Journal of Machine Learning Research 27 (2026) 1-82 Submitted 3/25; Revised 1/26; Published 5/26

Statistical Learning Theory for Neural Operators

Niklas Reinhardt NIKLAS.REINHARDT@IWR.UNI-HEIDELBERG.DE
Interdisziplinäres Zentrum für wissenschaftliches Rechnen
Universität Heidelberg
Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

Sven Wang SVEN.WANG@EPFL.CH
Institute of Mathematics
École Polytechnique Fédérale de Lausanne (EPFL)
Station 8, CH-1015 Lausanne, Switzerland

Jakob Zech JAKOB.ZECH@UNI-HEIDELBERG.DE
Interdisziplinäres Zentrum für wissenschaftliches Rechnen
Universität Heidelberg
Im Neuenheimer Feld 205, 69120 Heidelberg, Germany

Editor: Benjamin Guedj

Abstract

We present statistical convergence results for the learning of (possibly) non-linear mappings in infinite-dimensional spaces. Specifically, given a map $G_0 : \mathcal{X} \to \mathcal{Y}$ between two separable Hilbert spaces, we analyze the problem of recovering G_0 from $n \in \mathbb{N}$ noisy input-output pairs $(x_i, y_i)_{i=1}^n$ with $y_i = G_0(x_i) + \varepsilon_i$; here the $x_i \in \mathcal{X}$ represent randomly drawn "design" points, and the ε_i are assumed to be either i.i.d. white noise processes or subgaussian random variables in $\mathcal{Y}$. We provide general convergence results for least-squares-type empirical risk minimizers over compact regression classes $\mathbf{G} \subseteq L^\infty(\mathcal{X}, \mathcal{Y})$, in terms of their approximation properties and metric entropy bounds, which are derived using empirical process techniques. This generalizes classical results from finite-dimensional nonparametric regression to an infinite-dimensional setting. As a concrete application, we study an encoder-decoder based neural operator architecture termed FrameNet. Assuming G_0 to be holomorphic, we prove algebraic (in the sample size n) convergence rates in this setting, thereby overcoming the curse of dimensionality. To illustrate the wide applicability, as a prototypical example we discuss the learning of the non-linear solution operator to a parametric elliptic partial differential equation.

Keywords: nonparametric estimation, neural networks, operator learning, minimax convergence rates, empirical risk minimization, partial differential equations

1. Introduction

Learning non-linear relationships of high- and infinite-dimensional data is a fundamental problem in modern statistics and machine learning. In recent years, "Operator Learning" has emerged as a powerful tool for analyzing and approximating mappings G_0 between *infinite-dimensional* spaces (Li et al., 2020; Hesthaven and Ubbiali, 2018; Bhattacharya et al., 2021; Lu et al., 2021; Raonic et al., 2023; Anandkumar et al., 2019; Owhadi and Yoo, 2019; Nelsen and Stuart, 2024; Kovachki et al., 2024b). The primary motivation for

considering truly infinite-dimensional data stems from applications in the natural sciences, where inputs and outputs of operators are elements in function spaces. For instance, G_0 could be the operator relating an initial condition x of a dynamical system to the state $G_0(x)$ of the system after a certain time, or a coefficient-to-solution map of a parametric partial differential equation (PDE).

For finite-dimensional inputs and outputs, nonparametric regression is the standard framework for inferring general, non-linear relationships. There, one aims to reconstruct some "ground truth" $G_0 : \mathbb{R}^d \to \mathbb{R}^m$, d, $m \in \mathbb{N}$, from noisy data $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^m$, $i = 1, \dots, n$, generated via $y_i = G_0(x_i) + \varepsilon_i$, where x_i are called the "design points" and ε_i are typically independent and identically distributed (i.i.d.) noise variables. In the framework of empirical risk minimization (ERM), one chooses a suitable function class $\boldsymbol{G}$ of mappings from $\mathbb{R}^d$ to $\mathbb{R}^m$ and some loss function $L : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ measuring the discrepancy between predictions $G(x_i)$ and the data y_i. Statistical estimation is achieved by minimizing

$$\hat{G}_n \in \operatorname*{arg\,min}_{G \in \boldsymbol{G}} J_n(G), \quad J_n(G) := \frac{1}{n} \sum_{i=1}^{n} L(G(x_i), y_i), \tag{1.1}$$

assuming that minimizers exist. In the finite-dimensional setting, statistically optimal convergence rates for such estimators were established for least-squares, maximum likelihood, and more generally "minimum contrast" estimators (van de Geer, 2000; Barron et al., 1999; Birgé and Massart, 1993); see also Schmidt-Hieber (2020a) where such results are shown for ERMs over neural network classes. However, as is well-known, both—approximation rates (DeVore et al., 1989; DeVore and Lorentz, 1993) as well as statistical convergence rates (van de Geer, 2000; Giné and Nickl, 2016) over classical smoothness classes—deteriorate exponentially in terms of the dimension d. This renders computations practically infeasible for large d. This phenomenon is referred to as the *curse of dimensionality*, see also Section 4.1 ahead.

The framework for operator learning considered in this paper can be viewed as a direct extension of (1.1) to infinite dimensions. Given Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$, and a mapping $G_0 : \mathcal{X} \to \mathcal{Y}$, the goal is to reconstruct G_0 from "training data" $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ with

$$y_i = G_0(x_i) + \varepsilon_i,$$

where the regression class $\boldsymbol{G}$ is a suitable set of measurable mappings between $\mathcal{X}$ and $\mathcal{Y}$ and ε_i are centered noise variables, see Section 2 for details. This "supervised learning" setting underlies popular methods such as the PCA-Net (Hesthaven and Ubbiali, 2018; Bhattacharya et al., 2021).

We also mention the framework of "physics-informed learning" which is common in operator learning (relevant e.g. for the DeepONet Lu et al. 2021), but which is not considered in the present manuscript. Here, information on the ground truth G_0 is not known in the form of input-output pairs, but instead is implicitly described via

$$\mathcal{N}(x, G_0(x)) = 0,$$

where $\mathcal{N} : \mathcal{X} \times \mathcal{Y} \to \mathcal{Z}$ for a third vector space $\mathcal{Z}$. Typically, $\mathcal{N}(x, \cdot)$ encodes a family of differential operators parametrized by x which represent the underlying physical model. In

this case, the loss to minimize is a residual of the form $\sum_{i=1}^n \|\mathcal{N}(x_i, G(x_i))\|_{\mathcal{Z}}^2$, thus leading to an "unsupervised learning problem". The two cases, supervised and unsupervised learning, can also be combined. Various different (neural network based) architectures (i.e. regression classes $\boldsymbol{G}$) have been proposed in recent years for the purpose of supervised or unsupervised operator learning.

1.1 Outline and Contributions

We provide statistical convergence results for operator learning which do not suffer from the curse of dimensionality, and which can be applied to prototypical problems in the PDE literature. We first develop our theory in an abstract setting for ERMs over classes of mappings between separable Hilbert spaces, and later apply our theory to concrete examples. In doing so, we build upon and synthesize influential proof techniques from nonparametric statistics, in particular M-estimation (van de Geer, 2001; Nickl et al., 2020), approximation theory for parametric PDEs (Cohen et al., 2011; Cohen and DeVore, 2015; Herrmann et al., 2024), and empirical process theory (Talagrand, 2005; Dirksen, 2015).

To illustrate the scope of our contributions, we start by stating a convergence result for the elliptic "Darcy flow" problem on the d-dimensional torus. This is a standard example in PDE driven forward and inverse problems (Babuška et al., 2010; Chkifa et al., 2015b; Cliffe et al., 2011; Schwab and Stuart, 2012; Stuart, 2010; Nickl, 2023; Nickl et al., 2020). We aim for an informal exposition here, with full details given in Section 4: Denote by $\mathbb{T}^d$ the d-dimensional torus, fix a smooth source function $f : \mathbb{T}^d \to (0, \infty)$ and let $a_{\min} > 0$. For a sufficiently smooth and uniformly positive conductivity $a : \mathbb{T}^d \to \mathbb{R}$, denote by $G_0(a)$ the unique solution of the elliptic PDE

$$-\nabla \cdot (a \nabla u) = f \quad \text{on } \mathbb{T}^d \quad \text{and} \quad \int_{\mathbb{T}^d} u(x)dx = 0. \tag{1.2}$$

Now let γ be some probability distribution on $L^2(\mathbb{T}^d)$ such that

$$\operatorname{supp}(\gamma) \subseteq \big\{a \in H^{\mathfrak{s}}(\mathbb{T}^d) : \inf_{x \in \mathbb{T}^d} a(x) \geq a_{\min},\ \|a\|_{H^{\mathfrak{s}}(\mathbb{T}^d)} \leq R\big\}, \tag{1.3}$$

for some $R > 0$, $\mathfrak{s} > 2d + 1$. Suppose we observe noisy input-output pairs $(a_i, y_i)_{i=1}^n$ given by $y_i = G_0(a_i) + \varepsilon_i$, where the ε_i are independent $L^2(\mathbb{T}^d)$-Gaussian white noise processes (Section 2). The operator G_0 can then be learned from this data as stated in the next theorem (for details see Section 4.2); it regards empirical risk minimizers $\hat{G}_n$ over the so-called FrameNet $\boldsymbol{G}_{\mathrm{FN}}$, which corresponds to a neural network based class of measurable mappings $\mathcal{X} \to \mathcal{Y}$ (Section 3). Formally, $\hat{G}_n$ is defined as a minimizer of the least squares objective

$$\hat{G}_n \in \operatorname*{arg\,min}_{G \in \boldsymbol{G}_{\mathrm{FN}}} \frac{1}{n} \sum_{i=1}^n \|y_i - G(a_i)\|_{L^2(\mathbb{T}^d)}^2, \tag{1.4}$$

although a suitable modification is required to make this mathematically rigorous (Section 2.1).

Theorem 1 (Informal) *Consider the operator G_0 from the Darcy problem on the d-dimensional torus $\mathbb{T}^d$ ($d \geq 2$), and suppose that γ satisfies* (1.3) *for some $\mathfrak{s} > 3d/2 + 1$ and $a_{min} > 0$. Fix $\tau > 0$ (arbitrarily small).*

Then there exists a constant C such that for each $n \in \mathbb{N}$ there exists a FrameNet class $\boldsymbol{G}_{\mathrm{FN}}(n)$ and any empirical risk minimizer $\hat{G}_n$ in (1.4) satisfies[1]

$$\mathbb{E}_{G_0}\left[\int \|\hat{G}_n(a) - G_0(a)\|^2_{L^2(\mathbb{T}^d)} d\gamma(a)\right] \leq C n^{-\frac{2\mathfrak{s}+2-3d}{2\mathfrak{s}+2-d}+\tau}. \tag{1.5}$$

The most significant feature of the above statement is that, although G_0 maps between infinite-dimensional spaces, the convergence rate in (1.5) is algebraic in n. Thus, it does not suffer from a curse of dimensionality with respect to the operator input a. The strong dependence on the spatial dimension d remains, but this reflects a curse of dimensionality that is generally unavoidable due to finite spatial smoothness, e.g., DeVore et al. (1989). Moreover, for infinitely smooth input data ($\mathfrak{s} \to \infty$), the convergence rate gets arbitrarily close to n^{-1}. The classes $\boldsymbol{G}_{\mathrm{FN}}$, whose existence is postulated by the theorem, can be precisely characterized in terms of the sparsity, depth, width and other network class parameters, which are chosen in terms of the statistical sample size n. We also note that the regularity assumption $\mathfrak{s} > 3d/2 + 1$ was made here for convenience and can be weakened to $\mathfrak{s} > 3d/2$, see Theorem 27 and Remark 28 below.

To achieve Theorem 1 and several other related results, we build our theory in multiple steps. In Section 2, a general regression framework for mappings between Hilbert spaces is considered. Our first main result, Theorem 5, gives a non-asymptotic concentration upper bound on the empirical risk between $\hat{G}_n$ and G_0, with respect to the design points a_i. The upper bound is quantified in terms of the metric entropy of the "regression class" $\boldsymbol{G}$ and the best approximation of G_0 from $\boldsymbol{G}$. Theorem 6 strengthens this statement to $L^2(\gamma)$-loss, for the case of random design points $a_i \sim \gamma$. These results provide an operator learning analogue to classical convergence rates in nonparametric regression. The proofs rely on probabilistic generic chaining techniques (Talagrand, 2014; Dirksen, 2015) and "slicing" arguments as introduced in van de Geer (2001), which we generalize to the current setting. We also note our proofs contrast existing nonparametric statistical analyses of neural networks (Schmidt-Hieber, 2020a) for real-valued regression, where generic chaining techniques were not required for obtaining optimal rates (up to log-factors).

In the second part of this work, we apply our statistical results to the specific deep operator network class $\boldsymbol{G}_{\mathrm{FN}}$ termed "FrameNet" and introduced in Herrmann et al. (2024). Together with their underlying decoder-encoder structure and feedforward neural network structure, FrameNet classes are defined in Section 3. These classes are known to satisfy good approximation properties for holomorphic operators, a property which is fulfilled for the Darcy problem (1.2) and more broadly a wide range of PDE based problems (Cohen et al., 2010, 2011; Cohen and DeVore, 2015; Jerez-Hanckes et al., 2017; Harbrecht et al., 2016; Henríquez and Schwab, 2021; Hiptmair et al., 2018; Spence and Wunsch, 2023; Cohen et al., 2018). In Section 3, we identify such operator holomorphy as a key regularity property which allows to derive "dimension-free" statistical convergence rates. By extending approximation theoretic results from Herrmann et al. (2024), see Theorem 18 below, as well as establishing metric entropy bounds for $\boldsymbol{G}_{\mathrm{FN}}$ based on Schmidt-Hieber (2020a), we obtain algebraic convergence results for ERMs over FrameNet classes, for reconstructing holomorphic operators G_0. Specifically, Theorem 23 bounds the L^2-risk $\mathbb{E}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}] \lesssim n^{-\kappa/(\kappa+1)}$,

1. Here and in the following $\mathbb{E}_{G_0}$ denotes the expectation w.r.t. the random data $(x_i, y_i)_i$ generated by the ground truth G_0. Similarly, we write $\mathbb{P}_{G_0}$ for corresponding probabilities.

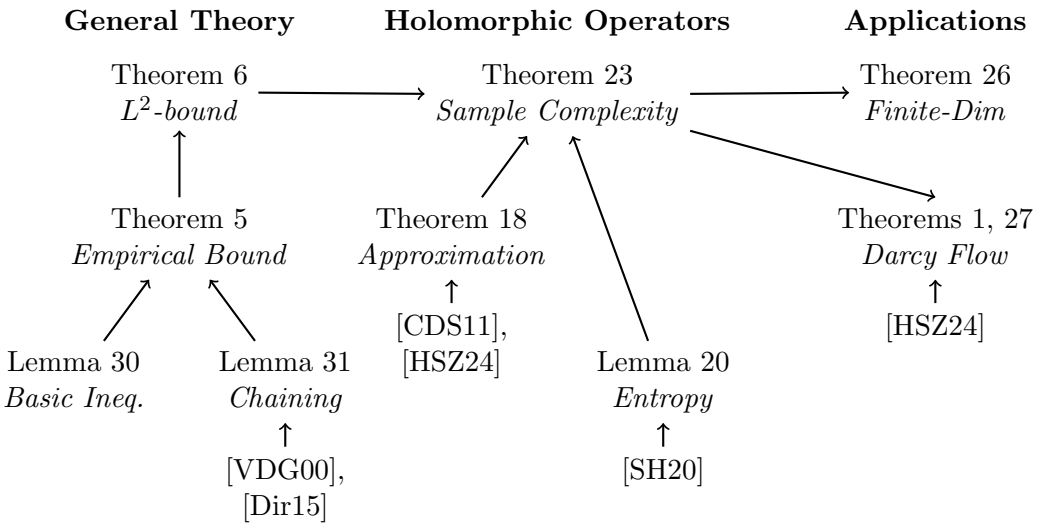

Figure 1: Dependency diagram of the main theorems, lemmas and references in this work.

where $\kappa > 0$ denotes the *approximation* rate established in Herrmann et al. (2024). We treat the case of ReLU and RePU (Li, 2020) activation functions for both sparse and fully-connected architectures.

In Section 4, we illustrate the usefulness of our general theory in two concrete settings. First, we show how our theory recovers well-known minimax-optimal convergence rates for real-valued regression (i.e., $\mathcal{Y} = \mathbb{R}$) on d-dimensional domains. This proves that our abstract results from Section 2 cannot be improved *in general*, although matching lower bounds are yet unknown in the infinite-dimensional setting. Thereafter, Section 4.2 demonstrates how our theory can be used to yield the first algebraic convergence rates for a non-linear operator arising from PDEs—see in particular Theorem 27 and Remark 28, which underlie Theorem 1. Figure 1 summarizes the dependencies between the main theorems, lemmas and key references underlying the results in this work.

1.2 Existing Results

The approximation of mappings between infinite-dimensional spaces has been studied extensively in the context of Uncertainty Quantification, where G_0 corresponds to the solution operator of a parameter dependent PDE. Various methodologies have been proposed and analyzed for this task, including for example compressed sensing (Doostan and Owhadi, 2011; Rauhut and Schwab, 2017), sparse-grid interpolation (Chkifa et al., 2013; Nobile et al., 2008), least-squares (Cohen et al., 2017; Chkifa et al., 2015a), and reduced basis methods (Quarteroni et al., 2016; Hesthaven et al., 2016). Recently, neural network approaches have become increasingly popular for this task as they provide a highly expressive and fast to evaluate parametrization of high-dimensional functions. These attributes make them particularly useful for learning surrogates in scientific applications (Hesthaven and Ubbiali, 2018; Lu et al., 2021; Li et al., 2020; Bhattacharya et al., 2021; Anandkumar et al.,

2019; O'Leary-Roseberry et al., 2022; O'Leary-Roseberry et al., 2024; Becker et al., 2024; Cicci et al., 2022; Dal Santo et al., 2020; Kröpfl et al., 2022).

Approximation Theory for Neural Operators. First theoretical results on operator learning focused on the approximation error, establishing the existence of neural network architectures capable of approximating G_0 up to a certain accuracy, with the error decreasing algebraic in terms of the number of learnable network parameters. For example, the works Schwab and Zech (2019); Kutyniok et al. (2022); Schwab and Zech (2023) showed that neural networks have sufficient expressivity to efficiently approximate certain (holomorphic) mappings G_0. Such results are based on the observation that the smoothness of G_0 implies the image of this operator to have moderate n-widths, i.e. to be well approximated in moderate-dimensional linear subspaces (Dung et al., 2023; Cohen et al., 2010, 2011; Cohen and DeVore, 2015; Hoang and Schwab, 2014; Bachmayr et al., 2017). Specifically for DeepONets (Lu et al., 2021), such a result was obtained in Lanthaler et al. (2022), and for the presently considered architecture in Herrmann et al. (2024). Moreover, Schwab et al. (2025) provided a statement of this type for Lipschitz continuous operators by exploiting so-called superexpressivity of certain classes of neural networks.

Statistical Theory for Neural Operators. The analysis of sample complexity has received less attention so far. In Lanthaler (2023), the authors analyzed in particular the error of PCA encoders and decoders used for PCA-Net, but did not analyze the statistical error for the full operator. The work de Hoop et al. (2023) provides such a result for the estimation of linear and diagonalizable mappings from noisy data; for lower bounds see for example Chagny et al. (2025). For other work on "functional regression", see, e.g., Greven and Scheipl (2017); Morris and Carroll (2006). An analysis for nonparametric regression of nonlinear mappings from noisy data in infinite dimensions was provided in Liu et al. (2024). There, the authors considered Lipschitz continuous mappings G_0, and proved consistency in the large data limit. Additionally they give convergence results, which, however, are subject to the curse of dimensionality. This is due to their very general assumption on the smoothness of G_0: It was shown very early (Mhaskar and Hahm, 1997), that the nonlinear n-width of Lipschitz operators in L^2 decays only logarithmically, i.e. the number of (exact) data points needed for the reconstruction of the functional is exponential in the desired accuracy. Recently, Kovachki et al. (2024a) generalized these results and showed a generic curse of dimensionality for the reconstruction of Lipschitz operators and C^k-operators from exact data. Moreover, the authors show that under the existence of some intrinsic low-dimensionality allowing for fast approximation, also the dependence on the data complexity improves.

Concerning the case of noisy and holomorphic operators, we also refer to the recent works Adcock et al. (2025, 2024) who consider a setup similar to ours, and, unlike us, treat the more general case of Banach space valued functions. The authors derive upper bounds and concentration inequalities for the L^2-error, and lower bounds for the approximation error, in terms of a neural network based architecture. Key differences to our work include in particular that the results of Adcock et al. (2025, 2024) do not address convergence in the noise-polluted large data limit $n \to \infty$, nor do they directly yield convergence rates for concrete PDE models; the latter typically requires to exploit PDE regularity theory to show holomorphy in spaces of higher spatial regularity, leading to a multilevel decomposition of the operator. Moreover, the generalization results in these works either focus on linearly

parametrized models (Adcock et al., 2025, 2024), or are restricted to specific subsets of minimizers (Adcock et al., 2024, Theorem 3.2). Our analysis aims to fill this gap by providing generalization bounds for arbitrary minimizers and nonlinear parameter dependence, as is standard for neural networks.

M-Estimation in Nonparametric Regression. Convergence theory of M-estimators and (penalized) empirical risk minimizers was investigated around the 2000s in foundational works (van de Geer, 2000; van de Geer, 2001; Birgé and Massart, 1993; Barron et al., 1999). These works build on concentration inequalities for empirical processes using "chaining" techniques which date back to seminal contributions, see Talagrand (2014); Giné and Nickl (2016) and references therein. These techniques are known to produce minimax-optimal rates $n^{-2\mathfrak{s}/(2\mathfrak{s}+d)}$ for ERMs over $\mathfrak{s}$-smooth Sobolev, Hölder and more generally, Besov smoothness classes of real-valued functions on bounded d-dimensional Euclidean domains. The analysis of neural-network based ERMs was initiated by the work Schmidt-Hieber (2020a), which considered regression over (compositional) Hölder classes on finite-dimensional domains, and was followed by several other works such as Suzuki (2019). We also mention Nickl et al. (2020); Nickl and Wang (2024); Agapiou and Wang (2024) which analyze ERMs in non-linear elliptic PDE-based inverse problems such as the "Darcy" flow problem studied here. The present Hilbert space setting falls outside the scope of such classical theory for scalar-valued functions. However, the derivation of our concentration inequalities for ERMs does build upon the same probabilistic empirical process machinery laid out above (Talagrand, 2014; Dirksen, 2015).

1.3 Notation

We write $\mathbb{N} = \{1, 2 \dots\}$ and $\mathbb{N}_0 = \{0, 1, 2, \dots\}$. We write $a_n \lesssim b_n, a_n \gtrsim b_n$ for real sequences $(a_n)_{n\in\mathbb{N}}$, $(b_n)_{n\in\mathbb{N}}$ if a_n is respectively upper or lower bounded by a positive multiplicative constant which does not depend on n (but may well depend on other ambient parameters which we make explicit whenever confusion may arise). By $a_n \simeq b_n$, we mean that both $a_n \lesssim b_n$ and $a_n \gtrsim b_n$.

For a pseudometric space (T, d) and any $\delta > 0$, let $N(T, d, \delta)$ be the δ-covering number of T, i.e. the minimal number of open δ-balls in d needed to cover T. We denote the metric entropy of T by

$$H(T, d, \delta) = \log N(T, d, \delta).$$

Given a Borel probability measure γ on $\mathcal{X}$ and a subset $D \subseteq \mathcal{X}$, we define the norms

$$\|G\|^2_{L^2(\mathcal{X},\gamma;\mathcal{Y})} \coloneqq \int_{\mathcal{X}} \|G(x)\|^2_{\mathcal{Y}} \, \mathrm{d}\gamma(x),$$
$$\|G\|_{\infty,D} \coloneqq \sup_{x\in D} \|G(x)\|_{\mathcal{Y}}$$

and also write $\|\cdot\|_{L^2(\gamma)}$ and $\|\cdot\|_\infty$ if the underlying spaces are clear from context. The space of real-valued, square summable sequences indexed over $\mathbb{N}$ is denoted by $\ell^2(\mathbb{N})$. The complexification of a real Hilbert space H is denoted by $H_{\mathbb{C}}$ (Kirwan, 1997; Muñoz et al., 1999).

2. Regression in Hilbert Spaces

In the following we formalize a regression framework for mappings between Hilbert spaces.

2.1 Problem Formulation

Throughout, let $\mathcal{X}$ and $\mathcal{Y}$ denote two separable (real) Hilbert spaces with respective inner products $\langle\cdot,\cdot\rangle_{\mathcal{X}}$, $\langle\cdot,\cdot\rangle_{\mathcal{Y}}$ and suppose

$$G_0 : \mathcal{X} \to \mathcal{Y}$$

is some non-linear (Borel measurable) operator which we aim to reconstruct. The observed data are assumed to be noisy "input-output pairs" $(x_i, y_i)_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$ given by

$$x_i \overset{\text{iid}}{\sim} \gamma, \qquad i = 1, \dots, n, \tag{2.1a}$$

and

$$y_i = G_0(x_i) + \sigma\varepsilon_i, \qquad i = 1, \dots, n, \tag{2.1b}$$

where $\sigma > 0$ denotes a scalar "noise level", ε_i are independent random noise variables and γ is a probability distribution on $\mathcal{X}$. The $x_i \in \mathcal{X}$ are also referred to as the "design points", and we write $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n$. We will both derive results which are *conditional* on the design $\boldsymbol{x}$, as well as results for *random design*. To avoid confusions, we will use the notations $P_{G_0}^{\boldsymbol{x}}$, $\mathbb{E}_{G_0}^{\boldsymbol{x}}$ to denote probabilities and expectations under the distribution (2.1) with fixed design $\boldsymbol{x}$, and we use $\mathbb{P}_{G_0}$, $\mathbb{E}_{G_0}$ to denote probabilities and expectations with random design $x_i \sim \gamma$.

Remark 2 *In practice, we will often deal with scenarios in which G_0 is only defined on some measurable subset $\mathcal{V} \subset \mathcal{X}$, see e.g. the solution operator for the Darcy flow in Section 4.2.1. In this case, our results from Section 2.2 can be applied to any measurable extension of G_0 onto $\mathcal{X}$. To apply the sample complexity results from Section 3.4 the extension additionally needs to be holomorphic in an open set containing $\mathcal{V}$, see Assumption 2 and Section 3 for the precise setup and statement.*

White Noise Model. We shall treat two different assumptions on the noise, the first being that the $(\varepsilon_i)_{i=1}^n$ in (2.1) are independent copies of a $\mathcal{Y}$-white noise process. Recall that for any given separable Hilbert space $\mathcal{Y}$, the $\mathcal{Y}$-Gaussian white noise process is defined as the mean-zero Gaussian process $\mathbb{W}_{\mathcal{Y}} = (\mathbb{W}_{\mathcal{Y}}(y) : y \in \mathcal{Y})$ indexed by $\mathcal{Y}$ with "iso-normal" covariance structure

$$\mathbb{W}_{\mathcal{Y}}(y) \sim \mathcal{N}(0, \|y\|_{\mathcal{Y}}^2), \quad \operatorname{Cov}(\mathbb{W}_{\mathcal{Y}}(y), \mathbb{W}_{\mathcal{Y}}(y')) = \langle y, y'\rangle_{\mathcal{Y}}, \quad \text{for all } y, y' \in \mathcal{Y}.$$

It is well-known that $\mathbb{W}_{\mathcal{Y}}$ does not take values in $\mathcal{Y}$ unless $\dim(\mathcal{Y}) < \infty$, but is interpreted as a stochastic process indexed by $\mathcal{Y}$, see Giné and Nickl (2016, p.19) for details. Nevertheless, we slightly abuse notation and use the common notation $\langle \mathbb{W}_{\mathcal{Y}}, y\rangle_{\mathcal{Y}} := \mathbb{W}_{\mathcal{Y}}(y)$.

Under this assumption, conditionally on x_i we interpret each observation y_i in (2.1) as a realisation of a Gaussian process $(y_i(f) : f \in \mathcal{Y})$ with

$$\mathbb{E}[y_i(f)] = \langle G_0(x_i), f\rangle_{\mathcal{Y}}, \quad \operatorname{Cov}(y_i(f), y_i(f')) = \langle f, f'\rangle_{\mathcal{Y}},$$

and we shall again use the notation $\langle y_i, f\rangle_{\mathcal{Y}}$ to denote $y_i(f)$ (see also Tsybakov 2009; Giné and Nickl 2016; Nickl et al. 2020 where this common viewpoint is explained in detail).

Example 1 *Let $\mathcal{O} \subseteq \mathbb{R}^d$ be a bounded, smooth domain. Then, for $\mathcal{Y} = L^2(\mathcal{O})$, one can show that draws of an $L^2(\mathcal{O})$-white noise process a.s. take values in negative Sobolev spaces $H^{-\kappa}$ for $\kappa > d/2$ (Nickl, 2020; Castillo and Nickl, 2013).*

Sub-Gaussian Noise Model. The second setting we consider is that of sub-Gaussian noise. We say that a random vector X taking values in $\mathcal{Y}$ is sub-Gaussian with parameter $\eta > 0$ if $\mathbb{E}[X] = 0$ and

$$\mathbb{P}(\|X\|_{\mathcal{Y}} \geq t) \leq 2\exp\left(-\frac{t^2}{2\eta^2}\right), \quad \text{for all } t \geq 0.$$

In the sub-Gaussian noise model, we assume that $(\varepsilon_i)_{i=1}^n$ in (2.1) are independent sub-Gaussian variables in $\mathcal{Y}$ with parameter $\eta = 1$.

