let $u : U \to \mathbb{R}$ be $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$ and $\varepsilon > 0$. For $\tau \in (0,1)$ let $\Lambda_\tau \subset \mathcal{F}$ be as in Theorem 3.7. Let $r \in \mathbb{N}$, $r \geq 2$.*

Then there exists $C > 0$ depending on $\boldsymbol{b}$, ε, u and r, such that for all $\tau \in (0,1)$ there exists a σ_r-NN $\tilde{u}_\tau$ with input variables $(y_j)_{j \in S_{\Lambda_\tau}}$ such that

$$\operatorname{size}(\tilde{u}_\tau) \leq C|\Lambda_\tau|, \qquad \operatorname{depth}(\tilde{u}_\tau) \leq C \log|\Lambda_\tau|$$

and $\tilde{u}_\tau$ satisfies the uniform error bound

$$\sup_{\boldsymbol{y} \in U} \left| u(\boldsymbol{y}) - \tilde{u}_\tau((y_j)_{j \in S_{\Lambda_\tau}}) \right| \leq C|\Lambda_\tau|^{-1/p+1}. \tag{34}$$

Proof. By Proposition 4.7, the $|S_{\Lambda_\tau}|$-variate polynomial $\sum_{\boldsymbol{\nu} \in \Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \in \mathbb{P}_{\Lambda_\tau}$ from Theorem 3.7 and Corollary 3.8 can be emulated exactly by a σ_r-NN satisfying

$$\operatorname{size}(\tilde{u}_\tau) \leq C|\Lambda_\tau|, \qquad \operatorname{depth}(\tilde{u}_\tau) \leq C \log(|\Lambda_\tau|),$$

for C independent of $|S_{\Lambda_\tau}|$. The error bound (34) holds by Theorem 3.7 (iii). □

Remarks 4.8 and 4.10 also apply here.

5 DNN surrogates of $\mathcal{X}$-valued functions

In this section, we address the DNN emulation of countably-parametric, holomorphic maps taking values in function spaces as typically arise in PDE UQ. In Section 5.1 we show DNN expression rate bounds for parametric PDE solution families, assuming the existence of suitable NN approximations of functions in the solution space of the PDE.

In Section 5.2.1 we review results on the *exact DNN emulation of Courant-type Finite Element spaces* on regular, simplicial triangulations. In Sections 5.2.2 and 5.2.3, we discuss Theorem 5.2 for the diffusion equation from Section 2.4.1 and the eigenvalue problem from Section 2.4.2.

5.1 ReLU DNN expression of $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic maps

So far, we considered the DNN expression of real-valued maps $u : U \to \mathbb{R}$. In applications to PDEs, often also the expression of maps $u : U \to \mathcal{X}$ is of interest. Here, the real Banach space $\mathcal{X}$ is a function space over a domain $\mathrm{D} \subset \mathbb{R}^d$ for $d \in \mathbb{N}$, and is interpreted as the solution space of the parametric forward model (1).

As it was shown for example in [26, 3, 85], for gpc coefficients $u_{\boldsymbol{\nu}}$, a $\boldsymbol{\nu}$-dependent degree of resolution in $\mathcal{X}$ of $u_{\boldsymbol{\nu}}$ is in general advantageous. We approach DNN expression of the parametric solution map through DNN emulation of multilevel gpc-FE approximations. To state these, a *regularity space* $\mathcal{X}^s \subset \mathcal{X}$ of functions with additional regularity will be required. We first present the result in an abstract setting, and subsequently detail it for an example in Sections 5.2.2 and 5.2.3.

For the DNN emulation of polynomials in the variables $\boldsymbol{y} \in U$, we use Lemma A.1, based on the networks constructed in the proof of Theorem 4.9. For the gpc coefficients, which we assume to be in $\mathcal{X}^s$, we allow sequences of NN approximations satisfying a mild bound on their L^∞-norm, as made precise in Assumption 5.1. This is needed to use the product networks from Proposition 4.3 to multiply NNs approximating the polynomials in $\boldsymbol{y}$ with NN approximations of gpc-coefficients.

Assumption 5.1. *Assume that there exist $\gamma > 0$, $\theta \geq 0$ and $C > 0$ such that for all $v \in \mathcal{X}^s$ and all $m \in \mathbb{N}$ there exists a NN Φ_m^v which satisfies*

$$\mathrm{depth}(\Phi_m^v) \leq C(1 + \log m), \qquad \mathrm{size}(\Phi_m^v) \leq Cm$$

and

$$\|v - \Phi_m^v\|_{\mathcal{X}} \leq C\|v\|_{\mathcal{X}^s} m^{-\gamma}, \quad \|\Phi_m^v\|_{\mathcal{X}} \leq C\|v\|_{\mathcal{X}}, \quad \left\|\Phi_m^v\right\|_{L^\infty(\mathrm{D})} \leq C\,\|v\|_{\mathcal{X}^s}\, m^{\theta}.$$

Let us consider an example. For a bounded polytope $\mathrm{D} \subset \mathbb{R}^d$, functions in the Kondratiev space $\mathcal{X}^s = \mathcal{K}^2_{1+\zeta}(\mathrm{D})$ with $\zeta \in (0, 1)$ (for a definition of $\mathcal{K}^2_{1+\zeta}(\mathrm{D})$ see Equation (39) ahead) can be approximated by continuous, piecewise affine functions on regular triangulations of D with convergence rate $\gamma = \frac{1}{d}$ (e.g. [2, 6, 48] for $d = 2$, [60] for $d > 2$). Continuous, piecewise affine functions on regular, simplicial partitions can be exactly emulated by ReLU networks, see Section 5.2.1. These NNs approximate functions in $\mathcal{X}^s = \mathcal{K}^2_{1+\zeta}(\mathrm{D})$ with (optimal) rate $\gamma = 1/d$. By the continuous embedding $\mathcal{X}^s \hookrightarrow L^\infty(\mathrm{D})$ ([53], [19, Theorem 27]), the last inequality in Assumption 5.1 is satisfied with $\theta = 0$. Here, the domain D may, but need not, be the physical domain of interest. The theorem below also applies to boundary integral equations, in which case D is the boundary of the physical domain. Holomorphic dependence of boundary integral operators on the shape of the domain ("shape-holomorphy") is shown in [34].

We obtain the following result, which generalizes [77, Theorem 4.8]. To state the theorem, we recall the notation $S_{\Lambda_\tau} = \cup_{\boldsymbol{\nu}\in\Lambda_\tau} \operatorname{supp}\boldsymbol{\nu} \subset \mathbb{N}$ introduced in (32).

Theorem 5.2. *Let $d \in \mathbb{N}$ and let $\mathcal{X} = W^{1,q}(\mathrm{D})$, $q \in [1,\infty]$,* [5] *$\mathcal{X}^s \subset \mathcal{X}$ be Banach spaces of functions $v : \mathrm{D} \to \mathbb{R}$ for some bounded domain $\mathrm{D} \subset \mathbb{R}^d$. Assume that Assumption 5.1 holds for some $\gamma > 0$ and $\theta \geq 0$. Let $u : U \to \mathcal{X}^s \subset \mathcal{X}$ be a $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic map, in the sense of Remark 3.1, for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$ and $\varepsilon > 0$. Let $(c_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \subset \mathcal{X}$ and $(\Lambda_\tau)_{\tau\in(0,1)}$ be as in Theorem 3.7. Assume that $(c_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \subset \mathcal{X}^s$ and that $(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p^s}$ for some $0 < p < p^s < 1$.*

Then, there exists a constant $C > 0$ depending on d, γ, θ, $\boldsymbol{b}$, (thus also on p), ε, p^s and u such that for all $\tau \in (0,1)$ there exists a ReLU NN $\tilde{u}_\tau$ with input variables $(x_1,\dots,x_d) = \boldsymbol{x} \in \mathrm{D}$ and $(y_j)_{j\in S_{\Lambda_\tau}}$ for $\boldsymbol{y} \in U$ and output dimension 1 such that for some $\mathcal{N}_\tau \in \mathbb{N}$ satisfying $\mathcal{N}_\tau \geq |\Lambda_\tau|$

$$\operatorname{size}(\tilde{u}_\tau) \leq C(1 + \mathcal{N}_\tau \cdot \log\mathcal{N}_\tau \cdot \log\log\mathcal{N}_\tau), \quad \operatorname{depth}(\tilde{u}_\tau) \leq C(1 + \log\mathcal{N}_\tau \cdot \log\log\mathcal{N}_\tau)$$

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

$$\sup_{\boldsymbol{y}\in U} \left\| u(\boldsymbol{y}) - \tilde{u}_\tau(\cdot, (y_j)_{j\in S_{\Lambda_\tau}}) \right\|_{\mathcal{X}} \leq C\mathcal{N}_\tau^{-r}, \qquad r := \gamma \min\left\{1, \frac{1/p - 1}{\gamma + 1/p - 1/p^s}\right\}. \tag{35}$$

The proof is given in Appendix A.4. Theorem 5.2 shows that for all $r^* < r$ there exists $C > 0$ (additionally depending on r^*) such that

$$\sup_{\boldsymbol{y}\in U} \left\| u(\boldsymbol{y}) - \tilde{u}_\tau(\cdot, (y_j)_{j\in S_{\Lambda_\tau}}) \right\|_{\mathcal{X}} \leq C(\operatorname{size}(\tilde{u}_\tau))^{-r^*}.$$

The limit r on the convergence rate in (35) is bounded from above by the gpc best n-term rate $1/p - 1$ for the truncation error of the gpc expansion and by the convergence rate γ of ReLU DNN approximations of functions in $\mathcal{X}^s$ from Assumption 5.1.

5.2 ReLU DNN expression of Courant Finite Elements

We now recall that any continuous, piecewise affine function on a locally convex, regular triangulation is representable by a ReLU network, e.g. [77, 33]. This is used in Section 5.2.2 to show an expression result for $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic functions, where $\mathcal{X}$ is a Sobolev space over a bounded domain.

5 Although $q = 2$ in all examples we consider, the theorem is stated slightly more generally for $q \in [1,\infty]$. In fact, the result also holds for weighted $W^{1,q}$-spaces.

5.2.1 Continuous, piecewise affine functions

In space dimension $d = 1$, any continuous, piecewise linear function on a partition $a = t_0 < t_1 < \cdots < t_N = b$ of a finite interval $[a, b]$ into N subintervals, can be expressed without error by a σ_1-NN with depth 1 and size $\mathcal{O}(N)$, e.g. [77, Lemma 4.5].

A similar result holds for $d \geq 2$. Consider a bounded polytope $G \subset \mathbb{R}^d$ with Lipschitz boundary ∂G being (the closure of) a finite union of plane $d-1$-faces. Let $\mathcal{T}$ be a regular, simplicial triangulation of G, i.e. the intersection of any two distinct closed simplices $\overline{T}, \overline{T}' \in \mathcal{T}$ is either empty or an entire k-simplex for some $0 \leq k < d$.[6] For the ReLU NN emulation of gpc-coefficients, we will use that also in space dimension $d \geq 2$, continuous, piecewise linear functions on a regular, simplicial mesh $\mathcal{T}$ can efficiently be emulated exactly by ReLU DNNs. For locally convex partitions, this was shown in [33], as we next recall in Proposition 5.3. The term locally convex refers to meshes $\mathcal{T}$ for which each patch, consisting of all elements attached to a fixed node of $\mathcal{T}$, is a convex set. See [33] for more details.

Set

$$S^1(G, \mathcal{T}) := \{v \in C^0(G) : v|_T \in \mathbb{P}_1,\ \forall T \in \mathcal{T}\}.$$

We denote by $\mathcal{N}(\mathcal{T})$ the set of nodes of the mesh $\mathcal{T}$ and by $k_{\mathcal{T}} := \max_{p \in \mathcal{N}} |\{T \in \mathcal{T} : p \in \overline{T}\}|$, the maximum number of elements sharing a node.

Proposition 5.3 ([33, Theorem 3.1])**.** *Let $\mathcal{T}$ be a regular, simplicial, locally convex triangulation of a bounded polytope G. Then every $v \in S^1(G, \mathcal{T})$ can be implemented exactly by a σ_1-NN of depth $1 + \log_2\lceil k_{\mathcal{T}} \rceil$ and size of the order $\mathcal{O}(|\mathcal{T}| k_{\mathcal{T}})$.*

Estimates on the network size for continuous, piecewise linear functions on general, regular simplicial partitions $\mathcal{T}$ are stated in [33, Theorem 5.2] based on [81], but are much larger than those in [33, Theorem 3.1].

5.2.2 Parametric diffusion problem

The standard example of a $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic parametric solution family is based on Section 2.4.1, i.e. the solution to an affine-parametric diffusion problem, see e.g. [13, 85]. In the setting of Section 2.4.1, we verify the assumptions of Theorem 5.2.

Let $\mathrm{D} \subset \mathbb{R}^2$ be a bounded polygonal Lipschitz domain (for details see [84, Remark 4.2.1]). We consider a linear, elliptic diffusion equation with uncertain

6 In other words, $\mathcal{T}$ is a cellular complex.

diffusion coefficient and with homogeneous Dirichlet boundary conditions. With $\mathcal{X} = \mathcal{Y} := H_0^1(\mathrm{D};\mathbb{C})$, $X := L^\infty(\mathrm{D};\mathbb{C})$ and for a fixed right-hand side $f \in \mathcal{Y}' = \mathcal{X}'$ the weak formulation reads: given $a \in X$, find $u(a) \in \mathcal{X}$ such that

$$\int_{\mathrm{D}} \nabla u(a)^\top a \nabla v \mathrm{d}\boldsymbol{x} = \langle f, v\rangle \qquad \forall v \in \mathcal{Y}. \tag{36}$$

The map $G : a \mapsto u(a) \in \mathcal{X}$ is then locally well-defined and holomorphic around every $a \in X$ for which $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a(\boldsymbol{x})) > 0$, see, e.g., [84, Example 1.2.38 and Equations (4.3.12) – (4.3.13)].

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 a nominal input $a_0 \in X$ and for a sequence of fluctuations $(\psi_j)_{j\in\mathbb{N}} \subseteq X$, define

$$a(\boldsymbol{y}) = a_0 + \sum_{j\in\mathbb{N}} y_j \psi_j. \tag{37}$$

Such expansions arise, for example, from Fourier-, Karhunen-Loève-, spline- or wavelet series representations of a.

If $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a_0(\boldsymbol{x})) = \gamma > 0$ then

$$\sum_{j\in\mathbb{N}} \|\psi_j\|_X < \gamma \tag{38}$$

ensures $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a(\boldsymbol{y})(\boldsymbol{x})) > 0$ for all $\boldsymbol{y} \in U$. This in turn implies that (36) admits a unique solution for all diffusion coefficients $a(\boldsymbol{y})$, $\boldsymbol{y} \in U$. Thus Lemma 3.5 yields $\boldsymbol{y} \mapsto u(\boldsymbol{y}) = G(a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j)$ to be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic for some $\varepsilon > 0$ and with $b_j := \|\psi_j\|_X$, $j \in \mathbb{N}$.

Next, we consider a smoothness space $\mathcal{X}^s$ and recall $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphy of $u : U \to \mathcal{X}^s : \boldsymbol{y} \mapsto u(\boldsymbol{y})$. First we recall the definition of Kondratiev spaces: Let $k \in \mathbb{N}_0$ and $\zeta \in \mathbb{R}$, and $r_{\mathrm{D}} : \mathrm{D} \to \mathbb{R}_{>0}$ be a smooth function which near vertices of D equals the distance to the closest vertex. Then,

$$\mathcal{K}^k_\zeta(\mathrm{D}) := \left\{ u : \mathrm{D} \to \mathbb{C} : r_{\mathrm{D}}^{|\boldsymbol{\xi}|-\zeta} \partial_x^{\boldsymbol{\xi}} u \in L^2(\mathrm{D}), \boldsymbol{\xi} \in \mathbb{N}_0^2, |\boldsymbol{\xi}| \le k \right\}. \tag{39}$$

To obtain the approximation rate $\gamma = \frac{1}{2}$ in Proposition 5.3, we consider $\mathcal{X}^s := \mathcal{K}^2_{\zeta+1}(\mathrm{D})$ for some $\zeta \in (0,1)$. By [5, Theorem 1.1] and [84, Example 1.2.38], there exists $\zeta \in (0,1)$ such that when $f \in \mathcal{K}^0_{\zeta-1}(\mathrm{D})$, $a \in W^{1,\infty}(\mathrm{D}) =: X^s$ and $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a(\boldsymbol{x})) > 0$, the map $G : a \mapsto u(a) \in \mathcal{X}^s$ is locally well-defined and holomorphic around every such a. We remark that the space from which we chose f satisfies $L^2(\mathrm{D}) \subset \mathcal{K}^0_{\zeta-1}(\mathrm{D}) \subset H^{-1}(\mathrm{D}) = \mathcal{Y}'$.

If in addition to previously made assumptions, $\{\psi_j\}_{j\in\mathbb{N}}$ satisfies

$$\sum_{j\in\mathbb{N}} \|\psi_j\|_{X^s} < \infty,$$

then Lemma 3.5 yields $\boldsymbol{y} \mapsto u(\boldsymbol{y}) = G(a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j)$ to be $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphic for some $\varepsilon^s > 0$ and with $b_j^s := \|\psi_j\|_{X^s}$, $j \in \mathbb{N}$. For a more detailed discussion of this example and more general advection-diffusion-reaction equations, see [84, Section 4.3].

Thus, for the map $U \to \mathcal{X}^s \subset \mathcal{X} : \boldsymbol{y} \mapsto u(\boldsymbol{y})$ to be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$- and $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphic for $\boldsymbol{b} \in \ell^p(\mathbb{N})$ and $\boldsymbol{b}^s \in \ell^{p^s}(\mathbb{N})$ for some $0 < p < p^s < 1$, we additionally need to assume that $(\|\psi_j\|_X)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ and $(\|\psi_j\|_{X^s})_{j\in\mathbb{N}} \in \ell^{p^s}(\mathbb{N})$. The $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphy and Theorem 3.7 give $(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s}\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p^s}$.

In summary, the assumptions on u in Theorem 5.2 hold when $f \in L^2(\mathrm{D})$ and $a_0, \{\psi_j\}_{j\in\mathbb{N}} \subset W^{1,\infty}(\mathrm{D})$ satisfy $\operatorname{ess\,inf}_{\boldsymbol{x}\in\mathrm{D}} \Re(a_0(\boldsymbol{x})) > 0$, Equation (38), $(\|\psi_j\|_X)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ and $(\|\psi_j\|_{X^s})_{j\in\mathbb{N}} \in \ell^{p^s}(\mathbb{N})$. Then, $u : U \to \mathcal{X}^s = \mathcal{K}^2_{\zeta+1}(\mathrm{D})$ for some $\zeta \in (0, 1)$. As mentioned below Assumption 5.1, the NN approximations in Section 5.2.1 satisfy Assumption 5.1 with $\theta = 0$ and approximation rate $\gamma = \frac{1}{2}$.

5.2.3 Parametric eigenvalue problem

We verify the assumptions of Theorem 5.2 for the parametric eigenvalue problem (12). To this end, we choose $\mathcal{X} := \mathbb{C} \times H^1_0(\mathrm{D}; \mathbb{C})$, $X := L^\infty(\mathrm{D}; \mathbb{C})$.