2.1.1 Empirical Risk Minimization

Let $\boldsymbol{G}$ be a class of (measurable) operators $\boldsymbol{G} \ni G : \mathcal{X} \to \mathcal{Y}$. We would like to study classical empirical risk minimizers of least-squares type over $\boldsymbol{G}$. Specifically, given regression data $(x_i, y_i)_{i=1}^n$, consider the empirical risk

$$\tilde{I}_n(G) := \frac{1}{n}\sum_{i=1}^n \big\|y_i - G(x_i)\big\|_{\mathcal{Y}}^2, \quad \tilde{I}_n : \boldsymbol{G} \to [0,\infty]. \tag{2.2}$$

However, this functional takes finite values almost surely only in the sub-Gaussian noise model. In the white noise model, since $y_i \notin \mathcal{Y}$, it holds $\tilde{I}_n(G) = \infty$ almost surely—we thus consider a modified definition of least-squares type estimators which is common in the literature on regression with white noise (Nickl et al., 2020; Giné and Nickl, 2016). Instead of (2.2), we consider

$$I_n(G) = \frac{1}{n}\sum_{i=1}^n -2\langle G(x_i), y_i\rangle_{\mathcal{Y}} + \big\|G(x_i)\big\|_{\mathcal{Y}}^2, \quad I_n : \boldsymbol{G} \to \mathbb{R}, \tag{2.3}$$

which takes finite values a.s. also in the white noise model. Note that the latter objective function can be obtained from (2.2) by formally subtracting the term $n^{-1}\sum_{i=1}^n \|y_i\|_{\mathcal{Y}}^2$ which exhibits no dependency on G. Therefore in the sub-Gaussian noise model the minimization of $\tilde{I}_n$ and I_n are equivalent which is why we consider (2.3) in the following. We will denote minimizers of $I_n(G)$ by $\hat{G}_n$.

Our assumptions on the class $\boldsymbol{G}$ in the ensuing theorems will ensure that a *measurable choice* of minimizers $\hat{G}_n$ of I_n exists, see Theorem 5 (i). However, the ERM $\hat{G}_n$ will in general not be unique, since we do not impose convexity on $\boldsymbol{G}$. The reason is that our main application, the NN-based FrameNet class $\boldsymbol{G}_{\mathrm{FN}}$, is non-convex.

Remark 3 (Connection to maximum likelihood) *In the white noise model, it follows from the Cameron-Martin theorem (see for example Giné and Nickl 2016, Theorem 2.6.13) that $-nI_n(G)/(2\sigma^2)$ constitutes the negative log-likelihood of the (dominated) statistical model arising from* (2.1) *with white noise. In this case $\hat{G}_n$ can also be interpreted as a (nonparametric) maximum likelihood estimator over the class $\boldsymbol{G}$.*

Remark 4 *Consider nonparametric regression of an unknown function $f : \mathcal{O} \to \mathbb{R}$ for some bounded, smooth domain $\mathcal{O}$. Here, it is well-known that the observation white noise error model, where data is given by $Y = f + \sigma\mathbb{W}$ (with $\mathbb{W}$ a $L^2(\mathcal{O})$-white noise process) is asymptotically equivalent in a Le Cam-sense to an observation model with m "equally spaced" (random or deterministic) observation points throughout $\mathcal{O}$,*

$$Y_i = f(z_i) + \eta_i, \qquad i = 1, ..., m,$$

with i.i.d. $N(0, 1)$ errors, where the equivalence holds for $\sigma \asymp 1/\sqrt{m}$, see Reiß (2008). Therefore, our observation model (2.1) *may be viewed as a simplified proxy.*

2.2 General Convergence Results

Let $\boldsymbol{G}$ be a class of operators mapping from $\mathcal{X}$ to $\mathcal{Y}$. For any fixed $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n$ and (Borel) measurable map $G : \mathcal{X} \to \mathcal{Y}$, we denote the empirical seminorm induced by $\boldsymbol{x}$ with

$$\|G\|_n^2 = \frac{1}{n}\sum_{i=1}^{n}\|G(x_i)\|_{\mathcal{Y}}^2. \tag{2.4}$$

For any element $G^* \in \boldsymbol{G}$ and $\delta > 0$, define the localized classes

$$\boldsymbol{G}_n^*(\delta) = \big\{G \in \boldsymbol{G} : \|G - G^*\|_n \le \delta\big\},$$

and denote its metric entropy integral by

$$J(\delta) = J(\boldsymbol{G}_n^*(\delta), \|\cdot\|_n) := \int_0^\delta H^{\frac{1}{2}}(\boldsymbol{G}_n^*(\delta), \|\cdot\|_n, \rho)d\rho. \tag{2.5}$$

The following result provides a general convergence theorem for empirical risk minimizers with high probability, which relates the empirical risk of ERMs over some operator class $\boldsymbol{G}$ to the metric entropy of $\boldsymbol{G}$. It can be viewed as a generalisation of classical convergence results for sieved M-estimators (van de Geer, 2000) to Hilbert space valued functions. The proof can be found in Appendix B.1.

Theorem 5 *For some measurable $G_0 : \mathcal{X} \to \mathcal{Y}$, let the data $(x_i, y_i)_{i=1}^n$ arise from* (2.1) *either with white noise or with sub-Gaussian noise. Let $\boldsymbol{G}$ be a class of measurable maps from $\mathcal{X} \to \mathcal{Y}$, let $G^* \in \boldsymbol{G}$, and let $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n$ be such that the following holds.*

(a) There exists a constant $C > 0$ s.t. $\boldsymbol{G}$ is compact with respect to some norm $\|\cdot\|$ satisfying $\|\cdot\|_n \le C\|\cdot\|$.

(b) There exists $\Psi_n : (0, \infty) \to [0, \infty)$ s.t. $\Psi_n(\delta) \ge J(\boldsymbol{G}_n^*(\delta), \|\cdot\|_n)$ for all $\delta > 0$ and

$$\delta \mapsto \frac{\Psi_n(\delta)}{\delta^2} \qquad \text{is non-increasing for } \delta \in (0, \infty).$$

Then the following holds.

(i) Minimizers $\hat{G}_n$ of the empirical risk (2.3) *exist, and there is a measurable selection (with respect to the data $(x_i, y_i)_{i=1}^n$) of such a minimizer.*

(ii) Fix any measurable selection $\hat{G}_n$ from part (i). Then there exists a universal constant $C_{\mathrm{Ch}} > 0$ (see Lemma 32) such that for any $G^* \in \boldsymbol{G}$ as above, any positive sequence $(\delta_n)_{n\in\mathbb{N}}$ satisfying

$$\sqrt{n}\delta_n^2 \geq 32 C_{ch}\sigma \Psi_n(\delta_n), \tag{2.6}$$

and for any

$$R_n \geq \max\Big\{\delta_n, \frac{4C_{ch}\sigma}{\sqrt{n}}, \sqrt{2}\|G^* - G_0\|_n\Big\},$$

it holds that

$$\mathbb{P}^{\boldsymbol{x}}_{G_0}\big(\|\hat{G}_n - G_0\|_n \geq R_n\big) \leq 2\exp\bigg(-\frac{nR_n^2}{16C_{ch}^2\sigma^2}\bigg). \tag{2.7}$$

The lower bound for R_n in (ii), which determines the convergence rate of $\|\hat{G}_n - G_0\|_n$, is typically optimized by balancing the "stochastic term" δ_n and the empirical approximation error $\|G^* - G_0\|_n$, see, e.g., Theorem 23 below. Note that Theorem 5 gives a convergence rate with respect to the *empirical seminorm* $\|\cdot\|_n$. Therefore, when the design points $\boldsymbol{x}$ are random, the norm itself is also random, and the assumptions (a) and (b) in Theorem 5 have to be understood conditional on possible realizations of $\boldsymbol{x}$. Theorem 6 below will give a corresponding concentration inequality on the $\|\cdot\|_{L^2(\gamma)}$-error under the assumption of i.i.d. random design $x_i \sim \gamma$. In this setting, a sufficient condition for (b) to be satisfied almost surely is to take Ψ_n as an upper bound for the entropy integral with uniform entropy $H(\boldsymbol{G}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}, \delta)$.

The compactness of $\boldsymbol{G}$ in part (a) is needed for the existence of the ERM $\hat{G}_n$ in (2.3). The existence result for $\hat{G}_n$ (see for example Nickl 2007, Proposition 5) requires compact *metric* spaces, which is why we assume compactness w.r.t. a *norm* $\|\cdot\|$ stronger than the empirical *seminorm* $\|\cdot\|_n$.

In particular, compactness implies separability of $\boldsymbol{G}$ with respect to $\|\cdot\|_n$, which in turn is needed to guarantee the measurability of certain suprema of empirical processes ranging over $\boldsymbol{G}$. See the proof of Theorem 5 below, in particular (B.4) and (B.6). The technical growth restriction in (b) on the function $\Psi_n(\delta)$ (i.e., our upper bound for the entropy integral) is required for the "peeling device" in (B.4). In particular, this assumption is satisfied in case that $H(\boldsymbol{G}, \|\cdot\|_\infty, \delta) \lesssim \delta^{-\alpha}$ for any $0 < \alpha < 2$, see Corollary 8 below for details.

Theorem 5 provides a concentration inequality for the empirical seminorm $\|\hat{G}_n - G_0\|_n$. Under the assumption of randomly chosen design points $x_i \sim \gamma$, $x_i \in \mathcal{X}$, this statement can be extended to a convergence result for the $L^2(\gamma)$-norm. To this end, we also need slightly stronger technical assumptions on the class $\boldsymbol{G}$ with respect to the $\|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$-norm.

Assumption 1 *For some probability measure γ on $\mathcal{X}$, assume $\boldsymbol{x} = (x_1, \ldots, x_n)$ arise from i.i.d. draws $x_i \sim \gamma$. Let $\boldsymbol{G}$ be a class of measurable maps $\mathcal{X} \to \mathcal{Y}$, $G^* \in \boldsymbol{G}$ and $\|G_0\|_{\infty,\mathrm{supp}(\gamma)} < \infty$. Suppose*

(a) $\boldsymbol{G}$ is compact with respect to $\|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$,

(b) There exists a (deterministic) upper bound $\Psi_n : (0,\infty) \to [0,\infty)$ *such that for a.e.* $\boldsymbol{x} \sim \gamma^n$*, it holds* $\Psi_n(\delta) \geq J(\boldsymbol{G}_n^*(\delta), \|\cdot\|_n)$ *for all* $\delta > 0$ *and*

$$\delta \mapsto \frac{\Psi_n(\delta)}{\delta^2} \qquad \text{is non-increasing for } \delta \in (0,\infty). \tag{2.8}$$

Note that Assumption 1 is strictly stronger than assumptions (a) and (b) in Theorem 5, since $\|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$ is stronger than $\|\cdot\|_n$ and Ψ_n in (2.8) does not depend on $\boldsymbol{x}$. In particular, Assumption 1 implies the existence of a measurable ERM $\hat{G}_n$ by Theorem 5 (i).

The uniform $\|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$ assumptions are used to control the concentration of the empirical seminorm $\|\cdot\|_n$ around the "population norm" $\|\cdot\|_{L^2(\gamma)}$, see Lemma 33 and also Lemma 34 below. Let us write $\boldsymbol{F} = \{G - G_0 : G \in \boldsymbol{G}\}$. As an immediate consequence of Assumption 1, there exists some $M_{\boldsymbol{F}} < \infty$ such that

$$\sup_{F\in\boldsymbol{F}} \|F\|_{\infty,\mathrm{supp}(\gamma)} = \sup_{G\in\boldsymbol{G}} \|G - G_0\|_{\infty,\mathrm{supp}(\gamma)} \leq M_{\boldsymbol{F}}. \tag{2.9}$$

We can now state our main concentration inequality for the convergence of $\|\hat{G}_n - G_0\|_{L^2(\gamma)}$, which in particular provides a bound for the mean squared error $\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}]$ as well. The proof of Theorem 6 can be found in Appendix B.2.

Theorem 6 ($L^2(\gamma)$-Concentration under Random Design) *Consider the nonparametric regression model* (2.1) *either with white noise or with sub-Gaussian noise and any measurable empirical risk minimizer* $\hat{G}_n$ *from* (2.3)*. Suppose that* G_0*,* $\boldsymbol{G}$*,* G^**,* γ *and* $\Psi_n(\cdot)$ *are such that Assumption 1 holds. Then there exists some universal constant* $C > 0$ *such that for any positive sequences* $(\delta_n)_{n\in\mathbb{N}}$ *and* $(\tilde{\delta}_n)_{n\in\mathbb{N}}$ *with*

$$\sqrt{n}\delta_n^2 \geq C\sigma\Psi_n(\delta_n) \quad \text{and} \quad n\tilde{\delta}_n^2 \geq CM_{\boldsymbol{F}}^2 H(\boldsymbol{G}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}, \tilde{\delta}_n), \tag{2.10}$$

all $G^* \in \boldsymbol{G}$ *as above and all*

$$R_n \geq C \max\left\{\delta_n, \tilde{\delta}_n, \|G^* - G_0\|_{\infty,\mathrm{supp}(\gamma)}, \frac{\sigma + M_{\boldsymbol{F}}}{\sqrt{n}}\right\} \tag{2.11}$$

we have,

$$\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R_n\right) \leq 2\exp\left(-\frac{nR_n^2}{C^2(\sigma^2 + M_{\boldsymbol{F}}^2)}\right). \tag{2.12}$$

To keep the presentation simple, we have left the numerical constants in the preceding theorem implicit. However, they can be made explicit, see the proof for details. The following bound on the mean squared error is obtained upon integration of the concentration inequalities from Theorem 5 and Lemma 33. Note that directly integrating the $L^2(\gamma)$-concentration, cf. (2.12), gives an approximation term in the uniform norm $\|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$ (following 2.11), which has weaker convergence properties in general, see Theorem 18. For the proof of Corollary 7, see Appendix B.3.

Corollary 7 ($L^2(\gamma)$-Mean Squared Error) *Consider the setting of Theorem 6 and assume in addition that Assumption 1 is fulfilled for all $G^* \in \boldsymbol{G}$ (with the same Ψ_n). Then, for some universal constant $C > 0$ and all $n \in \mathbb{N}$,*

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}\right] \leq C\Big(\delta_n^2 + \tilde{\delta}_n^2 + \frac{\sigma^2 + M_{\boldsymbol{F}}^2}{n}\Big) + 8 \inf_{G^* \in \boldsymbol{G}} \|G^* - G_0\|^2_{L^2(\gamma)}. \quad (2.13)$$

To demonstrate the typical use-cases of our abstract results, we summarize in the following corollary the rates which can be obtained under algebraic approximation properties of $\boldsymbol{G}$ and two concrete scalings of the metric entropy of $\boldsymbol{G}$. These correspond to the typical entropy bounds satisfied by (i) some fixed n-independent, infinite-dimensional regression class, and (ii) an N-dimensional approximation class, where N is chosen in terms of n. In the following corollary, $\lesssim$ refers to an inequality involving a constant independent of n, N and δ. For a proof of Corollary 8, see Appendix B.4.

Corollary 8 *Consider the setting of Corollary 7. Let $\boldsymbol{G} = \boldsymbol{G}(N)$, $N \in \mathbb{N}$, be a sequence of regression classes[2] such that $\inf_{G^* \in \boldsymbol{G}(N)} \|G^* - G_0\|^2_{L^2(\gamma)} \lesssim N^{-\beta}$ for some $\beta > 0$ and all $N \in \mathbb{N}$. Denote the entropy by $H(N, \delta) := H(\boldsymbol{G}(N), \|\cdot\|_{\infty, \mathrm{supp}(\gamma)}, \delta)$.*

(i) If $H(N, \delta) \lesssim \delta^{-\alpha}$ for some $0 < \alpha < 2$, then for all $n \in \mathbb{N}$ and all $N = N(n) = \lceil n^{\frac{2}{(2+\alpha)\beta}} \rceil$, it holds

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}\right] \lesssim n^{-\frac{2}{2+\alpha}}.$$

(ii) If $H(N, \delta) \lesssim N \log(\delta^{-1})$, then for all $n \in \mathbb{N}$, $n \geq 2$ and all $N(n) = \lceil n^{\frac{1}{\beta+1}} \rceil$, it holds

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}\right] \lesssim \log(n)\ n^{-\frac{\beta}{\beta+1}}.$$

Remark 9 (Effective smoothness) *In classical nonparametric regression over s-smooth function classes on $[0,1]^d$, the entropy assumption $H(N, \delta) \lesssim \delta^{-\alpha}$ from part (i) is fulfilled for $\alpha = d/\mathfrak{s}$, which yields the minimax-optimal rate $n^{-\frac{2\mathfrak{s}}{2\mathfrak{s}+d}}$, see Section 4.1 for details. Since the rate only depends on α (or equivalently α^{-1}), we can think of $\alpha^{-1} = \mathfrak{s}/d$ as the "effective smoothness" of the statistical model at hand.*

The sub-Gaussian noise model poses a more restrictive regularity assumption on ε_i than the assumption of white noise. It is possible to get $L^2(\gamma)$-convergence for sub-Gaussian noise in the case the entropy integral $J(\delta)$ in (2.5) is not finite, such that Assumption 1 (b) does not hold. The details are shown in the next theorem, its proof is deferred to Appendix B.5.

Theorem 10 *Consider the nonparametric regression model* (2.1) *with sub-Gaussian noise and the empirical risk from* (2.3)*. Let Assumption 1 (a) hold, that means suppose that*

2. We may think of N as the number of parameters of $\boldsymbol{G}$, e.g. the size of the FrameNet class $\boldsymbol{G}_{\mathrm{FN}}$ in Section 3 below.

$\|G_0\|_{\infty,\mathrm{supp}(\gamma)} < \infty$ *and* $\boldsymbol{G}$ *is compact with respect to* $\|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$. *Then there exists some universal constant* $C_1 > 0$, *such that for any positive sequences* $(\delta_n)_{n\in\mathbb{N}}$ *and* $(\tilde{\delta}_n)_{n\in\mathbb{N}}$ *with*

$$n\delta_n^4 \geq C_1^2\sigma^2 M_{\boldsymbol{F}}^2 H\left(\frac{\delta_n^2}{8\sigma^2+\delta_n^2}\right) \quad \text{and} \quad n\tilde{\delta}_n^2 \geq 6M_{\boldsymbol{F}}^2 H(\boldsymbol{G}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}, \tilde{\delta}_n), \tag{2.14}$$

all $G^* \in \boldsymbol{G}$ *and all*

$$R_n \geq \max\left\{\delta_n, \tilde{\delta}_n, \|G^* - G_0\|_{\infty,\mathrm{supp}(\gamma)}, \frac{M_{\boldsymbol{F}}}{\sqrt{n}}\right\},$$

we have

$$\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R_n\right) \leq 4\exp\left(-\frac{nR_n^4}{C_1^2\sigma^2(1+M_{\boldsymbol{F}}^2)}\right) + 2\exp\left(-\frac{nR_n^2}{C_1^2 M_{\boldsymbol{F}}^2}\right). \tag{2.15}$$

Furthermore, for all $n \in \mathbb{N}$ *there exists a universal constant* $C_2 > 0$ *such that*

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2\right] \leq C_2\Big(\delta_n^2 + \tilde{\delta}_n^2 + \frac{(1+\sigma)(1+M_{\boldsymbol{F}}^2)}{\sqrt{n}} + \inf_{G^*\in\boldsymbol{G}}\|G^* - G_0\|_{L^2(\gamma)}^2\Big). \tag{2.16}$$

Remark 11 *Consider* $\alpha \geq 2$ *in Corollary 8 (i). Then* $J(\delta)$ *in* (2.5) *is not necessarily finite, and hence Assumption 1 (b) need not be satisfied. Therefore Theorem 6 cannot be applied. In the sub-Gaussian noise case, we may still use the* $L^2(\gamma)$*-bound in* (2.16) *however. Similar to the proof of Corollary 8, it can then be shown that* $\delta_n^2 \lesssim n^{-\frac{1}{2+\alpha}}$ *and* $\tilde{\delta}_n^2 \lesssim n^{-\frac{2}{2+\alpha}}$ *satisfy* (2.14). *This yields*

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2\right] \lesssim n^{-\frac{1}{2+\alpha}},$$

i.e. half the convergence rate of the "chaining regime" considered in Corollary 8.

3. Learning Holomorphic Operators with FrameNet

In this section we first recall the NN-based operator class FrameNet from Herrmann et al. (2024, Section 2), see Sections 3.1 and 3.2. This will provide the regression class $\boldsymbol{G}_{\mathrm{FN}}$ over which to estimate G_0. Similar to for example PCA-Net (Hesthaven and Ubbiali, 2018), FrameNet consists of mappings

$$G = \mathcal{D}_{\mathcal{Y}} \circ g \circ \mathcal{E}_{\mathcal{X}}, \tag{3.1}$$

for a linear encoder $\mathcal{E}_{\mathcal{X}} : \mathcal{X} \to \ell^2(\mathbb{N})$, a linear decoder $\mathcal{D}_{\mathcal{Y}} : \ell^2(\mathbb{N}) \to \mathcal{Y}$, and a coefficient map $\ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$. The encoder maps an $x \in \mathcal{X}$ to its coefficients in some representation system of $\mathcal{X}$. Conversely, the decoder builds a $y \in \mathcal{Y}$ out of a coefficient series in $\ell^2(\mathbb{N})$. These representation systems consist of frames that are fixed a priori. The coefficient map g is represented by a feedforward neural network, that will be trained by ERM.

Subsequently, in Sections 3.3-3.4, we apply our analysis to the learning of *holomorphic* operators $G_0 : \mathcal{X} \to \mathcal{Y}$. For such mappings, the FrameNet architecture was shown in Herrmann et al. (2024) to be capable of overcoming the curse of dimensionality in terms of the *approximation error*. We generalize this property to the case of bounded network parameters in Subsection 3.3 and furthermore show metric entropy bounds. This allows us to prove that FrameNet can overcome the curse of dimensionality in the learning of holomorphic operators, both in terms of the *approximation capability* and in terms of *sample complexity*.

3.1 Representation Systems

We briefly recall basic definitions and properties of frames. For more details see for instance Christensen (2016).

3.1.1 Frames

Definition 12 *A family $\mathbf{\Psi} = \{\psi_j : j \in \mathbb{N}\} \subset \mathcal{X}$ is called a frame of $\mathcal{X}$, if the* analysis operator

$$T_{\mathbf{\Psi}} : \mathcal{X} \to \ell^2(\mathbb{N}), \qquad v \mapsto \big(\langle v, \psi_j \rangle_{\mathcal{X}}\big)_{j\in\mathbb{N}}$$

is bounded and boundedly invertible between $\mathcal{X}$ and $\operatorname{range}(F) \subset \ell^2(\mathbb{N})$.

Every orthonormal basis of $\mathcal{X}$ is trivially a frame. Since Definition 12 merely requires bounded invertibility on the range of $T_{\mathbf{\Psi}}$, $T_{\mathbf{\Psi}}$ need not be surjective, and in particular $\mathbf{\Psi}$ need not consist of linearly independent vectors. The *frame bounds* of $\mathbf{\Psi}$ are defined as

$$\Lambda_{\mathbf{\Psi}} \coloneqq \|T_{\mathbf{\Psi}}\|_{\mathcal{X}\to\ell^2} = \sup_{0\neq v\in\mathcal{X}} \frac{\|T_{\mathbf{\Psi}} v\|_{\ell^2}}{\|v\|_{\mathcal{X}}}, \quad \lambda_{\mathbf{\Psi}} \coloneqq \inf_{0\neq v\in\mathcal{X}} \frac{\|T_{\mathbf{\Psi}} v\|_{\ell^2}}{\|v\|_{\mathcal{X}}}, \tag{3.2}$$

the *synthesis operator* $T'_{\mathbf{\Psi}}$ as

$$T'_{\mathbf{\Psi}} : \ell^2(\mathbb{N}) \to \mathcal{X}, \quad \big(v_i\big)_{i\in\mathbb{N}} \mapsto \boldsymbol{v}^T\mathbf{\Psi} \coloneqq \sum_{i\in\mathbb{N}} v_i\psi_i,$$

and finally the *frame operator* as $Q_{\mathbf{\Psi}} \coloneqq T'_{\mathbf{\Psi}}T_{\mathbf{\Psi}} : \mathcal{X} \to \mathcal{X}$. The frame operator $Q_{\mathbf{\Psi}}$ is boundedly invertible, self-adjoint and positive, see Christensen (2016, Lemma 5.1.5). The family $\tilde{\mathbf{\Psi}} \coloneqq Q_{\mathbf{\Psi}}^{-1}\mathbf{\Psi}$ is a frame of $\mathcal{X}$, called the *(canonical) dual frame* of $\mathcal{X}$. The analysis operator of the dual frame is $\tilde{T}_{\mathbf{\Psi}} \coloneqq T_{\mathbf{\Psi}}(T'_{\mathbf{\Psi}}T_{\mathbf{\Psi}})^{-1}$ and its frame bounds are $\lambda_{\mathbf{\Psi}}^{-1}$ and $\Lambda_{\mathbf{\Psi}}^{-1}$.

Definition 13 *A family $\mathbf{\Psi} = \{\psi_j : j \in \mathbb{N}\} \subset \mathcal{X}$ is called a Riesz basis of $\mathcal{X}$ if there exists a bounded, bijective operator $A : \mathcal{X} \to \mathcal{X}$ and an orthonormal basis $(e_j)_{j\in\mathbb{N}}$ with $\psi_j = Ae_j$ for all $j \in \mathbb{N}$.*

A Riesz basis is a frame $\mathbf{\Psi}$ which is also a basis. Equivalently, a Riesz basis is a frame with $\ker(T'_{\mathbf{\Psi}}) = 0$ and therefore $\operatorname{range}(T_{\mathbf{\Psi}}) = \ell^2(\mathbb{N})$. Moreover, the dual frame $\tilde{\mathbf{\Psi}}$ of a Riesz basis is also a Riesz basis (Christensen, 2016, Section 5).

3.1.2 Encoder and Decoder

Throughout the rest of this paper we fix frames and their duals on $\mathcal{X}$, $\mathcal{Y}$ and denote them by

$$\mathbf{\Psi}_{\mathcal{X}} = (\psi_j)_{j\in\mathbb{N}}, \quad \tilde{\mathbf{\Psi}}_{\mathcal{X}} = (\tilde{\psi}_j)_{j\in\mathbb{N}}, \quad \mathbf{\Psi}_{\mathcal{Y}} = (\eta_j)_{j\in\mathbb{N}}, \quad \tilde{\mathbf{\Psi}}_{\mathcal{Y}} = (\tilde{\eta}_j)_{j\in\mathbb{N}}.$$

The corresponding analysis operators are $T_{\mathbf{\Psi}_{\mathcal{X}}}$, $T_{\tilde{\mathbf{\Psi}}_{\mathcal{X}}}$, $T_{\mathbf{\Psi}_{\mathcal{Y}}}$ and $T_{\tilde{\mathbf{\Psi}}_{\mathcal{Y}}}$. We then introduce encoder and decoder maps via

$$\mathcal{E}_{\mathcal{X}} \coloneqq T_{\mathbf{\Psi}_{\mathcal{X}}} = \begin{cases} \mathcal{X} \to \ell^2(\mathbb{N}), \\ x \mapsto (\langle x, \tilde{\psi}_j\rangle_{\mathcal{X}})_{j\in\mathbb{N}}, \end{cases} \qquad \mathcal{D}_{\mathcal{Y}} \coloneqq T'_{\mathbf{\Psi}_{\mathcal{Y}}} = \begin{cases} \ell^2(\mathbb{N}) \to \mathcal{Y}, \\ (y_j)_{j\in\mathbb{N}} \mapsto \sum_{j\in\mathbb{N}} y_j\eta_j. \end{cases} \tag{3.3}$$

In case $\mathbf{\Psi}_{\mathcal{X}}$ and $\mathbf{\Psi}_{\mathcal{Y}}$ are Riesz bases, the mappings in (3.3) are boundedly invertible.

3.1.3 Smoothness Scales

The encoder mapping $\mathcal{E}_{\mathcal{X}} : \mathcal{X} \to \ell^2(\mathbb{N})$ maps an element to its coefficients in the frame representation. For computational purposes, this coefficient sequence must be truncated, as only finitely many coefficients can be considered. Consequently, it is essential to control the error resulting from discarding higher-order frame coefficients. To formalize this we next introduce scales of subspaces of $\mathcal{X}$, $\mathcal{Y}$, whose elements exhibit a certain coefficient decay.

Definition 14 *Let $\boldsymbol{\theta} = (\theta_j)_{j\in\mathbb{N}}$ be a strictly positive, monotonically decreasing sequence such that $\boldsymbol{\theta}^{1+\varepsilon} \in \ell^1(\mathbb{N})$ for all $\varepsilon > 0$.*

For all r, $t \geq 0$ we introduce the subspaces $\mathcal{X}_r \subset \mathcal{X}$ and $\mathcal{Y}_t \subset \mathcal{Y}$ via

$$\mathcal{X}_r \coloneqq \{x \in \mathcal{X} : \|x\|_{\mathcal{X}_r} < \infty\} \qquad \text{and} \qquad \mathcal{Y}_t \coloneqq \{y \in \mathcal{Y} : \|y\|_{\mathcal{Y}_t} < \infty\}$$

where

$$\|x\|_{\mathcal{X}_r}^2 \coloneqq \sum_{j\in\mathbb{N}} \langle x, \tilde{\psi}_j\rangle_{\mathcal{X}}^2 \theta_j^{-2r} \qquad \text{and} \qquad \|y\|_{\mathcal{Y}_t}^2 \coloneqq \sum_{j\in\mathbb{N}} \langle y, \tilde{\eta}_j\rangle_{\mathcal{Y}}^2 \theta_j^{-2t}.$$

For every $r \geq 0$, $\mathcal{X}_r$ is a Hilbert space (Herrmann et al., 2024, Lemma 1).

3.2 FrameNet

In this subsection we recall the FrameNet architecture from Herrmann et al. (2024, Section 2). We start by formally introducing feedforward neural networks (NNs) following Opschoor et al. (2022a); Herrmann et al. (2024).