Then, the parametric first eigenpair $\{(\lambda_1(\boldsymbol{y}), w_1(\boldsymbol{y})) : \boldsymbol{y} \in U\} \subset \mathcal{X}$ admits a unique, holomorphic continuation $\{(\lambda_1(\boldsymbol{z}), w_1(\boldsymbol{z})) : \boldsymbol{z} \in V\} \subset \mathcal{X}$ to an open neighborhood V of U in $\mathbb{C}^{\mathbb{N}}$. The proof follows from the uniformity of the spectral gap of the parametric first and second eigenvalues, i.e. from $\lambda_2(\boldsymbol{y}) - \lambda_1(\boldsymbol{y}) > c_0$ for all $\boldsymbol{y} \in U$ and some $c_0 > 0$ which is shown in [29, Proposition 2.4]. Also see [1, Theorem 4] for a proof of analytic dependence on each y_j. Upon defining the parametric "right-hand side" $f(\boldsymbol{y}) := \lambda_1(\boldsymbol{y})w_1(\boldsymbol{y}) \in H^1_0(\mathrm{D}; \mathbb{R}) \subset L^2(\mathrm{D})$ for $\boldsymbol{y} \in U$, it follows that the map $u := (\lambda_1, w_1) \in \mathcal{X}$ satisfies $u : U \to \mathcal{X}^s = \mathbb{C} \times \mathcal{K}^2_{\zeta+1}(\mathrm{D})$ for some $\zeta \in (0, 1)$. It is, in addition, $(\boldsymbol{b}, \varepsilon, \mathcal{X})$- and $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphic for $\boldsymbol{b} \in \ell^p(\mathbb{N})$ and $\boldsymbol{b}^s \in \ell^{p^s}(\mathbb{N})$ for some $0 < p < p^s < 1$, $\varepsilon, \varepsilon^s > 0$, provided $(\|\psi_j\|_X)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ and $(\|\psi_j\|_{X^s})_{j\in\mathbb{N}} \in \ell^{p^s}(\mathbb{N})$. As before, $X^s = W^{1,\infty}(\mathrm{D})$. This $(\boldsymbol{b}, \varepsilon, \mathcal{X})$- and $(\boldsymbol{b}^s, \varepsilon^s, \mathcal{X}^s)$-holomorphy was proved in [1, Theorem 4 and Corollary 2] (where for simplicity the corollary was stated for the special case that D is convex).

6 Application to Bayesian inference

6.1 ReLU DNN approximations for inverse UQ

In this section we discuss how the results in Section 4.4 apply to Bayesian inverse problems from Sections 2.1 and 2.2.1.

In practice it is more convenient to work with measures on U, instead of their pushforwards under the map $\boldsymbol{y} \mapsto a(\boldsymbol{y}) := a_0 + \sum_{j\in\mathbb{N}} y_j\psi_j \in X$ on the Banach space X. For this reason, throughout this section we adopt the equivalent viewpoint of interpreting $\boldsymbol{y} \in U$ (instead of $a(\boldsymbol{y})$) as the unknown, μ_U as the prior, and $a^{-1}\sharp\mu^\delta$ as the posterior measure on U (which is the measure of the unknown $\boldsymbol{y} \in U$ conditioned on the data δ). Here, we assume that $a : \boldsymbol{y} \mapsto a(\boldsymbol{y})$ is invertible and that a^{-1} is measurable, and denote by $a^{-1}\sharp\mu^\delta$ the pushforward measure of μ^δ under a^{-1} (which is a measure on U).[7]

Corollary 6.1. *Let u be $(\boldsymbol{b},\varepsilon,\mathcal{X})$-holomorphic, $\boldsymbol{b} \in \ell^p$, $p \in (0,1)$, and assume the observation noise covariance $\Gamma \in \mathbb{R}^{K\times K}$ is symmetric, positive definite. Let the observation operator $\mathcal{O} : \mathcal{X} \to \mathbb{R}^K$ be deterministic, bounded and linear, let μ_U be the uniform measure on $U = [-1,1]^{\mathbb{N}}$, and let for a given data sample $\delta \in \mathbb{R}^K$*

$$\frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U}(\boldsymbol{y}) = \frac{1}{Z(\delta)}\exp(-\tfrac{1}{2}\|\delta - \mathcal{O}(u(\boldsymbol{y}))\|_\Gamma^2), \qquad \text{for all } \boldsymbol{y} \in U,$$

$$Z(\delta) = \int_U \exp(-\tfrac{1}{2}\|\delta - \mathcal{O}(u(\boldsymbol{y}))\|_\Gamma^2)\mathrm{d}\mu_U(\boldsymbol{y}).$$

Then also $\frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U}(\boldsymbol{y})$ is $(\boldsymbol{b},\varepsilon,\mathbb{R})$-holomorphic.

By Theorem 4.9 it can thus be uniformly approximated by ReLU NNs, with a convergence rate (in terms of the size of the network) arbitrarily close to $1/p - 1$.

Proof. The function $\frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U} : U \to \mathbb{R}$ can be expressed as the composition of the maps

$$\boldsymbol{y} \mapsto u(\boldsymbol{y}), \qquad u \mapsto \tfrac{1}{2}(\delta - \mathcal{O}(u))^\top\Gamma^{-1}(\delta - \mathcal{O}(u)), \qquad a \mapsto \exp(-a). \tag{40}$$

The first map is $(\boldsymbol{b},\varepsilon,\mathcal{X})$ holomorphic, the second map is a holomorphic mapping from $\mathcal{X} \to \mathbb{C}$, and the third map is holomorphic from $\mathbb{C} \to \mathbb{C}$. The composition is $(\boldsymbol{b},\varepsilon,\mathbb{R})$-holomorphic. The rest of the statement follows by Theorem 4.9. □

7 Alternative to looking for the unknown $a(\boldsymbol{y})$ in the Banach space X, we could interpret $\boldsymbol{y} \in U$ to be the unknown. In this case the posterior measure is defined on U (instead of X), and the assumption of invertibility of a, which is used to push forward μ^δ to a measure on U, would not be necessary.

In case the number of parameters $N \in \mathbb{N}$ is finite, *exponential convergence rates of ReLU DNN approximations* follow with [62, Theorem 3.6], but with the rate of convergence and other constants in the error bound depending on N.

For the approximation of the posterior expectation $Y \to Z : \delta \mapsto \mathbb{E}[Q \circ u|\delta]$, holomorphy of the posterior density implies holomorphy of the posterior expectation, but without control on the size of the domain of holomorphy. Thus [62, Theorem 3.6] gives exponential convergence with rate $C\exp(-b\mathcal{N}^{1/(K+1)})$, with possibly very small $b > 0$, in terms of the NN size $\mathcal{N}$. We remark that holomorphy of the data-to-QoI map is valid even for non-holomorphic input-to-response maps in the operator equation [37]. In [37], this was exploited by considering a *rational approximation of the Bayesian estimate* based on

$$\delta \mapsto \mathbb{E}[Q \circ u|\delta] = \int_{\tilde{X}} Q(u(a)) \frac{1}{Z(\delta)} \exp(-\Phi(a;\delta)) \mathrm{d}\mu_0(a) =: Z'(\delta)/Z(\delta),$$

where Z, Z' are entire functions of δ, i.e. they admit a holomorphic extension to $\mathbb{C}^K$. With that argument, convergence rates of the form $C\exp(-b\mathcal{N}^{1/(K+1)})$ with arbitrarily large $b > 0$ were obtained.

6.2 Posterior concentration

We consider the DNN expression of posterior densities in Bayesian inverse problems when the posterior density concentrates near a single point, the so-called *maximum a posteriori point (MAP point)*, at which the posterior density attains its maximum.

We consider in particular the case in which the posterior density exists, is unimodal, attaining its global maximum at the MAP point. In the mentioned scaling regimes, in the vicinity of the MAP point, the Bayesian posterior density is close to a Gaussian distribution with covariance matrix Γ, which arises in either the small noise or in the large data limits, cf. e.g. [74, 44]. We therefore study the behavior of the DNN expression rate bounds as $\Gamma \downarrow 0$. This limit applies to the situation of decreasing observation noise η or of increasing observation size $\dim(Y)$.

The results in Section 6.1 hold for all symmetric, positive definite covariance matrices Γ, but constants depend on Γ and may tend to infinity as $\Gamma \downarrow 0$. However, the concentration can be exploited for the approximation of the posterior density. As an example, we consider an inverse problem with $N < \infty$ parameters, with a holomorphic forward map $[-1,1]^N \to \mathcal{X} : \boldsymbol{y} \to u(\boldsymbol{y})$, a linear observation functional $\mathcal{O} : \mathcal{X} \to Y$ and a finite observation size $K := \dim(Y) < \infty$. In [73, Theorem 4.1], in case of a non-degenerate Hessian $\Phi_{\boldsymbol{y},\boldsymbol{y}}$ it was shown that after a Γ-dependent affine transformation the posterior density is analytic with polyradii of analyticity independent of Γ. Hence, by [62, Theorem 3.6], NN approximations of the posterior

density converge exponentially (albeit with constants depending exponentially on N).

Moreover, in [74, Appendix] it was shown that under suitable conditions a Gaussian distribution approximates the posterior density up to first order in Γ. This allows us to overcome the curse of dimensionality in terms of N for the unnormalized posterior density, by exploiting the radial symmetry of the Gaussian density function. By [63, Theorem 6.7], the Gaussian density function can be approximated by ReLU NNs with the network size growing polylogarithmically with the error, and the corresponding constants increasing at most quadratically in N. Thus, there is no curse of dimensionality for the approximation of the unnormalized posterior density when it concentrates near one point. Note that this ignores the consistency error of the posterior density with respect to this Gaussian approximation to the true posterior density. If the posterior concentrates near multiple well-separated points, and if it is close to a Gaussian near each of the points, then it can be approximated at the same rate by a sum of (localized) Gaussians.

The next proposition gives an approximation result for unnormalized Gaussian densities. We refer to Appendix A.5 for a proof.

Proposition 6.2. *For $N \in \mathbb{N}$, let $A : \mathbb{R}^N \to \mathbb{R}^N$ be a bijective linear map. For $\boldsymbol{x} \in \mathbb{R}^N$ set $\mathfrak{g}(\boldsymbol{x}) := \exp(-\frac{1}{2}\|A\boldsymbol{x}\|_2^2)$.*

Then, there exists $C > 0$ independent of A and N such that for every $\varepsilon \in (0,1)$ there exists a ReLU NN $\Phi_\varepsilon^{\mathfrak{g}}$ satisfying

$$\begin{aligned}\left\|\mathfrak{g} - \Phi_\varepsilon^{\mathfrak{g}}\right\|_{L^\infty(\mathbb{R}^N)} &\leq C\varepsilon = C\varepsilon \left\|\mathfrak{g}\right\|_{L^\infty(\mathbb{R}^N)},\\ \operatorname{depth}(\Phi_\varepsilon^{\mathfrak{g}}) &\leq C\left(\log(N)(1+\log(N/\varepsilon)) + 1 + \log(1/\varepsilon)\log\log(1/\varepsilon)\right),\\ \operatorname{size}(\Phi_\varepsilon^{\mathfrak{g}}) &\leq C\left((1+\log(1/\varepsilon))^2 + N\log(1/\varepsilon) + N^2\right).\end{aligned}$$

Remark 6.3. *The term CN^2 in the bound on the network size follows from bounding the number of nonzero coefficients in the linear map A by N^2. If A has at most CN nonzero coefficients, the network size is of the order $N\log(N)$.*

Densities of the type $\mathfrak{g}(\boldsymbol{x}) = \exp(-\frac{1}{2}\|A(\boldsymbol{x})\|_2^2)$ need to be normalized in order to become probability densities on $[-1,1]^N$. We now discuss an example to show the effect of the normalization constant on the approximation result, when the density concentrates.

Fix an observation noise covariance $\Gamma \in \mathbb{R}^{N\times N}$ symmetric positive definite, and for $n \in \mathbb{N}$ set $\Gamma_n := \Gamma/n$ and $\tilde{\mathfrak{g}}_n(\boldsymbol{x}) = \exp(-\frac{1}{2}\|\Gamma_n^{-1/2}\boldsymbol{x}\|_2^2)$ for $\boldsymbol{x} \in [-1,1]^N$. Given $\delta \in [-1,1]^N$, note that as $n \to \infty$, the unnormalized density $\tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)$ concentrates around $\delta \in [-1,1]^N$. For any $n \geq 1$, using the change of variables

$\boldsymbol{y} = \sqrt{n}\boldsymbol{x}$, we bound the normalization constant from below:

$$\begin{aligned}\int_{[-1,1]^N} \tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)\mathrm{d}\boldsymbol{x} &= \int_{[-1,1]^N} \exp\left(-\frac{1}{2}\|\sqrt{n}\Gamma^{-1/2}(\boldsymbol{x}-\delta)\|_2^2\right)\mathrm{d}\boldsymbol{x}\\ &= n^{-N/2}\int_{[-\sqrt{n},\sqrt{n}]^N} \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(\boldsymbol{y}-\sqrt{n}\delta)\|_2^2\right)\mathrm{d}\boldsymbol{y}\\ &\geq n^{-N/2}\inf_{\tilde{\delta}\in[-1,1]^N}\int_{[-1,1]^N} \exp\left(-\frac{1}{2}\|\Gamma^{-1/2}(\boldsymbol{y}-\tilde{\delta})\|_2^2\right)\mathrm{d}\boldsymbol{y}\\ &= n^{-N/2}C_0,\end{aligned}$$

with $C_0(\Gamma, N) > 0$ denoting the infimum in the second to last line, and where we used $\sqrt{n}\delta \in [-\sqrt{n}, \sqrt{n}]^N$.

Denote $Z_n(\delta) := \int_{[-1,1]^N} \tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)\mathrm{d}\boldsymbol{x} \geq C_0 n^{-N/2}$. Then, by Proposition 6.2 the normalized density $\mathfrak{g}_n(\boldsymbol{x}-\delta) := \tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)/Z_n(\delta) \leq C_0^{-1} n^{N/2}\tilde{\mathfrak{g}}_n(\boldsymbol{x}-\delta)$ can be uniformly approximated on $[-1,1]^N$ to accuracy $\varepsilon > 0$ with a ReLU network $\Phi_\varepsilon^{\mathfrak{g}_n}$ of size and depth bounded as follows, for $C(\Gamma, N) > 0$:

$$\begin{aligned}\mathrm{depth}(\Phi_\varepsilon^{\mathfrak{g}_n}) &\leq C\left(1+\left(\log(1/\varepsilon)+(1+\log(n))\right)\log\left(\log(1/\varepsilon)+(1+\log(n))\right)\right),\\ \mathrm{size}(\Phi_\varepsilon^{\mathfrak{g}_n}) &\leq C\left((1+\log(1/\varepsilon))^2+\log(1/\varepsilon)(1+\log_2(n))+(1+\log_2(n))^2\right).\end{aligned}$$

6.3 Posterior consistency

In Section 6.1 we proved $L^\infty(U)$-bounds on the approximation of the posterior density with NNs. Up to a constant, this immediately yields the same bounds for the Hellinger and total variation distances of the corresponding (normalized) Bayesian posterior measures as we show next.

Let λ be the Lebesgue measure on $[-1,1]$, and denote again by $\mu_U := \otimes_{j\in\mathbb{N}}\frac{\lambda}{2}$ the uniform probability measure on $U = [-1,1]^{\mathbb{N}}$ equipped with the product sigma algebra. Let $\mu \ll \mu_U$ and $\nu \ll \mu_U$ be two measures on U with Radon-Nikodym derivatives $\frac{\mathrm{d}\mu}{\mathrm{d}\mu_U} =: \pi_\mu : U \to \mathbb{R}$ and $\frac{\mathrm{d}\nu}{\mathrm{d}\mu_U} =: \pi_\nu : U \to \mathbb{R}$. Recall that the Hellinger distance (which we use here also for non-probability measures) is defined as

$$d_H(\mu,\nu) = \left(\frac{1}{2}\int_U\left(\sqrt{\pi_\mu(\boldsymbol{y})}-\sqrt{\pi_\nu(\boldsymbol{y})}\right)^2\mathrm{d}\mu_U(\boldsymbol{y})\right)^{1/2} = \tfrac{1}{\sqrt{2}}\|\sqrt{\pi_\mu}-\sqrt{\pi_\nu}\|_{L^2(U,\mu_U)}.$$

The total variation distance is defined as

$$d_{TV}(\mu,\nu) = \sup_B|\mu(B)-\nu(B)| \leq \int_U|\pi_\mu(\boldsymbol{y})-\pi_\nu(\boldsymbol{y})|\mathrm{d}\mu_U(\boldsymbol{y}) = \|\pi_\mu-\pi_\nu\|_{L^1(U,\mu_U)},$$

where the supremum is taken over all measurable $B \subseteq U$. Thus

$$d_{TV}(\mu,\nu) \le \|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}.$$

Since $|\sqrt{x}-\sqrt{y}| = \frac{|x-y|}{|\sqrt{x}+\sqrt{y}|}$ for all $x, y \ge 0$,

$$d_H(\mu,\nu) = \tfrac{1}{\sqrt{2}}\|\sqrt{\pi_\mu} - \sqrt{\pi_\nu}\|_{L^2(U,\mu_U)} \le \frac{\|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}}{\sqrt{2}\inf_{\boldsymbol{y}\in U}(\sqrt{\pi_\mu}(\boldsymbol{y}) + \sqrt{\pi_\nu}(\boldsymbol{y}))}.$$

Denote by $\overline{\mu} = \frac{\mu}{\mu(U)}$ and $\overline{\nu} = \frac{\nu}{\nu(U)}$ the normalized measures and by $\overline{\pi}_\mu$, $\overline{\pi}_\nu$ the corresponding densities (which are probability densities w.r.t. μ_U). Then for all $\boldsymbol{y} \in U$

$$\begin{aligned}|\overline{\pi}_\mu(\boldsymbol{y}) - \overline{\pi}_\nu(\boldsymbol{y})| &= \left|\frac{\pi_\mu(\boldsymbol{y})}{\mu(U)} - \frac{\pi_\nu(\boldsymbol{y})}{\nu(U)}\right| \\ &\le \frac{|\pi_\mu(\boldsymbol{y})\nu(U) - \pi_\nu(\boldsymbol{y})\nu(U)| + |\pi_\nu(\boldsymbol{y})\nu(U) - \pi_\nu(\boldsymbol{y})\mu(U)|}{\mu(U)\nu(U)}.\end{aligned}$$

Using $|\mu(U)-\nu(U)| \le \|\pi_\mu - \pi_\nu\|_{L^1(U,\mu_U)}$ we obtain for all $\boldsymbol{y} \in U$

$$|\overline{\pi}_\mu(\boldsymbol{y}) - \overline{\pi}_\nu(\boldsymbol{y})| \le \frac{\|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}\nu(U) + \|\pi_\nu\|_{L^\infty(U,\mu_U)}\|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}.$$

By symmetry this implies

$$d_{TV}(\overline{\mu},\overline{\nu}) \le \|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)} \min\left(\frac{\nu(U) + \|\pi_\nu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}, \frac{\mu(U) + \|\pi_\mu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}\right) \tag{41a}$$

and similarly as before

$$d_H(\overline{\mu},\overline{\nu}) \le \|\pi_\mu - \pi_\nu\|_{L^\infty(U,\mu_U)} \frac{\min\left(\frac{\nu(U)+\|\pi_\nu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}, \frac{\mu(U)+\|\pi_\mu\|_{L^\infty(U,\mu_U)}}{\nu(U)\mu(U)}\right)}{\sqrt{2}\inf_{\boldsymbol{y}\in U}(\sqrt{\overline{\pi}_\mu}(\boldsymbol{y}) + \sqrt{\overline{\pi}_\nu}(\boldsymbol{y}))}. \tag{41b}$$

Proposition 6.4. *Consider the setting of Corollary 6.1. Then for every $\tau \in (0,1)$ there exists a σ_1-NN $f_\tau : U \to [0,\infty)$ (with input variables $(y_j)_{j\in S_{\Lambda_\tau}}$) such that with Λ_τ as in Theorem 3.7*

$$\begin{aligned}\operatorname{size}(f_\tau) &\le C(1 + |\Lambda_\tau| \cdot \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|), \\ \operatorname{depth}(f_\tau) &\le C(1 + \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|)\end{aligned} \tag{42}$$

and the measure ν_τ on U with density $f_\tau = \frac{\mathrm{d}\nu_\tau}{\mathrm{d}\mu_U}$ satisfies

$$d_H\left(a^{-1}\sharp\mu^\delta, \bar{\nu}_\tau\right) \le C|\Lambda_\tau|^{-\frac{1}{p}+1}, \tag{43}$$

and the same bound holds w.r.t. d_{TV}.