3.2.1 Feedforward Neural Networks

Definition 15 *A function $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ is called a neural network, if there exist $\sigma : \mathbb{R} \to \mathbb{R}$, integers $p_1, \dots, p_{L+1} \in \mathbb{N}$, $L \in \mathbb{N}$ and real numbers $w_{i,j}^l$, $b_j^l \in \mathbb{R}$ such that for all $\boldsymbol{x} = (x_i)_{i=1}^{p_0} \in \mathbb{R}^{p_0}$*

$$z_j^1 = \sigma\Big(\sum_{i=1}^{p_0} w_{i,j}^1 x_i + b_j^1\Big), \quad j = 1, \dots, p_1, \tag{3.4a}$$

$$z_j^{l+1} = \sigma\Big(\sum_{i=1}^{p_l} w_{i,j}^{l+1} z_i^l + b_j^{l+1}\Big), \quad l = 1, \dots, L-1, \quad j = 1, \dots, p_{l+1},$$

$$f(x) = (z_j^{L+1})_{j=1}^{p_{L+1}} = \Big(\sum_{i=1}^{p_L} w_{i,j}^{L+1} z_i^L + b_j^{L+1}\Big)_{j=1}^{p_{L+1}}. \tag{3.4b}$$

We call σ the activation function*, L the* depth*, $p \coloneqq \max_{l=0,\dots,L+1} p_l$ the* width*, $w_{i,j}^l \in \mathbb{R}$ the* weights*, and $b_j^l \in \mathbb{R}$ the* biases *of the NN.*

While different NNs can realize the same function, for simplicity we refer to a function $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ as an NN of type (3.4), if it allows for (at least) one such representation. Additionally, our analysis will require some further terminology: For a NN $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ as in (3.4) its

- *size* is the number of nonzero parameters

$$\operatorname{size}(f) \coloneqq |\{(i,j,l) \,:\, w_{i,j}^l \neq 0\}| + |\{(j,l) \,:\, b_j^l \neq 0\}|,$$

- *maximum of parameters* is

$$\operatorname{mpar}(f) \coloneqq \max\Big\{ \max_{i,j,l} |w_{i,j}^l|, \max_{j,l} |b_j^l| \Big\},$$

- *maximum range* on $\Omega \subseteq \mathbb{R}^{p_0}$ is

$$\operatorname{mran}_\Omega(f) \coloneqq \sup_{x\in\Omega} \big\| f(x) \big\|_2 = \sup_{x\in\Omega} \left(\sum_{j=1}^{p_{L+1}} \big(z_j^{L+1}(x)\big)^2 \right)^{\frac{1}{2}}.$$

Throughout, for $q \in \mathbb{N}$, $q \geq 1$, we consider the activation function

$$\sigma_q(x) \coloneqq \max\{0, x\}^q, \qquad x \in \mathbb{R}.$$

For $q = 1$, σ_1 is the *rectified linear unit* (ReLU), for $q \geq 2$, σ_q is called *rectified power unit* (RePU).

Remark 16 *Definition 15 introduces a NN as a function $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$. Throughout, we also understand the realization of a NN as a map $f : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ via extension by zeros. This is equivalent to suitably padding the weight matrices $(w_{i,j}^l)_{i,j}$ and bias vectors $(b_j^l)_j$ in Definition 15 for $l \in \{0, L+1\}$ with infinitely many zeros, see Herrmann et al. (2024, Remark 13).*

3.2.2 The FrameNet Class

By definition of frames, the encoding operator $\mathcal{E}_\mathcal{X} : \mathcal{X} \to \ell^2(\mathbb{N})$ in (3.3) is injective, and the decoding operator $\mathcal{D}_\mathcal{Y} : \ell^2(\mathbb{N}) \to \mathcal{Y}$ is surjective. Thus for every mapping $G : \mathcal{X} \to \mathcal{Y}$, there exists a coefficient map $g : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ such that

$$G = \mathcal{D}_\mathcal{Y} \circ g \circ \mathcal{E}_X.$$

This motivates the introduction of the following function class.

Given $\sigma : \mathbb{R} \to \mathbb{R}$, positive integers L, p, $s \in \mathbb{N}$, and reals M, $B \in \mathbb{R}$, let

$$\begin{aligned} \boldsymbol{g}_{\mathrm{FN}}(\sigma, L, p, s, M, B) \coloneqq \big\{ & g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}} \text{ is NN with activation function } \sigma \text{ s.t.} \\ & \operatorname{depth}(g) \leq L,\ \operatorname{width}(g) \leq p,\ \operatorname{size}(g) \leq s, \\ & \operatorname{mpar}(g) \leq M,\ \operatorname{mran}_{[-1,1]^{p_0}}(g) \leq B,\ p_0, p_{L+1} \leq p \big\}. \end{aligned}$$

Instead of directly considering (3.1), for our analysis, contrary to Herrmann et al. (2024), with $\boldsymbol{\theta}$ from Definition 14, fixed $R > 0$, and $U \coloneqq [-1,1]^{\mathbb{N}}$, it will be convenient to introduce the linear scaling

$$S_r \coloneqq \begin{cases} \times_{j\in\mathbb{N}} [-R\theta_j^r, R\theta_j^r] \to U \\ (x_j)_{j\in\mathbb{N}} \mapsto \left(\frac{x_j}{R\theta_j^r}\right)_{j\in\mathbb{N}}. \end{cases} \tag{3.5}$$

The FrameNet class then consists of all operators

$$\boldsymbol{G}_{\mathrm{FN}}(\sigma, L, p, s, M, B) \coloneqq \big\{ G = \mathcal{D}_\mathcal{Y} \circ g \circ S_r \circ \mathcal{E}_\mathcal{X} : \ g \in \boldsymbol{g}_{\mathrm{FN}}(\sigma, L, p, s, M, B) \big\}. \tag{3.6}$$

3.3 Approximation of Holomorphic Operators

Our statistical theory established in Section 2 shows that sample complexity depends on

(i) the approximation quality of $\boldsymbol{G}$ w.r.t G_0, cf. the term $\inf_{G^* \in \boldsymbol{G}} \|G^* - G_0\|_{L^2(\gamma)}$ in (2.13),

(ii) the metric entropy of $\boldsymbol{G}$, cf. the terms δ_n^2 and $\tilde{\delta}_n^2$ in (2.13).

In this subsection we give results for both the approximation quality and the metric entropy of the FrameNet class $\boldsymbol{G}_{\mathrm{FN}}$.

3.3.1 Setting

Let us start by making the assumptions on G_0 and the sampling measure γ on $\mathcal{X}$ more precise. Denote in the following $U \coloneqq [-1,1]^{\mathbb{N}}$, let $r > \frac{1}{2}$, $R > 0$, and let the frames $(\psi_j)_{j\in\mathbb{N}}$, $(\eta_j)_{j\in\mathbb{N}}$ be as in Section 3, and $(\theta_j)_{j\in\mathbb{N}}$ as in Definition 14. Then

$$\sigma_R^r \coloneqq \begin{cases} U \to \mathcal{X}, \\ \boldsymbol{y} \mapsto R \sum_{j\in\mathbb{N}} \theta_j^r y_j \psi_j \end{cases} \tag{3.7}$$

yields a well-defined map, since by construction $(y_j \theta_j^r)_{j\in\mathbb{N}} \in \ell^2(\mathbb{N})$ for every $\boldsymbol{y} \in U$, and $(\psi_j)_{j\in\mathbb{N}}$ is a frame. Next, we introduce the "cubes"

$$C_R^r(\mathcal{X}) = \left\{ a \in \mathcal{X} : \sup_{j\in\mathbb{N}} \theta_j^{-r} |\langle a, \tilde{\psi}_j \rangle_{\mathcal{X}}| \le R \right\}. \tag{3.8}$$

Observe that with $\sigma_R^r(U) \coloneqq \{\sigma_R^r(\boldsymbol{y}) : \boldsymbol{y} \in U\}$, clearly

$$C_R^r(\mathcal{X}) \subseteq \sigma_R^r(U),$$

but equality holds in general only if the $(\psi_j)_{j\in\mathbb{N}}$ form a Riesz basis (Herrmann et al., 2024, Remark 10).

Example 2 *Let $\mathcal{X} = \mathbb{R}$, $r = 1$, $R = 1$, $\theta_1 = 3/2$ and $\theta_2 = 1/2$. Consider the frame $\boldsymbol{\Psi} = \{1, 1\}$ (which is not a basis) with dual analogue $\tilde{\boldsymbol{\Psi}} = \{1/2, 1/2\}$. Then with $U = [-1,1]^2$*

$$C_1^1(\mathbb{R}) = [-1,1] \subsetneq [-2,2] = \{\theta_1 y_1 \psi_1 + \theta_2 y_2 \psi_2 : \boldsymbol{y} \in U\}.$$

The holomorphy assumption on G_0 can now be formulated as follows (Herrmann et al., 2024, Assumption 1). Recall that for a real Hilbert space $\mathcal{X}$, we denote by $\mathcal{X}_{\mathbb{C}}$ its complexification.

Assumption 2 *For some $r > 1$, $R > 0$, $t > 0$, $C_{G_0} < \infty$ there exists an open set $O_{\mathbb{C}} \subset \mathcal{X}_{\mathbb{C}}$ containing $\sigma_R^r(U)$, such that*

(a) $\sup_{a\in O_{\mathbb{C}}} \|G_0(a)\|_{(\mathcal{Y}_{\mathbb{C}})_t} \le C_{G_0}$, *and $G_0 : O_{\mathbb{C}} \to \mathcal{Y}_{\mathbb{C}}$ is holomorphic,*

(b) γ is a probability measure on $\mathcal{X}$ with $\operatorname{supp}(\gamma) \subseteq C_R^r(\mathcal{X})$.

The assumption requires G_0 to be holomorphic on a superset of $\sigma_R^r(U)$. The probability measure γ is allowed to have support on $C_R^r(\mathcal{X})$ which is potentially smaller than $\sigma_R^r(U)$. In particular, G_0 is then holomorphic on a superset of the support of γ.

Example 3 *Denote by λ the Lebesgue measure. Then $\pi \coloneqq \otimes_{j\in\mathbb{N}} \frac{\lambda}{2}$ is the uniform probability measure on $U = [-1,1]^{\mathbb{N}}$, and*

$$\gamma \coloneqq (\sigma_R^r)_\sharp \pi \tag{3.9}$$

defines a probability measure on $\mathcal{X}$ with support $\sigma_R^r(U)$. If $\boldsymbol{\Psi}_{\boldsymbol{\chi}}$ is a Riesz basis, then $\operatorname{supp}(\gamma) = C_R^r(\mathcal{X})$*, so that* $\operatorname{supp}(\gamma) \subseteq C_R^r(\mathcal{X})$ *as required in Assumption 2.*

We emphasize that Assumption 2 does not require $\boldsymbol{\Psi}_\chi$ to be a Riesz basis, and all results below remain valid for general frames. However, the Riesz basis property yields sharper rates when combined with the specific choice of γ from Example 3.

3.3.2 Approximation Theory

Theorem 1 in Herrmann et al. (2024) establishes a convergence rate for approximating G_0 within $\boldsymbol{G}_{\mathrm{FN}}$ (with unbounded weights), in terms of the number of trainable network parameters. Our main results on sample complexity depend on bounds on the metric entropy, which require bounded network weights. To address this, we now extend Herrmann et al. (2024, Theorem 1) to FrameNet architectures with bounded weights, as introduced in Section 3.

Similar to Schwab and Zech (2019); Opschoor et al. (2022b); Herrmann et al. (2024), the analysis builds on polynomial chaos expansions (Xiu and Karniadakis, 2002). Denote by $(L_j)_{j\in\mathbb{N}}$ the univariate Legendre polynomials normalized such that $\frac{1}{2}\int_{-1}^{1} L_j(x)^2 \,\mathrm{d}x = 1$ for all $j \in \mathbb{N}$. Then (see Abramowitz and Stegun 1948, Chapter 22)

$$\sup_{x\in[-1,1]} |L_j(x)| \le \sqrt{2j+1} \qquad \forall j \in \mathbb{N}_0. \tag{3.10}$$

Next, let $\mathcal{F}$ be the set of infinite-dimensional multiindices with finite support, i.e.

$$\mathcal{F} \coloneqq \Big\{(\boldsymbol{\nu}_j)_{j\in\mathbb{N}_0} \in \mathbb{N}_0^{\mathbb{N}} : |\operatorname{supp}\boldsymbol{\nu}| < \infty\Big\}, \tag{3.11}$$

where $\operatorname{supp}\boldsymbol{\nu} \coloneqq \{j : \nu_j \neq 0\}$. For finite sets of multiindices $\Lambda \subseteq \mathcal{F}$, their effective dimension and maximal order is defined as

$$d(\Lambda) \coloneqq \sup\{|\operatorname{supp}\boldsymbol{\nu}| : \boldsymbol{\nu} \in \Lambda\} \qquad \text{and} \qquad m(\Lambda) \coloneqq \sup\{|\boldsymbol{\nu}| : \boldsymbol{\nu} \in \Lambda\},$$

where $|\boldsymbol{\nu}| \coloneqq \sum_{j\in\mathbb{N}} \nu_j$. Moreover, with $U \coloneqq [-1,1]^{\mathbb{N}}$ for all $\boldsymbol{y} \in U$ and $\boldsymbol{\nu} \in \mathcal{F}$, we let $L_{\boldsymbol{\nu}}(\boldsymbol{y}) \coloneqq \prod_{j\in\mathbb{N}} L_{\nu_j}(y_j)$ be the corresponding multivariate Legendre polynomial. This infinite product is well-defined, since for all but finitely many j holds $\nu_j = 0$ so that $L_{\nu_j} \equiv 1$. The next Proposition gives an approximation result for multivariate Legendre polynomials. It is an extension of Opschoor et al. (2022a, Proposition 2.13) to the case of bounded network parameters. The proof is provided in Appendix D.1.

Proposition 17 (σ_1-NN approximation of $L_{\boldsymbol{\nu}}$) *Let $\delta \in (0, 1/2)$ and $\Lambda \subset \mathcal{F}$ be finite. Then there exists a σ_1-NN $f_{\Lambda,\delta}$, such that its outputs $\{\tilde{L}_{\boldsymbol{\nu},\delta}\}_{\boldsymbol{\nu}\in\Lambda}$ satisfy*

$$\forall \boldsymbol{\nu} \in \Lambda : \qquad \sup_{\boldsymbol{y}\in U} |L_{\boldsymbol{\nu}}(\boldsymbol{y}) - \tilde{L}_{\boldsymbol{\nu},\delta}(\boldsymbol{y})| \le \delta. \tag{3.12}$$

Furthermore, there exists a constant $C > 0$ independent of Λ, $d(\Lambda)$, $m(\Lambda)$ and δ such that

$$\operatorname{depth}(f_{\Lambda,\delta}) \leq C\Big[\log(d(\Lambda))d(\Lambda)\log(m(\Lambda))m(\Lambda) + m(\Lambda)^2 + \log(\delta^{-1})\big(\log(d(\Lambda)) + m(\Lambda)\big)\Big]$$

$$\begin{aligned}\operatorname{width}\big(f_{\Lambda,\delta}\big) &\leq C|\Lambda|d(\Lambda)\\ \operatorname{size}\big(f_{\Lambda,\delta}\big) &\leq C\Big[|\Lambda|d(\Lambda)^2\log\big(m(\Lambda)\big) + m(\Lambda)^3 d(\Lambda)^2 + \log\big(\delta^{-1}\big)\big(m(\Lambda)^2 + |\Lambda|\big)d(\Lambda)\Big]\\ \operatorname{mpar}\big(f_{\Lambda,\delta}\big) &\leq 1.\end{aligned}$$

A similar result holds for RePU neural networks, see Proposition 51. Given $q \in \mathbb{N}$, and $N \in \mathbb{N}$, we introduce the two FrameNet classes

$$\begin{aligned}\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N) &:= \boldsymbol{G}_{\mathrm{FN}}(\sigma_q, \mathrm{width}_N, \mathrm{depth}_N, \mathrm{size}_N, M, B),\\ \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N) &:= \boldsymbol{G}_{\mathrm{FN}}(\sigma_q, \mathrm{width}_N, \mathrm{depth}_N, \infty, M, B)\end{aligned} \tag{3.14a}$$

where for certain constants C_L, C_p, C_s, M, $B \geq 1$ (to be determined later)

$$\mathrm{depth}_N = \max\{1, \lceil C_L \log(N)\rceil\}, \qquad \mathrm{width}_N = \lceil C_p N\rceil, \qquad \mathrm{size}_N = \lceil C_s N\rceil, \tag{3.14b}$$

and $\lceil x\rceil$ denotes the smallest integer larger or equal to $x \in \mathbb{R}$.

We emphasize that $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$ corresponds to a sparsely connected architecture, whereas $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$ represents a fully connected architecture (because there is no constraint on its size). In particular, since every linear transformation in between activation functions has at most $\mathrm{width}_N + \mathrm{width}_N^2$ parameters, we have

$$\mathrm{size}(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)) \leq (\mathrm{depth}_N + 1)(\mathrm{width}_N + \mathrm{width}_N^2) = O(\log(N)N^2) \qquad \text{as } N \to \infty. \tag{3.15}$$

Thus $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$ can essentially be quadratically larger than $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$.

The proof of Theorem 18 below, given in Appendix D.3, crucially uses Proposition 17 to construct FrameNets that approximate the multivariate Legendre polynomials corresponding to the 'most relevant' Legendre coefficients of G_0. Hereby Assumption 2 guarantees sufficient decay of the Legendre coefficients by using that algebraic decay in the input space, see Assumption 2 (b), is inherited by the Legendre coefficients of the output (see Theorem 54 and 55) if the respective operator is holomorphic, see Assumption 2 (a).

Theorem 18 is an extension of Herrmann et al. (2024, Theorems 1 and 2) to the case of bounded network parameters.

Theorem 18 (Sparse network approximation) *Let G_0, γ satisfy Assumption 2 with $r > 1$, $t > 0$. Let $q \geq 1$ be an integer and fix $\tau > 0$ (arbitrarily small).*

(i) There exists $C > 0$ s.t. for all $N \in \mathbb{N}$

$$\inf_{G \in \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)} \left[\big\|G - G_0\big\|_{\infty,\mathrm{supp}(\gamma)}^2\right] \leq C N^{-2\min\{r-1,t\}+\tau}.$$

(ii) Let $\boldsymbol{\Psi}_\chi$ be a Riesz basis and let γ be as in (3.9). Then there exists $C > 0$ s.t. for all $N \in \mathbb{N}$

$$\inf_{G \in \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)} \left[\big\|G - G_0\big\|_{L^2(\gamma)}^2\right] \leq C N^{-2\min\{r-\frac{1}{2},t\}+\tau}.$$

Remark 19 *Theorem 18 provides an approximation rate for* $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$ *in terms of* N*. Since* $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N) \supseteq \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$*, trivially the statement remains true for* $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$*. However, as the size of* $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$ *can be quadratically larger by* (3.15)*, the convergence rate in terms of* network size *is essentially halved for the fully connected architecture.*

3.3.3 Entropy Bounds for FrameNet

In the following we bound the metric entropy (see 1.3) of FrameNets for ReLU activation functions. Recall that $\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}$ is the frame constant in (3.2). The proof of Lemma 20 is given in Appendix D.4.

Lemma 20 (see Schmidt-Hieber 2020a, Lemma 5) *Let* L*,* p*,* s*,* M*,* $B \geq 1$*,* σ_R^r *be as in* (3.7) *and* $U = [-1,1]^{\mathbb{N}}$*. Then* $\boldsymbol{G}_{\mathrm{FN}} = \boldsymbol{G}_{\mathrm{FN}}(\sigma_1, L, p, s, M, B)$ *is compact with respect to* $\|\cdot\|_{\infty,\sigma_R^r(U)}$ *and* $\|\cdot\|_n$*. Furthermore,* $\boldsymbol{G}_{\mathrm{FN}}$ *satisfies for all* $\delta > 0$

$$H(\boldsymbol{G}_{\mathrm{FN}}, \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq (s+1)\log\Big(2^{L+6}\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}} L^2 M^{L+1} p^{L+4} \max\{1, \delta^{-1}\}\Big). \quad (3.16)$$

In particular there exist constants C_H^{SP}*,* $C_H^{FC} > 0$ *such that for the sparse and fully-connected FrameNet classes from* (3.14) *it holds*

$$H(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_1, N), \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq C_H^{SP} N \left(1 + \log(N)^2 + \log\left(\max\left\{1, \delta^{-1}\right\}\right)\right)$$
$$H(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_1, N), \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq C_H^{FC} N^2 \left(1 + \log(N)^3 + \log\left(\max\left\{1, \delta^{-1}\right\}\right)\right)$$

for all $\delta > 0$*.*

Remark 21 *The entropy bound is independent of the constant* B *bounding the maximum range of the network, see Subsection 3.2. However,* $B < \infty$ *will be necessary to apply Theorem 6.*

For RePU activation, as mentioned before, the metric entropy bounds exhibit a worse dependency on the network parameters, due to the lack of global Lipschitz continuity of σ_q if $q \geq 2$. The proof of Lemma 22 is given in Appendix D.5.

Lemma 22 *Let* $q \in \mathbb{N}$*,* $q \geq 2$*, and* L*,* p*,* s*,* M*,* $B \geq 1$*,* σ_R^r *be as in* (3.7) *and* $U = [-1,1]^{\mathbb{N}}$*. Then* $\boldsymbol{G}_{\mathrm{FN}} = \boldsymbol{G}_{\mathrm{FN}}(\sigma_q, L, p, s, M, B)$ *is compact with respect to* $\|\cdot\|_{\infty,\sigma_R^r(U)}$ *and* $\|\cdot\|_n$*. Furthermore,* $\boldsymbol{G}_{\mathrm{FN}}$ *satisfies for all* $\delta > 0$

$$H(\boldsymbol{G}_{\mathrm{FN}}, \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq (s+1)\log\Big(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}} L q^{L+q} (2pM)^{4q^{2L+2}} \max\{1, \delta^{-1}\}\Big). \quad (3.17)$$

Consider the constant C_L *from* (3.14)*. Then there exists* C_H^{SP}*,* $C_H^{FC} > 0$ *such that*

$$H(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N), \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq C_H^{SP} N^{1+2C_L\log(q)} \left(1 + \log(N)^2 + \log\left(\max\left\{1, \delta^{-1}\right\}\right)\right)$$
$$H(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N), \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq C_H^{FC} N^{2+2C_L\log(q)} \left(1 + \log(N)^2 + \log\left(\max\left\{1, \delta^{-1}\right\}\right)\right)$$

for all $\delta > 0$*.*

3.4 Statistical Theory

Using our statistical results from Theorem 6 and Corollaries 7-8 as well as the approximation result from Theorem 18, we can bound $\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}]$ solely in terms of the statistical sample size n. This is formalized in the next two theorems, for ReLU and RePU activation.

Our result distinguishes between the sparse and fully connected architectures $\boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$ and $\boldsymbol{G}^{\mathrm{full}}_{\mathrm{FN}}(\sigma_q, N)$ in (3.14). In practice, fully connected architectures are often preferred, due to their simpler implementation, and because training sparse NN architectures can run into problems like "bad" local minima, see for example Evci et al. (2019). Our theoretical upper bounds are sharper for sparse architectures; this is because the additional free parameters in the fully connected architecture increase the entropy of this class, but do not yield better approximation properties in our proofs.

Theorem 23 (ERM for white noise and ReLU) *Let $G_0 : \mathcal{X} \to \mathcal{Y}$ and γ satisfy Assumption 2 for some $r > 1$, $t > 0$. Fix $\tau > 0$ (arbitrarily small) and set (cp. Example 3)*

$$\kappa := \begin{cases} 2\min\{r - \frac{1}{2}, t\} & \text{if } \boldsymbol{\Psi}_{\mathcal{X}} \text{ is a Riesz basis and } \gamma = (\sigma_R^r)_{\#}\pi, \\ 2\min\{r-1, t\} & \text{otherwise.} \end{cases}$$

For every $n \in \mathbb{N}$, let $(x_i, y_i)_{i=1}^n$ be data generated by (2.1) either in the white noise model or in the sub-Gaussian noise model. Then there exist constants C_L, C_p, C_s, M, $B \geq 1$ in (3.14) and $C > 0$ (all independent of n) such that

(i) **Sparse FrameNet:** *with $N = N(n) = \lceil n^{\frac{1}{\kappa+1}} \rceil$ and $\boldsymbol{G} = \boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_1, N)$, there exists a measurable choice of an ERM $\hat{G}_n$ of (2.3). Any such $\hat{G}_n$ satisfies*

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}] \leq C n^{-\frac{\kappa}{\kappa+1}+\tau}, \tag{3.18}$$

(ii) **Fully connected FrameNet:** *with $N = N(n) = \lceil n^{\frac{1}{\kappa+2}} \rceil$ and $\boldsymbol{G} = \boldsymbol{G}^{\mathrm{full}}_{\mathrm{FN}}(\sigma_1, N)$, there exists a measurable choice of an ERM $\hat{G}_n$ of (2.3). Any such $\hat{G}_n$ satisfies*

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}] \leq C n^{-\frac{\kappa}{\kappa+2}+\tau}. \tag{3.19}$$

Proof The proof follows directly from the entropy bounds in Lemma 20 and Corollary 8 (ii): First, Lemma 20 in particular verifies Assumption 1 for all $G^* \in \boldsymbol{G}$, which is required for Corollary 8. Applying the corollary with $\beta = \kappa + \tau$ and absorbing the logarithmic term into τ then gives (3.18). With $\beta = \kappa/2 + \tau$ we obtain (3.19). Note that crucially $M_{\boldsymbol{F}}$, see (2.9), does not depend on N, because $\|G_0\|_{\infty,\operatorname{supp}\gamma} < \infty$ and $\boldsymbol{G}_{\mathrm{FN}}$ is universally bounded by the N-independent constant B, see (3.6). ■

Remark 24 *Consider the sparse FrameNet class from above. Similar to the proof of Corollary 8 (ii), one can show that the sequences*

$$(\delta_n)_{n\in\mathbb{N}}, \quad (\tilde{\delta}_n)_{n\in\mathbb{N}}, \quad \delta_n = \tilde{\delta}_n = N\log(N)^2(1+\log(n))/n$$

satisfy the entropy conditions (2.10) *for the sparse FrameNet class* $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_1, N)$. *By choosing* $N = \lceil n^{1/(\kappa+1)} \rceil$ *to balance the approximation term in* (2.11), *Theorem 6 gives the following high probability bound for* $\|\hat{G}_n - G_0\|_{L^2(\gamma)}$: *For all* $\tau > 0$ *there exists a constant* $C > 0$ *such that for all* $n \in \mathbb{N}$ *and all* $R_n \geq C n^{-\kappa/(\kappa+1)+\tau}$ *it holds*

$$\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R_n\right) \leq 2\exp\left(-\frac{nR_n^2}{C^2\left(\sigma^2 + M_{\boldsymbol{F}}^2\right)}\right).$$

Similar high-probability bounds hold for the fully connected FrameNet class and the RePU activation function.

Theorem 25 (ERM for white noise and RePU) *Consider the setting of Theorem 23 and let* $q \in \mathbb{N}$, $q \geq 2$. *There exist constants* C_L, C_p, C_s, M, $B \geq 1$ *in* (3.14) *and* $C > 0$ *(all independent of* N*) such that*

(i) **Sparse FrameNet:** *with* $N = N(n) = \lceil n^{\frac{1}{\kappa+1+4C_L\log(q)}} \rceil$ *and* $\boldsymbol{G} = \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$, *there exists a measurable choice of an ERM* $\hat{G}_n$ *of* (2.3). *Any such* $\hat{G}_n$ *satisfies*

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2] \leq C n^{-\frac{\kappa}{\kappa+1+4C_L\log(q)}+\tau}. \tag{3.20}$$

(ii) **Fully connected FrameNet:** *with* $N = N(n) = \lceil n^{\frac{1}{\kappa+2+4C_L\log(q)}} \rceil$, $\boldsymbol{G} = \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$, *there exists a measurable choice of ERM* $\hat{G}_n$ *of* (2.3). *Any such* $\hat{G}_n$ *satisfies*

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2] \leq C n^{-\frac{\kappa}{\kappa+2+4C_L\log(q)}+\tau}. \tag{3.21}$$

Proof Similar to the proof of Theorem 23, the proof of Theorem 25 follows from the entropy bounds in Lemma 22 and Corollary 8 (ii) with $\beta = \kappa/(1+2C_L\log(q))+\tau$ for (3.20) and $\beta = \kappa/(2+2C_L\log(q))+\tau/2$ for (3.21). Again the logarithmic term is absorbed into τ. ∎

A few remarks are in order. In the RePU case the activation function $\sigma_q(x) = \max\{0, x\}^q$ has no global Lipschitz condition for $q \geq 2$. As a result, the entropy bounds obtained for the corresponding FrameNet class are larger than for ReLU. This leads to worse convergence rates. Moreover, for the RePU case, the convergence rate depends on the constant C_L in (3.14b). The proof of Theorem 18 shows that C_L depends on the decay properties of the Legendre coefficients $(c_{\boldsymbol{\nu}_i,j})_{i,j\in\mathbb{N}}$ of the function $G_0 \circ \sigma_R^r$, i.e. C_L depends on G_0 (see (D.23) and (D.24) in Theorem 54). Explicit bounds on C_L are possible, see Zech (2018, Lemma 1.4.15).

4. Applications

We now present two applications of our results. First, in finite dimensional regression, our analysis recovers well-known minimax-optimal rates for standard smoothness classes. This indicates that our main statistical result in Theorem 6 is *in general optimal*. However, we do not claim optimality specifically for the approximation of holomorphic operators discussed in Section 3. Second, for an infinite dimensional problem we address the learning of a solution operator to a parameter dependent PDE.

4.1 Finite Dimensional Regression

Let $d \in \mathbb{N}$ and $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \mathbb{R}$. Moreover, let $D \subseteq \mathbb{R}^d$ be a bounded, open, smooth domain, and $G_0 : D \to \mathbb{R}$ a ground-truth regression function. Suppose that

$$x_i \overset{\text{iid}}{\sim} \gamma \quad \text{and} \quad \varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0,1) \qquad \forall i \in \mathbb{N}$$

are independent samples for some probability measure γ on D, and let

$$y_i = G_0(x_i) + \varepsilon_i \qquad \forall i \in \mathbb{N}.$$

Given a regression class $\boldsymbol{G}$ of measurable mappings from $D \to \mathbb{R}$, the least-squares problem is to determine

$$\hat{G}_n \in \operatorname*{arg\,min}_{G \in \boldsymbol{G}} \sum_{i=1}^{n} -2G(x_i)y_i + G(x_i)^2 = \operatorname*{arg\,min}_{G \in \boldsymbol{G}} \sum_{i=1}^{n} |G(x_i) - y_i|^2. \tag{4.2}$$

For $\mathfrak{s} \geq 0$, it is well-known that the minimax-optimal rate of recovering a ground-truth function G_0 in $\mathfrak{s}$-smooth function classes, such as the Sobolev space $H^{\mathfrak{s}}(D)$ or the Besov space $B^{\mathfrak{s}}_{\infty,\infty}(D)$ (see Edmunds and Triebel 1996 for definitions), equals $n^{-\frac{2\mathfrak{s}}{2\mathfrak{s}+d}}$ (Giné and Nickl, 2016; van de Geer, 2000).

Denote now by $\boldsymbol{G}^{\mathfrak{s}}_R$ the ball of radius $R > 0$ around the origin in either $H^{\mathfrak{s}}(D)$ or $B^{\mathfrak{s}}_{\infty,\infty}(D)$. Then,

$$H(\boldsymbol{G}^{\mathfrak{s}}_R, \|\cdot\|_{\infty,\operatorname{supp}(\gamma)}, \delta) \simeq \left(\frac{R}{\delta}\right)^{d/\mathfrak{s}} \qquad \forall \delta \in (0,1),$$

which holds for all $\mathfrak{s} > 0$ if $\boldsymbol{G}^{\mathfrak{s}}_R$ is the ball in $B^{\mathfrak{s}}_{\infty,\infty}(D)$, and for all $\mathfrak{s} > d/2$ in case $\boldsymbol{G}^{\mathfrak{s}}_R$ is the ball in $H^{\mathfrak{s}}(D)$, see Triebel (1995, Theorem 4.10.3). Corollary 8 (i) (with $\alpha = d/\mathfrak{s} < 2$) then directly yields the following theorem. It recovers the minimax optimal rate for nonparametric least squares/maximum likelihood estimators.

Theorem 26 *Let $R > 0$ and $\mathfrak{s} > d/2$. Then, there exists $C > 0$ such that for all $G_0 \in \boldsymbol{G}^{\mathfrak{s}}_R$, the estimator $\hat{G}_n$ in (4.2) with $\boldsymbol{G} = \boldsymbol{G}^{\mathfrak{s}}_R$ and data as in (4.1) satisfies*

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|^2_{L^2(D)}\right] \leq Cn^{-\frac{2}{2+\alpha}} = Cn^{-\frac{2\mathfrak{s}}{2\mathfrak{s}+d}} \qquad \forall n \in \mathbb{N}.$$

4.2 Parametric Darcy Flow

As a second application we apply Theorem 23 to the solution operator of the diffusion equation, extending the discussion of approximation errors in Herrmann et al. (2024, Section 7.1).

4.2.1 Setup

We recall the setup from Herrmann et al. (2024, Sections 7.1.1, 7.1.2).

Let $d \in \mathbb{N}$, and denote by $\mathbb{T}^d \simeq [0,1]^d$ the d-dimensional torus. In the following, all function spaces on $\mathbb{T}^d$ are understood to be one-periodic in each variable. Fix $\bar{a} \in L^\infty(\mathbb{T}^d)$ and $f \in H^{-1}(\mathbb{T}^d)/\mathbb{R}$ such that for some constant $a_{\min} > 0$

$$\operatorname*{ess\,inf}_{x \in \mathbb{T}^d} (\bar{a}(x) + a(x)) > a_{\min}. \tag{4.3}$$

We consider the ground truth $G_0 : a \mapsto u$, mapping $a \in L^\infty(\mathbb{T}^d)$ to the solution $u \in H^1(\mathbb{T}^d)$ of

$$-\nabla \cdot ((\bar{a} + a)\nabla u) = f \text{ on } \mathbb{T}^d \qquad \text{and} \qquad \int_{\mathbb{T}^d} u(x)\,\mathrm{d}x = 0. \tag{4.4}$$

Then $G_0 : \{a \in L^\infty(\mathbb{T}^d) : (4.3) \text{ holds}\} \to H^1(\mathbb{T}^d)$ is well-defined.

To represent a and u, we use Fourier expansions on $\mathbb{T}^d$. Denote for $j \in \mathbb{N}_0$ and $\boldsymbol{j} \in \mathbb{N}_0^d$, $d \geq 2$,

$$\xi_0 \coloneqq 1, \quad \xi_{2j}(x) \coloneqq \sqrt{2}\cos(2\pi j x), \quad \xi_{2j-1}(x) \coloneqq \sqrt{2}\sin(2\pi j x), \quad \xi_{\boldsymbol{j}}(x_1, \dots, x_d) \coloneqq \prod_{k=1}^{d} \xi_{j_k}(x_k).$$

Then for $r \geq 0$, $\{\max\{1, |\boldsymbol{j}|\}^r \xi_{\boldsymbol{j}} : \boldsymbol{j} \in \mathbb{N}_0^d\}$ forms an ONB of $H^r(\mathbb{T}^d)$ equipped with inner product

$$\langle u, v\rangle_{H^r(\mathbb{T}^d)} \coloneqq \sum_{\boldsymbol{j} \in \mathbb{N}_0^d} \langle u, \xi_{\boldsymbol{j}}\rangle_{L^2} \langle v, \xi_{\boldsymbol{j}}\rangle_{L^2} \max\{1, |\boldsymbol{j}|\}^{2r}.$$

In the following, fix r_0, $t_0 \geq 0$ and set

$$\begin{aligned} \mathcal{X} &\coloneqq H^{r_0}(\mathbb{T}^d), \qquad & \psi_{\boldsymbol{j}} &\coloneqq \max\{1, |\boldsymbol{j}|\}^{-r_0} \xi_{\boldsymbol{j}}, \\ \mathcal{Y} &\coloneqq H^{t_0}(\mathbb{T}^d), \qquad & \eta_{\boldsymbol{j}} &\coloneqq \max\{1, |\boldsymbol{j}|\}^{-t_0} \xi_{\boldsymbol{j}}, \end{aligned} \tag{4.5}$$

so that $\boldsymbol{\Psi}_{\mathcal{X}} \coloneqq (\psi_{\boldsymbol{j}})_{\boldsymbol{j} \in \mathbb{N}_0^d}$, $\boldsymbol{\Psi}_{\mathcal{Y}} \coloneqq (\eta_{\boldsymbol{j}})_{\boldsymbol{j} \in \mathbb{N}_0^d}$ form ONBs of $\mathcal{X}$, $\mathcal{Y}$ respectively. The encoder $\mathcal{E}_{\mathcal{X}}$ and decoder $\mathcal{D}_{\mathcal{Y}}$ are now as in (3.3). Direct calculation shows $\mathcal{X}^r = H^{r_0 + rd}$ and $\mathcal{Y}^t = H^{t_0 + td}$ for r, $t \geq 0$; for more details see Herrmann et al. (2024, Section 7.1.2).

4.2.2 Sample Complexity

We now analyze the sample complexity for learning the PDE solution operator G_0 in Section 4.2.1. For a proof of Theorem 27, see Appendix E.1.

Theorem 27 *Let $d \in \mathbb{N}$, $d \geq 2$, $\mathfrak{s} > \frac{3d}{2}$ and $t_0 \in [0, 1]$. Fix $\tau_1 > 0$, $\tau_2 \in (0, \min\{\mathfrak{s} - \frac{3d}{2}, \frac{\tau_1 d}{8}\})$ (both arbitrarily small), and set*

$$r_0 = \begin{cases} \frac{d}{2} + \tau_2 & \text{if } \mathfrak{s} \in (\frac{3d}{2}, 2d + 1 - t_0] \\ \frac{\mathfrak{s} + t_0 - 1}{2} & \text{if } \mathfrak{s} > 2d + 1 - t_0. \end{cases}$$

Moreover let $f \in C^\infty(\mathbb{T}^d)$, and let

- *(a)* **ground truth***: $G_0 : a \mapsto u$ be given through* (4.4)*,*
- *(b)* **representation system***: $\mathcal{E}_{\mathcal{X}}$, $\mathcal{D}_{\mathcal{Y}}$ be as in* (3.3) *with the orthonormal basis in* (4.5)*, and r_0, t_0 from above,*
- *(c)* **data***: γ be the measure defined in* (3.7) *and* (3.9) *with $r = \frac{\mathfrak{s} - r_0}{d}$ such that $\bar{a} + a$ satisfies* (4.3) *for all $a \in \mathrm{supp}(\gamma)$, and let $(x_i, y_i)_{i=1}^n$ be generated by* (2.1) *with the additive white noise model or the sub-Gaussian noise model,*

(d) **regression class***:* $\boldsymbol{G} = \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_1, N)$ *be the* n*-dependent sparse FrameNet architecture in* (3.14) *with* $N(n) = \lceil n^{\frac{1}{\kappa+1}} \rceil$.

Then there exists a constant $C > 0$ *such that for all* $n \in \mathbb{N}$ *there exists a measurable ERM* $\hat{G}_n \in \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_1, N(n))$ *in* (2.3)*, and any such* $\hat{G}_n$ *satisfies*

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(H^{r_0}(\mathbb{T}^d),\gamma;H^{t_0}(\mathbb{T}^d))}] \leq C n^{-\frac{\kappa}{\kappa+1}+\tau_1} \tag{4.6}$$

where

$$\kappa = \begin{cases} 2\min\{\frac{\mathfrak{s}}{d} - 1, \frac{1-t_0}{d}\} & \text{if } \mathfrak{s} \in (\frac{3d}{2}, 2d+1-t_0], \\ \frac{\mathfrak{s}+1-t_0}{d} - 1 & \text{if } \mathfrak{s} > 2d+1-t_0. \end{cases} \tag{4.7}$$

Remark 28 *Consider the setting of Theorem 27, and let* $\mathrm{supp}(\gamma) \subseteq C_R^r(\mathcal{X})$*. A slight modification of the proof of Theorem 27 similar to Herrmann et al. (2024), using the approximation bound in Theorem 18 (i) instead of (ii), then yields*

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(H^{r_0}(\mathbb{T}^d),\gamma;H^{t_0}(\mathbb{T}^d))}] \leq C n^{-\frac{\kappa}{\kappa+1}+\tau_1} \tag{4.8}$$

where for some (small) $\tau_2 > 0$

$$(r_0, \kappa) = \begin{cases} (\frac{d}{2} + \tau_2, \frac{2\mathfrak{s}}{d} - 3) & \text{if } \mathfrak{s} \in (\frac{3d}{2}, \frac{3d}{2} + 1 - t_0] \\ (\frac{\mathfrak{s}+t_0-\frac{d}{2}-1}{2}, \frac{\mathfrak{s}+1-t_0}{d} - \frac{3}{2}) & \text{if } \mathfrak{s} > \frac{3d}{2} + 1 - t_0. \end{cases}$$

Since

$$B_R(H^{\mathfrak{s}}(\mathbb{T}^d)) = B_R(\mathcal{X}^r) = \{x \in \mathcal{X} : \ \|x\|_{\mathcal{X}^r} \leq R\} = \left\{ x \in \mathcal{X} : \sum_{j\in\mathbb{N}} \langle x, \tilde{\psi}_j\rangle_{\mathcal{X}}^2 \theta_j^{-2r} \leq R^2 \right\}$$
$$\subseteq \left\{ a \in \mathcal{X} : \sup_{j\in\mathbb{N}} \theta_j^{-r} |\langle a, \tilde{\psi}_j\rangle_{\mathcal{X}}| \leq R \right\} = C_R^r(\mathcal{X}),$$

in particular, (4.8) *holds for any* γ *with* $\mathrm{supp}(\gamma) \subseteq B_R(H^{\mathfrak{s}}(\mathbb{T}^d))$*. This shows Theorem 1.*

Similar rates can also be obtained for this PDE model on a convex, polygonal domain $D \subset \mathbb{T}^2$ with Dirichlet boundary conditions. The argument uses the Riesz basis constructed in Davydov and Stevenson (2005), but is otherwise similar to the torus, for details see Herrmann et al. (2024, Section 7.2). Moreover, using the RePU activation function, (4.6) holds with convergence rate $\kappa/(\kappa + 1 + 4C_L \log(q))$, where κ is from (4.7) and C_L from (3.14b). The proof is similar to Theorem 27 using Theorem 25 instead of Theorem 23. Finally, rates for the fully-connected class $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$ can be established using Theorems 23 (ii) and 25 (ii).

5. Conclusions

In this work, we established convergence theorems for empirical risk minimization to learn mappings G_0 between infinite dimensional Hilbert spaces. Our setting assumes given data

in the form of n input-output pairs, with an additive noise model. We discuss both the case of Gaussian white noise and sub-Gaussian noise. Our main statistical result, Theorem 6, bounds the mean-squared L^2-error $\mathbb{E}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}]$ in terms of the approximation error, and algebraic rates in n depending only on the metric entropy of $\boldsymbol{G}$. This provides a general framework to study operator learning from the perspective of sample complexity.

In the second part of this work, we applied our statistical results to a specific operator learning architecture from Herrmann et al. (2024), termed FrameNet. As our main application, we showed that holomorphic operators G_0 can be learned with NN-based surrogates without suffering from the curse of dimensionality, cf. Theorem 23. Such results have wide applicability, as the required holomorphy assumption is well-established in the literature, and has been verified for a variety of models including for example general elliptic PDEs (Cohen et al., 2010, 2011; Cohen and DeVore, 2015; Harbrecht et al., 2016), Maxwell's equations (Jerez-Hanckes et al., 2017), the Calderon projector (Henríquez and Schwab, 2021), the Helmholtz equation (Hiptmair et al., 2018; Spence and Wunsch, 2023) and also nonlinear PDEs such as the Navier-Stokes equations (Cohen et al., 2018).

The main observation model considered in this work is white noise, which can be understood as an idealization of taking (equally spaced, noisy) point evaluations, see Remark 4. Our measurement model hence represents the process of gathering real-world data from physical systems rather than data generated from approximate solutions of PDEs, in which case the numerical error can be made arbitrarily small. Covering noise-free data with our statistical framework is left for future research.

Extending our results from Hilbert spaces to Banach spaces seems non-trivial. The white noise model is not well-defined in Banach spaces and would need to be replaced. Moreover, bounding the empirical risk in Theorem 5 uses the Hilbert space structure in an essential way: both the basic inequality (Lemma 30) and the chaining Lemma (Lemma 32) rely on the presence of an inner product and would need to be adapted in a Banach space setting.

We have also not discussed the training procedure for *finding* an empirical risk minimizer. The underlying optimization problem is in general non-convex and hence algorithms can typically only approximate the global minimizer up to some tolerance. However, we believe that large parts of our analysis extend straightforwardly to such approximate minimizers with the tolerance appearing as an additional additive term in the mean squared error bound. Full details are left for future work.

Acknowledgments

SW and JZ gratefully acknowledge support from the German Research Foundation (DFG) within the Priority Programme SPP 2298, *Theoretical Foundations of Deep Learning* (project number 543965776).

Appendix A. Auxiliary Probabilistic Lemmas

We recall the classical Bernstein inequality.

Lemma 29 (Bernstein's inequality) *Let $X_1, \dots, X_n$ be independent, centered RVs with finite second moments $\mathbb{E}[X_i^2] < \infty$ and uniform bound $|X_i| \leq M$ for $i = 1, \dots, n$. Then it holds*

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{i=1}^n \mathbb{E}[X_i^2] + \frac{2}{3}Mt}\right), \quad t \geq 0.$$

For a proof of Bernstein's inequality, see Pollard (1984, page 193).

Lemma 30 (Basic Inequality) *Let ε_i, $i = 1, \dots, n$ be i.i.d. white noise or sub-Gaussian noise. Then it holds for all $G^* \in \boldsymbol{G}$*

$$\|\hat{G}_n - G_0\|_n^2 \leq \|G^* - G_0\|_n^2 + \frac{2\sigma}{n}\sum_{i=1}^n \langle \varepsilon_i, \hat{G}_n(x_i) - G^*(x_i)\rangle_{\mathcal{Y}}.$$

Proof Let $G^* \in \boldsymbol{G}$ be arbitrary. Using the definition of $\hat{G}_n$ from (2.3), it holds

$$\begin{aligned}
&\|\hat{G}_n - G_0\|_n^2 \\
&= \frac{1}{n}\sum_{i=1}^n \|\hat{G}_n(x_i)\|_{\mathcal{Y}}^2 - 2\langle G_0(x_i), \hat{G}_n(x_i)\rangle_{\mathcal{Y}} + \|G_0(x_i)\|_{\mathcal{Y}}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \|\hat{G}_n(x_i)\|_{\mathcal{Y}}^2 - 2\langle G_0(x_i) + \sigma\varepsilon_i, \hat{G}_n(x_i)\rangle_{\mathcal{Y}} + 2\sigma\langle \varepsilon_i, \hat{G}_n(x_i)\rangle_{\mathcal{Y}} + \|G_0(x_i)\|_{\mathcal{Y}}^2 \\
&\leq \frac{1}{n}\sum_{i=1}^n \|G^*(x_i)\|_{\mathcal{Y}}^2 - 2\langle G_0(x_i) + \sigma\varepsilon_i, G^*(x_i)\rangle_{\mathcal{Y}} + 2\sigma\langle \varepsilon_i, \hat{G}_n(x_i)\rangle_{\mathcal{Y}} + \|G_0(x_i)\|_{\mathcal{Y}}^2 \\
&= \frac{1}{n}\sum_{i=1}^n \|G^*(x_i)\|_{\mathcal{Y}}^2 - 2\langle G_0(x_i), G^*(x_i)\rangle_{\mathcal{Y}} + \|G_0(x_i)\|_{\mathcal{Y}}^2 + 2\sigma\langle \varepsilon_i, \hat{G}_n(x_i) - G^*(x_i)\rangle_{\mathcal{Y}} \\
&= \|G^* - G_0\|_n^2 + \frac{2\sigma}{n}\sum_{i=1}^n \langle \varepsilon_i, \hat{G}_n(x_i) - G^*(x_i)\rangle_{\mathcal{Y}},
\end{aligned}$$

which shows the claim. ∎

Next, we state a generic chaining result from Dirksen (2015, Theorem 3.2), originally derived for finite index sets, for countable index sets T. We restrict ourselves to the case of real-valued stochastic processes.

Lemma 31 *Let T be a countable index set and $d : T \times T \to [0, \infty)$ a pseudometric. Furthermore, let $(X_t)_{t\in T}$ be an $\mathbb{R}$-valued stochastic process such that for some $\alpha > 0$ and all $s, t \in T$,*

$$\mathbb{P}\left(|X_t - X_s| \geq u d(t, s)\right) \leq 2\exp\left(-u^\alpha\right), \quad u \geq 0. \tag{A.1}$$

Then, there exists a constant $M > 0$ depending only on α, such that for all $t_0 \in T$ it holds that

$$\mathbb{P}\left(\sup_{t\in T}|X_t - X_{t_0}| \geq M\left(J_\alpha(T,d) + u\sup_{s,t\in T} d(s,t)\right)\right) \leq \exp\left(-\frac{u^\alpha}{\alpha}\right), \quad u \geq 1,$$

where J_α denotes the metric entropy integral

$$J_\alpha(T,d) = \int_0^\infty (\log N(T,d,u))^{\frac{1}{\alpha}}\, \mathrm{d}u.$$

Proof Since T is countable, we can write $T = \{t_j : j \in \mathbb{N}\}$. Using this, we define $T_n = \{t_j : j \leq n\}$ for $n \in \mathbb{N}$. Since T_n is finite (Dirksen, 2015, Theorem 3.2, Eq. 3.2 and its proof) gives $\tilde{M} > 0$ s.t.

$$\left(\mathbb{E}\left[\sup_{t\in T_n}|X_t - X_{t_0}|^p\right]\right)^{\frac{1}{p}} \leq \tilde{M}\left(J_\alpha(T_n,d) + \sup_{s,t\in T_n} d(s,t)p^{\frac{1}{\alpha}}\right) \tag{A.2}$$

for all $p \geq 1$, $t_0 \in T$ and $n \in \mathbb{N}$. In (A.2), we used Dirksen (2015, Eq. (2.3)) to upper bound the γ_α-functionals by the respective metric entropy integrals J_α. The monotone convergence theorem shows (A.2) for T in the limit $n \to \infty$. Applying Dirksen (2015, Lemma A.1) then gives the claim with $M = \exp(\alpha^{-1})\tilde{M}$. ■

We now use Lemma 31 to establish the following concentration bound, which is tailored towards the empirical processes appearing in our proofs, cf. Lemma 30. Note that this lemma can be viewed as a generalization of the key chaining Lemma 3.12 in Nickl and Wang (2024) to $\mathcal{Y}$-valued regression functions; the proof follows along the same lines.

Lemma 32 (Chaining Lemma) *Let $\mathcal{X}, \mathcal{Y}$ be separable Hilbert spaces, and suppose Θ is a (possibly uncountable) set parametrizing a class of maps*

$$\mathcal{H} = \{h_\theta : \mathcal{X} \to \mathcal{Y},\, \theta \in \Theta\}.$$

Consider an empirical process of the form

$$Z_n(\theta) = \frac{1}{n}\sum_{i=1}^n \langle h_\theta(x_i), \varepsilon_i\rangle_{\mathcal{Y}},$$

*where $x_1, \ldots, x_n \in \mathcal{X}$ are fixed elements and $\varepsilon_i, \ldots, \varepsilon_n$ are either **(i)** i.i.d. Gaussian white noise processes indexed by $\mathcal{Y}$, or **(ii)** i.i.d. sub-Gaussian random variables in $\mathcal{Y}$ with parameter 1.*

Recall the empirical seminorm $\|\cdot\|_n$. Suppose that

$$\sup_{\theta\in\Theta}\|h_\theta\|_n =: U < \infty, \tag{A.3}$$

and define the metric entropy integral

$$J(\mathcal{H}, d_n) = \int_0^U \sqrt{\log N(\mathcal{H}, d_n, \tau)}\, \mathrm{d}\tau, \quad d_n(\theta,\theta') = \|h_\theta - h_{\theta'}\|_n.$$

Let the space (Θ, d_n) be separable. Then $\sup_{\theta\in\Theta} Z_n(\theta)$ is measurable and there exists a universal constant $C_{\mathrm{Ch}} > 0$ such that for all $\delta > 0$ with

$$\sqrt{n}\delta \geq C_{\mathrm{Ch}} J(\mathcal{H}, d_n), \tag{A.4}$$

it holds

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |Z_n(\theta)| \geq \delta\right) \leq \exp\left(-\frac{8n\delta^2}{C_{\mathrm{Ch}}^2 U^2}\right).$$

Proof In both cases, we will apply Dirksen (2015, Theorem 3.2), which we stated in Lemma 31.

White noise case. Let $\theta, \theta' \in \Theta$ be arbitrary. Since ε_i, $i = 1, \ldots, n$ are independent white noise processes, we have $Z_n(\theta) - Z_n(\theta') \sim \mathcal{N}(0, n^{-1}\|h_\theta - h_{\theta'}\|_n^2)$, i.e. the increments of Z_n are normal (recall that $\boldsymbol{x}$ is regarded as fixed here). Thus,

$$\mathbb{P}\left(|Z_n(\theta) - Z_n(\theta')| \geq t\right) \leq 2\exp\left(-nt^2/(2\|h_\theta - h_{\theta'}\|_n^2)\right), \qquad t \geq 0,$$

and

$$\begin{aligned}\mathbb{P}\left(\left|Z_n(\theta) - Z_n(\theta')\right| \geq \frac{\sqrt{2}t d_n(\theta,\theta')}{\sqrt{n}}\right) &\leq 2\exp\left(-\frac{d_n(\theta,\theta')^2 t^2}{\|h_\theta - h_{\theta'}\|_n^2}\right) \\ &= 2\exp\left(-t^2\right), \qquad t \geq 0,\end{aligned} \tag{A.5}$$

which verifies the assumption (A.1) for $\alpha = 2$ and $\bar{d}_n := \sqrt{2}d_n/\sqrt{n}$.

Eq. (A.5) shows that the process $Z_n(\theta)$ is sub-Gaussian with respect to the pseudometric $\bar{d}_n(\theta,\theta') = \sqrt{2}\|h_\theta - h_{\theta'}\|_n/\sqrt{n}$. Therefore Giné and Nickl (2016, Theorem 2.3.7 (a)) yields that $Z_n(\theta)$ is sample bounded and uniformly sample continuous. Since (Θ, d_n) is separable, so is $(\Theta, \bar{d}_n)$. Thus it holds

$$\sup_{\theta\in\Theta} Z_n(\theta) = \sup_{\theta\in\Theta_0} Z_n(\theta) \quad \text{a.s.}, \tag{A.6}$$

where $\Theta_0 \subset \Theta$ denotes a countable, dense subset. The right hand side of (A.6) is measurable as a countable supremum. Therefore also the left hand side $\sup_{\theta\in\Theta} Z_n(\theta)$ is measurable. Applying Lemma 31 (to the countable set Θ_0) and using (A.6) gives that for some universal constant M and all $\theta_\dagger \in \Theta$,

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |Z_n(\theta) - Z_n(\theta_\dagger)| \geq M\left(J(\mathcal{H}, \bar{d}_n) + \frac{tU}{\sqrt{n}}\right)\right) \leq \exp\left(-\frac{t^2}{2}\right), \quad t \geq 1.$$

Due to (A.3) it holds $N(\mathcal{H}, d_n, \delta) = 1$ for all $\delta \geq U$, and thus

$$N(\mathcal{H}, \bar{d}_n, \tau) = N(\mathcal{H}, d_n, \sqrt{n}\tau/\sqrt{2}) = 1 \qquad \forall \tau \geq \frac{\sqrt{2}U}{\sqrt{n}}.$$

Substituting $\rho = \sqrt{n}\tau/\sqrt{2}$ we get

$$\begin{aligned} J(\mathcal{H}, \bar{d}_n) &= \int_0^{\sqrt{2}U/\sqrt{n}} \sqrt{\log N(\mathcal{H}, \bar{d}_n, \tau)}\, \mathrm{d}\tau \\ &= \frac{\sqrt{2}}{\sqrt{n}} \int_0^U \sqrt{\log N\Big(\mathcal{H}, \bar{d}_n, \frac{\sqrt{2}\rho}{\sqrt{n}}\Big)}\, \mathrm{d}\rho \\ &= \frac{\sqrt{2}}{\sqrt{n}} \int_0^U \sqrt{\log N(\mathcal{H}, d_n, \rho)}\, \mathrm{d}\rho = \frac{\sqrt{2} J(\mathcal{H}, d_n)}{\sqrt{n}} \end{aligned}$$

and therefore

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |Z_n(\theta) - Z_n(\theta_\dagger)| \geq \frac{\sqrt{2}M}{\sqrt{n}} \left(J(\mathcal{H}, d_n) + tU\right) \right) \leq \exp\left(-\frac{t^2}{2}\right), \quad t \geq 1. \tag{A.7}$$

Since $Z_n(\theta_\dagger) \sim \mathcal{N}(0, n^{-1}\|h_{\theta_\dagger}\|_n^2)$, it holds for all $\theta_\dagger \in \Theta$

$$\mathbb{P}\left(|Z_n(\theta_\dagger)| \geq \frac{MtU}{\sqrt{n}} \right) \leq \exp\left(\frac{-M^2U^2t^2}{2\|h_{\theta_\dagger}\|_n^2} \right) \leq \exp\left(-\frac{M^2t^2}{2}\right), \quad t \geq 0. \tag{A.8}$$

Combining (A.7) and (A.8) yields for $t \geq 1$

$$\begin{aligned} &\mathbb{P}\left(\sup_{\theta\in\Theta} |Z_n(\theta)| \geq \frac{3M}{\sqrt{n}} \left(J(\mathcal{H}, d_n) + tU\right) \right) \\ \leq\ &\mathbb{P}\left(\sup_{\theta\in\Theta} |Z_n(\theta) - Z_n(\theta_\dagger)| \geq \frac{\sqrt{2}M}{\sqrt{n}} \left(J(\mathcal{H}, d_n) + tU\right) \right) \\ &\qquad + \mathbb{P}\left(|Z_n(\theta_\dagger)| \geq \frac{M}{\sqrt{n}} \left(J(\mathcal{H}, d_n) + tU\right) \right) \\ \leq\ &\exp\left(\frac{-t^2}{2}\right) + \mathbb{P}\left(|Z_n(\theta_\dagger)| \geq \frac{MtU}{\sqrt{n}} \right) \\ \leq\ &\exp\left(\frac{-t^2}{2}\right) + \exp\left(\frac{-M^2t^2}{2}\right) \\ =\ &2\exp\left(\frac{-t^2}{2}\right), \end{aligned}$$

where we assumed without loss of generality that $M \geq 1$. Substitute $\delta = 3M/\sqrt{n}\,(J(\mathcal{H}, d_n) + tU)$, i.e. $t = (\sqrt{n}\delta/3M - J(\mathcal{H}, d_n))/U$. Because $N(\mathcal{H}, d_n, \tau) \geq 2$ for $\tau \leq U/2$, we have that $J(\mathcal{H}, d_n) \geq U/2\sqrt{\log(2)} > U/4$. Therefore $\sqrt{n}\delta \geq 15MJ(\mathcal{H}, d_n) \coloneqq C_{\mathrm{Ch}} J(\mathcal{H}, d_n)$ implies $t \geq 1$ and thus

$$\mathbb{P}\left(\sup_{\theta\in\Theta} |Z_n(\theta)| \geq \delta \right) \leq 2\exp\left(-\frac{\left(\sqrt{n}\delta/(3M) - J(\mathcal{H}, d_n)\right)^2}{2U^2} \right) \leq 2\exp\left(-\frac{8n\delta^2}{C_{\mathrm{Ch}}^2 U^2} \right), \tag{A.9}$$

which gives the claim for the white noise case.

Sub-Gaussian case. For $\theta, \theta' \in \Theta$ it holds

$$Z_n(\theta) - Z_n(\theta') = \frac{1}{n}\sum_{i=1}^{n} \langle \varepsilon_i, h_\theta(x_i) - h_{\theta'}(x_i)\rangle_{\mathcal{Y}}.$$

Since the centered RVs $n^{-1}\langle \varepsilon_i, h_\theta(x_i) - h_{\theta'}(x_i)\rangle_{\mathcal{Y}}$ are i.i.d. sub-Gaussian with parameter $n^{-1}\|h_\theta(x_i) - h_{\theta'}(x_i)\|_{\mathcal{Y}}$, the 'generalized' Hoeffding inequality for sub-Gaussian variables (see Vershynin 2018, Theorem 2.6.2) implies that for some universal constant $c > 0$ the increment $Z_n(\theta) - Z_n(\theta')$ is sub-Gaussian with parameter $c\|h_\theta - h_{\theta'}\|_n/\sqrt{n}$. Therefore

$$\mathbb{P}\left(\left| Z_n(\theta) - Z_n(\theta') \right| \geq \frac{\sqrt{2}tc}{\sqrt{n}} d_n(\theta, \theta') \right) \leq 2\exp\left(-t^2\right), \quad t \geq 0.$$

From here on, the proof is similar to the white noise case and we obtain (A.9) by absorbing c into C_{Ch}. ■

The following Lemma uses the 'peeling device' proof technique as developed in van de Geer (2000, Section 5) and is a generalization of van de Geer (2000, Lemma 5.13).