Proof. By Corollary 6.1 and Theorem 4.9 there exists a σ_1-NN $\tilde{f}_\tau : U \to \mathbb{R}$ satisfying (42) such that with $f(\boldsymbol{y}) := \frac{\mathrm{d}a^{-1}\sharp\mu^\delta}{\mathrm{d}\mu_U} = \frac{1}{Z(\delta)}\exp(-\frac{1}{2}\|\delta - \mathcal{O}(u(\boldsymbol{y}))\|_\Gamma^2)$, where u is $(\boldsymbol{b},\varepsilon)$-holomorphic, holds

$$\|f - \tilde{f}_\tau\|_{L^\infty(U)} \leq C|\Lambda_\tau|^{-\frac{1}{p}+1}. \tag{44}$$

Let $f_\tau := \sigma_1(\tilde{f}_\tau)$. Then $f_\tau : U \to [0,\infty)$ and the bound (44) remains true for f_τ since $f(\boldsymbol{y}) \geq 0$ for all $\boldsymbol{y} \in U$.

Since any $(\boldsymbol{b},\varepsilon)$-holomorphic function is continuous on U and because $f(\boldsymbol{y}) > 0$ for all $\boldsymbol{y} \in U$, we have $\inf_{\boldsymbol{y}\in U} f(\boldsymbol{y}) > 0$ and $\sup_{\boldsymbol{y}\in U} f(\boldsymbol{y}) < \infty$. Thus (41) implies (43) for d_{TV} and d_H. □

7 Conclusions and further directions

In this paper we presented dimension independent expression rates for the approximation of infinite-parametric functions occurring in forward and inverse UQ by deep neural networks. Our results are based on multilevel gpc expansions, and generalize the statements of [77] in that they do not require analytic extensions of the target function to complex polydiscs, but merely to complex polyellipses. Additionally, while for $\mathcal{X}$-valued functions [77] only treated the case of $\mathcal{X} = H^1([0,1])$, here we considered $\mathcal{X} = W^{1,q}(\mathrm{D})$, with D being a bounded polytope, for example. It was shown that our theory also comprises analyticity of parametric maps in scales of corner-weighted Sobolev spaces in D, allowing to retain optimal convergence rates of FEM in the presence of corner singularities of the PDE solution. These generalizations allow to treat much broader problem classes, comprising for example a forward operator mapping inputs to the solution of the parametric (nonlinear) Navier-Stokes equations [17]. Another instance includes domain uncertainty, which typically does not yield forward operators with holomorphic parameter dependence on polydiscs, see e.g. [38].

As one possible application of our results, we treated in more detail the approximation of posterior densities in Bayesian inference. Having cheaply evaluable surrogates of this density (in the form of a DNN) can be a powerful tool, as any inference technique could require thousands of evaluations of the posterior density. On top of that, in case of MCMC, arguably the most widely used inference algorithm, these evaluations are inherently sequential and not parallel. Each such evaluation requires a (time-consuming, approximate) computation of a PDE solution, which can render MCMC infeasible in practice. Variational inference, on the other hand, where sampling from the posterior is replaced by an optimization problem, does not necessarily require sequential computation of (approximate) PDE solutions,

however it still demands a high number of evaluations of the posterior, which may be significantly sped up if this posterior is replaced by a cheap surrogate. We refer for example to transport based methods such as [52].

As already indicated in the introduction of the present article, the idea of using DNNs for expressing the input-to-response map (i.e., the "forward" map) for PDE models has been proposed repeatedly in recent years. The motivation for this is the nonlinearity of such maps, even for linear PDEs, and the often high regularity (e.g. holomorphy) of such maps. Here, DNNs are a computational tool alongside other reduction methods, such as reduced basis (RB for short) or Model Order Reduction methods (MOR for short). Indeed, in [45, Remark 4.6] it has been suggested that under the provision that reduced bases for a compact solution manifold of a linear, elliptic parametric PDE admit an efficient DNN expression, so does the input-to-solution map of this PDE. The abstract, Lipschitz dependence result Theorem 2.8 (which is [21, Theorem 18]) will imply with the present results and the DNN expression results of RB/MOR approximations for forward PDE problems as developed in [45] analogous results also for the corresponding Bayesian inverse problems considered in the present paper. MOR and RB approaches can be developed along the lines of [11], where BIP subject RB/MOR approximation of the forward, input-to-response maps were considered in conjunction with Bayesian inverse problems of the type considered here. Should reduced bases admit good DNN expression rates, the analysis of [11] would imply with the present results corresponding improved DNN expression rates, along the lines of [45].

We remark that the DNN expression rate bounds for the posterior densities are obtained from DNN reapproximation of gpc surrogates. DNN expression rate bounds follow from the corresponding approximation rates of N-term truncated gpc expansions. These, in turn, are based on gpc coefficient estimates which were obtained as e.g. in [77] by analytic continuation of parametric solution families into the complex domain. Analytic continuation can be avoided if, instead, real-variable induction arguments for bounding derivatives of parametric solutions are employed. We refer to [32] for forward UQ in an elliptic control problem, and to [35, Section 7] for a proof of derivative bounds for the Bayesian posterior with Gaussian prior. As in [77], the present DNN expression rate analysis relies on "intermediate" polynomial chaos approximations of the posterior density, assuming a prior given by the uniform probability measure on $U = [-1, 1]^{\mathbb{N}}$. The emulation of the posterior density by DNNs can leverage, however, the *compositional structure* of DNNs to accommodate changes of (prior) probability, with essentially the same expression rates, as long as the changes of measure can be emulated efficiently by DNNs. This may include nonanalytic / nonholomorphic densities. We refer to [62, Section 4.3.5] for an example.

We also showed in Section 6.2 that ReLU DNN expression rates are either independent of or depend only logarithmically on concentration in the posterior density, provided the concentration happens only in a finite number of 'informed' variables, and the posterior density is of 'MAP' type, in particular (locally) unimodal. While important, this is only a rather particular special case in applications, where oftentimes posterior concentration occurs along smooth submanifolds. In such cases, ReLU DNNs can also be expected to exhibit robust expression rates, according to the expression rate bounds in [67, Section 5]. Details are to be developed elsewhere.

A Proofs

A.1 Proof of Theorem 3.7

Proof. Since $(b_j)_{j\in\mathbb{N}} \in \ell^p(\mathbb{N})$ it holds $b_j \to 0$. Thus we can find $\kappa > 1$ so small and $J \in \mathbb{N}$ so large that with $C_0 = 4/9$

$$\begin{aligned} &\sup_{j>J} b_j^{1-p} < \frac{1}{2}, \\ &(\kappa-1)\sum_{j\in\mathbb{N}} b_j + C_0^{-1}\max\left\{3, \frac{2e}{\varepsilon}\right\}\max\left\{\sum_{j>J} b_j, \sum_{j>J} b_j^p\right\} < \min\left\{1, \frac{\varepsilon}{2}\right\}. \end{aligned} \tag{45}$$

We fix such values for J and κ throughout the proof.

Step 1. We give an upper bound for $\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}$. First, recall that by Cauchy's integral formula, for any holomorphic function $f : B_r^{\mathbb{C}} \to \mathcal{X}$ we have for any $0 < \tilde{r} < r$ and any $k \in \mathbb{N}_0$

$$f^{(k)}(0) = \frac{k!}{2\pi \mathrm{i}} \int\limits_{\{\zeta\in\mathbb{C}\,:\,|\zeta|=\tilde{r}\}} \frac{f(\zeta)}{\zeta^{1+k}} \mathrm{d}\zeta,$$

where the circle $\{\zeta \in \mathbb{C} : |\zeta| = \tilde{r}\}$ in the line integral is oriented positively. Therefore

$$\frac{\|f^{(k)}(0)\|_{\mathcal{X}}}{k!} \leq \frac{1}{r^k} \sup_{z\in B_r^{\mathbb{C}}} \|f(z)\|_{\mathcal{X}}. \tag{46}$$

Similarly, as shown in [22, Section 12.4] (also see the proof of [13, Theorem 2.2]), for any $r > 1$ and any $k \in \mathbb{N}_0$ and for a holomorphic function $f : \mathcal{E}_r \to \mathcal{X}$

$$\left\| \int\limits_{-1}^{1} f(y) L_k(y) \frac{\mathrm{d}y}{2} \right\|_{\mathcal{X}} \leq \frac{\pi(1+2k)}{2(r-1)} \frac{1}{r^k} \sup_{x\in\mathcal{E}_r} \|f(x)\|_{\mathcal{X}}. \tag{47}$$

We will now use these estimates to obtain an upper bound for $\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}$.

Fix $\boldsymbol{\nu} = (\boldsymbol{\nu}_E, \boldsymbol{\nu}_F) \in \mathcal{F}$ and define

$$\rho_j := \begin{cases} \kappa & \text{if } j \le J, \\ \max\left\{3, \frac{\varepsilon}{2}\frac{\nu_j}{b_j|\boldsymbol{\nu}_F|}\right\} & \text{if } j > J, \end{cases}$$

where $\nu_j/|\boldsymbol{\nu}_F| := 0$ if $|\boldsymbol{\nu}_F| = 0$. Then by (45)

$$\sum_{j\in\mathbb{N}}(\rho_j - 1)b_j \le (\kappa-1)\sum_{j=1}^{J} b_j + 3\sum_{j>J} b_j + \sum_{j>J}\frac{\varepsilon}{2}\frac{\nu_j}{b_j|\boldsymbol{\nu}_F|}b_j < \varepsilon,$$

so that $\boldsymbol{\rho} = (\rho_j)_{j\in\mathbb{N}}$ is $(\boldsymbol{b}, \varepsilon)$-admissible in the sense of Definition 3.3. Thus, by Definition 3.3, u allows a separately holomorphic extension to $\bigtimes_{j=1}^{J}\mathcal{E}_\kappa \times \bigtimes_{j>J}\mathcal{E}_{\rho_j}$ which contains the set $\bigtimes_{j=1}^{J}\mathcal{E}_\kappa \times \bigtimes_{j>J} B^{\mathbb{C}}_{C_0\rho_j}$ by Lemma 3.6, and it holds

$$\sup_{(\boldsymbol{y}_E,\boldsymbol{y}_F)\in\bigtimes_{j=1}^{J}\mathcal{E}_\kappa\times\bigtimes_{j>J}B^{\mathbb{C}}_{C_0\rho_j}} \|u(\boldsymbol{y}_E,\boldsymbol{y}_F)\|_{\mathcal{X}} \le M, \tag{48}$$

for M as in Definition 3.3.

To find an upper bound for $\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}$, we use that $\|\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E, \boldsymbol{0})\|_{\mathcal{X}}$ is uniformly bounded for all $\boldsymbol{y}_E$ in the compact set U_E (due to the continuous dependence on $\boldsymbol{y}_E$), so that an application of Fubini's theorem (for Bochner integrals) yields

$$\begin{aligned} c_{\boldsymbol{\nu}} &= \int_{U_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E,\boldsymbol{0})}{\boldsymbol{\nu}_F!}\mathrm{d}\mu_E(\boldsymbol{y}_E) \\ &= \int_{-1}^{1} L_{\nu_1}(y_1)\cdots\int_{-1}^{1} L_{\nu_J}(y_J)\frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E,\boldsymbol{0})}{\boldsymbol{\nu}_F!}\frac{\mathrm{d}y_J}{2}\cdots\frac{\mathrm{d}y_1}{2}. \end{aligned}$$

Hence by repeated application of (47)

$$\begin{aligned} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} &\le \frac{\pi(1+2\nu_1)}{2(\kappa-1)}\kappa^{-\nu_1}\sup_{y_1\in\mathcal{E}_\kappa}\left\|\int_{-1}^{1} L_{\nu_2}(y_2)\cdots\int_{-1}^{1} L_{\nu_J}(y_J)\frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E,\boldsymbol{0})}{\boldsymbol{\nu}_F!}\frac{\mathrm{d}y_J}{2}\cdots\frac{\mathrm{d}y_2}{2}\right\|_{\mathcal{X}} \\ &\le \cdots \le \left(\prod_{j=1}^{J}(1+2\nu_j)\right)\left(\frac{\pi}{2(\kappa-1)}\right)^{J}\kappa^{-|\boldsymbol{\nu}_E|}\sup_{\boldsymbol{y}_E\in\bigtimes_{j=1}^{J}\mathcal{E}_\kappa}\left\|\frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E,\boldsymbol{0})}{\boldsymbol{\nu}_F!}\right\|_{\mathcal{X}}. \end{aligned} \tag{49}$$

Next, we bound the last supremum in (49). Using that u allows a separately holomorphic extension satisfying (48), repeated application of (46) gives

$$\sup_{\boldsymbol{y}_E\in\bigtimes_{j=1}^{J}\mathcal{E}_\kappa}\left\|\frac{\partial^{\boldsymbol{\nu}_F}_{\boldsymbol{y}_F} u(\boldsymbol{y}_E,\boldsymbol{0})}{\boldsymbol{\nu}_F!}\right\|_{\mathcal{X}} = \sup_{\boldsymbol{y}_E\in\bigtimes_{j=1}^{J}\mathcal{E}_\kappa}\left\|\frac{\partial^{\nu_{J+1}}_{y_{J+1}}\cdots u(\boldsymbol{y}_E,\boldsymbol{0})}{\prod_{j>J}\nu_j!}\right\|_{\mathcal{X}}$$

$$\leq \sup_{\boldsymbol{y}_E\in\times_{j=1}^J \mathcal{E}_\kappa} \sup_{\boldsymbol{y}_F\in\times_{j>J} B^{\mathbb{C}}_{C_0\rho_j}} \|u(\boldsymbol{y}_E,\boldsymbol{y}_F)\|_{\mathcal{X}} \prod_{j>J}(C_0\rho_j)^{-\nu_j}$$
$$\leq M\prod_{j>J}(C_0\rho_j)^{-\nu_j}. \tag{50}$$

Due to $n^n \geq n! \geq e^{-n}n^n$ for all $n\in\mathbb{N}$ and using $\rho_j \geq \varepsilon\nu_j/(2b_j|\boldsymbol{\nu}_F|)$,

$$\prod_{j>J}(C_0\rho_j)^{-\nu_j} \leq \prod_{j\in\operatorname{supp}\boldsymbol{\nu}_F}\left(\frac{C_0\varepsilon}{2}\frac{\nu_j}{b_j|\boldsymbol{\nu}_F|}\right)^{-\nu_j} = \left(\frac{2}{C_0\varepsilon}\right)^{|\boldsymbol{\nu}_F|}\frac{|\boldsymbol{\nu}_F|^{|\boldsymbol{\nu}_F|}}{\boldsymbol{\nu}_F^{\boldsymbol{\nu}_F}}\boldsymbol{b}_F^{\boldsymbol{\nu}_F}$$
$$\leq \left(\frac{2e}{C_0\varepsilon}\right)^{|\boldsymbol{\nu}_F|}\frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!}\boldsymbol{b}_F^{\boldsymbol{\nu}_F}. \tag{51}$$

Altogether, there exists a constant C such that for any $\boldsymbol{\nu}\in\mathcal{F}$

$$\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \leq C\left(\prod_{j=1}^J(1+2\nu_j)\right)\kappa^{-|\boldsymbol{\nu}_E|}\frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!}\left(\frac{2e}{C_0\varepsilon}\right)^{|\boldsymbol{\nu}_F|}\boldsymbol{b}_F^{\boldsymbol{\nu}_F}. \tag{52}$$

Step 2. We show (i) and the first part of (iii). Fix $\gamma_1,\gamma_2\in(1,2)$ such that $1<\gamma_1\gamma_2<\kappa$. By (16) it holds $\|L_n\|_{L^\infty([-1,1])}\leq(1+2n)^{1/2}$ for all $n\in\mathbb{N}_0$. Thus there exists a constant $C<\infty$ such that for all $\boldsymbol{\nu}_E\in\mathcal{F}_E$

$$\left(\prod_{j=1}^J(1+2\nu_j)\right)\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \leq \gamma_1^{|\boldsymbol{\nu}_E|}\sup_{\boldsymbol{\mu}\in\mathcal{F}_E}\frac{\prod_{j=1}^J(1+2\mu_j)^{3/2}}{\gamma_1^{|\boldsymbol{\mu}|}} \leq C\gamma_1^{|\boldsymbol{\nu}_E|}.$$

Next set

$$\delta_j := \begin{cases}\gamma_2 & \text{if } j\leq J\\ \min\{b_j^{p-1}, j^{2/p}\} & \text{if } j>J.\end{cases} \tag{53}$$

By (45) it holds $b_j^{p-1}>2$ for all $j>J$ and since $\gamma_2<2$ by definition, $(\delta_j)_{j\in\mathbb{N}}$ is monotonically increasing. Furthermore $(\delta_j^{-1})\in\ell^{p/(1-p)}(\mathbb{N})$ since $(b_j)_{j\in\mathbb{N}}\in\ell^p(\mathbb{N})$. Moreover, by definition $\delta_j\leq C_1 j^{2/p}$ for $C_1:=\gamma_2$ and all $j\in\mathbb{N}$. Thus $\boldsymbol{\delta}=(\delta_j)_{j\in\mathbb{N}}$ satisfies the properties stated in (iii).