Lemma 33 *Let $x_i \overset{iid}{\sim} \gamma$ and consider the entropy $H(\delta) = H(\boldsymbol{G}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}, \delta)$. Let $(\tilde{\delta}_n)_{n\in\mathbb{N}}$ be a positive sequence with*

$$n\tilde{\delta}_n^2 \geq 6M_{\boldsymbol{F}}^2 H(\tilde{\delta}_n).$$

Then for $R \geq \max\{8\tilde{\delta}_n, 18M_{\boldsymbol{F}}/\sqrt{n}\}$, it holds that

$$\mathbb{P}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R,\ \|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq 2\|\hat{G}_n - G_0\|_n \right) \leq 2\exp\left(-\frac{nR^2}{320M_{\boldsymbol{F}}^2} \right).$$

Proof In the following we write $\|\cdot\|_\infty = \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$. Let $\boldsymbol{F} = \{G - G_0 : G \in \boldsymbol{G}\}$. Then for $R \geq 0$ it holds that

$$\begin{aligned} &\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R,\ \|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq 2\|\hat{G}_n - G_0\|_n \right) \\ &\leq \mathbb{P}_{G_0}\left(\sup_{F\in\boldsymbol{F},\, \|F\|_{L^2(\gamma)}\geq R} \|F\|_{L^2(\gamma)} \geq 2\|F\|_n \right). \end{aligned} \tag{A.10}$$

For $s \in \mathbb{N}$, we define $\boldsymbol{F}_s = \{F \in \boldsymbol{F} : sR \leq \|F\|_{L^2(\gamma)} \leq (s+1)R\}$. As a union of disjoint events, it holds

$$\begin{aligned} \mathbb{P}_{G_0}\left(\sup_{F\in\boldsymbol{F},\, \|F\|_{L^2(\gamma)}\geq R} \|F\|_{L^2(\gamma)} \geq 2\|F\|_n \right) &= \sum_{s=1}^{\infty} \mathbb{P}_{G_0}\left(\sup_{F\in\boldsymbol{F}_s} \|F\|_{L^2(\gamma)} \geq 2\|F\|_n \right) \\ &\leq \sum_{s=1}^{\infty} \mathbb{P}_{G_0}\left(\sup_{F\in\boldsymbol{F}_s} \left| \|F\|_{L^2(\gamma)} - \|F\|_n \right| \geq \frac{sR}{2} \right), \end{aligned}$$

where we used for $F \in \boldsymbol{F}_s$ and $s \in \mathbb{N}$

$$\|F\|_{L^2(\gamma)} \geq 2\|F\|_n \implies \|F\|_{L^2(\gamma)} - \|F\|_n \geq \frac{\|F\|_{L^2(\gamma)}}{2} \geq \frac{sR}{2}.$$

Now consider some covering $\boldsymbol{F}_s^* = \{F_{s,j}\}_{j=1}^N$ with $N = N(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8)$. Thus for arbitrary $s \in \mathbb{N}$ and $F \in \boldsymbol{F}_s$ there exists $F_{s,j} \in \boldsymbol{F}_s^*$ s.t. $\|F - F_{s,j}\|_\infty \leq R/8$. We get

$$\begin{aligned}\left|\|F\|_{L^2(\gamma)} - \|F\|_n\right| &\leq \left|\|F\|_{L^2(\gamma)} - \|F_{s,j}\|_{L^2(\gamma)}\right| + \left|\|F_{s,j}\|_{L^2(\gamma)} - \|F_{s,j}\|_n\right| \\ &\quad + \left|\|F_{s,j}\|_n - \|F\|_n\right| \\ &\leq \frac{R}{4} + \left|\|F_{s,j}\|_{L^2(\gamma)} - \|F_{s,j}\|_n\right|.\end{aligned}$$

Therefore

$$\begin{aligned}&\sum_{s=1}^\infty \mathbb{P}_{G_0}\left(\sup_{F\in\boldsymbol{F}_s}\left|\|F\|_{L^2(\gamma)} - \|F\|_n\right| \geq \frac{sR}{2}\right) \\ &\leq \sum_{s=1}^\infty \mathbb{P}_{G_0}\left(\max_{j=1,\dots,N}\left|\|F_{s,j}\|_{L^2(\gamma)} - \|F_{s,j}\|_n\right| \geq \frac{sR}{4}\right) \\ &\leq \sum_{s=1}^\infty N(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) \max_{j=1,\dots,N} \mathbb{P}_{G_0}\left(\left|\|F_{s,j}\|_{L^2(\gamma)} - \|F_{s,j}\|_n\right| \geq \frac{sR}{4}\right), \qquad \text{(A.11)}\end{aligned}$$

where we used $sR/2 - R/4 \geq sR/4$ for $s \in \mathbb{N}$. Since $F_{s,j} \in \boldsymbol{F}_s$, we have $\|F_{s,j}\|_{L^2(\gamma)} \geq sR$ and therefore

$$\begin{aligned}\left|\|F_{s,j}\|_{L^2(\gamma)} - \|F_{s,j}\|_n\right| &= \frac{\left|\|F_{s,j}\|_{L^2(\gamma)}^2 - \|F_{s,j}\|_n^2\right|}{\|F_{s,j}\|_{L^2(\gamma)} + \|F_{s,j}\|_n} \\ &\leq \frac{1}{sR}\left|\|F_{s,j}\|_{L^2(\gamma)}^2 - \|F_{s,j}\|_n^2\right|.\end{aligned}$$

Inserting into (A.11) gives

$$\begin{aligned}&\sum_{s=1}^\infty N(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) \max_{j=1,\dots,N} \mathbb{P}_{G_0}\left(\left|\|F_{s,j}\|_{L^2(\gamma)} - \|F_{s,j}\|_n\right| \geq \frac{sR}{4}\right) \\ &\leq \sum_{s=1}^\infty N(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) \max_{j=1,\dots,N} \mathbb{P}_{G_0}\left(\left|\|F_{s,j}\|_{L^2(\gamma)}^2 - \|F_{s,j}\|_n^2\right| \geq \frac{s^2R^2}{4}\right).\end{aligned}$$

Define the variables

$$Y_i = \frac{1}{n}\left(\|F_{s,j}\|_{L^2(\gamma)}^2 - \|F_{s,j}(x_i)\|_y^2\right), \quad i = 1, \dots, n.$$

It holds for all $i = 1, \dots, n$

$$\begin{aligned}
\mathbb{E}[Y_i] &= 0, \\
|Y_i| &\leq \frac{2M_{\boldsymbol{F}}^2}{n}, \\
\mathbb{E}[Y_i^2] &= \frac{1}{n^2}\mathbb{E}\left[\left(\|F_{s,j}\|_{L^2(\gamma)}^2 - \|F_{s,j}(x_i)\|_y^2\right)^2\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\left(\|F_{s,j}\|_{L^2(\gamma)}^4 - 2\|F_{s,j}\|_{L^2(\gamma)}^2\|F_{s,j}(x_i)\|_y^2 + \|F_{s,j}(x_i)\|_y^4\right)\right] \\
&= \frac{1}{n^2}\mathbb{E}\left[\|F_{s,j}(x_i)\|_y^4\right] - \|F_{s,j}\|_{L^2(\gamma)}^4 \\
&\leq \frac{M_{\boldsymbol{F}}^2}{n^2}\|F_{s,j}\|_{L^2(\gamma)}^2.
\end{aligned}$$

Applying Bernstein's inequality (Lemma 29) for the variables Y_i yields

$$\begin{aligned}
&\sum_{s=1}^{\infty} N(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) \max_{j=1,\dots,N} \mathbb{P}_{G_0}\left(\left|\|F_{s,j}\|_{L^2(\gamma)}^2 - \|F_{s,j}\|_n^2\right| \geq \frac{s^2R^2}{4}\right) \\
\leq &\sum_{s=1}^{\infty} N(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) \max_{j=1,\dots,N} \exp\left(-\frac{ns^4R^4}{32M_{\boldsymbol{F}}^2(\|F_{s,j}\|_{L^2(\gamma)}^2 + \frac{s^2R^2}{6})}\right) \\
\leq &\sum_{s=1}^{\infty} \exp\left(H(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) - \frac{ns^2R^2}{160M_{\boldsymbol{F}}^2}\right),
\end{aligned}$$

where we used $\|F_{s,j}\|_{L^2(\gamma)}^2 \leq (s+1)^2R^2 \leq 4s^2R^2$ for all $j = 1, \dots, N$ and $s \in \mathbb{N}$, since $F_{j,s} \in \boldsymbol{F}_s$.

Since $H(\delta)/\delta^2$ is non-increasing in δ and $R \geq 8\tilde{\delta}_n$, we have

$$\frac{ns^2R^2}{160M_{\boldsymbol{F}}^2} \geq \frac{n}{3M_{\boldsymbol{F}}^2}\left(\frac{R}{8}\right)^2 \geq 2H(\boldsymbol{G}, \|\cdot\|_\infty, R/8) \geq 2H(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8).$$

Therefore we get

$$\begin{aligned}
&\sum_{s=1}^{\infty} \exp\left(H(\boldsymbol{F}_s, \|\cdot\|_\infty, R/8) - \frac{ns^2R^2}{160M_{\boldsymbol{F}}^2}\right) \\
\leq &\sum_{s=1}^{\infty} \exp\left(-\frac{ns^2R^2}{320M_{\boldsymbol{F}}^2}\right) \leq \sum_{s=1}^{\infty} \frac{1}{s^2}\exp\left(-\frac{nR^2}{320M_{\boldsymbol{F}}^2}\right) < 2\exp\left(-\frac{nR^2}{320M_{\boldsymbol{F}}^2}\right). \qquad \text{(A.12)}
\end{aligned}$$

Note that for $x, y \geq 1$, we have $\exp(-xy) \leq \exp(-y)/x$. Applying this in the case $x = s^2 \geq 1$ and $y = nR^2/(320M_{\boldsymbol{F}}^2) \geq 1$ for $R \geq 18M_{\boldsymbol{F}}/\sqrt{n}$, together with $\sum_{s=1}^{\infty} 1/s^2 = \pi^2/6 < 2$, gives (A.12). Combining (A.10)–(A.12) shows the claim. ∎

In the case of sub-Gaussian noise, standard bounds on the maximum of sub-Gaussian RVs suffice to bound the generalisation error.

Lemma 34 *Consider the sub-Gaussian noise model, i.e. suppose $\|\varepsilon_i\|_{\mathcal{Y}}$ are i.i.d. sub-Gaussian with parameter 1. Abbreviate $H(\delta) = H(\boldsymbol{G}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}, \delta)$. Then there exists a universal constant $C > 0$ s.t. for all positive sequences $(\delta_n)_{n\in\mathbb{N}}$ with*

$$n\delta_n^4 \geq C^2\sigma^2 M_{\boldsymbol{F}}^2 H\left(\frac{\delta_n^2}{8\sigma^2+\delta_n^2}\right),$$

all $G^* \in \boldsymbol{G}$ and $R \geq \max\{\delta_n, \sqrt{2}\|G^* - G_0\|_n\}$, it holds for every $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n$

$$\mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\|\hat{G}_n - G_0\|_n \geq R\right) \leq 4\exp\left(-\frac{nR^4}{C^2\sigma^2(1+M_{\boldsymbol{F}}^2)}\right). \tag{A.13}$$

Proof In the following we write $\|\cdot\|_\infty = \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}$. For $R^2 \geq 2\|G^* - G_0\|_n^2$, we use the basic inequality (Lemma 30) to obtain

$$\begin{aligned}\mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\|\hat{G}_n - G_0\|_n^2 \geq R^2\right) &\leq \mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2 \geq \frac{R^2}{2}\right). \\ &\leq \mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, \hat{G}_n(x_i) - G^*(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{4\sigma}\right). \end{aligned} \tag{A.14}$$

For $R > 0$, let $\boldsymbol{G}^* = \{G - G^*,\ G \in \boldsymbol{G}\}$ and $(G_j)_{j=1}^N$ with $N = N(\boldsymbol{G}^*, \|\cdot\|_\infty, R^2/(8\sigma\mathbb{E}[\|\varepsilon_i\|_{\mathcal{Y}}] + R^2))$ denote a minimal $\|\cdot\|_\infty$-cover of $\boldsymbol{G}^*$.

It holds for $R > 0$

$$\begin{aligned}&\mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, \hat{G}_n(x_i) - G^*(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{4\sigma}\right) \\ &\leq \mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\sup_{G\in\boldsymbol{G}}\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, G(x_i) - G^*(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{4\sigma}\right) \\ &\leq \mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\sup_{G\in\boldsymbol{G}}\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, G(x_i) - G^*(x_i) - G_{j^*}(x_i)\rangle_{\mathcal{Y}}\right| + \max_{j=1,...N}\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, G_j(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{4\sigma}\right) \\ &\leq \underbrace{\mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\sup_{G\in\boldsymbol{G}}\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, G(x_i) - G^*(x_i) - G_{j^*}(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{8\sigma}\right)}_{(i)} \\ &\quad + \underbrace{\mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\max_{j=1,...N}\left|\frac{1}{n}\sum_{i=1}^n \langle \varepsilon_i, G_j(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{8\sigma}\right)}_{(ii)}.\end{aligned}$$

We estimate the terms (i) and (ii) separately. For (i), use $\|G-G^*-G_{j^*}\|_\infty \leq R^2/(8\sigma\mathbb{E}[\|\varepsilon_i\|_{\mathcal{Y}}] + R^2)$ for all $G \in \boldsymbol{G}$ and estimate

$$\begin{aligned}
&\mathbb{P}^{\boldsymbol{x}}_{G_0}\left(\sup_{G\in\boldsymbol{G}}\left|\frac{1}{n}\sum_{i=1}^n\langle\varepsilon_i, G(x_i)-G^*(x_i)-G_{j^*}(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{8\sigma}\right)\\
&\leq \mathbb{P}^{\boldsymbol{x}}_{G_0}\left(\left|\frac{1}{n}\sum_{i=1}^n\|\varepsilon_i\|_{\mathcal{Y}}\frac{R^2}{8\sigma\mathbb{E}[\|\varepsilon_i\|_{\mathcal{Y}}]+R^2}\right| \geq \frac{R^2}{8\sigma}\right)\\
&\leq \mathbb{P}^{\boldsymbol{x}}_{G_0}\left(\left|\frac{1}{n}\sum_{i=1}^n\|\varepsilon_i\|_{\mathcal{Y}}-\mathbb{E}[\|\varepsilon_i\|_{\mathcal{Y}}]\right| \geq \frac{R^2}{8\sigma}\right) \leq 2\exp\left(-\frac{nR^4}{C^2\sigma^2}\right),
\end{aligned}$$

where we used the Hoeffding inequality for sub-Gaussian random variables from Vershynin (2018, Theorem 2.6.2).

For (ii), it holds

$$\begin{aligned}
&\mathbb{P}^{\boldsymbol{x}}_{G_0}\left(\max_{j=1,\dots N}\left|\frac{1}{n}\sum_{i=1}^n\langle\varepsilon_i, G_j(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{8\sigma}\right)\\
&\leq N\max_{j=1,\dots N}\mathbb{P}^{\boldsymbol{x}}_{G_0}\left(\left|\frac{1}{n}\sum_{i=1}^n\langle\varepsilon_i, G_j(x_i)\rangle_{\mathcal{Y}}\right| \geq \frac{R^2}{8\sigma}\right)\\
&\leq 2N\exp\left(-\frac{2nR^4}{C^2\sigma^2M_{\boldsymbol{F}}^2}\right) \leq 2\exp\left(H\left(\frac{R^2}{8\sigma^2+R^2}\right)-\frac{2nR^4}{C^2\sigma^2M_{\boldsymbol{F}}^2}\right),
\end{aligned}$$

where we used $\mathbb{E}[\|\varepsilon_i\|_{\mathcal{Y}}] \leq \sigma$ (Cauchy–Schwarz) and Vershynin (2018, Theorem 2.6.2) for $Y_i = n^{-1}\langle\varepsilon_i, G_j(x_i)\rangle_{\mathcal{Y}}$, $i = 1,\dots,n$.

Since $H(\delta)/\delta^2$ is non-increasing, also $H(\delta/(8\sigma^2+\delta))/\delta^2$ is non-increasing and thus it holds for $R \geq \delta_n$

$$\frac{nR^4}{C^2\sigma^2M_{\boldsymbol{F}}^2} \geq H\left(\frac{R^2}{8\sigma^2+R^2}\right).$$

This gives for $R \geq \delta_n$

$$(ii) \leq 2\exp\left(-\frac{nR^4}{C^2\sigma^2M_{\boldsymbol{F}}^2}\right). \tag{A.15}$$

Combining (A.14)–(A.15) gives the result. ∎

Appendix B. Proofs of Section 2

B.1 Proof of Theorem 5

B.1.1 Existence and Measurability of $\hat{G}_n$

White noise case. For $i = 1,\dots,n$, denote the probability space of the RVs x_i and the white noise processes ε_i as $(\Omega, \Sigma, \mathbb{P})$. Furthermore, equip the Hilbert spaces $\mathcal{X}$ and $\mathcal{Y}$ with

Borel σ-algebras $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\mathcal{Y}$ and the space $\boldsymbol{G}$ with the Borel σ-algebra $\mathcal{B}_{\boldsymbol{G}}$. Then, there exists an orthonormal basis $(\psi_j)_{j\in\mathbb{N}}$ of $\mathcal{Y}$ and i.i.d. Gaussian variables $Z_j \sim \mathcal{N}(0,1)$ s.t.

$$\begin{aligned}\varepsilon_i &: \Omega \times \mathcal{Y} \to \mathbb{R},\\ \varepsilon_i(\omega, y) &= \sum_{j=1}^{\infty} \langle y, \psi_j\rangle_\mathcal{Y} Z_j(\omega)\end{aligned}$$

for $i = 1, \dots, n$, see also Giné and Nickl (2016, Example 2.1.11)

Recall the noise level $\sigma > 0$. For $i = 1, \dots, n$, we define

$$\begin{aligned}u_i &: \Omega \times \boldsymbol{G} \to \mathbb{R},\\ u_i(\omega, G) &= 2\sigma\varepsilon_i\left(\omega, G(x_i(\omega))\right) + 2\langle G_0(x_i(\omega)), G(x_i(\omega))\rangle_\mathcal{Y} - \|G(x_i(\omega))\|_\mathcal{Y}^2. \qquad \text{(B.1)}\end{aligned}$$

We aim to apply Nickl (2007, Proposition 5) to $u \coloneqq 1/n \sum_{i=1}^n u_i$ in order to get existence and measurability of $\hat{G}_n$ in (2.3).

Per assumption, the metric space $(\boldsymbol{G}, \|\cdot\|)$ is compact. We show that u_i from (B.1) is measurable in the first component and continuous in the second component for all $i = 1, \dots, n$. Then Nickl (2007, Proposition 5) shows the claim. Consider an arbitrary $G \in \boldsymbol{G}$. We show that $u_i(\,.\,, G)$ is $(\Sigma, \mathcal{B}_\mathbb{R})$-measurable, where $\mathcal{B}_\mathbb{R}$ is the Borel σ-algebra in $\mathbb{R}$.

The RVs x_i are $(\Sigma, \mathcal{B}_\mathcal{X})$-measurable by definition. The maps G and G_0 are assumed to be $(\mathcal{B}_\mathcal{X}, \mathcal{B}_\mathcal{Y})$-measurable. Furthermore, because of their continuity, the scalar product $\langle\,.\,,.\,\rangle_\mathcal{Y}$ is $(\mathcal{B}_\mathcal{Y}, \mathcal{B}_\mathbb{R})$-measurable in both components and also the norm $\|\,.\,\|_\mathcal{Y}$ is $(\mathcal{B}_\mathcal{Y}, \mathcal{B}_\mathbb{R})$-measurable. Therefore, since the composition of measurable functions is measurable, the latter two summands in (B.1) are $(\Sigma, \mathcal{B}_\mathbb{R})$-measurable.

Proceeding with the first summand, the RVs Z_j are $(\Sigma, \mathcal{B}_\mathbb{R})$-measurable by definition for all $j \in \mathbb{N}$. Therefore the products $\langle\,.\,, \psi_j\rangle_\mathcal{Y} Z_j$ are $(\Sigma \otimes \mathcal{B}_\mathcal{Y}, \mathcal{B}_\mathbb{R})$-measurable for all $j \in \mathbb{N}$. Thus ε_i is, as the pointwise limit, $(\Sigma \otimes \mathcal{B}_\mathcal{Y}, \mathcal{B}_\mathbb{R})$-measurable. Then, as the composition of measurable functions, the first summand in (B.1) is $(\Sigma, \mathcal{B}_\mathbb{R})$-measurable. Therefore $u_i(\,.\,, G)$, $i = 1, \dots, n$, and thus $u(\,.\,, G)$ is $(\Sigma, \mathcal{B}_\mathbb{R})$-measurable for all $G \in \boldsymbol{G}$.

We proceed and show that $u(\omega, \,.\,)$ is continuous w.r.t. $\|\cdot\|$. Therefore choose $G, G' \in \boldsymbol{G}$ and $\omega \in \Omega$. Then it holds for $i = 1, \dots, n$ and $x_i = x_i(\omega)$

$$\begin{aligned}\left|u_i(\omega, G) - u_i(\omega, G')\right| &\le 2\sigma\left|\varepsilon_i\left(\omega, G(x_i) - G'(x_i)\right)\right| + 2\left|\langle G_0(x_i), G(x_i) - G'(x_i)\rangle_\mathcal{Y}\right|\\ &\quad + \left|\|G(x_i)\|_\mathcal{Y} - \|G'(x_i)\|_\mathcal{Y}\right| \left|\|G(x_i)\|_\mathcal{Y} + \|G'(x_i)\|_\mathcal{Y}\right|\\ &\le 2\sigma\left|\varepsilon_i\left(\omega, G(x_i) - G'(x_i)\right)\right|\\ &\quad + \sqrt{n}\left(2\|G_0(x_i)\|_\mathcal{Y} + \|G(x_i)\|_\mathcal{Y} + \|G'(x_i)\|_\mathcal{Y}\right) \|G - G'\|_n\\ &\le 2\sigma\left|\varepsilon_i\left(\omega, G(x_i) - G'(x_i)\right)\right|\\ &\quad + C\sqrt{n}\left(2\|G_0(x_i)\|_\mathcal{Y} + \|G(x_i)\|_\mathcal{Y} + \|G'(x_i)\|_\mathcal{Y}\right) \|G - G'\|, \qquad \text{(B.2)}\end{aligned}$$

where we used that $\|G(x_i) - G'(x_i)\|_\mathcal{Y} \le \sqrt{n}\|G - G'\|_n$ for $i = 1, \dots, n$ and $\|\cdot\|_n \le C\|\cdot\|$ at the last inequality. Furthermore, Giné and Nickl (2016, page 40 and Proposition 2.3.7(a)) yields that the white noise processes ε_i are a.s. sample d_{ε_i}-continuous w.r.t. their intrinsic

pseudometrics $d_{\varepsilon_i} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, $d_{\varepsilon_i}(y, y') = \mathbb{E}[\langle \varepsilon_i, y - y' \rangle_{\mathcal{Y}}^2]^{1/2} = \|y - y'\|_{\mathcal{Y}}$. Since for all G, $G' \in \boldsymbol{G}$ we have

$$\|G(x_i) - G'(x_i)\|_{\mathcal{Y}} \leq \sqrt{n}\|G - G'\|_n \leq C\sqrt{n}\|G - G'\|,$$

the white noise processes are also a.s. sample d-continuous, where d is the metric induced by $\|\cdot\|$.

Together with (B.2) this shows that there exists a null-set $\Omega_0 \subset \Omega$ such that for all $\omega \in \Omega \backslash \Omega_0$, $u(\omega, .)$ is continuous w.r.t. $\|\cdot\|$. Now we choose versions $\tilde{x}_i$ and $\tilde{\varepsilon}_i$ s.t. $u(\omega, .) = 0$ for all $\omega \in \Omega_0$. Then $G \mapsto u(\omega, G) : (\boldsymbol{G}, \|\cdot\|) \to \mathbb{R}$ is continuous for all $\omega \in \Omega$. Applying Nickl (2007, Proposition 5) gives an $(\Sigma, \mathcal{B}_{\boldsymbol{G}})$-measurable MLSE $\hat{G}_n$ in (2.3) with the desired minimization property.

Sub-Gaussian case. For $i = 1, \ldots, n$, consider the functions u_i from (B.1). The measurability of $u_i(\,.\,, G)$, $i = 1, \ldots, n$ follows from the measurability of the scalar product $\langle .\,, .\rangle_{\mathcal{Y}}$, the norm $\|\cdot\|_{\mathcal{Y}}$, G_0 and all $G \in \boldsymbol{G}$. Note that in contrast to the white noise case, ε_i are RVs in $\mathcal{Y}$ for all $i = 1, \ldots, n$ and therefore measurable without any further investigation. Also, since $\varepsilon_i(\omega) \in \mathcal{Y}$ for all ω, Cauchy–Schwarz immediately shows that $u_i(\omega, .)$ and therefore $u(\omega, .)$ is continuous w.r.t. $\|\cdot\|$ for all $\omega \in \Omega \backslash \Omega^0$. Choosing versions $\tilde{x}_i$ and $\tilde{\varepsilon}_i$ and applying Nickl (2007, Proposition 5) as above gives the existence of an $(\Sigma, \mathcal{B}_{\boldsymbol{G}})$-measurable LSE $\hat{G}_n$ in (2.3) and therefore finishes the proof.

B.1.2 Concentration Inequality for $\hat{G}_n$

The proof follows ideas developed in van de Geer (2000, Section 10.3) as well as Nickl and Wang (2024), where generic chaining bounds from Dirksen (2015) were used to bound the relevant empirical processes appearing below.

Slicing argument. Recall the definition (2.4) of the empirical norm. For $R_n^2 \geq 2\|G^* - G_0\|_n^2$, we have

$$\mathbb{P}_{G_0}^{\boldsymbol{x}}\big(\|\hat{G}_n - G_0\|_n^2 \geq R_n^2\big) \leq \mathbb{P}_{G_0}^{\boldsymbol{x}}\big(2\big(\|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2\big) \geq R_n^2\big). \tag{B.3}$$

As a union of disjoint events, it holds

$$\begin{aligned} &\mathbb{P}_{G_0}^{\boldsymbol{x}}\big(2\big(\|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2\big) \geq R_n^2\big) \\ &= \sum_{s=0}^{\infty} \mathbb{P}_{G_0}^{\boldsymbol{x}}\Big(2^{2s}R_n^2 \leq 2\big(\|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2\big) < 2^{2s+2}R_n^2\Big). \end{aligned}$$

Applying the basic inequality (Lemma 30) and defining the empirical process $(X_G : G \in \boldsymbol{G})$ indexed by the operator class $\boldsymbol{G}$ as

$$X_G = \frac{1}{n}\sum_{i=1}^{n} \langle \varepsilon_i, G(x_i) - G^*(x_i)\rangle_{\mathcal{Y}},$$

gives

$$\begin{aligned}
&\sum_{s=0}^{\infty} \mathbb{P}_{G_0}^{\boldsymbol{x}}\Big(2^{2s}R_n^2 \le 2\big(\|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2\big) < 2^{2s+2}R_n^2\Big) \\
&\le \sum_{s=0}^{\infty} \mathbb{P}_{G_0}^{\boldsymbol{x}}\Big(\big|X_{\hat{G}_n}\big| \ge \frac{2^{2s-2}R_n^2}{\sigma},\ \|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2 < 2^{2s+1}R_n^2\Big) \\
&\le \sum_{s=0}^{\infty} \mathbb{P}_{G_0}^{\boldsymbol{x}}\Big(\sup_{G\in \boldsymbol{G}_n^*(2^{s+3/2}R_n)} \big|X_G\big| \ge \frac{2^{2s-2}R_n^2}{\sigma}\Big). \qquad \text{(B.4)}
\end{aligned}$$

In (B.4), we additionally used that if $2\|G^* - G_0\|_n^2 \le R_n^2$ and

$$\big|\|\hat{G}_n - G_0\|_n^2 - \|G^* - G_0\|_n^2\big| < 2^{2s+1}R_n^2$$

then $\|\hat{G}_n - G_0\|_n^2 \le 2^{2s+1}R_n^2 + R_n^2/2$ and thus

$$\|\hat{G}_n - G^*\|_n^2 \le 2\big(\|\hat{G}_n - G_0\|_n^2 + \|G^* - G_0\|_n^2\big) \le 2^{2s+2}R_n^2 + 2R_n^2 \le 2^{2s+3}R_n^2.$$

Concentration inequality for each slice. We wish to apply Lemma 32 to bound the probabilities in (B.4). Let C_{Ch} be the generic constant from this lemma and let δ_n satisfy (2.6). Then, due to $\delta \mapsto \Psi_n(\delta)/\delta^2$ being non-increasing, (2.6) gives for all $R_n \ge \delta_n$ and $s \in \mathbb{N}_0$

$$\sqrt{n}(2^{2s+3}R_n^2) \ge 32C_{\mathrm{Ch}}\sigma\Psi_n(2^{s+3/2}R_n)$$

so that with $\Theta := \boldsymbol{G}_n^*(2^{s+3/2}R_n)$

$$\sqrt{n}\frac{2^{2s-2}R_n^2}{\sigma} \ge C_{\mathrm{Ch}}\Psi_n(2^{s+3/2}R_n) \ge C_{\mathrm{Ch}}J(\Theta, \|\cdot\|_n). \qquad \text{(B.5)}$$

With $h_G := G - G^*$ and $\mathcal{H} := \{h_G : G \in \Theta\}$ we have $J(\Theta, \|\cdot\|_n) = J(\mathcal{H}, \|\cdot\|_n)$ and thus (B.5) verifies assumption (A.4) of Lemma 32 for $\delta = 2^{2s-2}R_n^2/\sigma$.

Furthermore, since $\Theta = \boldsymbol{G}_n^*(2^{s+3/2}R_n) \subset \boldsymbol{G}$ for $s \in \mathbb{N}_0$, $R_n > 0$ and $(\boldsymbol{G}, \|\cdot\|_n)$ is compact, the space $(\Theta, \|\cdot\|_n)$ is separable, which verifies the last assumption of Lemma 32. Applying this lemma with $U = 2^{s+3/2}R_n$ shows that the (uncountable) suprema in (B.4) are measurable and that for all $R_n \ge \delta_n$,

$$\begin{aligned}
\sum_{s=0}^{\infty} \mathbb{P}_{G_0}^{\boldsymbol{x}}\Big(\sup_{G\in \boldsymbol{G}_n^*(2^{s+3/2}R_n)} \big|X_G\big| \ge \frac{2^{2s-2}R_n^2}{\sigma}\Big) &\le \sum_{s=0}^{\infty} \exp\Big(-\frac{8n2^{4s-4}R_n^4}{C_{\mathrm{Ch}}^2\sigma^2 2^{2s+3}R_n^2}\Big) \\
&\le \sum_{s=0}^{\infty} \exp\Big(-\frac{2^{2s-4}nR_n^2}{C_{\mathrm{Ch}}^2\sigma^2}\Big) \\
&\le \sum_{s=0}^{\infty} 2^{-2s}\exp\Big(-\frac{nR_n^2}{16C_{\mathrm{Ch}}^2\sigma^2}\Big) \\
&< 2\exp\Big(-\frac{nR_n^2}{16C_{\mathrm{Ch}}^2\sigma^2}\Big). \qquad \text{(B.6)}
\end{aligned}$$

In (B.6) we additionally used $\exp(-xy) \le \exp(-x)/y$, which holds for all $x, y \ge 1$, i.e. for $R_n \ge 4C_{\mathrm{Ch}}\sigma/\sqrt{n}$. Combining (B.3)–(B.4) and (B.6) gives (2.7) and therefore shows the claim.