Now, by (52) and (53)

$$\sum_{\boldsymbol{\nu}\in\mathcal{F}}\boldsymbol{\delta}^{\boldsymbol{\nu}}\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)}\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}$$
$$\leq C\sum_{\boldsymbol{\nu}\in\mathcal{F}}\left(\frac{\kappa}{\gamma_1\gamma_2}\right)^{-|\boldsymbol{\nu}_E|}\frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!}\left(\frac{2e}{C_0\varepsilon}\right)^{|\boldsymbol{\nu}_F|}\boldsymbol{b}_F^{\boldsymbol{\nu}_F}\boldsymbol{b}_F^{(p-1)\boldsymbol{\nu}_F}$$
$$= C\left(\sum_{\boldsymbol{\nu}_E\in\mathcal{F}_E}\left(\frac{\kappa}{\gamma_1\gamma_2}\right)^{-|\boldsymbol{\nu}_E|}\right)\left(\sum_{\boldsymbol{\nu}_F\in\mathcal{F}_F}\frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!}\prod_{j>J}\left(\frac{2eb_j^p}{C_0\varepsilon}\right)^{\nu_j}\right).$$

Due to $\kappa/(\gamma_1\gamma_2) > 1$, the first series is finite according to [15, Lemma 7.1], and the second series is finite according to [15, Theorem 7.2] since

$$\sum_{j>J} \frac{2eb_j^p}{C_0\varepsilon} < 1$$

by (45). This shows (21).

To show (i), we point out that due to $(\delta_j^{-1})_{j\in\mathbb{N}} \in \ell^{p/(1-p)}(\mathbb{N})$ and $\sup_{j\in\mathbb{N}} \delta_j^{-1} \leq \gamma_2^{-1} < 1$, [15, Lemma 7.1] implies $(\boldsymbol{\delta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^{p/(1-p)}(\mathcal{F})$. Hence applying Hölder's inequality

$$\begin{aligned}
&\sum_{\boldsymbol{\nu}\in\mathcal{F}} (\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)}\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})^p \\
&= \sum_{\boldsymbol{\nu}\in\mathcal{F}} (\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)}\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}\boldsymbol{\delta}^{\boldsymbol{\nu}}\boldsymbol{\delta}^{-\boldsymbol{\nu}})^p \\
&\leq \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)}\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}\boldsymbol{\delta}^{\boldsymbol{\nu}}\right)^p \left(\sum_{\boldsymbol{\nu}\in\mathcal{F}} (\boldsymbol{\delta}^{-\boldsymbol{\nu}})^{\frac{p}{1-p}}\right)^{1-p} < \infty.
\end{aligned}$$

Step 3. We show (ii). Fix $\boldsymbol{y}_E \in U_E$. Then, since $(\kappa-1)\sum_{j=1}^J b_j + 3\sum_{j>J} b_j < \varepsilon$ by (45), for every $\boldsymbol{y}_E \in U_E$, the map $\boldsymbol{y}_F \mapsto u(\boldsymbol{y}_E, \boldsymbol{y}_F)$ is separately holomorphic as a function of $\boldsymbol{y}_F \in \bigtimes_{j>J} B_{3C_0}^{\mathbb{C}}$ by Definition 3.3. Note that $3C_0 = 12/9 > 1$, and by (45) we can find $\theta \in (1, 3C_0)$ such that

$$\sum_{j>J} \frac{2e\theta b_j}{C_0\varepsilon} < 1.$$

Then, again by [15, Theorem 7.2] and (50), (51) it holds

$$\sum_{\boldsymbol{\nu}_F\in\mathcal{F}_F} \theta^{|\boldsymbol{\nu}_F|} \left\| \frac{\partial_{\boldsymbol{y}_F}^{\boldsymbol{\nu}_F} u(\boldsymbol{y}_E, \mathbf{0})}{\boldsymbol{\nu}_F!} \right\|_{\mathcal{X}} \leq \sum_{\boldsymbol{\nu}_F\in\mathcal{F}_F} \frac{|\boldsymbol{\nu}_F|!}{\boldsymbol{\nu}_F!} \prod_{j>J} \left(\frac{2e\theta b_j}{C_0\varepsilon}\right)^{\nu_j} < \infty. \tag{54}$$

This and the fact that $u : U \to \mathcal{X}$ is continuous by Definition 3.3 implies by [84, Proposition 2.1.5] and [84, Remark 2.1.7] that for all $\boldsymbol{y}_F \in U_F$

$$u(\boldsymbol{y}_E, \boldsymbol{y}_F) = \sum_{\boldsymbol{\nu}_F\in\mathcal{F}_F} \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \frac{\partial_{\boldsymbol{y}_F}^{\boldsymbol{\nu}_F} u(\boldsymbol{y}_E, \mathbf{0})}{\boldsymbol{\nu}_F!}$$

with uniform and absolute (i.e. the norms are summable) convergence for all $\boldsymbol{y}_F \in U_F$.

Next fix $\boldsymbol{y}_F \in U_F$. Then, since $(\kappa-1)\sum_{j\in\mathbb{N}} b_j < \varepsilon$, the map $\boldsymbol{y}_E \mapsto u(\boldsymbol{y}_E, \boldsymbol{y}_F)$ is separately holomorphic on $\boldsymbol{y}_E \in \bigtimes_{j=1}^J \mathcal{E}_\kappa$ and with $\sup_{\boldsymbol{y}_E\in\times_{j=1}^J \mathcal{E}_\kappa} \|u(\boldsymbol{y}_E, \boldsymbol{y}_F)\|_{\mathcal{X}} \leq$

M. As in (49) this allows us to show that there exists a constant C (not depending on $\boldsymbol{y}_F$) such that

$$\left\| \int_{U_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) u(\boldsymbol{y}_E, \boldsymbol{y}_F) \mathrm{d}\mu_E(\boldsymbol{y}_E) \right\|_{\mathcal{X}} \leq C\kappa^{-|\boldsymbol{\nu}_E|} \left(\prod_{j=1}^{J} (1+2\nu_j) \right).$$

By similar arguments as in Step 1 we then get

$$\sum_{\boldsymbol{\nu}_E \in \mathcal{F}_E} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \left\| \int_{U_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) u(\boldsymbol{y}_E, \boldsymbol{y}_F) \mathrm{d}\mu_E(\boldsymbol{y}_E) \right\|_{\mathcal{X}} \leq \sum_{\boldsymbol{\nu}_E \in \mathcal{F}_E} C\kappa^{-|\boldsymbol{\nu}_E|} \prod_{j=1}^{J} (1+\nu_j)^{3/2} < \infty.$$

It then follows, e.g. by a finite dimensional version of [84, Proposition 2.1.13], that there holds the uniformly and absolutely convergent expansion

$$\begin{aligned} u(\boldsymbol{y}_E, \boldsymbol{y}_F) &= \sum_{\boldsymbol{\nu}_E \in \mathcal{F}_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \int_{U_E} L_{\boldsymbol{\nu}_E}(\tilde{\boldsymbol{y}}_E) u(\tilde{\boldsymbol{y}}_E, \boldsymbol{y}_F) \mathrm{d}\mu_E(\tilde{\boldsymbol{y}}_E) \\ &= \sum_{\boldsymbol{\nu}_E \in \mathcal{F}_E} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \int_{U_E} L_{\boldsymbol{\nu}_E}(\tilde{\boldsymbol{y}}_E) \sum_{\boldsymbol{\nu}_F \in \mathcal{F}_F} \frac{\partial_{\boldsymbol{y}_F}^{\boldsymbol{\nu}_F} u(\tilde{\boldsymbol{y}}_E, \boldsymbol{0})}{\boldsymbol{\nu}_F!} \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \mathrm{d}\mu_E(\tilde{\boldsymbol{y}}_E). \end{aligned}$$

By (54) (recall that $\theta > 1$) it holds $\sup_{\tilde{\boldsymbol{y}}_E \in U_E} \sum_{\boldsymbol{\nu}_F \in \mathcal{F}_F} \|\partial_{\boldsymbol{y}_F}^{\boldsymbol{\nu}_F} u(\tilde{\boldsymbol{y}}_E, \boldsymbol{0})\|_{\mathcal{X}} / \boldsymbol{\nu}_F! < \infty$, so that by Lebesgue dominated convergence we can interchange the integration with the summation to get

$$u(\boldsymbol{y}_E, \boldsymbol{y}_F) = \sum_{\boldsymbol{\nu}_E \in \mathcal{F}_E} \sum_{\boldsymbol{\nu}_F \in \mathcal{F}_F} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \int_{U_E} L_{\boldsymbol{\nu}_E}(\tilde{\boldsymbol{y}}_E) \frac{\partial_{\boldsymbol{y}_F}^{\boldsymbol{\nu}_F} u(\tilde{\boldsymbol{y}}_E, \boldsymbol{0})}{\boldsymbol{\nu}_F!} \mathrm{d}\mu_E(\tilde{\boldsymbol{y}}_E),$$

with absolute and uniform convergence for all $\boldsymbol{y} \in U$. This shows (ii).

Step 4. We complete the proof of (iii). Fix $\tau \in (0,1)$, so that $|\Lambda_\tau| > 0$. In Step 2 we verified (21) and showed that $(\boldsymbol{\delta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}} \in \ell^{p/(1-p)}(\mathcal{F})$. Denote by $(x_j)_{j \in \mathbb{N}}$ a monotonically decreasing arrangement of $(\boldsymbol{\delta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}}$, i.e. there is a bijection $\pi : \mathbb{N} \to \mathcal{F}$ such that $x_i = \boldsymbol{\delta}^{-\pi(i)}$ for all $i \in \mathbb{N}$, and additionally $(x_i)_{i \in \mathbb{N}}$ is monotonically decreasing. Then $x_n^{p/(1-p)} \leq n^{-1} \sum_{j=1}^{n} x_j^{p/(1-p)}$ and thus $x_n \leq n^{-1/p+1} \|(\boldsymbol{\delta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}}\|_{\ell^{p/(1-p)}(\mathcal{F})}$ for all $n \in \mathbb{N}$. Since Λ_τ corresponds to the $|\Lambda_\tau|$ multiindices $\boldsymbol{\nu} \in \mathcal{F}$ with the largest values of $\boldsymbol{\delta}^{-\boldsymbol{\nu}}$, we get $\sup_{\boldsymbol{\nu} \in \mathcal{F} \setminus \Lambda_\tau} \boldsymbol{\delta}^{-\boldsymbol{\nu}} \leq \|(\boldsymbol{\delta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu} \in \mathcal{F}}\|_{\ell^{p/(1-p)}(\mathcal{F})} |\Lambda_\tau|^{-1/p+1}$. Thus

$$\sup_{\boldsymbol{y} \in U} \left\| u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu} \in \Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \right\|_{\mathcal{X}} \leq \sum_{\boldsymbol{\nu} \in \mathcal{F} \setminus \Lambda_\tau} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}$$

$$\leq |\Lambda_\tau|^{-\frac{1}{p}+1} \|(\boldsymbol{\delta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^{p/(1-p)}(\mathcal{F})} \sum_{\boldsymbol{\nu}\in\mathcal{F}} \boldsymbol{\delta}^{\boldsymbol{\nu}} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}},$$

which concludes the proof, since the final sum is finite by (21) as we showed already. □

A.2 Proof of Lemma 4.1

Proof. When $L = 0$ the properties of the lemma are satisfied by the parallelization defined in Section 4.2.1. In the remainder of the proof, we assume $L > 0$.

We first describe the structure of $(f_1, \ldots, f_k)_{\mathrm{s}}$, and then define its weights explicitly. We denote for $t = 1, \ldots, k$ the depth of f_t by L_t and the number of computation nodes of f_t in layer $\ell = 1, \ldots, L_t + 1$ by $N_\ell^{(t)} \in \mathbb{N}_0$, with the (unusual) convention that $N_\ell^{(t)} := 0$ for $\ell \leq 0$.

We construct $(f_1, \ldots, f_k)_{\mathrm{s}}$ out of $k + 1$ parallel networks, namely an identity network with input dimension n and $f_1, \ldots, f_k$, such that the $L + 1$'st layer of $(f_1, \ldots, f_k)_{\mathrm{s}}$ is the output layer of $f_1, \ldots, f_k$, but it does not contain the output of the identity network. As a result, for $t = 1, \ldots, k$ the $\ell = 1, \ldots, L_t + 1$'th layer of f_t is part of the $\ell + L - L_t$'th layer of $(f_1, \ldots, f_k)_{\mathrm{s}}$, and $2n + \sum_{t=1}^k N_{\ell+L_t-L}^{(t)}$ is the number of computation nodes of $(f_1, \ldots, f_k)_{\mathrm{s}}$ in layer $\ell = 1, \ldots, L$.

For the construction of $(f_1, \ldots, f_k)_{\mathrm{s}}$, it remains to discuss how $f_1, \ldots, f_k$ receive their input. The identity network and the NNs f_t, $t = 1, \ldots, k$ whose depth equals L directly take their input from the input of $(f_1, \ldots, f_k)_{\mathrm{s}}$. For the other f_t, $t = 1, \ldots, k$, we replace the one input weight in the input layer of f_t by two weights, as for each component $x_i \in \mathbb{R}$, $i = 1, \ldots, n$ of the input it holds that $x_i = \sigma_1(x_i) - \sigma_1(-x_i)$, where $\sigma_1(x_i)$ and $\sigma_1(-x_i)$ are computed by the hidden layers of the identity network and can thus be used as input for f_t in layer $1 + L - L_t$ of $(f_1, \ldots, f_k)_{\mathrm{s}}$.

We will denote the weights of $(f_1, \ldots, f_k)_{\mathrm{s}}$ by $w_{i,j}^\ell$ and those of f_t, $t = 1, \ldots, k$ by $w_{i,j}^{(t),\ell}$. Moreover, we write $M_\ell^{(t)} := \sum_{s=1}^{t-1} N_{\ell+L_s-L}^{(s)}$ for all $t = 1, \ldots, k$ and $\ell = 1, \ldots, L + 1$ for the number of computational nodes in layer ℓ of $(f_1, \ldots, f_k)_{\mathrm{s}}$ used to emulate $f_1, \ldots, f_{t-1}$. With this notation, the network weights of $(f_1, \ldots, f_k)_{\mathrm{s}}$ are

$$\begin{aligned}
w_{i,i}^1 &= 1 && i = 1, \ldots, n, \\
w_{i,n+i}^1 &= -1 && i = 1, \ldots, n, \\
w_{i,i}^\ell &= 1 && i = 1, \ldots, 2n, \\
& && \ell = 2, \ldots, L, \\
w^1_{i,2n+M_1^{(t)}+j} &= w_{i,j}^{(t),1} && i = 1, \ldots, n, && j = 1, \ldots, N_1^{(t)}, \\
& && && t = 1, \ldots, k \text{ satisfying } L_t = L, \\
w^\ell_{i,2n+M_\ell^{(t)}+j} &= w_{i,j}^{(t),1} && i = 1, \ldots, n, && j = 1, \ldots, N_1^{(t)},
\end{aligned}$$

$$
\begin{aligned}
& & & \ell = 1 + L - L_t, && t = 1, \dots, k \text{ satisfying } 0 < L_t < L, \\
w^{\ell}_{n+i, 2n+M^{(t)}_{\ell}+j} &= -w^{(t),1}_{i,j} && i = 1, \dots, n, && j = 1, \dots, N^{(t)}_1, \\
& & & \ell = 1 + L - L_t, && t = 1, \dots, k \text{ satisfying } 0 < L_t < L, \\
w^{\ell}_{2n+M^{(t)}_{\ell-1}+i, 2n+M^{(t)}_{\ell}+j} &= w^{(t),\ell+L_t-L}_{i,j} && i = 1, \dots, N^{(t)}_{\ell-1+L_t-L}, && j = 1, \dots, N^{(t)}_{\ell+L_t-L}, \\
& & & \ell = 2 + L - L_t, \dots, L, && t = 1, \dots, k, \\
w^{L+1}_{2n+M^{(t)}_{L}+i, M^{(t)}_{L+1}+j} &= w^{(t),L_t+1}_{i,j} && i = 1, \dots, N^{(t)}_{L_t}, && j = 1, \dots, N^{(t)}_{L_t+1}, \\
& & & && t = 1, \dots, k \text{ satisfying } 0 < L_t, \\
w^{L+1}_{i, M^{(t)}_{L+1}+j} &= w^{(t),1}_{i,j} && i = 1, \dots, n, && j = 1, \dots, N^{(t)}_1, \\
& & & && t = 1, \dots, k \text{ satisfying } L_t = 0, \\
w^{L+1}_{n+i, M^{(t)}_{L+1}+j} &= -w^{(t),1}_{i,j} && i = 1, \dots, n, && j = 1, \dots, N^{(t)}_1, \\
& & & && t = 1, \dots, k \text{ satisfying } L_t = 0, \\
w^{\ell}_{i,j} &= 0 && \text{otherwise}, \\
b^{\ell}_{2n+M^{(t)}_{\ell}+j} &= b^{(t),\ell+L_t-L}_{j} && && j = 1, \dots, N^{(t)}_{\ell+L_t-L}, \\
& & & \ell = 1 + L - L_t, \dots, L, && t = 1, \dots, k, \\
b^{L+1}_{M^{(t)}_{L+1}+j} &= b^{(t),L_t+1}_{j} && && j = 1, \dots, N^{(t)}_{L_t+1}, \\
& & & && t = 1, \dots, k, \\
b^{\ell}_{j} &= 0 && \text{otherwise}.
\end{aligned}
$$

The first three equations describe the first L layers of an identity network. The output layer of the identity network is not included, because it is not desired that the input of $(f_1, \dots, f_k)_{\mathrm{s}}$ is part of the output of $(f_1, \dots, f_k)_{\mathrm{s}}$. The fourth, fifth and sixth equation describe how the input of the network is connected to the parts emulating f_t, for $t = 1, \dots, k$ that satisfy $L_t > 0$. The seventh equation describes the remaining hidden layer weights of f_t, $t = 1, \dots, k$. The weights of the output layer, indexed by $L + 1$, are described in the eighth, ninth and tenth equation. The only remaining nonzero weights are the biases of $f_1, \dots, f_k$, described in the twelfth and thirteenth equation.

The expressions for the input dimension, the output dimension and the depth follow directly from the construction. The bound on the network size is obtained by noting that all biases $b^{(t),\ell}_j$ appear exactly once, the first three equations involve $2nL$ nonzero weights, that in the expression for the network weights $w^{(t),1}_{i,j}$ appears exactly once if $L_t = L$ and exactly twice if $L_t < L$, and that $w^{(t),\ell}_{i,j}$ appears exactly once for $\ell > 1$. The bound on the first layer size follows from the first, second, fourth (for t such that $L_t = L$) and twelfth equation (for $L_t = L$). Likewise, the bound on the output layer size follows from the eighth, ninth, tenth and thirteenth equation. □

A.3 Proof of Theorem 4.9

Proof. If $|\Lambda_\tau| = 1$, then Proposition 3.8 item (iv) implies $\Lambda_\tau = \{\mathbf{0}\}$. Hence, $\sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F}$ is constant in $\boldsymbol{y}\in U$. Therefore, it is emulated exactly by a σ_1-NN of depth 0 and size 1.

We use that $|\Lambda_\tau| \geq 2$. The proof is given in several steps. In the first step, we define the approximation $\tilde{u}_\tau$ of u. Then, we estimate its error. In the third step we construct a network which emulates $\tilde{u}_\tau$, the depth and size of which are estimated in the fourth and last step.