B.2 Proof of Theorem 6

We use the concentration inequality from Theorem 5 for the empirical error and combine it with a key concentration result for the empirical norm around the $L^2(\gamma)$-norm, proved in Lemma 33. For any $R > 0$,

$$\begin{aligned}\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R\right) \leq \mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_n \geq \frac{R}{2}\right) \\ + \mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R,\ \|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq 2\|\hat{G}_n - G_0\|_n\right).\end{aligned} \tag{B.7}$$

Now suppose that R_n satisfies

$$R_n \geq 2\max\left\{\delta_n, 4\tilde{\delta}_n, \sqrt{2}\|G^* - G_0\|_{\infty,\mathrm{supp}(\gamma)}, \frac{4C_{ch}\sigma}{\sqrt{n}}, \frac{9M_{\boldsymbol{F}}}{\sqrt{n}}\right\}.$$

This is implied by (2.11) for an appropriate choice of C. In particular, $R_n \geq 2\sqrt{2}\|G^* - G_0\|_n$ for all $\boldsymbol{x} \in \mathcal{X}^n$. Since Theorem 5 holds for γ^n-almost every $\boldsymbol{x} \in \mathcal{X}^n$, taking expectations over $\boldsymbol{x} \sim \gamma^n$ gives

$$\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_n \geq \frac{R_n}{2}\right) \leq 2\exp\left(-\frac{nR_n^2}{64C_{\mathrm{Ch}}^2\sigma^2}\right). \tag{B.8}$$

Furthermore, Lemma 33 gives

$$\begin{aligned}&\mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq R_n,\ \|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq 2\|\hat{G}_n - G_0\|_n\right) \\ \leq\ & 2\exp\left(-\frac{nR_n^2}{320M_{\boldsymbol{F}}^2}\right),\end{aligned} \tag{B.9}$$

since we have assumed $R_n \geq \max\{8\tilde{\delta}_n, 18M_{\boldsymbol{F}}/\sqrt{n}\}$. Combining (B.8) and (B.9) shows (2.12) for some C.

B.3 Proof of Corollary 7

It remains to show (2.13). We use (B.7) to estimate

$$\begin{aligned}&\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2\right] \\ =\ & \int_0^\infty \mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq \sqrt{R_n}\right) \mathrm{d}R_n \\ \leq\ & \int_0^\infty \int_{\mathcal{X}^n} \mathbb{P}_{G_0}^{\boldsymbol{x}}\left(\|\hat{G}_n - G_0\|_n \geq \frac{\sqrt{R_n}}{2}\right) \mathrm{d}\gamma(\boldsymbol{x})\, \mathrm{d}R_n \\ &+ \int_0^\infty \mathbb{P}_{G_0}\left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq \sqrt{R_n},\ \|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq 2\|\hat{G}_n - G_0\|_n\right) \mathrm{d}R_n.\end{aligned} \tag{B.10}$$

For $G^* \in \boldsymbol{G}$ set $R_{n,1} = 2\max\{\delta_n, \sqrt{2}\|G^* - G_0\|_n, \frac{4C_{\mathrm{Ch}}\sigma}{\sqrt{n}}\}$, where C_{Ch} is from Lemma 32. Then Theorem 5 gives

$$\begin{aligned}
&\int_0^\infty \int_{\mathcal{X}^n} \mathbb{P}_{G_0}^{\boldsymbol{x}} \left(\|\hat{G}_n - G_0\|_n \geq \frac{\sqrt{R_n}}{2} \right) \mathrm{d}\gamma(x)\, \mathrm{d}R_n \\
&\leq \int_{\mathcal{X}^n} R_{n,1}^2 \, \mathrm{d}\gamma(\boldsymbol{x}) + 2\int_{R_{n,1}^2}^\infty \exp\left(-\frac{nR_n}{64C_{\mathrm{Ch}}^2\sigma^2}\right) \mathrm{d}R_n \\
&\leq 4\delta_n^2 + 8\|G^* - G_0\|_{L^2(\gamma)}^2 + \frac{16C_{\mathrm{Ch}}^2\sigma^2}{n} + 2\int_{R_{n,1}^2}^\infty \exp\left(-\frac{nR_n}{64C_{\mathrm{Ch}}^2\sigma^2}\right) \mathrm{d}R_n \\
&\leq 4\delta_n^2 + 8\|G^* - G_0\|_{L^2(\gamma)}^2 + \frac{80C_{\mathrm{Ch}}^2\sigma^2}{n}.
\end{aligned}$$

Lemma 33 gives with $R_{n,2} = \max\{8\tilde{\delta}_n, 18M_{\boldsymbol{F}}/\sqrt{n}\}$

$$\begin{aligned}
&\int_0^\infty \mathbb{P}_{G_0} \left(\|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq \sqrt{R_n},\ \|\hat{G}_n - G_0\|_{L^2(\gamma)} \geq 2\|\hat{G}_n - G_0\|_n \right) \mathrm{d}R_n \\
&\leq R_{n,2}^2 + 2\int_{R_{n,1}^2}^\infty \exp\left(-\frac{nR_n}{320M_{\boldsymbol{F}}^2}\right) \mathrm{d}R_n \\
&\leq 64\tilde{\delta}_n^2 + \frac{964M_{\boldsymbol{F}}^2}{n}. \qquad \text{(B.11)}
\end{aligned}$$

Combining (B.10)–(B.11) and taking an infimum over $G^* \in \boldsymbol{G}$ shows (2.13) for some C and thus finishes the proof of Corollary 7.

B.4 Proof of Corollary 8

We construct sequences $(\delta_n)_{n\in\mathbb{N}}$ and $(\tilde{\delta}_n)_{n\in\mathbb{N}}$ satisfying (2.10) and balance the approximation term and the entropy terms in (2.12) by choosing $N(n)$ appropriately.

Proof of (i): Choosing $\tilde{\delta}_n^2 \simeq n^{-2/(2+\alpha)}$ immediately shows the second part of (2.10). For $J(\delta)$ from (2.5) it holds

$$J(\delta) \leq \int_0^\delta \sqrt{H(\rho, N)}\, \mathrm{d}\rho \lesssim \delta^{1-\alpha/2} =: \Psi_n(\delta).$$

Hence $\delta_n^2 \simeq n^{-2/(2+\alpha)}$ satisfies the first part of (2.10). Therefore Corollary 7 shows

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2\right] \lesssim N^{-\beta} + n^{-\frac{2}{2+\alpha}}.$$

Choosing $N = \lceil n^{\frac{2}{(2+\alpha)\beta}} \rceil$ gives the claim.

Proof of (ii): Choosing $\tilde{\delta}_n^2 \simeq N(1+\log(n))/n$ shows in particular $\log(n) \gtrsim \log(\tilde{\delta}_n^{-1})$, which in turn shows that $\tilde{\delta}_n$ satisfies the second part of (2.10). For $J(\delta)$ from (2.5) it holds

$$J(\delta) \lesssim \int_0^\delta \sqrt{H(\rho, N)}\, \mathrm{d}\rho \lesssim \sqrt{N}\delta\left(1 + \log(\delta^{-1})\right) =: \Psi_n(\delta).$$

Hence $\delta_n^2 \simeq N(1+\log(n))/n$ satisfies the first part of (2.10). Therefore Corollary 7 shows

$$\mathbb{E}_{G_0}\left[\|\hat{G}_n - G_0\|_{L^2(\gamma)}^2\right] \lesssim N^{-\beta} + \frac{N(1+\log(n))}{n}.$$

Choosing $N(n) = \lceil n^{\frac{1}{\beta+1}} \rceil$ for all $n \geq 1$ gives the claim.

B.5 Proof of Theorem 10

The concentration inequality (2.15) follows immediately from (B.7), (B.9) (which holds for sub-Gaussian noise) and (A.13) from Lemma 34. Application of Lemma 34 yields

$$\begin{aligned}
&\int_0^\infty \int_{\mathcal{X}^n} \mathbb{P}_{G_0}^{\boldsymbol{x}} \left(\|\hat{G}_n - G_0\|_n \geq \frac{\sqrt{\delta}}{2} \right) \, \mathrm{d}\gamma(\boldsymbol{x}) \, \mathrm{d}\delta \\
&\leq 4\delta_n^2 + 8\|G^* - G_0\|_{L^2(\gamma)}^2 + 4 \int_0^\infty \exp\left(-\frac{n\delta^2}{16 C_1^2 \sigma^2 (1 + M_{\boldsymbol{F}}^2)} \right) \, \mathrm{d}\delta \\
&\leq 4\delta_n^2 + 8\|G^* - G_0\|_{L^2(\gamma)}^2 + 16 C_1 \sigma (1 + M_{\boldsymbol{F}}) \sqrt{\frac{\pi}{n}}.
\end{aligned}$$

Using (B.10) and (B.11), and minimizing over $G^* \in \boldsymbol{G}$ gives the bound on the mean-squared error (2.16) for some C_2.

Appendix C. Neural Network Theory

In this section we recap elementary operations and approximation theory for neural networks, based on Petersen and Voigtlaender (2018). For a NN $f : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}$ (see Definition 15) we denote by $\mathrm{size}_{\mathrm{in}}(f)$ and $\mathrm{size}_{\mathrm{out}}(f)$ the number of nonzero weights and biases of the first and last layer respectively.

C.1 Operations on Neural Networks

In this subsection, we recap elementary operations on NNs. We start with the parallelization of two NNs following Opschoor et al. (2022a, Section 2.2.1).

Definition 35 (Parallelization) *Let $q \in \mathbb{N}$, $q \geq 2$ and $\sigma \in \{\sigma_1, \sigma_q\}$. Let f and g be two σ-NNs realizing the functions f and g with the same depth L. Furthermore, define the input dimensions of f and g as n_f and n_g and the output dimensions as m_f and m_g.*[3] *Then there exists a σ-NN (f, g), called the* parallelization *of f and g, which simultaneously realizes f and g, i.e.*

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

It holds

$$\begin{aligned}
\mathrm{size}\left((f, g)\right) &= \mathrm{size}(f) + \mathrm{size}(g), \\
\mathrm{depth}\left((f, g)\right) &= \mathrm{depth}(f) = \mathrm{depth}(g), \\
\mathrm{width}\left((f, g)\right) &= \mathrm{width}(f) + \mathrm{width}(g), \\
\mathrm{mpar}\left((f, g)\right) &= \max\left\{\mathrm{mpar}(f), \mathrm{mpar}(g)\right\}, \\
\mathrm{mran}_\Omega\left((f, g)\right)^2 &= \mathrm{mran}_\Omega(f)^2 + \mathrm{mran}_\Omega(g)^2.
\end{aligned} \tag{C.1}$$

3. Using the syntax from (3.4a)–(3.4b), it holds $n_f = p_0(f)$, $n_g = p_0(g)$, $m_f = p_{L+1}(f)$ and $m_g = p_{L+1}(g)$.

Let $N \in \mathbb{N}$, $N \geq 3$. We extend Definition 35 to parallelize N σ-neural networks f_i, $i = 1, \dots N$ with equal depth and denote the resulting σ-NN as $(\{f_i\}_{i=1}^N)$. It holds that

$$\begin{aligned}
\operatorname{size}\left(\left(\{f_i\}_{i=1}^N\right)\right) &= \sum_{i=1}^N \operatorname{size}(f_i), && \text{(C.2)}\\
\operatorname{size_{in}}\left(\left(\{f_i\}_{i=1}^N\right)\right) &= \sum_{i=1}^N \operatorname{size_{in}}(f_i),\\
\operatorname{size_{out}}\left(\left(\{f_i\}_{i=1}^N\right)\right) &= \sum_{i=1}^N \operatorname{size_{out}}(f_i),\\
\operatorname{depth}\left(\left(\{f_i\}_{i=1}^N\right)\right) &= \operatorname{depth}(f_1),\\
\operatorname{width}\left(\left(\{f_i\}_{i=1}^N\right)\right) &= \sum_{i=1}^N \operatorname{width}(f_i), && \text{(C.3)}\\
\operatorname{mpar}\left(\left(\{f_i\}_{i=1}^N\right)\right) &= \max_{i=1,\dots,N} \operatorname{mpar}(f_i),\\
\operatorname{mran_\Omega}\left(\left(\{f_i\}_{i=1}^N\right)\right)^2 &= \sum_{i=1}^N \operatorname{mran_\Omega}(f_i)^2.
\end{aligned}$$

Next, we recall the concatenation of NNs, see Petersen and Voigtlaender (2018, Definition 2.2).

Lemma 36 (Concatenation) *Let $q \in \mathbb{N}$, $q \geq 2$ and $\sigma \in \{\sigma_1, \sigma_q\}$. Let f and g be two σ-NNs. Furthermore, let the output dimension m_g of g equal the input dimension n_f of f. Then there exists a σ-NN $f \bullet g$ realizing the composition $f \circ g : x \mapsto f(g(x))$ of the functions f and g. It holds*

$$\begin{aligned}
\operatorname{depth}(f \bullet g) &= \operatorname{depth}(f) + \operatorname{depth}(g),\\
\operatorname{width}(f \bullet g) &= \max\{\operatorname{width}(f), \operatorname{width}(g)\},\\
\operatorname{mran_\Omega}(f \bullet g) &= \operatorname{mran}_{g(\Omega)}(f).
\end{aligned}$$

There is no simple control over the size and the weight bound of the concatenation $f \bullet g$ in Definition 36. The reason is that the network $f \bullet g$ multiplies network weights and biases of the NNs f and g at layer $l = \operatorname{depth}(g) + 1$ (for details see Petersen and Voigtlaender 2018, Definition 2.2). In the following we use *sparse concatenation* to get control over the size and the weights. We first introduce the realization of the identity map and separate the analysis for the σ_1- and σ_q-case. The following lemma is proven in Petersen and Voigtlaender (2018, Remark 2.4).

Lemma 37 (σ_1-realization of identity map) *Let $d \in \mathbb{N}$ and $L \in \mathbb{N}$. Then there exists a σ_1-*identity network $\operatorname{Id}_{\mathbb{R}^d}$ *of depth L, which exactly realizes the identity map $\operatorname{Id}_{\mathbb{R}^d} : \mathbb{R}^d \to \mathbb{R}^d$, $\boldsymbol{x} \mapsto \boldsymbol{x}$. It holds*

$$\begin{aligned}
\operatorname{size}(\operatorname{Id}_{\mathbb{R}^d}) &\leq 2d(L+1), && \text{(C.4)}\\
\operatorname{width}(\operatorname{Id}_{\mathbb{R}^d}) &\leq 2d, && \text{(C.5)}\\
\operatorname{mpar}(\operatorname{Id}_{\mathbb{R}^d}) &\leq 1.
\end{aligned}$$

We proceed with the analogous result for the RePU activation function. The following lemma follows from the construction in Li (2020, Theorem 2.5 part 2) with a parallelization argument.

Lemma 38 (σ_q-realization of identity map) *Let $q \in \mathbb{N}$ with $q \geq 2$. Further let $d \in \mathbb{N}$ and $L \in \mathbb{N}$ be arbitrary. Then there exists a σ_q-NN $\mathrm{Id}_{\mathbb{R}^d}$ of depth L, which exactly realizes the identity map $\mathrm{Id}_{\mathbb{R}^d}$. It holds*

$$\begin{aligned} \mathrm{size}(\mathrm{Id}_{\mathbb{R}^d}) &\leq C_q d L, \\ \mathrm{width}(\mathrm{Id}_{\mathbb{R}^d}) &\leq C_q d, \\ \mathrm{mpar}(\mathrm{Id}_{\mathbb{R}^d}) &\leq C_q, \end{aligned} \tag{C.6}$$

where C_q is independent of d (but does depend on q).

We next define the sparse concatenation of two NNs, see Petersen and Voigtlaender (2018, Definition 2.5).

Definition 39 (σ_1-sparse concatenation) *Let f and g be two σ_1-NNs. Furthermore, let the output dimension m_g of g equal the input dimension n_f of f. Then there exists a σ_1-NN $f \circ g$ with*

$$f \circ g \coloneqq f \bullet \mathrm{Id}_{\mathbb{R}_{m_g}} \bullet g$$

realizing the composition $f \circ g : x \mapsto f(g(x))$ of the functions f and g.[4] *It holds*

$$\mathrm{size}(f \circ g) \leq \mathrm{size}(g) + \mathrm{size}_{\mathrm{out}}(g) + \mathrm{size}_{\mathrm{in}}(f) + \mathrm{size}(f) \leq 2\mathrm{size}(f) + 2\mathrm{size}(g), \tag{C.7}$$

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

$$\mathrm{size}_{\mathrm{out}}(f \circ g) \leq \begin{cases} \mathrm{size}_{\mathrm{out}}(f) & \mathrm{depth}(f) \geq 1, \\ 2\mathrm{size}_{\mathrm{out}}(f) & \mathrm{depth}(f) = 0, \end{cases} \tag{C.8}$$

$$\mathrm{depth}(f \circ g) = \mathrm{depth}(f) + \mathrm{depth}(g) + 1, \tag{C.9}$$

$$\mathrm{width}(f \circ g) \leq 2 \max\{\mathrm{width}(f), \mathrm{width}(g)\}, \tag{C.10}$$

$$\mathrm{mpar}(f \circ g) \leq \max\left\{\mathrm{mpar}(f), \mathrm{mpar}(g)\right\}, \tag{C.11}$$

$$\mathrm{mran}_{\Omega}\left(f \circ g\right) = B_{g(\Omega)}\left(f\right). \tag{C.12}$$

Proof The bounds in (C.9), (C.10) and (C.12) follow from Definition 37 with the NN calculus from Definition 36. The bounds on the sizes in (C.7)–(C.8) and the weight and bias bound (C.11) follow from the specific structure of the σ_1-identity network, see Petersen and Voigtlaender (2018, Remark 2.6). ■

We proceed with the sparse concatenation of σ_q-NNs.

4. The symbol $\circ$ does mean either the functional concatenation of f and g or the sparse concatenation of the NN f and g (which realizes the function $f \circ g$).

Definition 40 (σ_q-sparse concatenation) *Let $q \in \mathbb{N}$ with $q \geq 2$. Let f and g be two σ_q-NNs. Furthermore, let the output dimension m_g of g equal the input dimension n_f of f. Then there exists a σ_q-NN $f \circ g$ with*

$$f \circ g \coloneqq f \bullet \mathrm{Id}_{\mathbb{R}^{m_g}} \bullet g$$

realizing the composition $f \circ g : x \mapsto f(g(x))$ of the functions f and g. It holds

$$\begin{aligned}
\mathrm{size}(f \circ g) &\leq \mathrm{size}(g) + (C_q - 1)\mathrm{size}_{\mathrm{out}}(g) + (C_q - 1)\mathrm{size}_{\mathrm{in}}(f) + \mathrm{size}(f) \\
&\leq C_q \mathrm{size}(f) + C_q \mathrm{size}(g), && \text{(C.13)} \\
\mathrm{size}_{\mathrm{in}}(f \circ g) &\leq \begin{cases} \mathrm{size}_{\mathrm{in}}(g) & \mathrm{depth}(g) \geq 1, \\ C_q \mathrm{size}_{\mathrm{in}}(g) & \mathrm{depth}(g) = 0, \end{cases} \\
\mathrm{size}_{\mathrm{out}}(f \circ g) &\leq \begin{cases} \mathrm{size}_{\mathrm{out}}(f) & \mathrm{depth}(f) \geq 1, \\ C_q \mathrm{size}_{\mathrm{out}}(f) & \mathrm{depth}(f) = 0, \end{cases} && \text{(C.14)} \\
\mathrm{depth}(f \circ g) &= \mathrm{depth}(f) + \mathrm{depth}(g) + 1, && \text{(C.15)} \\
\mathrm{width}(f \circ g) &\leq C_q \max\{\mathrm{width}(f), \mathrm{width}(g)\}, && \text{(C.16)} \\
\mathrm{mpar}(f \circ g) &\leq C_q \max\left\{\mathrm{mpar}(f), \mathrm{mpar}(g)\right\}, && \text{(C.17)} \\
\mathrm{mran}_{\Omega}\left(f \circ g\right) &= B_{g(\Omega)}\left(f\right) && \text{(C.18)}
\end{aligned}$$

with a constant $C_q > 1$ depending only on q.

Proof The bounds in (C.15), (C.16) and (C.18) follow from Definition 38 with the NN calculus from Definition 36. The bounds on the sizes in (C.13)–(C.14) and the weight and bias bound in (C.17) hold because of the specific structure of the σ_q-identity network $\mathrm{Id}_{\mathbb{R}^n}$, see Li (2020, Eq. 2.57). ■

Definitions 39 and 40 show that one can control the size, as well as the weights and biases of the concatenation of two NN by inserting one additional identity layer between the two networks. We end this subsection by introducing summation and scalar multiplication networks.

Definition 41 (Summation networks) *Let $q \in \mathbb{N}$, $q \geq 2$, $\sigma \in \{\sigma_1, \sigma_q\}$ and $d, m \in \mathbb{N}$. Then there exists a σ-NN Σ_m such that for $x_1, \ldots, x_m \in \mathbb{R}^d$*

$$\Sigma_m(x_1, \ldots, x_m) = \sum_{i=1}^{m} x_i,$$

with $\mathrm{depth}(\Sigma_m) = 0$, $\mathrm{width}(\Sigma_m) = md$, $\mathrm{size}(\Sigma_m) = md$ and $\mathrm{mpar}(\Sigma_m) = 1$.

Proof Set $\boldsymbol{w}^1 = (\mathbb{1}_d, \ldots, \mathbb{1}_d)$ and $\boldsymbol{b}^1 = 0$ with the $d \times d$ identity matrices $\mathbb{1}_d$. ■

Definition 42 (Scalar multiplication networks) *Let $q \in \mathbb{N}$, $q \geq 2$ and $\sigma \in \{\sigma_1, \sigma_q\}$. Let $\alpha \in \mathbb{R}$ and $d \in \mathbb{N}$. Then there exists a σ-NN SM_α with*

$$SM_\alpha(x) = \alpha x, \qquad x \in \mathbb{R}^d.$$

Furthermore, there exists a constant C_q only depending on q such that

$$\begin{aligned}\operatorname{depth}(SM_\alpha) &\leq C_q \max\{1, \log(|\alpha|)\},\\ \operatorname{width}(SM_\alpha) &\leq C_q d,\\ \operatorname{size}(SM_\alpha) &\leq C_q d \max\{1, \log(|\alpha|)\},\\ \operatorname{mpar}(SM_\alpha) &\leq C_q.\end{aligned}$$

Proof For the proof of the σ_1-case, see Elbrachter et al. (2021, Lemma A.1). We prove the RePU case in the following.

Without loss of generality we can choose $\alpha > 1$. For $\alpha < 0$ we set $SM_\alpha = -SM_{-\alpha}$. For $0 < \alpha < 1$ we set $SM_\alpha = \alpha \operatorname{Id}_{\mathbb{R}^d}$, where we directly multiply the weights and biases of the identity network (with depth $L = 1$) with α.

Therefore let $\alpha > 1$. Let K be the maximum integer smaller than $\log_2(\alpha)$, and set $\tilde{\alpha} = 2^{-K+1}\alpha < 1$. Furthermore, set $A_1 = (\operatorname{Id}_{\mathbb{R}^d}, \operatorname{Id}_{\mathbb{R}^d})$ and $A_2 = \Sigma_2$ with the one-layered identity network $\operatorname{Id}_{\mathbb{R}^d}$ and the summation network Σ_2 from Definition 41. We notice that

$$A_2 \bullet A_1 x = 2x \quad \forall x \in \mathbb{R}^d.$$

Using the bounds for Σ_2 from Definition 41 and $\operatorname{Id}_{\mathbb{R}^d}$ from Definition 38, we have $\operatorname{depth}(A_2 \bullet A_1) = 1$, $\operatorname{width}(A_2 \bullet A_1) \leq C_q d$, $\operatorname{size}(A_2 \bullet A_1) \leq C_q d$ and $\operatorname{mpar}(A_2 \bullet A_1) \leq C_q$. Setting $A_{2k+1} = A_1$, $A_{2k+2} = A_2$ for $k = 1, \ldots, K$ and $A_{2K+3} = \tilde{\alpha} \operatorname{Id}_{\mathbb{R}^d}$, we get

$$A_{2K+3} \circ A_{2K+2} \bullet A_{2K+1} \circ A_{2K} \bullet A_{2K-1} \circ \ldots.. \circ A_2 \bullet A_1 x = \alpha x \quad \forall x \in \mathbb{R}^d.$$

Applying the σ_q-NN calculus for concatenation (Definition 36) and sparse concatenation (Definition 40) gives the desired bounds for SM_α. ■

C.2 Neural Network Approximation Theory

In this subsection, we summarize approximation results for ReLU and RePU neural networks. In recent years, the expressivity and approximation properties of neural network architectures have been extensively studied in the literature (Mhaskar, 1996; Pinkus, 1999; Yarotsky, 2017; Poggio et al., 2017; Petersen and Voigtlaender, 2018; Elbrachter et al., 2021; Opschoor et al., 2022a; De Ryck et al., 2021). However, with few exceptions (Schmidt-Hieber, 2020a; De Ryck et al., 2021; Elbrachter et al., 2021), most of these works do not provide bounds on the size of the weights, which are crucial for controlling the entropy. Therefore, we revisit some of these arguments to provide complete proofs of our results.

We start with the well-known result that ReLU-NNs can approximate the multiplication map exponentially fast. The following proposition was shown in Elbrachter et al. (2021, Proposition III.3).

Proposition 43 (σ_1-NN approximation of multiplication) *Let $D \in \mathbb{R}$, $D \geq 1$ and $\delta \in (0, 1/2)$. Then there exists a σ_1-NN $\tilde{\times}_{\delta,D} : [-D, D]^2 \to \mathbb{R}$ satisfying*

$$\sup_{x,y \in [-D,D]} \left| \tilde{\times}_{\delta,D}(x, y) - xy \right| \leq \delta.$$

Furthermore, there exists a constant C, independent of D and δ, such that

$$\operatorname{depth}\left(\tilde{\times}_{\delta,D}\right) \leq C\left(\log(D) + \log\left(\delta^{-1}\right)\right), \tag{C.19}$$
$$\operatorname{width}\left(\tilde{\times}_{\delta,D}\right) \leq 5, \tag{C.20}$$
$$\operatorname{size}\left(\tilde{\times}_{\delta,D}\right) \leq C\left(\log(D) + \log\left(\delta^{-1}\right)\right), \tag{C.21}$$
$$\operatorname{mpar}\left(\tilde{\times}_{\delta,D}\right) \leq 1. \tag{C.22}$$

The next proposition shows that the multiplication map can be exactly realized by a σ_q-NN, which follows directly from Li (2020, Theorem 2.5 and Eq. 2.59).

Proposition 44 (σ_q-NN approximation of multiplication) *Let $q \in \mathbb{N}$, $q \geq 2$. There exists a σ_q-NN $\tilde{\times} : \mathbb{R}^2 \to \mathbb{R}$ with $\operatorname{depth}(\tilde{\times}) = 1$ exactly realizing the multiplication of two numbers, i.e.*

$$\tilde{\times}(x, y) = xy \quad \forall x, y \in \mathbb{R}.$$

We can extend the above results to the multiplication of N numbers, see Opschoor et al. (2022a, Proposition 2.6).

Proposition 45 (σ_1-NN multiplication of N numbers) *Let $N \in \mathbb{N}$ with $N \geq 2$. Furthermore, let $D \in \mathbb{R}$, $D \geq 1$ and $\delta \in (0, 1/2)$. Then there exists a σ_1-NN $\widetilde{\prod}_{\delta,D}$: $[-D, D]^N \to \mathbb{R}$ such that*

$$\sup_{(y_i)_{i=1}^N \in [-D,D]^N} \left| \prod_{j=1}^{N} y_j - \widetilde{\prod}_{\delta,D}(y_1, \ldots, y_N) \right| \leq \delta. \tag{C.23}$$

Furthermore, there exists a constant C independent of N, δ and D such that

$$\operatorname{depth}\left(\widetilde{\prod}_{\delta,D}\right) \leq C \log(N)\left(\log(N) + N\log(D) + \log\left(\delta^{-1}\right)\right),$$
$$\operatorname{width}\left(\widetilde{\prod}_{\delta,D}\right) \leq 5N, \tag{C.24}$$
$$\operatorname{size}\left(\widetilde{\prod}_{\delta,D}\right) \leq CN\left(\log(N) + N\log(D) + \log\left(\delta^{-1}\right)\right), \tag{C.25}$$
$$\operatorname{mpar}\left(\widetilde{\prod}_{\delta,D}\right) \leq 1.$$

Proof Analogous to Opschoor et al. (2022a, Proposition 2.6) we construct $\widetilde{\prod}_{\delta,D}$ as a binary tree of $\tilde{\times}_{\cdot,\cdot}$-networks from Proposition 43. We modify the proof of Opschoor et al. (2022a) to get a construction with bounded weights.