Step 1. For all $\boldsymbol{\nu}\in\mathcal{F}$ let $(j_{i;\boldsymbol{\nu}_F})_{i=1}^{|\boldsymbol{\nu}_F|_1} \subset \mathbb{N}$ be such that $\prod_{i=1}^{|\boldsymbol{\nu}_F|_1} y_{j_{i;\boldsymbol{\nu}_F}} = \boldsymbol{y}_F^{\boldsymbol{\nu}_F}$ for all $\boldsymbol{y}\in U$. In addition, we define $\Lambda_{\tau,E} := \{\boldsymbol{\nu}_E \in \mathcal{F}_E : \boldsymbol{\nu}\in\Lambda_\tau\}$. As shown in Proposition 3.8 item (iii), $|\Lambda_{\tau,E}| \leq C(1+\log|\Lambda_\tau|)^J$.

For all $\boldsymbol{\nu}\in\Lambda_\tau$, we define

$$f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) := \tilde{\times}_{\delta_{\boldsymbol{\nu}},R}\left(\tilde{L}_{\boldsymbol{\nu}_E,\delta}(\boldsymbol{y}_E), \tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\left(\{y_{j_{i;\boldsymbol{\nu}_F}}\}_{i=1}^{|\boldsymbol{\nu}_F|_1}\right)\right), \qquad \boldsymbol{y}\in U,$$

where $\tilde{\times}_{\delta_{\boldsymbol{\nu}},R}$ is as in Proposition 4.3, $\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}$ as in Proposition 4.4 and $\tilde{L}_{\boldsymbol{\nu}_E,\delta}$ as in Proposition 4.6. We choose the accuracy of all tensor product Legendre polynomials to be $\delta := \frac{1}{3}\min\left\{1, \|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|^{-1}_{\ell^1(\mathcal{F})}|\Lambda_\tau|^{-1/p+1}\right\}$. By choosing δ independent of $\boldsymbol{\nu}_E\in\Lambda_{\tau,E}$, we can use $\tilde{L}_{\boldsymbol{\nu}_E,\delta}$ for multiple different $\boldsymbol{\nu}$ (there may be multiple $\boldsymbol{\nu}\in\Lambda_\tau$ with the same $\boldsymbol{\nu}_E$). For the accuracy of $\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\left(\{y_{j_{i;\boldsymbol{\nu}_F}}\}_{i=1}^{|\boldsymbol{\nu}_F|_1}\right)$, we choose $\varepsilon_{\boldsymbol{\nu},F} := (2m(\Lambda_\tau)+2)^{-J}\frac{1}{3}\min\left\{1, |c_{\boldsymbol{\nu}}|^{-1}|\Lambda_\tau|^{-1/p}\right\}$. For $\tilde{\times}_{\delta_{\boldsymbol{\nu}},R}$, we choose accuracy $\delta_{\boldsymbol{\nu}} := \frac{1}{3}\min\left\{1, |c_{\boldsymbol{\nu}}|^{-1}|\Lambda_\tau|^{-1/p}\right\}$, and note that the absolute values of its inputs are bounded by $R := (2m(\Lambda_{\tau,E})+2)^J$.

Finally, we define

$$\tilde{u}_\tau := \sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}} f_{\boldsymbol{\nu},\tau}.$$

Step 2. The error can be estimated as follows:

$$\begin{aligned}
&\sup_{\boldsymbol{y}\in U}\left|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})\right| \\
&\leq \sup_{\boldsymbol{y}\in U}\left|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - \tilde{L}_{\boldsymbol{\nu}_E,\delta}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F}\right| \\
&\quad + \sup_{\boldsymbol{y}\in U}\left|\tilde{L}_{\boldsymbol{\nu}_E,\delta}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - \tilde{L}_{\boldsymbol{\nu}_E,\delta}(\boldsymbol{y}_E)\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\left(\{y_{j_{i;\boldsymbol{\nu}_F}}\}_{i=1}^{|\boldsymbol{\nu}_F|_1}\right)\right| \\
&\quad + \sup_{\boldsymbol{y}\in U}\left|\tilde{L}_{\boldsymbol{\nu},\delta}(\boldsymbol{y}_E)\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\left(\{y_{j_{i;\boldsymbol{\nu}_F}}\}_{i=1}^{|\boldsymbol{\nu}_F|_1}\right) - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})\right| \\
&\leq \delta + (2m(\Lambda_\tau)+2)^J\varepsilon_{\boldsymbol{\nu},F} + \delta_{\boldsymbol{\nu}} \\
&\leq \tfrac{1}{3}\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|^{-1}_{\ell^1(\mathcal{F})}|\Lambda_\tau|^{-1/p+1} + \tfrac{2}{3}|c_{\boldsymbol{\nu}}|^{-1}|\Lambda_\tau|^{-1/p}.
\end{aligned} \tag{55}$$

To estimate the first two terms of the three, we used Propositions 4.4 and 4.6. For the third term, we used Proposition 4.3. As a result, we find

$$\begin{aligned}
&\sup_{\boldsymbol{y}\in U}\left|\sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - \tilde{u}_\tau((y_j)_{j\in S_{\Lambda_\tau}})\right| \\
&\leq \sum_{\boldsymbol{\nu}\in\Lambda_\tau} |c_{\boldsymbol{\nu}}| \sup_{\boldsymbol{y}\in U}\left|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})\right| \\
&\leq \sum_{\boldsymbol{\nu}\in\Lambda_\tau} |c_{\boldsymbol{\nu}}| \cdot \left(\tfrac{1}{3}\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^1(\mathcal{F})}^{-1}|\Lambda_\tau|^{-1/p+1} + \tfrac{2}{3}|c_{\boldsymbol{\nu}}|^{-1}|\Lambda_\tau|^{-1/p}\right) \\
&\leq |\Lambda_\tau|^{-1/p+1}.
\end{aligned}$$

Together with Theorem 3.7 item (iii), which states that

$$\sup_{\boldsymbol{y}\in U}\left|u(\boldsymbol{y}) - \sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}} L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F}\right| \leq C_2|\Lambda_\tau|^{-1/p+1},$$

we get Equation (33).

Step 3. We now construct a network which emulates $\tilde{u}_\tau$. It consists of four concatenated subnetworks:

$$\tilde{u}_\tau := \tilde{u}_\tau^{(1)} \circ \tilde{u}_\tau^{(2)} \circ \tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}.$$

The first subnetwork $\tilde{u}_\tau^{(4)}$ has input dimension $|S_{\Lambda_\tau}|$, output dimension $|\Lambda_{\tau,E}|+|\Lambda_\tau|$ and in parallel emulates approximations of $\{L_{\boldsymbol{\nu}_E}\}_{\boldsymbol{\nu}_E\in\Lambda_{\tau,E}}$ and $\{\boldsymbol{y}_F^{\boldsymbol{\nu}_F}\}_{\boldsymbol{\nu}\in\Lambda_\tau}$:

$$\tilde{u}_\tau^{(4)} := \left(\mathrm{Id}_{\mathbb{R}^{|\Lambda_{\tau,E}|}} \circ f_{\Lambda_{\tau,E},\delta}, \left\{\mathrm{Id}_{\mathbb{R}^1} \circ \tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\left(\left(y_{j_{i;\boldsymbol{\nu}_F}}\right)_{i=1}^{|\boldsymbol{\nu}_F|_1}\right)\right\}_{\boldsymbol{\nu}\in\Lambda_\tau}\right), \tag{56}$$

where $f_{\Lambda_{\tau,E},\delta}$ is as constructed in Proposition 4.6 and where the depth of the σ_1-identity networks is such that

$$\operatorname{depth}\left(\tilde{u}_\tau^{(4)}\right) \leq 1 + \max\{\operatorname{depth}(f_{\Lambda_{\tau,E},\delta})\} \cup \left\{\operatorname{depth}\left(\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\right)\right\}_{\boldsymbol{\nu}\in\Lambda_\tau}.$$

The second subnetwork $\tilde{u}_\tau^{(3)}$ has zero depth, i.e. it consists of an affine transformation only. It has input dimension $|\Lambda_{\tau,E}|+|\Lambda_\tau|$ and output dimension $2|\Lambda_\tau|$. For a fixed but arbitrary enumeration $\left(\boldsymbol{\nu}^{(j)}\right)_{j=1}^{|\Lambda_\tau|}$ of Λ_τ, the output of $\tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}$ is

$$\left.\begin{aligned}
\left(\tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}((y_j)_{j\in S_{\Lambda_\tau}})\right)_{2k-1} &= \tilde{L}_{\boldsymbol{\nu}_E^{(k)},\delta}(\boldsymbol{y}_E), \\
\left(\tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}((y_j)_{j\in S_{\Lambda_\tau}})\right)_{2k} &= \tilde{\prod}_{\varepsilon_{\boldsymbol{\nu}^{(k)},F},1}\left(\left(y_{j_{i;\boldsymbol{\nu}_F^{(k)}}}\right)_{i=1}^{|\boldsymbol{\nu}_F^{(k)}|}\right),
\end{aligned}\right\} \quad \begin{aligned}&\forall \boldsymbol{y}\in U, \\ &\forall k \leq |\Lambda_\tau|.\end{aligned} \tag{57}$$

The third subnetwork $\tilde{u}_\tau^{(2)}$ is defined to be the parallelization of networks from Proposition 4.3 concatenated with σ_1-identity networks:

$$\tilde{u}_\tau^{(2)} := \left(\left\{ \mathrm{Id}_{\mathbb{R}^1} \circ \tilde{\times}_{\delta_{\boldsymbol{\nu}^{(j)}}, R} \right\}_{j=1}^{|\Lambda_\tau|} \right)_{\mathrm{d}},$$

where the identity networks are such that

$$\mathrm{depth}\left(\tilde{u}_\tau^{(2)}\right) \leq 1 + \max_{\boldsymbol{\nu} \in \Lambda_\tau} C\left(1 + \log_2\left(R/\delta_{\boldsymbol{\nu}}\right)\right).$$

Its output is given by

$$\left(\tilde{u}_\tau^{(2)} \circ \tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}((y_j)_{j \in S_{\Lambda_\tau}})\right)_k = f_{\boldsymbol{\nu}^{(k)}, \tau}((y_j)_{j \in S_{\Lambda_\tau}}), \qquad \forall \boldsymbol{y} \in U, k \leq |\Lambda_\tau|. \tag{58}$$

Finally, the last subnetwork $\tilde{u}_\tau^{(1)}$ has depth 0, input dimension $|\Lambda_\tau|$ and output dimension 1, and emulates a linear combination of its inputs, with weight $c_{\boldsymbol{\nu}^{(j)}}$ in coordinate j, and without bias. As a result,

$$\left(\tilde{u}_\tau^{(1)} \circ \tilde{u}_\tau^{(2)} \circ \tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}\right)((y_j)_{j \in S_{\Lambda_\tau}}) = \tilde{u}_\tau((y_j)_{j \in S_{\Lambda_\tau}}), \qquad \forall \boldsymbol{y} \in U.$$

Step 4. We now give estimates on the depth of the subnetworks and the network itself. We use that $m(\Lambda_{\tau,E}) \leq m(\Lambda_\tau) \leq C(1 + \log|\Lambda_\tau|)$, where the second inequality is Proposition 3.8 item (ii). We get, using Propositions 4.6, 4.4 and 4.3:

$$\begin{aligned}
\mathrm{depth}\left(\tilde{u}_\tau^{(4)}\right) &\leq 1 + \max\{\mathrm{depth}(f_{\Lambda_{\tau,E},\delta})\} \cup \left\{ \mathrm{depth}\left(\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1} \right) \right\}_{\boldsymbol{\nu} \in \Lambda_\tau} \\
&\leq \max\left\{ C\big(1 + \log m(\Lambda_{\tau,E})\big)\big(m(\Lambda_{\tau,E}) + \log_2(1/\delta)\big) \right\} \\
&\quad \cup \left\{ C\big(1 + \log(|\boldsymbol{\nu}_F|_1)\log(|\boldsymbol{\nu}_F|_1/\varepsilon_{\boldsymbol{\nu},F})\big) \right\}_{\boldsymbol{\nu} \in \Lambda_\tau} \\
&\leq \max\Big\{ C\big(1 + \log m(\Lambda_{\tau,E})\big)\Big(m(\Lambda_{\tau,E}) + \log 3 \\
&\quad + \max\left\{0, \log\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu} \in \mathcal{F}}\|_{\ell^1(\mathcal{F})} + \tfrac{1-p}{p}\log|\Lambda_\tau|\right\}\Big), C\Big(1 + \log(m(\Lambda_\tau)) \\
&\quad \log\Big(m(\Lambda_\tau)(2m(\Lambda_\tau)+2)^J 3 \max\left\{1, \|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu} \in \mathcal{F}}\|_{\ell^1(\mathcal{F})}|\Lambda_\tau|^{1/p}\right\}\Big)\Big)\Big\} \\
&\leq C\big(1 + \log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big), \\
\mathrm{depth}\left(\tilde{u}_\tau^{(3)}\right) &= 0, \\
\mathrm{depth}\left(\tilde{u}_\tau^{(2)}\right) &\leq 1 + \max_{\boldsymbol{\nu} \in \Lambda_\tau} C\left(1 + \log_2\left((2m(\Lambda_\tau)+2)^J 3 \max\left\{1, \|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu} \in \mathcal{F}}\|_{\ell^1(\mathcal{F})}|\Lambda_\tau|^{1/p}\right\}\right)\right) \\
&\leq C(1 + \log|\Lambda_\tau|), \\
\mathrm{depth}\left(\tilde{u}_\tau^{(1)}\right) &= 0, \\
\mathrm{depth}\left(\tilde{u}_\tau\right) &= \mathrm{depth}\left(\tilde{u}_\tau^{(1)}\right) + 1 + \mathrm{depth}\left(\tilde{u}_\tau^{(2)}\right) + 1 + \mathrm{depth}\left(\tilde{u}_\tau^{(3)}\right) + 1 + \mathrm{depth}\left(\tilde{u}_\tau^{(4)}\right)
\end{aligned}$$

$$\leq C\big(1+\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big).$$

For the bounds on the network size, we use that the depth of the identity networks in $\tilde{u}_\tau^{(4)}$ is less than depth $\left(\tilde{u}_\tau^{(4)}\right)$. There is one identity network with input dimension $|\Lambda_{\tau,E}|$ and there are $|\Lambda_\tau|$ identity networks with input dimension 1. The sum of the network sizes is bounded by

$$2(|\Lambda_{\tau,E}|+|\Lambda_\tau|)\operatorname{depth}\left(\tilde{u}_\tau^{(4)}\right)\leq C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big).$$

The depth of identity networks in $\tilde{u}_\tau^{(2)}$ is less than depth $\left(\tilde{u}_\tau^{(2)}\right)$, their input dimension is 1 and their number is $|\Lambda_\tau|$. Hence, the sum of their sizes is bounded by

$$2|\Lambda_\tau|\operatorname{depth}\left(\tilde{u}_\tau^{(2)}\right)\leq C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\big).$$

We find using (27):

$$\begin{aligned}
\operatorname{size}\left(\tilde{u}_\tau^{(4)}\right)&\leq 2\operatorname{size}\left(f_{\Lambda_{\tau,E},\delta}\right)+2\sum_{\boldsymbol{\nu}_F\in\Lambda_\tau}\operatorname{size}\left(\tilde{\prod}_{\varepsilon_{\boldsymbol{\nu},F},1}\right)\\
&\quad+2C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big)\\
&\leq C\Big(m(\Lambda_{\tau,E})^2+m(\Lambda_{\tau,E})\log_2(1/\delta)+|\Lambda_{\tau,E}|\big(1+\log_2 m(\Lambda_{\tau,E})+\log_2(1/\delta)\big)\Big)\\
&\quad+\sum_{\boldsymbol{\nu}\in\Lambda_\tau}C\big(1+|\boldsymbol{\nu}_F|_1\log(|\boldsymbol{\nu}_F|_1/\varepsilon_{\boldsymbol{\nu},F})\big)+C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big)\\
&\leq\Big(C(1+\log|\Lambda_\tau|)^2+C(1+\log|\Lambda_\tau|)\\
&\quad\cdot\log_2\left(3\max\left\{1,\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^1(\mathcal{F})}|\Lambda_\tau|^{1/p-1}\right\}\right)+C(1+\log|\Lambda_\tau|)^J\\
&\quad\cdot\left(1+\log\log|\Lambda_\tau|+\log_2\left(3\max\left\{1,\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^1(\mathcal{F})}|\Lambda_\tau|^{1/p-1}\right\}\right)\right)\Big)\\
&\quad+\sum_{\boldsymbol{\nu}\in\Lambda_\tau}C\left(1+m(\Lambda_\tau)\log m(\Lambda_\tau)\right)\\
&\quad+\sum_{\boldsymbol{\nu}\in\Lambda_\tau}C\left(1+m(\Lambda_\tau)\log\left(3\max\left\{1,|c_{\boldsymbol{\nu}}|\,|\Lambda_\tau|^{1/p}\right\}\right)\right)\\
&\quad+C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big)\\
&\overset{(*)}{\leq}C(1+\log|\Lambda_\tau|)^{J+1}+C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big)\\
&\leq C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big).
\end{aligned}$$

At (*) we used the following estimate, which uses that $\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^p(\mathcal{F})} < \infty$ by Theorem 3.7 item (i) for $\mathcal{X} = \mathbb{R}$, and that $\log(\max\{1, x\}) \leq x$ for all $x > 0$:

$$\begin{aligned}
&\sum_{\boldsymbol{\nu}\in\Lambda_\tau} C\left(1 + m(\Lambda_\tau)\log\left(3\max\left\{1, |c_{\boldsymbol{\nu}}|\,|\Lambda_\tau|^{1/p}\right\}\right)\right)\\
&\leq C(1+\log|\Lambda_\tau|)\sum_{\boldsymbol{\nu}\in\Lambda_\tau}\log\left(3\max\left\{1, |c_{\boldsymbol{\nu}}|\,|\Lambda_\tau|^{1/p}\right\}\right)\\
&\leq C(1+|\Lambda_\tau|\log|\Lambda_\tau|) + C(1+\log|\Lambda_\tau|)\sum_{\boldsymbol{\nu}\in\Lambda_\tau}\tfrac{1}{p}\log\left(\max\left\{1, |c_{\boldsymbol{\nu}}|^p\,|\Lambda_\tau|\right\}\right)\\
&\leq C(1+|\Lambda_\tau|\log|\Lambda_\tau|) + C(1+\log|\Lambda_\tau|)\sum_{\boldsymbol{\nu}\in\Lambda_\tau}|c_{\boldsymbol{\nu}}|^p\,|\Lambda_\tau|\\
&\leq C(1+|\Lambda_\tau|\log|\Lambda_\tau|) + C(1+\log|\Lambda_\tau|)\cdot\|(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}\|^p_{\ell^p(\mathcal{F})}\cdot|\Lambda_\tau|\\
&\leq C(1+|\Lambda_\tau|\log|\Lambda_\tau|).
\end{aligned} \tag{59}$$

The number of nonzero weights of $\tilde{u}_\tau^{(3)}$ is at most $2|\Lambda_\tau|$, because each output depends on at most one input. We can hence estimate

$$\operatorname{size}\left(\tilde{u}_\tau^{(3)}\right) \leq 2|\Lambda_\tau|.$$

Again using Equations (27) and (59), we find

$$\begin{aligned}
\operatorname{size}\left(\tilde{u}_\tau^{(2)}\right) &\leq 2\sum_{\boldsymbol{\nu}\in\Lambda_\tau} C\left(1+\log_2\left((2m(\Lambda_{\tau,E})+2)^J 3\max\left\{1, |c_{\boldsymbol{\nu}}|\,|\Lambda_\tau|^{1/p}\right\}\right)\right)\\
&\quad + 2C(1+|\Lambda_\tau|\log|\Lambda_\tau|)\\
&\leq C(1+|\Lambda_\tau|\log|\Lambda_\tau|),\\
\operatorname{size}\left(\tilde{u}_\tau^{(1)}\right) &\leq |\Lambda_\tau|,\\
\operatorname{size}\left(\tilde{u}_\tau\right) &\leq 4\operatorname{size}\left(\tilde{u}_\tau^{(1)}\right) + 4\operatorname{size}\left(\tilde{u}_\tau^{(2)}\right) + 4\operatorname{size}\left(\tilde{u}_\tau^{(3)}\right) + 4\operatorname{size}\left(\tilde{u}_\tau^{(4)}\right)\\
&\leq C\big(1+|\Lambda_\tau|\log(|\Lambda_\tau|)\log\log(|\Lambda_\tau|)\big).
\end{aligned}$$

□

Most of the network constructed in the proof of Theorem 4.9 will also be used in the proof of Theorem 5.2 in Section A.4 ahead, namely the part of the network which in parallel emulates the gpc basis polynomials $\{U \ni \boldsymbol{y} \mapsto L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F}\}_{\boldsymbol{\nu}\in\Lambda_\tau}$. Therefore, we state the properties of that part of the network as a lemma. We state the lemma for the general case of a $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic function $u : U \to \mathcal{X}$. The construction of the neural network is the same as for a $(\boldsymbol{b}, \varepsilon, \mathbb{R})$-holomorphic function $u : U \to \mathbb{R}$, except that we now use the sequence $(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}$ instead of $(|c_{\boldsymbol{\nu}}|)_{\boldsymbol{\nu}\in\mathcal{F}}$ to define the accuracy.