Define $\tilde{N} := \min\{2^k : k \in \mathbb{N}, 2^k \geq N\}$. We now consider the multiplication of $\tilde{N}$ numbers with $y_{N+1}, \ldots, y_{\tilde{N}} := 1$. This can be implemented by a zero-layer network with

$$w_{i,j}^1 = \begin{cases} 1 & i = j \leq N, \\ 0 & \text{otherwise}, \end{cases}$$
$$b_j^1 = \begin{cases} 0 & j \leq N, \\ 1 & N < j \leq \tilde{N}. \end{cases}$$

For $l = 0, \dots, \log_2 \tilde{N} - 1$ we define the mapping R^l

$$\begin{aligned} R^l &: [-D_l, D_l]^{\frac{\tilde{N}}{2^l}} \mapsto [-D_{l+1}, D_{l+1}]^{\frac{\tilde{N}}{2^{l+1}}}, \\ R^l(y_1^l, \dots, y_{2^{\log_2(\tilde{N})-l}}^l) &:= \left(\tilde{\times}_{\delta', D^l}(y_1^l, y_2^l), \dots, \tilde{\times}_{\delta', D^l}(y_{2^{\log_2(\tilde{N})-l}-1}^l, y_{2^{\log_2(\tilde{N})-l}}^l) \right) \end{aligned} \tag{C.26}$$

with $\delta' := \delta / (\tilde{N}^2 D^{2\tilde{N}})$ and $D_l := 2^l D^{2^l}$. We now set

$$\widetilde{\prod}_{\delta, D} := R^{\log_2(\tilde{N})-1} \circ \dots \circ R^0. \tag{C.27}$$

Eq. (C.27) shows that the map R^l describes the multiplications on level l of the binary tree $\widetilde{\prod}_{\delta, D}$. In order for (C.27) to be well-defined, we have to show that the outputs of the NN R^l are admissible inputs for the NN R^{l+1}.

We therefore denote with y_j^l, $j = 1, \dots, 2^{\log_2(\tilde{N})-l}$, $l = 1, \dots, \log_2(\tilde{N}) - 1$, the output of the network $R^{l-1} \circ \dots \circ R^0$ applied to the input $y_k^0 = y_k$, $k = 1, \dots, \tilde{N}$. Then we have to show $|y_j^l| \leq D_l$ for $l = 0, \dots, \log_2(\tilde{N}) - 1$ and $j = 1, \dots, 2^{\log_2(\tilde{N})-l}$. We will show this claim by induction. For $l = 0$ it holds $|y_j^l| \leq D = D_0$. Now assume $|y_j^l| \leq D_l$ for arbitrary but fixed $l \in \{0, \dots, \log_2(\tilde{N}) - 2\}$ and all $j = 1, \dots, 2^{\log_2(\tilde{N})-l}$. Then it holds

$$\begin{aligned} |y_j^{l+1}| &= |\tilde{\times}_{\delta', D^l}(y_{2j-1}^l, y_{2j}^l)| \\ &= |y_{2j-1}^l \cdot y_{2j}^l + \delta'| \leq D_l^2 + \underbrace{\delta'}_{\leq 1 \leq D_l^2} \leq 2D_l^2 = D_{l+1} \end{aligned}$$

for all $j = 1, \dots, 2^{\log_2(\tilde{N})-(l+1)}$, which shows the claim. We proceed by showing the error bound in (C.23). Therefore define

$$z_j^l := \prod_{k=1}^{2^l} y_{k+2^l(j-1)}$$

for $l = 0, \dots, \log_2(\tilde{N})$ and $j = 1, \dots, 2^{\log_2(\tilde{N})-l}$. The quantites z_j^l describe the exact computations up to level l of the binary tree, i.e. the output of level $l - 1$, if one uses standard multiplication instead of the multiplication networks $\tilde{\times}$ in the first $l - 1$ levels. We now prove

$$|y_j^l - z_j^l| \leq 4^l D^{2^{l+1}} \delta', \quad j = 1, \dots, 2^{\log_2(\tilde{N})-l} \tag{C.28}$$

by induction over $l = 0, \dots, \log_2 \tilde{N}$. Inserting $l = \log_2(\tilde{N})$ then shows the error bound in (C.23) using the definition of δ'.

We have $y_j^0 = y_j = z_j^0$ for $j = 1, \dots, \tilde{N}$, therefore (C.28) holds for $l = 0$. Now assume (C.28) for arbitrary but fixed $l \in \{0, \dots, \log_2(\tilde{N}) - 1\}$. For $j = 1, \dots, 2^{\log_2(\tilde{N})-(l+1)}$ it holds

$$\begin{aligned}
|y_j^{l+1} - z_j^{l+1}| &= \left| \tilde{\times}_{D_l,\delta'}(y_{2j-1}^l, y_{2j}^l) - z_{2j-1}^l z_{2j}^l \right| \\
&= \left| y_{2j-1}^l y_{2j}^l + \delta' - z_{2j-1}^l z_{2j}^l \right| \\
&= \left| (y_{2j-1}^l - z_{2j-1}^l + z_{2j-1}^l) \cdot (y_{2j}^l - z_{2j}^l + z_{2j}^l) + \delta' - z_{2j-1}^l z_{2j}^l \right| \\
&= \big| (y_{2j-1}^l - z_{2j-1}^l) \cdot (y_{2j}^l - z_{2j}^l) + (y_{2j-1}^l - z_{2j-1}^l) z_{2j}^l \\
&\qquad + (y_{2j}^l - z_{2j}^l) z_{2j-1}^l + \delta' \big| \\
&\leq \underbrace{\left| y_{2j-1}^l - z_{2j-1}^l \right|}_{\text{use(C.28)}} \cdot \underbrace{\left| y_{2j}^l - z_{2j}^l \right|}_{\leq 1} + \underbrace{\left| y_{2j-1}^l - z_{2j-1}^l \right|}_{\text{use(C.28)}} \cdot \underbrace{\left| z_{2j}^l \right|}_{\leq D^{2^l}} \\
&\qquad + \underbrace{\left| y_{2j}^l - z_{2j}^l \right|}_{\text{use(C.28)}} \cdot \underbrace{\left| z_{2j-1}^l \right|}_{\leq D^{2^l}} + |\delta'| \\
&\leq \delta' \left(4^l D^{2^{l+1}} \left(1 + 2D^{2^l} \right) + 1 \right) \\
&\leq 4\, 4^l D^{2^{l+1}} D^{2^{l+1}} \delta' \leq 4^{l+1} D^{2^{l+2}} \delta',
\end{aligned}$$

which shows (C.28) for $l + 1$ and therefore the claim.

We proceed by calculating the depth of $\widetilde{\prod}_{\delta,D}$. Since $\widetilde{\prod}_{\delta,D}$ concatenates the maps $\tilde{\times}_{\delta',D_l}$, we can repeatedly use (C.9) and get

$$\operatorname{depth}\left(\widetilde{\prod}_{\delta,D} \right) \leq \sum_{j=0}^{\log_2(\tilde{N})-1} \operatorname{depth}\left(\tilde{\times}_{\delta',D_j} \right) + \log_2(\tilde{N}) - 1.$$

We use the depth bound for $\tilde{\times}$ from (C.19) and calculate

$$\begin{aligned}
\operatorname{depth}\left(\widetilde{\prod}_{\delta,D} \right) &\leq C \sum_{j=0}^{\log_2(\tilde{N})-1} \log\left(D_j \delta'^{-1} \right) + \log_2(\tilde{N}) \\
&= C \sum_{j=0}^{\log_2(\tilde{N})-1} \log\left(2^j D^{2^j} \delta'^{-1} \right) + \log_2(\tilde{N}) \\
&= C \log\left(\prod_{j=0}^{\log_2(\tilde{N})-1} 2^j D^{2^j} \delta'^{-1} \right) + \log_2(\tilde{N}) \\
&\leq C \log\left(2^{\frac{\log_2(\tilde{N}) \cdot (\log_2(\tilde{N})-1)}{2}} D^{2^{\log_2(\tilde{N})}} \left(\delta' \right)^{-\log_2(\tilde{N})} \right) + \log_2(\tilde{N}) \\
&\leq C \log\left(2^{\left(\log_2(\tilde{N})\right)^2} D^{\tilde{N}} \tilde{N}^{2\log_2(\tilde{N})} D^{2\tilde{N}\log_2(\tilde{N})} \delta^{-\log_2(\tilde{N})} \right) + \log_2(\tilde{N}) \\
&\leq C \log\left(2^{\left(\log_2(\tilde{N})\right)^2} \tilde{N}^{2\log_2(\tilde{N})} D^{3\tilde{N}\log_2(\tilde{N})} \delta^{-\log_2(\tilde{N})} \right) + \log_2(\tilde{N}) \\
&\leq C \log(N) \left(\log(N) + N \log(D) + \log\left(\delta^{-1} \right) \right). \qquad \text{(C.29)}
\end{aligned}$$

The constant C changes from line to line in (C.29).

For a bound on the width we use the fact that $\widetilde{\prod}_{\delta,D}$ is a parallelization of at most $\tilde{N}/2 \leq N$ networks $\tilde{\times}_{\delta',D_l}$ in each layer $l \in \{1,\dots,\log_2(\tilde{N})\}$. With (C.3) and the width bound of $\tilde{\times}$ in (C.20) it holds

$$\text{width}\left(\widetilde{\prod}_{\delta,D_l}\right) \leq N \,\text{width}\left(\tilde{\times}_{\delta',D_l}\right) \leq 5N.$$

For a bound on the size $\text{size}(\widetilde{\prod}_{\delta,D})$ we observe that level l, $l = 0,\dots,\log_2(\tilde{N})-1$, of the binary tree $\tilde{\prod}$ consists of $2^{\log_2(\tilde{N})-l-1}$ product networks $\tilde{\times}_{\delta',D_l}$. We calculate

$$\begin{aligned}
\text{size}\left(\widetilde{\prod}_{\delta,D}\right) &\leq \sum_{l=0}^{\log_2(\tilde{N})-1} 2^{\log_2(\tilde{N})-l-1}\left(s_{in}\left(\tilde{\times}_{\delta',D_l}\right) + s_{out}\left(\tilde{\times}_{\delta',D^l}\right) + \text{size}\left(\tilde{\times}_{\delta',D^l}\right)\right)\\
&\leq \sum_{l=0}^{\log_2(\tilde{N})-1} 2^{\log_2(\tilde{N})-l-1} 3C\log\left(D_l\delta'^{-1}\right)\\
&\leq C\sum_{l=0}^{\log_2(\tilde{N})-1} 2^{\log_2(\tilde{N})-l-1}\log\left(2^l D^{2^l}\tilde{N}^2 D^{2\tilde{N}}\delta^{-1}\right)\\
&\leq C\sum_{l=0}^{\log_2(\tilde{N})-1} 2^{\log_2(\tilde{N})-l-1}\left(l + 2^l\log(D) + \log\left(\tilde{N}\right) + \tilde{N}\log(D) + \log\left(\delta^{-1}\right)\right)\\
&\leq C\left(\tilde{N}\log_2(\tilde{N}) + \tilde{N}\log_2(\tilde{N})\log(D) + \tilde{N}\log(\tilde{N}) + \tilde{N}^2\log(D) + \tilde{N}\log\left(\delta^{-1}\right)\right)\\
&\leq CN\left(\log(N) + N\log(D) + \log\left(\delta^{-1}\right)\right). \qquad \text{(C.30)}
\end{aligned}$$

In (C.30) we used (C.7) to bound the size of a sparse concatenation and (C.21) for the size of the product network $\tilde{\times}_{\delta,D}$.

For the bound on the weights and biases, we get $\text{mpar}(\widetilde{\prod}_{\delta,D}) \leq 1$ because of $\text{mpar}(\tilde{\times}_{\delta,D}) \leq 1$, see (C.22), and the NN calculus for sparse concatenation in (C.11) and parallelization in (C.1). ∎

We continue with the RePU-case.

Proposition 46 (σ_q-NN of multiplication of N numbers) *Let $N, q \in \mathbb{N}$ with $N, q \geq 2$. Then there exists a σ_q-NN $\widetilde{\prod}\colon \mathbb{R}^N \to \mathbb{R}$ such that*

$$\widetilde{\prod}(y_1,\dots,y_N) = \prod_{j=1}^{N} y_j.$$

Furthermore, there exists a constant C_q independent of N such that

$$\operatorname{depth}\left(\widetilde{\prod}_{\delta,D}\right) \leq C_q \log(N), \tag{C.31}$$

$$\operatorname{width}\left(\widetilde{\prod}_{\delta,D}\right) \leq C_q N, \tag{C.32}$$

$$\operatorname{size}\left(\widetilde{\prod}_{\delta,D}\right) \leq C_q N, \tag{C.33}$$

$$\operatorname{mpar}\left(\widetilde{\prod}_{\delta,D}\right) \leq C_q. \tag{C.34}$$

Proof The construction is similar to the ReLU case. We define $\tilde{\prod}$ as a binary tree of product networks $\tilde{\times}$, see (C.26) and (C.27). The binary tree has a maximum of $2N$ binary networks $\tilde{\times}$, a maximum height of $\log_2(2N)$ and a maximum width of N. Therefore (C.31)–(C.34) follow with the NN calculus rules from Definition 40. ■

We proceed and state the approximation results for univariate polynomials. We start with the ReLU case. The following proposition was shown in Elbrachter et al. (2021, Proposition III.5).

Proposition 47 (σ_1-NN approximation of polynomials) *Let $m \in \mathbb{N}$ and $a = (a_i)_{i=0}^m \in \mathbb{R}^{m+1}$. Further let $D \in \mathbb{R}$, $D \geq 1$ and $\delta \in (0, 1/2)$. Define $a_\infty = \max\{1, \|a\|_\infty\}$. Then there exists a σ_1-NN $\tilde{p}_{\delta,D} : [-D, D] \to \mathbb{R}$ satisfying*

$$\sup_{x\in[-D,D]} \left| \tilde{p}_{\delta,D}(x) - \sum_{i=0}^{m} a_i x^i \right| \leq \delta.$$

Furthermore, there exists a constant C independent of m, a_i, D and δ such that

$$\begin{aligned}
\operatorname{depth}(\tilde{p}_{\delta,D}) &\leq Cm\left(m\log\left(D\right) + \log\left(\delta^{-1}\right) + \log(m) + \log(a_\infty)\right), \\
\operatorname{width}\left(\tilde{p}_{\delta,D}\right) &\leq 9, \\
\operatorname{size}\left(\tilde{p}_{\delta,D}\right) &\leq Cm\left(m\log\left(D\right) + \log\left(\delta^{-1}\right) + \log(m) + \log(a_\infty)\right), \\
\operatorname{mpar}\left(\tilde{p}_{\delta,D}\right) &\leq 1.
\end{aligned}$$

In the RePU-case we get the well-known result that polynomials can be exactly realized by σ_q-NNs, see Li et al. (2019).

Proposition 48 (σ_q-NN realization of polynomials) *Let $m, q \in \mathbb{N}$, $q \geq 2$ and $a = (a_i)_{i=0}^m \in \mathbb{R}^{m+1}$. Set $a_\infty = \max\{1, \max_{i=0,\dots,m} a_i\}$. Then there exists a σ_q-NN $\tilde{p} : \mathbb{R} \to \mathbb{R}$ satisfying*

$$\tilde{p}(x) = \sum_{i=0}^{m} a_i x^i \quad \forall x \in \mathbb{R}.$$

Furthermore, there exists a constant C_q only depending on q such that

$$\begin{aligned}\operatorname{depth}(\tilde{p}) &\leq C_q\left(\log(a_\infty)+m\right),\\ \operatorname{width}(\tilde{p}) &\leq C_q,\\ \operatorname{size}(\tilde{p}) &\leq C_q\left(\log(a_\infty)+m\right),\\ \operatorname{mpar}(\tilde{p}) &\leq C_q.\end{aligned}$$

Proof We use Horner's method for polynomial evaluation and write

$$\sum_{i=0}^{m} a_i x^i = a_\infty\left(\frac{a_0}{a_\infty}+x\left(\frac{a_1}{a_\infty}+\cdots+x\left(\frac{a_{m-1}}{a_\infty}+x\frac{a_m}{a_\infty}\right)\ldots\right)\right). \tag{C.35}$$

Following (C.35), we build $\tilde{p}$ via

$$\tilde{p} = SM_{a_\infty}\circ\Sigma_2\left(\frac{a_0}{a_\infty},\tilde{\times}\left(\mathrm{Id}_{\mathbb{R}},\Sigma_2\left(\frac{a_1}{a_\infty},\ldots,SM_{a_m a_\infty^{-1}}\left(\mathrm{Id}_{\mathbb{R}}\right)\right)\ldots\right)\right).$$

The bounds for $\tilde{p}$ follow from the respective bounds for Σ_2 from Definition 41, $\mathrm{Id}_{\mathbb{R}}$ from Lemma 38 and SM_α from Definition 42. ■

We now use Propositions 47 and 48 to get an approximation result for univariate Legendre polynomials.

Corollary 49 (σ_1-NN approximation of L_j) *Let $j\in\mathbb{N}_0$ and $\delta\in(0,1/2)$. Then there exists a σ_1-NN $\tilde{L}_{j,\delta}:[-1,1]\to\mathbb{R}$ with*

$$\sup_{x\in[-1,1]}|\tilde{L}_{j,\delta}(x)-L_j(x)|\leq\delta.$$

Furthermore, there exists a constant C such that it holds

$$\operatorname{depth}\left(\tilde{L}_{j,\delta}\right) \leq Cj\left(j+\log(\delta^{-1})\right), \tag{C.36}$$

$$\operatorname{width}\left(\tilde{L}_{j,\delta}\right) \leq 9, \tag{C.37}$$

$$\operatorname{size}\left(\tilde{L}_{j,\delta}\right) \leq Cj\left(j+\log\left(\delta^{-1}\right)\right), \tag{C.38}$$

$$\operatorname{mpar}\left(\tilde{L}_{j,\delta}\right) \leq 1.$$

Proof For $j\in\mathbb{N}$, $l\in\mathbb{N}_0$, $l\leq j$, denote the coefficients of L_j with c_l^j. In Opschoor et al. (2020, Eq. (4.17)) the bound $\sum_{l=0}^{j}|c_l^j|\leq 4^j$ is derived. With $c^j=(c_l^j)_{l=0}^j$ it holds $\|c^j\|_\infty\leq\sum_{l=0}^{j}|c_l^j|\leq 4^j$. The result now follows with Proposition 47. ■

We continue with the σ_q-case.

Corollary 50 (σ_q-NN approximation of L_j) *Let $j \in \mathbb{N}_0$. Then there exists a σ_q-NN $\tilde{L}_j : \mathbb{R} \to \mathbb{R}$ with*

$$\tilde{L}_j(x) = L_j(x) \quad \forall x \in \mathbb{R}.$$

Furthermore, there exists a constant C_q only depending on q such that it holds

$$\operatorname{depth}\left(\tilde{L}_j\right) \leq C_q j, \tag{C.39}$$

$$\operatorname{width}\left(\tilde{L}_j\right) \leq C_q, \tag{C.40}$$

$$\operatorname{size}\left(\tilde{L}_j\right) \leq C_q j, \tag{C.41}$$

$$\operatorname{mpar}\left(\tilde{L}_j\right) \leq C_q.$$

Proof The bounds follow similar to the σ_1-case using Proposition 48. ■

Appendix D. Proofs of Section 3

D.1 Proof of Proposition 17

We proceed analogously to the proof of Opschoor et al. (2022a, Proposition 2.13). We define $f_{\Lambda,\delta}$ as a composition of two subnetworks, $f_{\Lambda,\delta} \coloneqq f^{(1)}_{\Lambda,\delta} \circ f^{(2)}_{\Lambda,\delta}$. Corollary 49 ensures the existence of (arbitrarily good) σ_1-NN approximations of univariate Legendre polynomials. All relevant approximations are evaluated, in parallel, by subnetwork $f^{(2)}_{\Lambda,\delta}$, i.e.

$$f^{(2)}_{\Lambda,\delta}(\boldsymbol{y}) \coloneqq \left(\left\{\mathrm{Id}_{\mathbb{R}} \circ \tilde{L}_{\nu_j,\delta'}(y_j)\right\}_{(j,\nu_j)\in T}\right), \tag{D.1}$$

where we used

$$\begin{aligned} T &\coloneqq \left\{(j,\nu_j) \in \mathbb{N}^2 : \boldsymbol{\nu} \in \Lambda,\, j \in \operatorname{supp}\boldsymbol{\nu}\right\}, \\ \delta' &\coloneqq (2d(\Lambda))^{-1} \left(2m(\Lambda)+2\right)^{-d(\Lambda)+1} \delta \end{aligned} \tag{D.2}$$

and $\boldsymbol{y} = (y_j)_{(j,\nu_j)\in T}$. In (D.1) the big round brackets denote a parallelization and we use the identity networks to synchronize the depth. The subnetwork $f^{(1)}_{\Lambda,\delta}$ takes the output of $f^{(2)}_{\Lambda,\delta}$ as input and computes, in parallel, tensorized Legendre polynomials using the multiplication networks $\widetilde{\prod}_{\cdot,\cdot}$ introduced in Proposition 45. With $M_{\boldsymbol{\nu}} \coloneqq 2|\boldsymbol{\nu}|_1 + 2$ we define

$$\begin{aligned} f^{(1)}_{\Lambda,\delta}\left((z_k)_{k\leq|T|}\right) &= f^{(1)}_{\Lambda,\delta}\left(f^{(2)}_{\Lambda,\delta}(\boldsymbol{y})\right) \\ &\coloneqq \left(\left\{\mathrm{Id}_{\mathbb{R}} \circ \widetilde{\prod}_{\delta/2,M_{\boldsymbol{\nu}}}\left(\left\{\mathrm{Id}_{\mathbb{R}} \circ \tilde{L}_{\nu_j,\delta'}(y_j)\right\}_{j\in\operatorname{supp}\boldsymbol{\nu}}\right)\right\}_{\boldsymbol{\nu}\in\Lambda}\right). \end{aligned} \tag{D.3}$$

The multiplication networks in (D.3) are well-defined, since

$$\sup_{y_j\in[-1,1]} |\tilde{L}_{\nu_j,\delta'}(y_j)| \leq 2\nu_j + 2 \leq 2|\boldsymbol{\nu}|_1 + 2 = M_{\boldsymbol{\nu}}, \tag{D.4}$$

where we used (3.10) and $\delta' < 1$.

We will first show the error bound in (3.12). Let $\boldsymbol{\nu} \in \Lambda$ be arbitrary. We use the shorthand notation $\|\cdot\| \coloneqq \|\cdot\|_{L^\infty([-1,1]^{|T|})}$ and calculate

$$\begin{aligned}
&\left\| L_{\boldsymbol{\nu}} - \tilde{L}_{\boldsymbol{\nu},\delta} \right\| \\
\leq &\left\| L_{\boldsymbol{\nu}} - \prod_{j \in \operatorname{supp} \boldsymbol{\nu}} \tilde{L}_{\nu_j,\delta'} \right\| + \left\| \prod_{j \in \operatorname{supp} \boldsymbol{\nu}} \tilde{L}_{\nu_j,\delta'} - \widetilde{\prod}_{\delta/2, M_{\boldsymbol{\nu}}} \left(\left\{ \tilde{L}_{\nu_j,\delta'} \right\}_{j \in \operatorname{supp} \boldsymbol{\nu}} \right) \right\| \\
\leq &\sum_{k \in \operatorname{supp} \boldsymbol{\nu}} \left\| \prod_{\substack{j \in \operatorname{supp} \boldsymbol{\nu}: \\ j<k}} \tilde{L}_{\nu_j,\delta'} \right\| \cdot \left\| L_{\nu_k} - \tilde{L}_{\nu_k,\delta'} \right\| \cdot \left\| \prod_{\substack{j \in \operatorname{supp} \boldsymbol{\nu}: \\ j>k}} L_{\nu_j} \right\| + \frac{\delta}{2} \\
\leq &d(\Lambda) M_{\boldsymbol{\nu}}^{d(\Lambda)-1} \delta' + \frac{\delta}{2} \leq \left(\frac{M_{\boldsymbol{\nu}}}{2m(\Lambda)+2} \right)^{d(\Lambda)-1} \frac{\delta}{2} + \frac{\delta}{2} \leq \delta,
\end{aligned}$$

where we used (D.4), $M_{\boldsymbol{\nu}} \leq 2m(\Lambda)+2$ and the definition of δ'.

We proceed and calculate the depth L of $f_{\Lambda,\delta}$. Since $f_{\Lambda,\delta} = f_{\Lambda,\delta}^{(1)} \circ f_{\Lambda,\delta}^{(2)}$, it holds $\operatorname{depth}(f_{\Lambda,\delta}) \leq \operatorname{depth}(f_{\Lambda,\delta}^{(1)}) + \operatorname{depth}(f_{\Lambda,\delta}^{(2)}) + 1$, see (C.9). We start with a depth bound of $f_{\Lambda,\delta}^{(2)}$. Denoting by C a universal multiplicative constant that is allowed to change from line to line, it holds that

$$\begin{aligned}
\operatorname{depth}\left(f_{\Lambda,\delta}^{(2)}\right) &= 1 + \max_{\substack{\boldsymbol{\nu} \in \Lambda \\ j \in \operatorname{supp} \boldsymbol{\nu}}} \operatorname{depth}\left(\tilde{L}_{\nu_j,\delta'}\right) \\
&\leq C \max_{\substack{\boldsymbol{\nu} \in \Lambda \\ j \in \operatorname{supp} \boldsymbol{\nu}}} \nu_j \left(\nu_j + \log\left(\delta'^{-1}\right)\right) \\
&\leq Cm(\Lambda)\left(m(\Lambda) + \log\left(\delta'^{-1}\right)\right) \\
&\leq Cm(\Lambda)\left(\log\left(d(\Lambda)\right) + d(\Lambda)\log\left(m(\Lambda)\right) + m(\Lambda) + \log\left(\delta^{-1}\right)\right) \\
&\leq Cm(\Lambda)\left(\log\left(d(\Lambda)\right) + d(\Lambda)\log\left(m(\Lambda)\right) + m(\Lambda) + \log\left(\delta^{-1}\right)\right). \qquad \text{(D.5)}
\end{aligned}$$

In (D.5) we used the depth bound for univariate Legendre polynomials, (C.36), at the first inequality. Furthermore, we used $\nu_j \leq m(\Lambda)$. For the depth of $f_{\Lambda,\delta}^{(1)}$ it holds

$$\begin{aligned}
\operatorname{depth}\left(f_{\Lambda,\delta}^{(1)}\right) &= 1 + \max_{\boldsymbol{\nu} \in \Lambda} \operatorname{depth}\left(\widetilde{\prod}_{\delta/2, M_{\boldsymbol{\nu}}}\right) \\
&\leq 1 + C \max_{\boldsymbol{\nu} \in \Lambda} \log\left(|\operatorname{supp} \boldsymbol{\nu}|\right)\left(\log\left(|\operatorname{supp} \boldsymbol{\nu}|\right) + |\operatorname{supp} \boldsymbol{\nu}| \log\left(M_{\boldsymbol{\nu}}\right) + \log\left(\delta^{-1}\right)\right) \\
&\leq 1 + C\log\left(d(\Lambda)\right)\left(\log\left(d(\Lambda)\right) + d(\Lambda)\log\left(m(\Lambda)\right) + \log\left(\delta^{-1}\right)\right), \qquad \text{(D.6)}
\end{aligned}$$

where we used $|\operatorname{supp} \boldsymbol{\nu}| \leq d(\Lambda)$ for all $\boldsymbol{\nu} \in \Lambda$, $M_\nu \leq 4m(\Lambda)$ and the depth bound for σ_1-multiplication networks from Proposition 45. Combining the two depth bounds (D.5) and

(D.6), we get

$$\begin{aligned}\operatorname{depth}(f_{\Lambda,\delta}) &= 1+\operatorname{depth}\left(f^{(1)}_{\Lambda,\delta}\right)+\operatorname{depth}\left(f^{(2)}_{\Lambda,\delta}\right)\\ &\leq Cm(\Lambda)\left(\log\left(d(\Lambda)\right)+d(\Lambda)\log\left(m(\Lambda)\right)+m(\Lambda)+\log\left(\delta^{-1}\right)\right)\\ &+C\log\left(d(\Lambda)\right)\left(\log\left(d(\Lambda)\right)+d(\Lambda)\log\left(m(\Lambda)\right)+\log\left(\delta^{-1}\right)\right)\\ &\leq C\Big[\log(d(\Lambda))d(\Lambda)\log(m(\Lambda))m(\Lambda)+m(\Lambda)^2+\log(\delta^{-1})\big(\log(d(\Lambda))+m(\Lambda)\big)\Big].\end{aligned}$$

For the width $\operatorname{width}(f_{\Lambda,\delta})$ we use $\operatorname{width}(f_{\Lambda,\delta})\leq 2\max\{\operatorname{width}(f^{(1)}_{\Lambda,\delta}),\operatorname{width}(f^{(2)}_{\Lambda,\delta})\}$, see (C.10). This leaves us to calculate $\operatorname{width}(f^{(2)}_{\Lambda,\delta})$ and $\operatorname{width}(f^{(1)}_{\Lambda,\delta})$. It holds

$$\begin{aligned}\operatorname{width}\left(f^{(2)}_{\Lambda,\delta}\right)&\leq\sum_{(j,\nu_j)\in T}\operatorname{width}\left(\mathrm{Id}_{\mathbb{R}}\circ\tilde{L}_{\nu_j,\delta'}\right)\\ &\leq 2\sum_{(j,\nu_j)\in T}\operatorname{width}\left(\tilde{L}_{\nu_j,\delta'}\right)\leq 18|T|,\end{aligned}\tag{D.7}$$

where we used (C.5) for the width of the σ_1-identity network and (C.37) for the width of $\tilde{L}_{\nu_j,\delta'}$. For $\operatorname{width}(f^{(1)}_{\Lambda,\delta})$ it holds

$$\begin{aligned}\operatorname{width}\left(f^{(1)}_{\Lambda,\delta}\right)&\leq\sum_{\boldsymbol{\nu}\in\Lambda}\operatorname{width}\left(\mathrm{Id}_{\mathbb{R}}\circ\widetilde{\prod}_{\delta/2,M_{\boldsymbol{\nu}}}\right)\\ &\leq 2\sum_{\boldsymbol{\nu}\in\Lambda}\operatorname{width}\left(\widetilde{\prod}_{\delta/2,M_{\boldsymbol{\nu}}}\right)\\ &\leq\sum_{\boldsymbol{\nu}\in\Lambda}10d(\Lambda)=10|\Lambda|d(\Lambda),\end{aligned}\tag{D.8}$$

again using (C.5) and (C.24) for the width of the multiplication network $\widetilde{\prod}$. Combining (D.7) and (D.8) gives

$$\operatorname{width}(f_{\Lambda,\delta})\leq 36|\Lambda|d(\Lambda),$$

where $|T|\leq|\Lambda|d(\Lambda)$ was used.