Lemma A.1. *Let $u : U \to \mathcal{X}$ be $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphic for some $\boldsymbol{b} \in \ell^p(\mathbb{N})$, $p \in (0,1)$ and $\varepsilon > 0$. Let $J \in \mathbb{N}$, $(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \mathbb{R}^{\mathcal{F}}$ and $\emptyset \neq \Lambda_\tau \subset \mathcal{F}$ for $\tau \in (0,1)$ be as in Theorem 3.7.*

Then, the σ_1-NN $\tilde{f}_{\Lambda_\tau} := \tilde{u}_\tau^{(2)} \circ \tilde{u}_\tau^{(3)} \circ \tilde{u}_\tau^{(4)}$ has input dimension $|S_{\Lambda_\tau}|$ and output dimension $|\Lambda_\tau|$. The components of its output are

$$\big(\tilde{f}_{\Lambda_\tau}((y_j)_{j\in S_{\Lambda_\tau}})\big)_k = f_{\boldsymbol{\nu}^{(k)},\tau}((y_j)_{j\in S_{\Lambda_\tau}}), \qquad \text{for all } \boldsymbol{y} \in U, k \leq |\Lambda_\tau|, \tag{60}$$

for an arbitrary but fixed enumeration $(\boldsymbol{\nu}^{(k)})_{k=1}^{|\Lambda_\tau|}$ of Λ_τ. They satisfy the uniform error bound

$$\begin{aligned}&\sup_{\boldsymbol{y}\in U} \big|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in\operatorname{supp}\boldsymbol{\nu}})\big| \\ &\leq \tfrac{1}{3}\|(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^1(\mathcal{F})}^{-1}|\Lambda_\tau|^{-1/p+1} + \tfrac{2}{3}\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}^{-1}|\Lambda_\tau|^{-1/p}, \qquad \text{for all } \boldsymbol{\nu} \in \Lambda_\tau.\end{aligned} \tag{61}$$

The depth and size are bounded as follows:

$$\begin{aligned}\operatorname{size}(\tilde{f}_{\Lambda_\tau}) &\leq C(1 + |\Lambda_\tau| \cdot \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|), \\ \operatorname{depth}(\tilde{f}_{\Lambda_\tau}) &\leq C(1 + \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|).\end{aligned}$$

It follows from Proposition 4.3 and the definitions of R and $\delta_{\boldsymbol{\nu}}$ in Step 1 of the proof of Theorem 4.9 that

$$\begin{aligned}\sup_{\boldsymbol{y}\in U}\big|f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}})\big| &\leq \sup_{\boldsymbol{y}\in U}\big|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}^{\boldsymbol{\nu}_F}\big| + \sup_{\boldsymbol{y}\in U}\big|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}})\big| \\ &\leq R + \delta_{\boldsymbol{\nu}} \leq R + 1 = (2m(\Lambda_{\tau,E}) + 2)^J + 1.\end{aligned}$$

A.4 Proof of Theorem 5.2

Proof. Throughout the proof, we fix $\tau \in (0,1)$, and thereby Λ_τ. The proof consists of 5 steps. In Step 1, we construct the networks which approximate the gpc-coefficients $\{c_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu}\in\Lambda_\tau}$ and the polynomials in $\boldsymbol{y} \in U$. In Step 2, we construct $\tilde{u}_\tau$. In Step 3, the error is estimated. In Step 4, a NN emulating $\tilde{u}_\tau$ is discussed in detail. In Step 5, the NN depth and size are estimated.

Step 1. We first construct a subnetwork which approximates the gpc coefficients $\{c_{\boldsymbol{\nu}}\}_{\boldsymbol{\nu}\in\Lambda_\tau}$. Let $\boldsymbol{\delta}^{-1} \in \ell^{p/(1-p)}(\mathbb{N})$ be as in Theorem 3.7 based on $(\boldsymbol{b}, \varepsilon, \mathcal{X})$-holomorphy of u. To optimize the choice of network size used for the emulation of each gpc coefficient, we use [77, Lemma 4.7], which in turn is based on [3, Section 3] and [30, Section 2]. We apply the result for $a_{\boldsymbol{\nu}} := \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s}\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \in (0,\infty)$, $b_{\boldsymbol{\nu}} := \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}\|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \in (0,\infty)$ for all $\boldsymbol{\nu} \in \mathcal{F}$, $\boldsymbol{\beta} := \boldsymbol{\delta}^{-1} \in (0,1)^{\mathbb{N}}$, $p_a := p^s$, $p_b := p$, $n := |\Lambda_\tau|$ and $\Lambda_n := \Lambda_\tau$. Instead of the assumption that $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in$

$\ell^2(\mathcal{F})$ and $\boldsymbol{\beta} \in \ell^{2p/(2-p)}$, we have $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$ and $\boldsymbol{\beta} \in \ell^{p/(1-p)}$ (Theorem 3.7 item (iii)). Under the current assumption we obtain the same result as in [77, Lemma 4.7], as in both cases [77, Lemma 2.8] implies that (in the notation of [77])

$$\sum_{\mathcal{F}\setminus\Lambda_n} b_{\boldsymbol{\nu}} \leq Cn^{-1/p_b+1}. \tag{62}$$

The rest of the proof of [77, Lemma 4.7] only uses (62), hence the conclusion of [77, Lemma 4.7] also holds when $(b_{\boldsymbol{\nu}}\boldsymbol{\beta}^{-\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^1(\mathcal{F})$ and $\boldsymbol{\beta} \in \ell^{p/(1-p)}$.

Thus, it follows from [77, Lemma 4.7] that there exists a constant $C > 0$ and a sequence $(m_{n;\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_n} \in \mathbb{N}^{|\Lambda_n|}$ (in the notation of [77]), which we denote by $(m_{\tau;\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\Lambda_\tau}$, such that with $\mathcal{N}_\tau := \sum_{\boldsymbol{\nu}\in\Lambda_\tau} m_{\tau;\boldsymbol{\nu}} \geq |\Lambda_\tau|$ it holds

$$\begin{aligned}|\Lambda_\tau|^{-1/p+1} + \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} m_{\tau;\boldsymbol{\nu}}^{-\gamma} + \sum_{\boldsymbol{\nu}\in\mathcal{F}\setminus\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \\ \leq C\mathcal{N}_\tau^{-r}\end{aligned} \tag{63}$$

for r as in Equation (35).

For all $\boldsymbol{\nu} \in \Lambda_\tau$, let $\tilde{c}_{\boldsymbol{\nu},\tau} := \Phi_{m_{\tau;\boldsymbol{\nu}}}^{c_{\boldsymbol{\nu}}}$ be as provided by Assumption 5.1. Then, we consider the parallelization with shared identity operator $\tilde{g}_{\Lambda_\tau} := (\{\tilde{c}_{\boldsymbol{\nu},\tau}\}_{\boldsymbol{\nu}\in\Lambda_\tau})_{\mathrm{s}}$ introduced in Lemma 4.1. With Assumption 5.1, it follows that

$$\begin{aligned}\operatorname{depth}(\tilde{g}_{\Lambda_\tau}) &= \max_{\boldsymbol{\nu}\in\Lambda_\tau} \operatorname{depth}(\tilde{c}_{\boldsymbol{\nu},\tau}) \leq \max_{\boldsymbol{\nu}\in\Lambda_\tau} C(1+\log(m_{\tau;\boldsymbol{\nu}})) \leq C(1+\log\mathcal{N}_\tau),\\ \operatorname{size}(\tilde{g}_{\Lambda_\tau}) &\leq 2d\operatorname{depth}(\tilde{g}_{\Lambda_\tau}) + 2\sum_{\boldsymbol{\nu}\in\Lambda_\tau} \operatorname{size}(\tilde{c}_{\boldsymbol{\nu},\tau})\\ &\leq C(1+\log\mathcal{N}_\tau) + 2\sum_{\boldsymbol{\nu}\in\Lambda_\tau} Cm_{\tau;\boldsymbol{\nu}} \leq C\mathcal{N}_\tau.\end{aligned}$$

For the approximation of the polynomials in $\boldsymbol{y} \in U$, we use the DNN $\tilde{f}_{\Lambda_\tau}$ from Lemma A.1. We denote the components of its output by $f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}})$, for all $\boldsymbol{\nu} \in \Lambda_\tau$ and $\boldsymbol{y} \in U$.

Step 2. In this step we define $\tilde{u}_\tau$, combining the components of $\tilde{g}_{\Lambda_\tau}$ and $\tilde{f}_{\Lambda_\tau}$.

First, we note that by Assumption 5.1, it holds that for all $\boldsymbol{\nu} \in \Lambda_\tau$

$$\|\tilde{c}_{\boldsymbol{\nu},\tau}\|_{L^\infty(\mathrm{D})} \leq C\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s} m_{\tau;\boldsymbol{\nu}}^\theta \leq C\|(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^{p^s}(\mathcal{F})} m_{\tau;\boldsymbol{\nu}}^\theta = Cm_{\tau;\boldsymbol{\nu}}^\theta.$$

With Proposition 3.8, item (ii), this implies with $R := (2m(\Lambda_{\tau,E}) + 2)^J$ that

$$\begin{aligned}R'_{\boldsymbol{\nu}} &:= \max\{\{R+1\} \cup \{\|\tilde{c}_{\boldsymbol{\nu},\tau}\|_{L^\infty(\mathrm{D})}\}_{\boldsymbol{\nu}\in\Lambda_\tau}\} \leq \max\{R+1, Cm_{\tau;\boldsymbol{\nu}}^\theta\}\\ &\leq C\max\{(1+\log|\Lambda_\tau|)^J, m_{\tau;\boldsymbol{\nu}}^\theta\},\end{aligned}$$

for some constant C which is independent of Λ_τ.

We define the NN $\tilde{u}_\tau$ approximating u: for $\lambda := \mathcal{N}_\tau^{-r-1}$, $\boldsymbol{x} \in \mathrm{D}$ and for $\boldsymbol{y} \in U$, we set

$$\tilde{u}_\tau(\boldsymbol{x}, (y_j)_{j\in S_{\Lambda_\tau}}) := \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}} \left(\tilde{c}_{\boldsymbol{\nu},\tau}(\boldsymbol{x}), f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}})\right).$$

Step 3. We estimate the NN expression error.

$$\begin{aligned}
&\sup_{\boldsymbol{y}\in U} \left\| u(\boldsymbol{y}) - \tilde{u}_\tau(\cdot, (y_j)_{j\in S_{\Lambda_\tau}}) \right\|_{\mathcal{X}} \\
&\le \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\mathcal{F}} c_{\boldsymbol{\nu}}(\cdot) L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} - \sum_{\boldsymbol{\nu}\in\Lambda_\tau} c_{\boldsymbol{\nu}}(\cdot) L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \right\|_{\mathcal{X}} \\
&\quad + \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \left(c_{\boldsymbol{\nu}}(\cdot) L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} - c_{\boldsymbol{\nu}}(\cdot) f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right) \right\|_{\mathcal{X}} \\
&\quad + \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \left(c_{\boldsymbol{\nu}}(\cdot) f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) - \tilde{c}_{\boldsymbol{\nu},\tau}(\cdot) f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right) \right\|_{\mathcal{X}} \\
&\quad + \sup_{\boldsymbol{y}\in U} \left\| \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \left(\tilde{c}_{\boldsymbol{\nu},\tau}(\cdot) f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) - \tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}} \left(\tilde{c}_{\boldsymbol{\nu},\tau}(\cdot), f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right) \right) \right\|_{\mathcal{X}} \\
&\le \sum_{\boldsymbol{\nu}\in\mathcal{F}\setminus\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \sup_{\boldsymbol{y}\in U} \left| L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \right| \\
&\quad + \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \sup_{\boldsymbol{y}\in U} \left| L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right| \\
&\quad + \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}} - \tilde{c}_{\boldsymbol{\nu},\tau}\|_{\mathcal{X}} \left(\sup_{\boldsymbol{y}\in U} \left| L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} \right| + \sup_{\boldsymbol{y}\in U} \left| L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E) \boldsymbol{y}_F^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right| \right) \\
&\quad + \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \Bigg(\left\| \tilde{c}_{\boldsymbol{\nu},\tau}(\cdot) f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) - \tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}} \left(\tilde{c}_{\boldsymbol{\nu},\tau}(\cdot), f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right) \right\|_{L^q(\mathrm{D})}^q \\
&\quad + \left\| \left(f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) - [D\tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}}]_1 \left(\tilde{c}_{\boldsymbol{\nu},\tau}(\cdot), f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}}) \right) \right) \nabla \tilde{c}_{\boldsymbol{\nu},\tau}(\cdot) \right\|_{L^q(\mathrm{D})^d}^q \Bigg)^{1/q} \\
&\overset{(*)}{\le} C \sum_{\boldsymbol{\nu}\in\mathcal{F}\setminus\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)} \\
&\quad + \tfrac{1}{3} \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \|(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}\|_{\ell^1(\mathcal{F})}^{-1} |\Lambda_\tau|^{-1/p+1} + \tfrac{2}{3} \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}}^{-1} |\Lambda_\tau|^{-1/p} \\
&\quad + \sum_{\boldsymbol{\nu}\in\Lambda_\tau} C \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s} m_{\tau;\boldsymbol{\nu}}^{-\gamma} \|L_{\boldsymbol{\nu}_E}\|_{L^\infty(U_E)}
\end{aligned}$$

$$
\begin{aligned}
&+ \tfrac{1}{3} \sum_{\boldsymbol{\nu}\in\Lambda_\tau} C \left\|c_{\boldsymbol{\nu}}\right\|_{\mathcal{X}} \left\|(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}\right\|_{\ell^1(\mathcal{F})}^{-1} |\Lambda_\tau|^{-1/p+1} + \tfrac{2}{3} \sum_{\boldsymbol{\nu}\in\Lambda_\tau} C \left\|c_{\boldsymbol{\nu}}\right\|_{\mathcal{X}} \left\|c_{\boldsymbol{\nu}}\right\|_{\mathcal{X}}^{-1} |\Lambda_\tau|^{-1/p} \\
&+ \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \left(\lambda^q \left\|1\right\|_{L^q(\mathrm{D})}^q + \lambda^q \left\|\nabla \tilde{c}_{\boldsymbol{\nu},\tau}\right\|_{L^q(\mathrm{D})^d}^q\right)^{1/q} \\
&\leq C\left[\mathcal{N}_\tau^{-r} + |\Lambda_\tau|^{-1/p+1} + \mathcal{N}_\tau^{-r} + |\Lambda_\tau|^{-1/p+1} + \mathcal{N}_\tau^{-r}\right] \leq C\mathcal{N}_\tau^{-r}.
\end{aligned}
$$

In case $q = \infty$, the ℓ^q-sums have to be replaced by a maximum.

At (*), the first term can be estimated with Equation (63). To obtain the second and third term, we used Lemma A.1. To obtain the fourth term, we used $\|c_{\boldsymbol{\nu}} - \tilde{c}_{\boldsymbol{\nu},\tau}\|_{\mathcal{X}} \leq C\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}^s} m_{\tau,\boldsymbol{\nu}}^{-\gamma}$ from Assumption 5.1 and we used Equation (63) to estimate $\sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}} - \tilde{c}_{\boldsymbol{\nu},\tau}\|_{\mathcal{X}} \sup_{\boldsymbol{y}\in U} \left|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F}\right|$. To obtain the fifth and sixth term, we used that by Assumption 5.1 $\|c_{\boldsymbol{\nu}} - \tilde{c}_{\boldsymbol{\nu},\tau}\|_{\mathcal{X}} \leq \|c_{\boldsymbol{\nu}}\|_{\mathcal{X}} + \|\tilde{c}_{\boldsymbol{\nu},\tau}\|_{\mathcal{X}} \leq C\left\|c_{\boldsymbol{\nu}}\right\|_{\mathcal{X}}$ to estimate $\sum_{\boldsymbol{\nu}\in\Lambda_\tau} \|c_{\boldsymbol{\nu}} - \tilde{c}_{\boldsymbol{\nu},\tau}\|_{\mathcal{X}} \sup_{\boldsymbol{y}\in U} \left|L_{\boldsymbol{\nu}_E}(\boldsymbol{y}_E)\boldsymbol{y}_F^{\boldsymbol{\nu}_F} - f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}})\right|$ using Lemma A.1. To obtain the seventh term, Proposition 4.3 was used, and to estimate it, we again used Assumption 5.1 and $|\Lambda_\tau| \leq \mathcal{N}_\tau$ to obtain $|\Lambda_\tau|\mathcal{N}_\tau^{-r-1} \leq \mathcal{N}_\tau^{-r}$:

$$
\begin{aligned}
\sum_{\boldsymbol{\nu}\in\Lambda_\tau} \left(\lambda^q \left\|1\right\|_{L^q(\mathrm{D})}^q + \lambda^q \left\|\nabla \tilde{c}_{\boldsymbol{\nu},\tau}\right\|_{L^q(\mathrm{D})^d}^q\right)^{1/q} &\leq \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \left(\lambda \left\|1\right\|_{L^q(\mathrm{D})} + \lambda C \left\|c_{\boldsymbol{\nu}}\right\|_{\mathcal{X}}\right) \\
&\leq C|\Lambda_\tau|\lambda + C\lambda \left\|(\|c_{\boldsymbol{\nu}}\|_{\mathcal{X}})_{\boldsymbol{\nu}\in\mathcal{F}}\right\|_{\ell^1(\mathcal{F})} \\
&\leq C\mathcal{N}_\tau^{-r}.
\end{aligned}
$$

Step 4. We now construct a network emulating $\tilde{u}_\tau$. It is the concatenation of four subnetworks, $\tilde{u}_\tau := \tilde{u}_\tau^{(5)} \circ \tilde{u}_\tau^{(6)} \circ \tilde{u}_\tau^{(7)} \circ \tilde{u}_\tau^{(8)}$. The first NN $\tilde{u}_\tau^{(8)}$ has input dimension $d + |S_{\Lambda_\tau}|$, output dimension $2|\Lambda_\tau|$ and is defined as

$$
\tilde{u}_\tau^{(8)} := \left(\tilde{g}_{\Lambda_\tau} \circ \mathrm{Id}_{\mathbb{R}^d}, \tilde{f}_{\Lambda_\tau} \circ \mathrm{Id}_{\mathbb{R}^{|S_{\Lambda_\tau}|}}\right)_{\mathrm{d}},
$$

where the depth of the identity networks is such that $\mathrm{depth}(\tilde{u}_\tau^{(8)}) = 1 + \max\{\mathrm{depth}(\tilde{g}_{\Lambda_\tau}), \mathrm{depth}(\tilde{f}_{\Lambda_\tau})\}$. The second NN $\tilde{u}_\tau^{(7)}$ emulates an affine map. It has depth 0, and its input dimension and output dimension both equal $2|\Lambda_\tau|$. For a fixed but arbitrary enumeration $(\boldsymbol{\nu}^{(j)})_{j=1}^{|\Lambda_\tau|}$, the NN $\tilde{u}_\tau^{(7)}$ is defined such that

$$
\left.
\begin{aligned}
\left(\tilde{u}_\tau^{(7)} \circ \tilde{u}_\tau^{(8)}(\boldsymbol{x}, (y_j)_{j\in S_{\Lambda_\tau}})\right)_{2k-1} &= \tilde{c}_{\boldsymbol{\nu}^{(k)},\tau}(\boldsymbol{x}), \\
\left(\tilde{u}_\tau^{(7)} \circ \tilde{u}_\tau^{(8)}(\boldsymbol{x}, (y_j)_{j\in S_{\Lambda_\tau}})\right)_{2k} &= f_{\boldsymbol{\nu}^{(k)},\tau}((y_j)_{j\in S_{\Lambda_\tau}}),
\end{aligned}
\right\}
\quad
\begin{aligned}
&\forall \boldsymbol{x} \in \mathrm{D},\ \forall \boldsymbol{y} \in U, \\
&\forall k = 1, \ldots, |\Lambda_\tau|.
\end{aligned}
$$

The third NN $\tilde{u}_\tau^{(6)}$ is a parallelization of NNs from Proposition 4.3:

$$
\tilde{u}_\tau^{(6)} := \left(\left\{\mathrm{Id}_{\mathbb{R}} \circ \tilde{\times}_{\lambda, R'_{\boldsymbol{\nu}^{(k)}}}\right\}_{k=1}^{|\Lambda_\tau|}\right)_{\mathrm{d}},
$$

where the depth of the identity networks is such that all components of the parallelization have equal depth, so that the parallelization has depth $\max_{\boldsymbol{\nu}\in\Lambda_\tau} 1 + \mathrm{depth}(\tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}})$. For all $k = 1, \dots, |\Lambda_\tau|$, the k'th component of the output of $\tilde{u}_\tau^{(6)}$ is

$$\left(\tilde{u}_\tau^{(6)} \circ \tilde{u}_\tau^{(7)} \circ \tilde{u}_\tau^{(8)}(\boldsymbol{x}, (y_j)_{j\in S_{\Lambda_\tau}})\right)_k = \tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}^{(k)}}} \left(\tilde{c}_{\boldsymbol{\nu}^{(k)},\tau}(\boldsymbol{x}), f_{\boldsymbol{\nu}^{(k)},\tau}((y_j)_{j\in S_{\Lambda_\tau}})\right), \qquad \forall \boldsymbol{x} \in \mathrm{D},\ \forall \boldsymbol{y} \in U.$$

Finally, $\tilde{u}_\tau^{(5)}$ has depth 0, input dimension $|\Lambda_\tau|$, output dimension 1 and computes the sum of its inputs. As a result, it holds that

$$\tilde{u}_\tau(\boldsymbol{x}, (y_j)_{j\in S_{\Lambda_\tau}}) = \sum_{\boldsymbol{\nu}\in\Lambda_\tau} \tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}} \left(\tilde{c}_{\boldsymbol{\nu},\tau}(\boldsymbol{x}), f_{\boldsymbol{\nu},\tau}((y_j)_{j\in S_{\Lambda_\tau}})\right), \qquad \forall \boldsymbol{x} \in \mathrm{D},\ \forall \boldsymbol{y} \in U.$$

Step 5. Finally, we bound the NN depth and size of $\tilde{u}_\tau$.

We first estimate the network depth. It follows from Assumption 5.1 and Lemma A.1 that

$$\begin{aligned}
\mathrm{depth}(\tilde{u}_\tau^{(8)}) &= 1 + \max\{\mathrm{depth}(\tilde{g}_{\Lambda_\tau}), \mathrm{depth}(\tilde{f}_{\Lambda_\tau})\} \\
&\leq 1 + \max\{C(1 + \log \mathcal{N}_\tau), C(1 + \log|\Lambda_\tau| \cdot \log\log|\Lambda_\tau|)\} \\
&\leq C(1 + \log \mathcal{N}_\tau \cdot \log\log \mathcal{N}_\tau).
\end{aligned}$$

In addition, it holds that

$$\begin{aligned}
\mathrm{depth}(\tilde{u}_\tau^{(7)}) &= 0, \\
\mathrm{depth}(\tilde{u}_\tau^{(6)}) &= \max_{\boldsymbol{\nu}\in\Lambda_\tau} 1 + \mathrm{depth}(\tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}}) \leq \max_{\boldsymbol{\nu}\in\Lambda_\tau} C(1 + \log(R'_{\boldsymbol{\nu}}/\lambda)) \\
&\leq C \max_{\boldsymbol{\nu}\in\Lambda_\tau} (1 + J\log\log(|\Lambda_\tau|) + \theta\log(m_{\tau;\boldsymbol{\nu}}) + (r+1)\log(\mathcal{N}_\tau)) \\
&\leq C(1 + \log \mathcal{N}_\tau), \\
\mathrm{depth}(\tilde{u}_\tau^{(5)}) &= 0, \\
\mathrm{depth}(\tilde{u}_\tau) &\leq \mathrm{depth}(\tilde{u}_\tau^{(5)}) + 1 + \mathrm{depth}(\tilde{u}_\tau^{(6)}) + 1 + \mathrm{depth}(\tilde{u}_\tau^{(7)}) + 1 + \mathrm{depth}(\tilde{u}_\tau^{(8)}) \\
&\leq C(1 + \log \mathcal{N}_\tau \cdot \log\log \mathcal{N}_\tau).
\end{aligned}$$

We now estimate the network size. By Proposition 3.8 item (i), it follows that $|S_{\Lambda_\tau}| \leq |\Lambda_\tau|$. As a result, the sizes of the identity networks in $\tilde{u}_\tau^{(8)}$ can be estimated as follows:

$$\begin{aligned}
\mathrm{size}(\mathrm{Id}_{\mathbb{R}^d}) &\leq 2d(1 + \mathrm{depth}(\tilde{u}_\tau^{(8)})) \leq C(1 + \log \mathcal{N}_\tau \cdot \log\log \mathcal{N}_\tau), \\
\mathrm{size}(\mathrm{Id}_{\mathbb{R}^{|S_{\Lambda_\tau}|}}) &\leq 2|S_{\Lambda_\tau}|(1 + \mathrm{depth}(\tilde{u}_\tau^{(8)})) \leq C(1 + \mathcal{N}_\tau \cdot \log \mathcal{N}_\tau \cdot \log\log \mathcal{N}_\tau).
\end{aligned}$$

We find:

$$\mathrm{size}(\tilde{u}_\tau^{(8)}) \leq 2\,\mathrm{size}(\tilde{g}_{\Lambda_\tau}) + 2\,\mathrm{size}(\mathrm{Id}_{\mathbb{R}^d}) + 2\,\mathrm{size}(\tilde{f}_{\Lambda_\tau}) + 2\,\mathrm{size}(\mathrm{Id}_{\mathbb{R}^{|S_{\Lambda_\tau}|}})$$

$$\begin{aligned}
&\leq 2C\mathcal{N}_\tau + 2C(1+\log\mathcal{N}_\tau\cdot\log\log\mathcal{N}_\tau) + 2C(1+|\Lambda_\tau|\cdot\log|\Lambda_\tau|\cdot\log\log|\Lambda_\tau|)\\
&\quad + 2C(1+\mathcal{N}_\tau\cdot\log\mathcal{N}_\tau\cdot\log\log\mathcal{N}_\tau)\\
&\leq 2C(1+\mathcal{N}_\tau\cdot\log\mathcal{N}_\tau\cdot\log\log\mathcal{N}_\tau).
\end{aligned}$$

Because each component of the output of $\tilde{u}_\tau^{(7)}$ only depends on one component of its input, it holds that $\operatorname{size}(\tilde{u}_\tau^{(7)}) \leq 2|\Lambda_\tau|$. Furthermore, it holds that

$$\begin{aligned}
\operatorname{size}(\tilde{u}_\tau^{(6)}) &\leq \sum_{\boldsymbol{\nu}\in\Lambda_\tau} 2\operatorname{size}(\mathrm{Id}_{\mathbb{R}}) + 2\operatorname{size}(\tilde{\times}_{\lambda,R'_{\boldsymbol{\nu}}})\\
&\leq \sum_{\boldsymbol{\nu}\in\Lambda_\tau} 4(1+\operatorname{depth}(\tilde{u}_\tau^{(6)})) + C(1+\log(R'_{\boldsymbol{\nu}}/\lambda)) \leq C(1+\mathcal{N}_\tau\cdot\log\mathcal{N}_\tau),\\
\operatorname{size}(\tilde{u}_\tau^{(5)}) &\leq |\Lambda_\tau|,\\
\operatorname{size}(\tilde{u}_\tau) &\leq 4\operatorname{size}(\tilde{u}_\tau^{(5)}) + 4\operatorname{size}(\tilde{u}_\tau^{(6)}) + 4\operatorname{size}(\tilde{u}_\tau^{(7)}) + 4\operatorname{size}(\tilde{u}_\tau^{(8)})\\
&\leq C(1+\mathcal{N}_\tau\cdot\log\mathcal{N}_\tau\cdot\log\log\mathcal{N}_\tau).
\end{aligned}$$

This finishes the proof. □

A.5 Proof of Proposition 6.2

To prove Proposition 6.2, we will use [63, Theorem 6.7]. In the following lemma, we verify the assumptions of that result concerning the approximation of the Gaussian density function, using [63, Theorem 5.15], and cutting off the NN approximation sufficiently far away from zero.

Lemma A.2. *Let $g:\mathbb{R}\to\mathbb{R}: x\mapsto \exp(-\frac{1}{2}x^2)$.*

For all $\beta\in(0,1]$ there exists a σ_1-NN Φ_β^g with input dimension 1 and output dimension 1 and an absolute constant $C>0$ such that

$$\left\|g-\Phi_\beta^g\right\|_{L^\infty(\mathbb{R})} \leq \beta = \beta\left\|g\right\|_{L^\infty(\mathbb{R})},$$
$$\operatorname{depth}(\Phi_\beta^g) \leq C(1+\log(1/\beta)\log\log(1/\beta)),\ \operatorname{size}(\Phi_\beta^g) \leq C(1+\log(1/\beta))^2.$$

Proof. For arbitrary $\beta\in(0,1]$, we first construct a ReLU NN approximation $\Phi^g_{\beta/3,[-R,R]}$ of g satisfying $\left\|g-\Phi^g_{\beta/3,[-R,R]}\right\|_{L^\infty([-R,R])} \leq \beta/3$, for $R := 1+\sqrt{2\log(3/\beta)}$. Here, $R>1$ is chosen such that $g(R-1)=\beta/3=\|g\|_{L^\infty(-\infty,-R+1)} = \|g\|_{L^\infty(R-1,\infty)}$. Let $h:\mathbb{R}\to\mathbb{R}: x\mapsto\exp(-\frac{1}{2}x)$, so that $h(x^2)=g(x)$, $x\in\mathbb{R}$. For the approximation of h on $[0,R^2]$, ReLU NNs obtain exponential convergence, with network size independent of R. Applying [63, Theorem 5.15] (see also the remark after that result) to $h((R^2+2)(x+1)/2)$, $x\in[-1,1]$ with accuracy $\beta/(6\exp(1/2))$

and with the weights in the output layer multiplied by $\exp(1/2)$, we obtain that for all $R > 1$, for $h_R(x) := \exp(1/2)h((R^2+2)(x+1)/2) = h((R^2+2)(x+1)/2-1)$, $x \in [-1,1]$, there exists a NN $\Phi^{h_R}_{\beta/6,[-1,1]}$ satisfying, for an absolute constant $C > 0$ *independent of* R

$$\begin{aligned}\left\|h_R - \Phi^{h_R}_{\beta/6,[-1,1]}\right\|_{L^\infty([-1,1])} &\le \beta/6,\\ \operatorname{depth}(\Phi^{h_R}_{\beta/6,[-1,1]}) &\le C(1+\log(1/\beta)\log\log(1/\beta)),\\ \operatorname{size}(\Phi^{h_R}_{\beta/6,[-1,1]}) &\le C(1+\log(1/\beta))^2.\end{aligned}$$

Here, we applied [63, Theorem 5.15] with $\eta := 2(2+R^2)^{-1}$ and $1/\eta \le C(1+\log(1/\beta))$, for η as defined in that reference.

Let T be the affine transformation $\mathbb{R} \to \mathbb{R} : x \mapsto 2(x+1)/(2+R^2) - 1$ satisfying $T([-1, R^2+1]) = [-1,1]$ and $h = h_R \circ T$. Then, the NN $\Phi^{h_R}_{\beta/6,[-1,1]} \circ T$ approximates h on $[-1, R^2+1]$ with network size bounded as stated above. The map g can be approximated as

$$\Phi^g_{\beta/3,[-R,R]}(x) := \Phi^{h_R}_{\beta/6,[-1,1]} \circ T \circ \tilde{\times}_{\beta/6,R}(x,x), \quad x \in [-R,R].$$

To bound the error, we use that for all $x \in [-R,R]$ it holds $\tilde{\times}_{\beta/6,R}(x,x) \in [-1, R^2+1]$ and thus $T(\tilde{\times}_{\beta/6,R}(x,x)) \in [-1,1]$. Note that we approximate h on $[-1,R^2+1]$ rather than $[0,R^2]$ because $\tilde{\times}_{\beta/6,R}(x,x)$ need not be in $[0,R^2]$ for all $x \in [-R,R]$. We obtain the following error estimate, for all $R > 1$, using that $|h|_{W^{1,\infty}([-1,\infty))} = \frac{1}{2}\exp(1/2) < 1$:

$$\begin{aligned}&\left\|g - \Phi^g_{\beta/3,[-R,R]}\right\|_{L^\infty([-R,R])}\\ &\le \left\|h((\cdot)^2) - h\left(\tilde{\times}_{\beta/6,R}(\cdot,\cdot)\right)\right\|_{L^\infty([-R,R])}\\ &\quad + \left\|h_R \circ T\left(\tilde{\times}_{\beta/6,R}(\cdot,\cdot)\right) - \Phi^{h_R}_{\beta/6,[-1,1]} \circ T\left(\tilde{\times}_{\beta/6,R}(\cdot,\cdot)\right)\right\|_{L^\infty([-R,R])}\\ &\le |h|_{W^{1,\infty}([-1,R^2+1])}\left\|(\cdot)^2 - \tilde{\times}_{\beta/6,R}(\cdot,\cdot)\right\|_{L^\infty([-R,R])}\\ &\quad + \left\|h_R - \Phi^{h_R}_{\beta/6,[-1,1]}\right\|_{L^\infty([-1,1])}\\ &\le \tfrac{\beta}{6} + \tfrac{\beta}{6} = \tfrac{\beta}{3}.\end{aligned}$$

We estimate the NN depth and size as

$$\begin{aligned}\operatorname{depth}(\Phi^g_{\beta/3,[-R,R]}) &\le \operatorname{depth}\left(\Phi^{h_R}_{\beta/6,[-1,1]}\right) + 1 + \operatorname{depth}(T) + 1 + \operatorname{depth}\left(\tilde{\times}_{\beta/6,R}(\cdot,\cdot)\right)\\ &\le C(1+\log(6/\beta)\log\log(6/\beta)) + 1 + 0 + 1 + C(1+\log(6R/\beta))\\ &\le C(1+\log(1/\beta)\log\log(1/\beta)),\end{aligned}$$

$$\begin{aligned}\operatorname{size}(\Phi^g_{\beta/3,[-R,R]}) &\le 2\operatorname{size}\left(\Phi^{h_R}_{\beta/6,[-1,1]}\right) + 4\operatorname{size}(T) + 4\operatorname{size}\left(\tilde{\times}_{\beta/6,R}(\cdot,\cdot)\right)\\ &\le C(1+\log(6/\beta))^2 + 8 + C(1+\log(6R/\beta))\\ &\le C(1+\log(1/\beta))^2,\end{aligned}$$

for C independent of R, using that $R \le C(1+\log(1/\beta))^{1/2}$.

Based on $\Phi^g_{\beta/3,[-R,R]}$, we define the following ReLU NN approximation of g on $\mathbb{R}$:

$$\Phi^g_\beta(x) := \tilde{\times}_{\beta/3,2}\left(\Phi^g_{\beta/3,[-R,R]}(x), \max\{0, R-|x|\} - \max\{0, R-1-|x|\}\right).$$

This can be emulated exactly by the network

$$\begin{aligned}\tilde{\times}_{\beta/3,2} \circ B \circ \Big(&\Phi^g_{\beta/3,[-R,R]}, \sigma_1(\cdot + R)\circ \mathrm{Id}_{\mathbb{R}}, \sigma_1(\cdot + R - 1)\circ \mathrm{Id}_{\mathbb{R}},\\ &\sigma_1(R-1-\cdot)\circ \mathrm{Id}_{\mathbb{R}}, \sigma_1(R-\cdot)\circ \mathrm{Id}_{\mathbb{R}}\Big),\end{aligned}$$

where $B : \mathbb{R}^5 \to \mathbb{R}^2 : (x_1, x_2, x_3, x_4, x_5) \mapsto (x_1, x_2 - x_3 - x_4 + x_5)$ and where the depth of the identity networks is $\operatorname{depth}(\Phi^g_{\beta/3,[-R,R]}) - 2$, such that all components of the parallelization have equal depth.

We estimate the NN depth and size as

$$\begin{aligned}\operatorname{depth}(\Phi^g_\beta) &\le \operatorname{depth}(\tilde{\times}_{\beta/3,2}) + 1 + \operatorname{depth}(B) + 1 + \operatorname{depth}(\Phi^g_{\beta/3,[-R,R]})\\ &\le C(1+\log(3/\beta)) + 1 + 0 + 1 + C(1+\log(1/\beta)\log\log(1/\beta))\\ &\le C(1+\log(1/\beta)\log\log(1/\beta)),\\ \operatorname{size}(\Phi^g_\beta) &\le 4\operatorname{size}(\tilde{\times}_{\beta/3,2}) + 4\operatorname{size}(B) + 2\operatorname{size}(\Phi^g_{\beta/3,[-R,R]}) + 4\operatorname{size}(\sigma_1(\cdot+R))\\ &\quad + 4\operatorname{size}(\mathrm{Id}_{\mathbb{R}}) + 4\operatorname{size}(\sigma_1(\cdot+R-1)) + 4\operatorname{size}(\mathrm{Id}_{\mathbb{R}})\\ &\quad + 4\operatorname{size}(\sigma_1(R-1-\cdot)) + 4\operatorname{size}(\mathrm{Id}_{\mathbb{R}}) + 4\operatorname{size}(\sigma_1(R-\cdot)) + 4\operatorname{size}(\mathrm{Id}_{\mathbb{R}})\\ &\le C(1+\log(3/\beta)) + 20 + 2\big(C(1+\log(1/\beta))^2\big)\\ &\quad + 4\big(12 + 8\big(C(1+\log(1/\beta)\log\log(1/\beta))\big)\big)\\ &\le C(1+\log(1/\beta))^2.\end{aligned}$$

On $[0, R-1]$ and $[R-1, R]$, respectively, it holds that

$$\begin{aligned}&\left\|g - \Phi^g_\beta\right\|_{L^\infty([0,R-1])}\\ &\le \left\|g - \Phi^g_{\beta/3,[-R,R]}\right\|_{L^\infty([0,R-1])}\\ &\quad + \left\|\Phi^g_{\beta/3,[-R,R]}(\cdot) - \tilde{\times}_{\beta/3,2}\left(\Phi^g_{\beta/3,[-R,R]}(\cdot), 1\right)\right\|_{L^\infty([0,R-1])}\\ &\le \beta/3 + \beta/3 < \beta,\end{aligned}$$

$$
\begin{aligned}
&\left\|g-\Phi_\beta^g\right\|_{L^\infty([R-1,R])}\\
&\le \|g(\cdot)-(R-\cdot)g(\cdot)\|_{L^\infty([R-1,R])}\\
&\quad+\left\|(R-\cdot)g(\cdot)-(R-\cdot)\Phi^g_{\beta/3,[-R,R]}(\cdot)\right\|_{L^\infty([R-1,R])}\\
&\quad+\left\|(R-\cdot)\Phi^g_{\beta/3,[-R,R]}(\cdot)-\tilde{\times}_{\beta/3,2}(\Phi^g_{\beta/3,[-R,R]}(\cdot),(R-\cdot))\right\|_{L^\infty([R-1,R])}\\
&\le \beta/3+\beta/3+\beta/3=\beta.
\end{aligned}
$$

On (R,∞), it holds that $\Phi_\beta^g\equiv 0$ and hence $\left\|g-\Phi_\beta^g\right\|_{L^\infty([R,\infty))}\le\beta/3<\beta$. The same estimates hold on $(-\infty,0]$, which finishes the proof of the lemma. □

Using [66, Lemma 3.5] instead of [63, Theorem 5.15] for the approximation of h, the bound on the network size would be $C(1+\log(1/\beta))^2\lceil R\rceil\le C(1+\log(1/\beta))^{5/2}$.

Proof of Proposition 6.2. We apply [63, Theorem 6.7], for g and Φ_β^g as in the Lemma A.2 above. With $\beta:=\varepsilon/2$, $R=1+\sqrt{2\log(3/\beta)}$ and $D:=\{\boldsymbol{x}\in\mathbb{R}^N:\|A\boldsymbol{x}\|_2\le R\}$, we obtain $\Phi_\varepsilon^{\mathfrak{g}}$ satisfying

$$
\begin{aligned}
\left\|\mathfrak{g}-\Phi_\varepsilon^{\mathfrak{g}}\right\|_{L^\infty(D)}&\le\varepsilon\|g\|_{W^{1,\infty}(D)}\le\varepsilon,\\
\operatorname{depth}(\Phi_\varepsilon^{\mathfrak{g}})&\le C(1+\log(2/\varepsilon)\log\log(2/\varepsilon))+\log(N)\log_2(10\pi NR(2/\varepsilon))+1\\
&\le C\log(N)(1+\log(N/\varepsilon))+C(1+\log(1/\varepsilon)\log\log(1/\varepsilon)),\\
\operatorname{size}(\Phi_\varepsilon^{\mathfrak{g}})&\le 2C(1+\log(2/\varepsilon))^2+4N^2+64(N-1)\log_2(10\pi NR(2/\varepsilon))+4N\\
&\le C(1+\log(1/\varepsilon))^2+CN\log(1/\varepsilon)+CN^2.
\end{aligned}
$$

On $\mathbb{R}^N\backslash D$, it holds that $\Phi_\varepsilon^{\mathfrak{g}}=0$, which follows from the fact that the network Φ_β^g constructed in Lemma A.2 vanishes on (R,∞). We recall from the proof of the lemma that R was defined such that $\left\|\mathfrak{g}-\Phi_\varepsilon^{\mathfrak{g}}\right\|_{L^\infty(\mathbb{R}^N\backslash D)}\le\|g\|_{L^\infty((-\infty,-R+1)\cup(R-1,\infty))}=\beta/3=\varepsilon/6$. Combined with the estimate above, it holds that $\left\|\mathfrak{g}-\Phi_\varepsilon^{\mathfrak{g}}\right\|_{L^\infty(\mathbb{R}^N)}\le\varepsilon$. □

Acknowledgment: Supported in part by SNSF Grant No. 159940. Work performed in part during visit of CS and JZ at the CRM, Montreal, Canada in March 2019, and by a visit of JZ to the FIM, Department of Mathematics, ETH Zürich, in October 2019. JZ acknowledges support by the Swiss National Science Foundation under Early Postdoc.Mobility Fellowship 184530. This paper was written during the postdoctoral stay of JZ at MIT. This extend version of the published version [64] additionally contains appendices with several proofs of results from the main text.

References

[1] Roman Andreev and Christoph Schwab. Sparse tensor approximation of parametric eigenvalue problems. In *Numerical Analysis of Multiscale Problems*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 203–241. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[2] I. Babuška, R. B. Kellogg, and J. Pitkäranta. Direct and inverse error estimates for finite elements with mesh refinements. *Numerische Mathematik*, 33(4):447–471, 1979.

[3] Markus Bachmayr, Albert Cohen, Dinh Dũng, and Christoph Schwab. Fully discrete approximation of parametric and stochastic elliptic PDEs. *SIAM J. Numer. Anal.*, 55(5):2151–2186, 2017.

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

[5] Constantin Băcuţă, Hengguang Li, and Victor Nistor. Differential operators on domains with conical points: precise uniform regularity estimates. *Rev. Roumaine Math. Pures Appl.*, 62(3):383–411, 2017.

[6] Constantin Băcuţă, Victor Nistor, and Ludmil T. Zikatanov. Improving the rate of convergence of 'high order finite elements' on polygons and domains with cusps. *Numerische Mathematik*, 100(2):165–184, 2005.

[7] Andrew 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.

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

[9] J. Berg and K. Nyström. Neural network augmented inverse problems for PDEs. 2017. ArXiv:1712.09685.

[10] Helmut Bölcskei, Philipp Grohs, Gitta Kutyniok, and Philipp Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM Journal on Mathematics of Data Science*, 1(1):8–45, 2019.

[11] Peng Chen and Christoph Schwab. Sparse-grid, reduced-basis Bayesian inversion: Nonaffine-parametric nonlinear equations. *Journal of Computational Physics*, 316:470–503, 2016.

[12] Abdellah Chkifa, Albert Cohen, and Christoph Schwab. High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. *Journ. Found. Comp. Math.*, 14(4):601–633, 2013.

[13] Albert Cohen, Abdellah Chkifa, and Christoph Schwab. Breaking the curse

of dimensionality in sparse polynomial approximation of parametric PDEs. *Journ. Math. Pures et Appliquees*, 103(2):400–428, 2015.

[14] Albert Cohen and Ronald DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015.

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

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

[17] Albert Cohen, Christoph Schwab, and Jakob Zech. Shape holomorphy of the stationary Navier-Stokes equations. *SIAM J. Math. Analysis*, 50(2):1720–1752, 2018.

[18] Nadav Cohen, Or Sharir, and Amnon 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.

[19] Stephan Dahlke, Markus Hansen, Cornelia Schneider, and Winfried Sickel. Properties of Kondratiev spaces, 2019. ArXiv:1911.01962.

[20] Masoumeh Dashti, Stephen Harris, and Andrew Stuart. Besov priors for Bayesian inverse problems. *Inverse Probl. Imaging*, 6(2):183–200, 2012.

[21] Masoumeh Dashti and Andrew M. Stuart. The Bayesian approach to inverse problems. In *Handbook of uncertainty quantification. Vol. 1, 2, 3*, pages 311–428. Springer, Cham, 2017.

[22] P.J. Davis. *Interpolation and Approximation*. Dover Books on Mathematics. Dover Publications, 1975.

[23] Joseph Daws Jr. and Clayton G. Webster. A polynomial-based approach for architectural design and learning with deep neural networks. 2019. ArXiv:1905.10457.

[24] Persi Diaconis and David Freedman. On the consistency of Bayes estimates. *Ann. Statist.*, 14(1):1–67, 1986. With a discussion and a rejoinder by the authors.

[25] Josef Dick, Robert N. Gantner, Quoc T. Le Gia, and Christoph Schwab. Multilevel higher-order quasi-Monte Carlo Bayesian estimation. *Math. Models Methods Appl. Sci.*, 27(5):953–995, 2017.

[26] Dinh Dũng. Linear collective collocation and Galerkin approximations for parametric and stochastic elliptic PDEs, 2015. ArXiv:1511.03377.

[27] Dennis Elbrächter, Philipp Grohs, Arnulf Jentzen, and Christoph Schwab. DNN expression rate analysis of high-dimensional PDEs: Application to option pricing. *Constructive Approximation*, 55(1):3–71, 2022.

[28] Ken-Ichi Funahashi. On the approximate realization of continuous mappings by neural networks. *Neural networks*, 2(3):183–192, 1989.

[29] A. D. Gilbert, I. G. Graham, F. Y. Kuo, R. Scheichl, and I. H. Sloan. Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients. *Numer. Math.*, 142(4):863–915, 2019.

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

[31] Philipp Grohs, Thomas Wiatowski, and Helmut Bölcskei. Deep convolutional neural networks on cartoon functions. In *2016 IEEE International Symposium on Information Theory (ISIT)*, pages 1163–1167, 2016.

[32] Philipp A. Guth, Vesa Kaarnioja, Frances Y. Kuo, Claudia Schillings, and Ian H. Sloan. A quasi-Monte Carlo method for optimal control under uncertainty. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):354–383, 2021.

[33] Juncai He, Lin Li, Jinchao Xu, and Chunyue Zheng. ReLU deep neural networks and linear finite elements. *Journal Sci. Comp.*, 38(3):502–527, 2020.

[34] Fernando Henríquez and Christoph Schwab. Shape holomorphy of the Calderón projector for the Laplacian in $\mathbb{R}^2$. *Integral Equations and Operator Theory*, 93(4):43, 2021.

[35] Lukas Herrmann, Magdalena Keller, and Christoph Schwab. Quasi-Monte Carlo Bayesian estimation under Besov priors in elliptic inverse problems. *Mathematics of Computation*, 90:1831–1860, 2021.

[36] Lukas Herrmann and Christoph Schwab. Multilevel quasi-Monte Carlo uncertainty quantification for advection-diffusion-reaction. *Proc. MCQMC2018*, 324:31–67, 2020.

[37] Lukas Herrmann, Christoph Schwab, and Jakob Zech. Deep neural network expression of posterior expectations in Bayesian PDE inversion. *Inverse Problems*, 36(12):125011, 2020.

[38] R. Hiptmair, L. Scarabosio, C. Schillings, and Ch. Schwab. Large deformation shape uncertainty quantification in acoustic scattering. *Adv. Comput. Math.*, 44(5):1475–1518, 2018.

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

[40] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[41] Bamdad Hosseini and Nilima Nigam. Well-posed Bayesian inverse problems: priors with exponential tails. *SIAM/ASA J. Uncertain. Quantif.*, 5(1):436–465, 2017.

[42] Thomas Y. Hou, Ka Chun Lam, Pengchuan Zhang, and Shumao Zhang. Solving Bayesian inverse problems from the perspective of deep generative networks. *Comput. Mech.*, 64(2):395–408, 2019.

[43] Carlos Jerez-Hanckes, Christoph Schwab, and Jakob Zech. Electromagnetic wave scattering by random surfaces: Shape holomorphy. *Math. Mod. Meth. Appl. Sci.*, 27(12):2229–2259, 2017.

[44] B. T. Knapik, A. W. van der Vaart, and J. H. van Zanten. Bayesian inverse problems with Gaussian priors. *Ann. Statist.*, 39(5):2626–2657, 2011.

[45] Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *Constructive Approximation*, 55(1):73–125, 2022.

[46] Matti Lassas, Eero Saksman, and Samuli Siltanen. Discretization-invariant Bayesian inversion and Besov space priors. *Inverse Probl. Imaging*, 3(1):87–122, 2009.

[47] Bo Li, Shanshan Tang, and Haijun Yu. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Commun. Comput. Phys.*, 27(2):379–411, 2020.

[48] Hengguang Li, Anna Mazzucato, and Victor Nistor. Analysis of the finite element method for transmission/mixed boundary value problems on general polygonal domains. *Electron. Trans. Numer. Anal.*, 37:41–69, 2010.

[49] Liang Liang, Fanwei Kong, Caitlin Martin, Thuy Pham, Qian Wang, James Duncan, and Wei 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.

[50] Liang Liang, Minlian Liu, Caitlin Martin, and Wei 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.

[51] Shiyu Liang and Roger Srikant. Why deep neural networks for function approximation? In *Proc. of ICLR 2017*, 2017. ArXiv:1610.04161.

[52] Youssef Marzouk, Tarek Moselhy, Matthew Parno, and Alessio Spantini. Sampling via measure transport: an introduction. In *Handbook of uncertainty quantification. Vol. 1, 2, 3*, pages 785–825. Springer, Cham, 2017.

[53] Vladimir Maz'ya and Jürgen Rossmann. *Elliptic equations in polyhedral domains*, volume 162 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2010.

[54] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1(1):61–80, Feb 1993.

[55] H. N. Mhaskar. Neural networks for localized approximation of real functions.

In *Neural Networks for Signal Processing III - Proceedings of the 1993 IEEE-SP Workshop*, pages 190–196. IEEE, 1993.

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

[57] James R. Munkres. *Topology*. Prentice Hall, Inc., Upper Saddle River, NJ, second edition, 2000.

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

[59] F. Nobile, R. Tempone, and C. G. Webster. An anisotropic sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM J. Numer. Anal.*, 46(5):2411–2442, 2008.

[60] Ricardo H. Nochetto and Andreas Veeser. Primer of adaptive finite element methods. In *Multiscale and adaptivity: modeling, numerics and applications*, volume 2040 of *Lecture Notes in Math.*, pages 125–225. Springer, Heidelberg, 2012.

[61] Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, editors. *NIST handbook of mathematical functions*. U.S. Department of Commerce, National Institute of Standards and Technology, Washington, DC; Cambridge University Press, Cambridge, 2010. With 1 CD-ROM (Windows, Macintosh and UNIX).

[62] J. A. A. Opschoor, Ch. Schwab, and J. Zech. Exponential ReLU DNN expression of holomorphic maps in high dimension. *Constructive Approximation*, 55(1):537–582, 2022.

[63] Joost A. A. Opschoor, Philipp C. Petersen, and Christoph Schwab. Deep ReLU networks and high-order finite element methods. *Analysis and Applications*, 18(05):715–770, 2020.

[64] Joost A. A. Opschoor, Christoph Schwab, and Jakob Zech. Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion. In Roland Herzog, Matthias Heinkenschloss, Dante Kalise, Georg Stadler, and Emmanuel Trélat, editors, *Optimization and Control for Partial Differential Equations: Uncertainty quantification, open and closed-loop control, and shape optimization*, pages 419–462. De Gruyter, 2022.

[65] Guofei Pang, Lu Lu, and George Em Karniadakis. fPINNs: fractional physics-informed neural networks. *SIAM J. Sci. Comput.*, 41(4):A2603–A2626, 2019.

[66] Dmytro Perekrestenko, Philipp Grohs, Dennis Elbrächter, and Helmut Bölcskei. The universal approximation power of finite-width deep ReLU networks. Jun 2018. ArXiv: 1806.01528.

[67] Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise

smooth functions using deep ReLU neural networks. *Neural Networks*, 108:296–330, 2018.

[68] Allan 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.

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

[70] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: a deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.*, 378:686–707, 2019.

[71] Sebastian Reich and Colin Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, New York, 2015.

[72] David Rolnik and Max Tegmark. The power of deeper networks for expressing natural functions. In *International Conference on Learning Representations*, 2018.

[73] Claudia Schillings and Christoph Schwab. Sparsity in Bayesian inversion of parametric operator equations. *Inverse Problems*, 30(6), 2014.

[74] Claudia Schillings and Christoph Schwab. Scaling limits in computational Bayesian inversion. *ESAIM: M2AN*, 50(6):1825–1856, 2016.

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

[76] Christoph Schwab and Radu Alexandru Todor. Convergence rates of sparse chaos approximations of elliptic problems with stochastic coefficients. *IMA Journal of Numerical Analysis*, 44:232–261, 2007.

[77] Christoph Schwab and Jakob Zech. Deep learning in high dimension: neural network expression rates for generalized polynomial chaos expansions in UQ. *Anal. Appl. (Singap.)*, 17(1):19–55, 2019.

[78] Justin Sirignano and Konstantinos Spiliopoulos. DGM: a deep learning algorithm for solving partial differential equations. *J. Comput. Phys.*, 375:1339–1364, 2018.

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

[80] T.J. Sullivan. Well-posed Bayesian inverse problems and heavy-tailed stable quasi-Banach space priors. *Inverse Probl. Imaging*, 11(5):857–874, 2017.

[81] J.M Tarela and M.V Martinez. Region configurations for realizability of lattice piecewise-linear models. *Mathematical and Computer Modelling*, 30(11-12):17–27, 1999.

[82] Rohit K. Tripathy and Ilias Bilionis. Deep UQ: learning deep neural network

surrogate models for high dimensional uncertainty quantification. *J. Comput. Phys.*, 375:565–588, 2018.

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

[84] Jakob Zech. *Sparse-grid approximation of high-dimensional parametric PDEs*. PhD thesis, 2018. Dissertation 25683, ETH Zürich.

[85] Jakob Zech, Dinh Dung, and Christoph Schwab. Multilevel approximation of parametric and stochastic PDEs. *M3AS*, 29(9):1753–1817, 2019.

[86] Jakob Zech and Christoph Schwab. Convergence rates of high dimensional Smolyak quadrature. *ESAIM Math. Model. Numer. Anal.*, 54(4):1259–1307, 2020.

[87] Dongkun Zhang, Lu Lu, Ling Guo, and George Em Karniadakis. Quantifying total uncertainty in physics-informed neural networks for solving forward and inverse stochastic problems. *J. Comput. Phys.*, 397:108850, 19, 2019.