To estimate $\operatorname{size}(f_{\Lambda,\delta})$, we use (C.7) and find $\operatorname{size}(f_{\Lambda,\delta})\leq 2\operatorname{size}(f^{(1)}_{\Lambda,\delta})+2\operatorname{size}(f^{(2)}_{\Lambda,\delta})$. We calculate

$$\begin{aligned}\operatorname{size}\left(f^{(2)}_{\Lambda,\delta}\right)&=\operatorname{size}\left(\left\{\mathrm{Id}_{\mathbb{R}}\circ\tilde{L}_{\nu_j,\delta'}(y_j)\right\}_{(j,\nu_j)\in T}\right)\\ &=\sum_{(j,\nu_j)\in T}\operatorname{size}\left(\mathrm{Id}_{\mathbb{R}}\circ\tilde{L}_{\nu_j,\delta'}(y_j)\right)\\ &\leq 2m(\Lambda)d(\Lambda)\max_{(j,\nu_j)\in T}\left(\operatorname{size}\left(\mathrm{Id}_{\mathbb{R}}\right)+\operatorname{size}\left(\tilde{L}_{\nu_j,\delta'}(y_j)\right)\right)\\ &\leq 10m(\Lambda)d(\Lambda)\max_{(j,\nu_j)\in T}\operatorname{size}\left(\tilde{L}_{\nu_j,\delta'}(y_j)\right)\\ &\leq Cd(\Lambda)m(\Lambda)^2\left(m(\Lambda)+\log\left(\delta'^{-1}\right)\right)\\ &\leq Cd(\Lambda)m(\Lambda)^2\left(\log\left(d(\Lambda)\right)+d(\Lambda)\log\left(m(\Lambda)\right)+m(\Lambda)+\log\left(\delta^{-1}\right)\right).\end{aligned}\tag{D.9}$$

In (D.9) we used the NN calculus rules for the sizes of a sparse concatenation in (C.7) and a parallelization in (C.2). Furthermore, we used $|T| \leq m(\Lambda)d(\Lambda)$ at the first equality and

$$\operatorname{size}(\mathrm{Id}_{\mathbb{R}}) \leq 4 \max_{(j,\nu_j)\in T} \operatorname{depth}\left(\tilde{L}_{\nu_j,\delta'}(y_j)\right) \leq 4 \max_{(j,\nu_j)\in T} \operatorname{size}\left(\tilde{L}_{\nu_j,\delta'}(y_j)\right), \tag{D.10}$$

which follows from (C.4). At the third inequality in (D.9) we used the size bound for the univariate Legendre polynomials from (C.38).

For $\operatorname{size}(f^{(1)}_{\Lambda,\delta})$ it holds

$$\begin{aligned}
\operatorname{size}\left(f^{(1)}_{\Lambda,\delta}\right) &= \sum_{\boldsymbol{\nu}\in\Lambda} \operatorname{size}\left(\mathrm{Id}_{\mathbb{R}} \circ \widetilde{\prod}_{\delta/2,M_{\boldsymbol{\nu}}}\right) \\
&\leq 2\sum_{\boldsymbol{\nu}\in\Lambda}\left(\operatorname{size}(\mathrm{Id}_{\mathbb{R}}) + \operatorname{size}\left(\widetilde{\prod}_{\delta/2,M_{\boldsymbol{\nu}}}\right)\right) \\
&\leq 10|\Lambda| \max_{\boldsymbol{\nu}\in\Lambda} \operatorname{size}\left(\widetilde{\prod}_{\delta/2,M_{\boldsymbol{\nu}}}\right) \\
&\leq C|\Lambda| \max_{\boldsymbol{\nu}\in\Lambda} |\operatorname{supp}\boldsymbol{\nu}|\left(\log(|\operatorname{supp}\boldsymbol{\nu}|) + |\operatorname{supp}\boldsymbol{\nu}|\log(M_{\boldsymbol{\nu}}) + \log\left(\delta^{-1}\right)\right) \\
&\leq C|\Lambda|d(\Lambda)\left(\log(d(\Lambda)) + d(\Lambda)\log(m(\Lambda)) + \log\left(\delta^{-1}\right)\right).
\end{aligned} \tag{D.11}$$

In (D.11), we used the size bound for $\tilde{\prod}$ from (C.25) and the argument from (D.10). Additionally we used $M_{\boldsymbol{\nu}} = 2|\boldsymbol{\nu}|_1 + 2 \leq 4m(\Lambda)$. Combining (D.9) and (D.11) shows the size bound for $f_{\Lambda,\delta}$.

The network $f_{\Lambda,\delta}$ consists of sparse concatenations and parallelizations of the networks $\tilde{\prod}$ and $\tilde{L}_j$. Because we have $\operatorname{mpar}(\tilde{\prod}) \leq 1$ and $\operatorname{mpar}(\tilde{L}_j) \leq 1$, the NN calculus rules (C.11) and (C.1) yield $\operatorname{mpar}(f_{\Lambda,\delta}) \leq 1$. This finishes the proof.

D.2 RePU-Realization of Tensorized Legendre Polynomials

We show a result analogous to Proposition 17 for the RePU-realization of tensorized Legendre polynomials. The construction is similar to Opschoor et al. (2022a, Proposition 2.13).

Proposition 51 (σ_q-NN approximation of $L_{\boldsymbol{\nu}}$) *Consider the setting of Proposition 17. Let $q \in \mathbb{N}$, $q \geq 2$. Then there exists a σ_q-NN f_Λ such that the outputs $\{\tilde{L}_{\boldsymbol{\nu}}\}_{\nu\in\Lambda}$ of f_Λ satisfy*

$$\forall \nu \in \Lambda,\ \forall \boldsymbol{y} \in U: \qquad \tilde{L}_{\boldsymbol{\nu}}(\boldsymbol{y}) = L_{\boldsymbol{\nu}}(\boldsymbol{y}).$$

Furthermore, there exists a constant $C_q > 0$ depending only on q such that

$$\begin{aligned}
\operatorname{depth}\big(f_\Lambda\big) &\leq C_q\Big(m(\Lambda) + \log\big(d(\Lambda)\big)\Big), \\
\operatorname{width}\big(f_\Lambda\big) &\leq C_q|\Lambda|d(\Lambda), \\
\operatorname{size}\big(f_{\Lambda,\delta}\big) &\leq C_q d(\Lambda)\Big(|\Lambda| + m(\Lambda)^2\Big), \\
\operatorname{mpar}\big(f_{\Lambda,\delta}\big) &\leq C_q.
\end{aligned}$$

Proof Similar to the proof of Proposition 17 we define f_Λ as a composition of two subnetworks $f_\Lambda^{(1)}$ and $f_\Lambda^{(2)}$. It holds

$$f_\Lambda^{(2)}(\boldsymbol{y}) \coloneqq \left(\left\{ \mathrm{Id}_\mathbb{R} \circ \tilde{L}_{\nu_j}(y_j) \right\}_{(j,\nu_j)\in T} \right)$$

and

$$f_\Lambda^{(1)}\left((z_k)_{k\leq |T|}\right) = f_\Lambda^{(1)}\left(f_\Lambda^{(2)}(\boldsymbol{y})\right) \coloneqq \left(\left\{ \mathrm{Id}_\mathbb{R} \circ \widetilde{\prod} \left(\left\{ \mathrm{Id}_\mathbb{R} \circ \tilde{L}_{\nu_j}(y_j) \right\}_{j\in \operatorname{supp} \boldsymbol{\nu}} \right) \right\}_{\nu\in\Lambda} \right)$$

with T from (D.2) and $\boldsymbol{y} = (y_j)_{(j,\nu_j)\in T}$. Furthermore, we use the σ_q-NNs $\tilde{L}_j$ from Corollary 50 and $\widetilde{\prod}$ from Proposition 46. The calculations are similar to the proof of Proposition 17. It holds

$$\operatorname{depth}\left(f_\Lambda^{(2)}\right) = 1 + \max_{\substack{\boldsymbol{\nu}\in\Lambda \\ j\in\operatorname{supp}\boldsymbol{\nu}}} \operatorname{depth}\left(\tilde{L}_{\nu_j}\right) \leq C_q \max_{\substack{\boldsymbol{\nu}\in\Lambda \\ j\in\operatorname{supp}\boldsymbol{\nu}}} \nu_j \leq C_q m(\Lambda). \tag{D.12}$$

In (D.12) we used the depth bound for univariate Legendre polynomials, (C.39). Furthermore, we used $\nu_j \leq m(\Lambda)$ for all $\boldsymbol{\nu} \in \Lambda$ and $j \in \operatorname{supp} \boldsymbol{\nu}$. For the depth of $f_\Lambda^{(1)}$ it holds

$$\begin{aligned} \operatorname{depth}\left(f_\Lambda^{(1)}\right) = 1 + \max_{\boldsymbol{\nu}\in\Lambda} \operatorname{depth}\left(\widetilde{\prod}\right) &\leq 1 + C_q \max_{\boldsymbol{\nu}\in\Lambda} \log\left(|\operatorname{supp}\boldsymbol{\nu}|\right) \\ &\leq 1 + C_q \log\left(d(\Lambda)\right), \end{aligned} \tag{D.13}$$

where we used $|\operatorname{supp}\boldsymbol{\nu}| \leq d(\Lambda)$ for all $\boldsymbol{\nu} \in \Lambda$ and the depth bound for σ_q-multiplication networks from Proposition 46. Combining the two depth bounds from (D.12) and (D.13), we get

$$\operatorname{depth}(f_\Lambda) \leq C_q \left(m(\Lambda) + \log(d(\Lambda))\right).$$

For the width $\operatorname{width}(f_\Lambda)$ we calculate

$$\operatorname{width}\left(f_\Lambda^{(2)}\right) = \sum_{(j,\nu_j)\in T} \operatorname{width}\left(\mathrm{Id}_\mathbb{R} \circ \tilde{L}_{\nu_j}\right) \leq C_q \sum_{(j,\nu_j)\in T} \operatorname{width}\left(\tilde{L}_{\nu_j}\right) \leq C_q |T|, \tag{D.14}$$

where we used (C.16) for the width of a σ_q-sparse concatenation and (C.40) for the width of $\tilde{L}_{\nu_j}$. For $\operatorname{width}(f_\Lambda^{(1)})$ it holds

$$\begin{aligned} \operatorname{width}\left(f_\Lambda^{(1)}\right) \leq \sum_{\boldsymbol{\nu}\in\Lambda} \operatorname{width}\left(\mathrm{Id}_\mathbb{R} \circ \widetilde{\prod}\right) &\leq C_q \sum_{\boldsymbol{\nu}\in\Lambda} \operatorname{width}\left(\widetilde{\prod}\right) \\ &\leq C_q \sum_{\boldsymbol{\nu}\in\Lambda} d(\Lambda) = C_q |\Lambda| d(\Lambda) \end{aligned} \tag{D.15}$$

using (C.16) and (C.32) for the width of the multiplication network $\widetilde{\prod}$. Combining (D.14) and (D.15) gives

$$\operatorname{width}\left(f_\Lambda\right) \leq C_q |\Lambda| d(\Lambda),$$

where $|T| \leq |\Lambda| d(\Lambda)$ was used.

To estimate $\operatorname{size}(f_\Lambda)$, we use (C.13) and find $\operatorname{size}(f_\Lambda) \leq C_q(\operatorname{size}(f_\Lambda^{(1)}) + \operatorname{size}(f_\Lambda^{(2)}))$. We calculate

$$\begin{aligned}
\operatorname{size}\left(f_\Lambda^{(2)}\right) &= \operatorname{size}\left(\left\{\mathrm{Id}_{\mathbb{R}} \circ \tilde{L}_{\nu_j}(y_j)\right\}_{(j,\nu_j)\in T}\right) \\
&= \sum_{(j,\nu_j)\in T} \operatorname{size}\left(\mathrm{Id}_{\mathbb{R}} \circ \tilde{L}_{\nu_j}(y_j)\right) \\
&\leq C_q m(\Lambda) d(\Lambda) \max_{(j,\nu_j)\in T} \left(\operatorname{size}\left(\mathrm{Id}_{\mathbb{R}}\right) + \operatorname{size}\left(\tilde{L}_{\nu_j}(y_j)\right)\right) \\
&\leq C_q m(\Lambda) d(\Lambda) \max_{(j,\nu_j)\in T} \operatorname{size}\left(\tilde{L}_{\nu_j}(y_j)\right) \\
&\leq C_q d(\Lambda) m(\Lambda)^2. \qquad \text{(D.16)}
\end{aligned}$$

In (D.16) we used $|T| \leq m(\Lambda) d$ at the first inequality and

$$\operatorname{size}\left(\mathrm{Id}_{\mathbb{R}}\right) \leq C_q \max_{(j,\nu_j)\in T} \operatorname{depth}\left(\tilde{L}_{\nu_j}(y_j)\right) \leq C_q \max_{(j,\nu_j)\in T} \operatorname{size}\left(\tilde{L}_{\nu_j}(y_j)\right), \qquad \text{(D.17)}$$

which follows from (C.6). At the third inequality in (D.16) we used the size bound for the univariate Legendre polynomials from (C.41).

For $\operatorname{size}(f_\Lambda^{(1)})$ it holds

$$\begin{aligned}
\operatorname{size}\left(f_\Lambda^{(1)}\right) &= \sum_{\boldsymbol{\nu}\in\Lambda} \operatorname{size}\left(\mathrm{Id}_{\mathbb{R}} \circ \widetilde{\prod}\right) \\
&\leq C_q \sum_{\boldsymbol{\nu}\in\Lambda} \left(\operatorname{size}(\mathrm{Id}_{\mathbb{R}}) + \operatorname{size}\left(\widetilde{\prod}\right)\right) \\
&\leq C_q |\Lambda| \max_{\boldsymbol{\nu}\in\Lambda} \operatorname{size}\left(\widetilde{\prod}\right) \\
&\leq C_q |\Lambda| \max_{\boldsymbol{\nu}\in\Lambda} |\operatorname{supp} \boldsymbol{\nu}| \\
&\leq C_q |\Lambda| d(\Lambda). \qquad \text{(D.18)}
\end{aligned}$$

In (D.18) we used the size bound for $\tilde{\prod}$ from (C.33) and the argument from (D.17). Combining (D.16) and (D.18) shows the size bound for f_Λ.

The network f_Λ consists of sparse concatenations and parallelizations of the networks $\tilde{\prod}$ and $\tilde{L}_j$. Because we have $\operatorname{mpar}(\tilde{\prod}) \leq C_q$ and $\operatorname{mpar}(\tilde{L}_j) \leq C_q$, the NN calculus rules (C.17) and (C.1) yield $\operatorname{mpar}(f_\Lambda) \leq C_q$. This finishes the proof. ∎

D.3 Proof of Theorem 18

The following two theorems are similar to Herrmann et al. (2024, Theorem 5) and will be required for the proof of Theorem 18.

Theorem 52 *For $N, q \in \mathbb{N}$, consider the sparse FrameNet class $\boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$. Let Assumption 2 be satisfied with $r > 1$ and $t > 0$. Fix $\tau > 0$ (arbitrary small). Then there exists a constant $C > 0$ independent of N, such that there exists $\Gamma_N \in \boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$ with*

$$\sup_{a \in C^r_R(\mathcal{X})} \|\Gamma_N(a) - G_0(a)\|_{\mathcal{Y}} \leq C N^{-\min\{r-1,t\}+\tau}. \tag{D.19}$$

Theorem 53 *Consider the setting of Theorem 52. Let $\boldsymbol{\Psi}_{\mathcal{X}}$ be a Riesz basis. Additionally, let γ be as in (3.9). Fix $\tau > 0$ (arbitrary small). Then there exists a constant $C > 0$ independent of N, such that there exists $\Gamma_N \in \boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$ with*

$$\|\Gamma_N - G_0\|_{L^2(C^r_R(\mathcal{X}),\gamma;\mathcal{Y})} \leq C N^{-\min\{r-\frac{1}{2},t\}+\tau}. \tag{D.20}$$

We first show that Theorems 52 and 53 imply Theorem 18.

Proof [Proof of Theorem 18] First consider the setting of Theorem 52. Let $\tau > 0$. Then there exists a constant C independent of N and a FrameNet $\Gamma_N \in \boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$ such that for all $N \in \mathbb{N}$

$$\|\Gamma_N - G_0\|^2_{\infty,\mathrm{supp}(\gamma)} \leq \sup_{a \in C^r_R(\mathcal{X})} \|\Gamma_N(a) - G_0(a)\|^2_{\mathcal{Y}} \leq C N^{-2\min\{r-1,t\}+\tau},$$

where we used (D.19) with $\tau/2$ and $\mathrm{supp}(\gamma) \subseteq C^r_R(\mathcal{X})$ by Assumption 2.

Now consider the setting of Theorem 53. Let $\tau > 0$. Then there exists a constant C independent of N and a FrameNet $\boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$ with

$$\|\Gamma_N - G_0\|^2_{L^2(\gamma)} \leq \|\Gamma_N - G_0\|^2_{L^2(C^r_R(\mathcal{X}),\gamma;\mathcal{Y})} \leq C N^{-2\min\{r-\frac{1}{2},t\}+\tau},$$

where we used $\mathrm{supp}(\gamma) \subset C^r_R(\mathcal{X})$ (Assumption 2) and (D.20) with $\tau/2$. ■

We are left to prove Theorems 52 and 53. We need some auxiliary results.

D.3.1 Auxiliary Results

For $r > 1$, $R > 0$, $U = [-1,1]^{\mathbb{N}}$ and σ^r_R from (3.7) we define

$$u : U \to \mathcal{Y}, \quad u(\boldsymbol{y}) \coloneqq (G_0 \circ \sigma^r_R)(\boldsymbol{y}).$$

For the proofs of Theorems 52 and 53 we do a $\mathcal{Y}$-valued tensorized Legendre expansion of u in the frame $(\eta_j L_{\boldsymbol{\nu}}(\boldsymbol{y}))_{j,\boldsymbol{\nu}}$ of $L^2(U, \pi; \mathcal{Y})$, which reads

$$u(\boldsymbol{y}) = G_0(\sigma^r_R(\boldsymbol{y})) = \sum_{j \in \mathbb{N}} \sum_{\boldsymbol{\nu} \in \mathcal{F}} c_{\boldsymbol{\nu},j} \eta_j L_{\boldsymbol{\nu}}(\boldsymbol{y}) \tag{D.21}$$

with Legendre coefficients

$$c_{\boldsymbol{\nu},j} \coloneqq \int_U L_{\boldsymbol{\nu}}(\boldsymbol{y}) \, \langle u(\boldsymbol{y}), \tilde{\eta}_j \rangle_{\mathcal{Y}} \, \mathrm{d}\pi(\boldsymbol{y}). \tag{D.22}$$

Our aim is to construct the network Γ_N out of the tensorized Legendre polynomials with the "most important" contributions to the expansion. This contribution is quantified via

the Legendre coefficients $c_{\boldsymbol{\nu},j}$ in (D.22). We therefore have to examine bounds on $c_{\boldsymbol{\nu},j}$ and analyze their respective structure. Therefore consider the following order relation on multi-indices in $\mathcal{F}$ from (3.11). For $\boldsymbol{\mu},\boldsymbol{\nu} \in \mathcal{F}$ we write $\boldsymbol{\mu} \leq \boldsymbol{\nu}$ if and only if $\mu_j \leq \nu_j$ for all $j \in \mathbb{N}$. We call a set $\Lambda \subset \mathcal{F}$ *downward closed* if and only if $\boldsymbol{\nu} \in \mathcal{F}$ implies $\boldsymbol{\mu} \in \mathcal{F}$ for all $\boldsymbol{\mu} \leq \boldsymbol{\nu}$. Furthermore, for $\boldsymbol{\nu} \in \mathcal{F}$, define

$$\omega_{\boldsymbol{\nu}} \coloneqq \prod_{j=1}^{\infty}(1+2\nu_j).$$

The following theorem is a special case of Zech (2018, Theorem 2.2.10). The formulation is similar to Herrmann et al. (2024, Theorem 4).

Theorem 54 (Herrmann et al. 2024, Theorem 4) *Let Assumption 2 be satisfied with $r > 1$ and $t > 0$. Fix $\tau > 0$, $p \in (\frac{1}{r}, 1]$ and $t' \in [0,t]$. Consider $\mathcal{F}$ from (3.11), and let $\pi = \otimes_{j\in\mathbb{N}} \frac{\lambda}{2}$ be the infinite product (probability) measure on $U = [-1,1]^{\mathbb{N}}$, where λ denotes the Lebesgue measure on $[-1,1]$. Then there exists $C > 0$ and a sequence $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$ of positive numbers such that*

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

$$\omega_{\boldsymbol{\nu}}^{\tau} \left\| \int_U L_{\boldsymbol{\nu}}(\boldsymbol{y}) u(\boldsymbol{y}) \, \mathrm{d}\pi(\boldsymbol{y}) \right\|_{\mathcal{Y}^{t'}} \leq C a_{\boldsymbol{\nu}},$$

(ii) there exists an enumeration $(\boldsymbol{\nu_i})_{i\in\mathbb{N}}$ of $\mathcal{F}$ such that $(a_{\boldsymbol{\nu_i}})_{i\in\mathbb{N}}$ is monotonically decreasing, the set $\Lambda_N \coloneqq \{\boldsymbol{\nu_i} : i \leq N\} \subseteq \mathcal{F}$ is downward closed for each $N \in \mathbb{N}$, and additionally

$$m(\Lambda_N) \coloneqq \max_{i=1,\dots,N} |\boldsymbol{\nu_i}| = \mathcal{O}\left(\log(|\Lambda_N|)\right), \tag{D.23}$$

$$d(\Lambda_N) \coloneqq \max_{i=1,\dots,N} |\operatorname{supp} \boldsymbol{\nu_i}| = o\left(\log(|\Lambda_N|)\right) \tag{D.24}$$

for $N \to \infty$,

(iii) the following expansion holds with absolute and uniform convergence:

$$\forall \boldsymbol{y} \in U: \qquad u(\boldsymbol{y}) = \sum_{\boldsymbol{\nu}\in\mathcal{F}} L_{\boldsymbol{\nu}}(\boldsymbol{y}) \int_U L_{\boldsymbol{\nu}}(\boldsymbol{x}) u(\boldsymbol{x}) \, \mathrm{d}\pi(\boldsymbol{x}) \in \mathcal{Y}^{t'}.$$

The following proposition reformulates Theorem 54 (i) into a bound for $c_{\boldsymbol{\nu},j}$. It was shown in Herrmann et al. (2024, Proposition 2). Recall that θ_j denote the weights to define the spaces $\mathcal{Y}^{t'}$, $t' > 0$, see Definition 14.

Proposition 55 (Herrmann et al. 2024, Proposition 2) *Consider the setting of Theorem 54. Then for each $\boldsymbol{\nu} \in \mathcal{F}$*

$$\omega_{\boldsymbol{\nu}}^{2\tau} \sum_{j\in\mathbb{N}} \theta_j^{-2t'} c_{\boldsymbol{\nu},j}^2 \leq C^2 a_{\boldsymbol{\nu}}^2.$$

Proposition 55 gives decay of the coefficients $c_{\boldsymbol{\nu},j}$ in both j and $\boldsymbol{\nu}$. Since $\theta_j = \mathcal{O}(j^{-1+\tau})$ for all $\tau > 0$ we have $c_{\boldsymbol{\nu},j}^2 = \mathcal{O}(j^{-1-2t'+\tilde{\tau}})$ for $\tilde{\tau} < 2\tau t'$ and every $\boldsymbol{\nu} \in \Lambda_N$. Furthermore, since $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}} \in \ell^p(\mathcal{F})$ the Legendre coefficients $c_{\boldsymbol{\nu},j}$ decay algebraically in $\boldsymbol{\nu}$. We continue with a technical lemma, which was shown in Herrmann et al. (2024, Lemma 4).

Lemma 56 (Herrmann et al. 2024, Lemma 4) *Let $\alpha > 1$, $\beta > 0$ and assume two sequences $(a_i)_{i\in\mathbb{N}}$ and $(d_j)_{j\in\mathbb{N}}$ in $\mathbb{R}$ with $a_i \lesssim i^{-\alpha}$ and $d_j \lesssim j^{-\beta}$ for all $i, j \in \mathbb{N}$. Additionally assume that $(d_j)_{j\in\mathbb{N}}$is monotonically decreasing. Suppose that there exists a constant $C < \infty$ such that the sequence $(c_{i,j})_{i,j\in\mathbb{N}}$ satisfies*

$$\forall i \in \mathbb{N}: \quad \sum_{j\in\mathbb{N}} c_{i,j}^2 d_j^{-2} \leq C^2 a_i^2.$$

Then for every $\tau > 0$

(i) for all $N \in \mathbb{N}$ there exists $(m_i)_{i\in\mathbb{N}} \subseteq \mathbb{N}_0^{\mathbb{N}}$ monotonically decreasing s.t. $\sum_{i\in\mathbb{N}} m_i \leq N$ and

$$\sum_{i\in\mathbb{N}} \left(\sum_{j>m_i} c_{i,j}^2 \right)^{\frac{1}{2}} \lesssim N^{-\min\{\alpha-1,\beta\}+\tau},$$

(ii) for all $N \in \mathbb{N}$ there exists $(m_i)_{i\in\mathbb{N}} \subseteq \mathbb{N}_0^{\mathbb{N}}$ monotonically decreasing s.t. $\sum_{i\in\mathbb{N}} m_i \leq N$ and

$$\left(\sum_{i\in\mathbb{N}} \sum_{j>m_i} c_{i,j}^2 \right)^{\frac{1}{2}} \lesssim N^{-\min\left\{\alpha-\frac{1}{2},\beta\right\}+\tau}.$$

In the following, we use Lemma 56 to get a decay property for the Legendre coefficients $c_{\boldsymbol{\nu_i},j}$ with the enumeration $\boldsymbol{\nu_i}$ of Λ_N from Theorem 54. The sequence $\boldsymbol{m} = (m_i)_{i\in\mathbb{N}}$ quantifies which coefficients of the Legendre expansion are "important" and are therefore used to define the surrogate Γ_N.

We first show that Theorem 54 yields sufficient decay on the Legendre coefficients $c_{\boldsymbol{\nu_i},j}$ s.t. the assumptions of Lemma 56 are satisfied.

Lemma 57 *Consider the setting of Theorem 54. Let $\tilde{\tau} > 0$ such that $1/p > r - \tilde{\tau}/2$. Then the assumptions of Lemma 56 are fulfilled for $\alpha = r - \tilde{\tau}/2$, $\beta = t - \tilde{\tau}/2$, $a_i = a_{\boldsymbol{\nu_i}}$, $d_j = \theta_j^{t'}$ and $c_{i,j} = \omega_{\boldsymbol{\nu_i}}^{1/2} c_{\boldsymbol{\nu_i},j}$ for $i, j \in \mathbb{N}$.*

Proof Proposition 55 with $\tau = \frac{1}{2}$ gives

$$\omega_{\boldsymbol{\nu_i}}^{\frac{1}{2}} \left(\sum_{j\in\mathbb{N}} \theta_j^{-2t'} c_{\boldsymbol{\nu_i},j}^2 \right)^{\frac{1}{2}} = \mathcal{O}(a_{\boldsymbol{\nu_i}}) = \mathcal{O}\left(i^{-r+\frac{\tilde{\tau}}{2}}\right). \tag{D.25}$$

The last equality in (D.25) holds because $i a_{\boldsymbol{\nu_i}}^p \leq \sum_{j\in\mathbb{N}} a_{\boldsymbol{\nu_j}}^p < \infty$ (since $a_{\boldsymbol{\nu_i}}$ is monotonically decreasing) implies $a_{\boldsymbol{\nu_i}} = \mathcal{O}(i^{-1/p}) = \mathcal{O}(i^{-r+\tilde{\tau}/2})$. Since $(\theta_j^{t'})_{j\in\mathbb{N}} \in \ell^{1/(t-\tilde{\tau}/2)}$ (see Definition 14) it holds

$$\theta_j^{t'} = \mathcal{O}(j^{-t+\tilde{\tau}/2})$$

with the same argument. ■

D.3.2 Proofs of Theorems 52 and 53

The proof of Theorems 52 and 53 is similar to Herrmann et al. (2024, Sections 4.2-4.4, Proofs of Theorems 1,2 and 5).

Proof [Proof of Theorem 52] Let $(a_{\boldsymbol{\nu}})_{\boldsymbol{\nu}\in\mathcal{F}}$ be the enumeration $(\boldsymbol{\nu_i})_{i\in\mathbb{N}}$ from Theorem 54, where we use the case $\tau = \frac{1}{2}$. Therefore $(a_{\boldsymbol{\nu_i}})_{i\in\mathbb{N}}$ is monotonically decreasing and belongs to ℓ^p with $p \in (\frac{1}{r}, 1]$. We further fix $\tilde{\tau} > 0$ and demand $\frac{1}{p} > r - \frac{\tilde{\tau}}{2}$. Fix $\tilde{N} \in \mathbb{N}$ and set $\Lambda_{\tilde{N}} \coloneqq \{\boldsymbol{\nu_j} : j \leq \tilde{N}\} \subset \mathcal{F}$, which is downward closed by Theorem 54. Now we approximate the tensorized Legendre polynomials $L_{\boldsymbol{\nu}}$ on the index set $\Lambda_{\tilde{N}}$. Let $\rho \in (0, \frac{1}{2})$. In the ReLU case, Proposition 17 gives a NN $f_{\Lambda_{\tilde{N}},\rho}$ with outputs $\{\tilde{L}_{\boldsymbol{\nu},\rho}\}_{\boldsymbol{\nu}\in\Lambda_{\tilde{N}}}$ s.t.

$$\sup_{\boldsymbol{y}\in U}\max_{\boldsymbol{\nu}\in\Lambda_{\tilde{N}}}\left|L_{\boldsymbol{\nu}}(\boldsymbol{y}) - \tilde{L}_{\boldsymbol{\nu},\rho}(\boldsymbol{y})\right| \leq \rho.$$

Using $|\Lambda_{\tilde{N}}| = \tilde{N}$, (D.23) and (D.24), it holds for $\tilde{N} \geq 2$

$$\begin{aligned}
\operatorname{depth}(f_{\Lambda_{\tilde{N}},\rho}) &= \mathcal{O}\left(\log(\tilde{N})^2\log(\log(\tilde{N}))^2 + \log(\tilde{N})\log\left(\rho^{-1}\right)\right),\\
\operatorname{width}(f_{\Lambda_{\tilde{N}},\rho}) &= \mathcal{O}(\tilde{N}\log(\tilde{N})),\\
\operatorname{size}(f_{\Lambda_{\tilde{N}},\rho}) &= \mathcal{O}\left(\tilde{N}\log(\tilde{N})^2\log(\log(\tilde{N})) + \tilde{N}\log(\tilde{N})\log\left(\rho^{-1}\right)\right),\\
\operatorname{mpar}(f_{\Lambda_{\tilde{N}},\rho}) &= 1.
\end{aligned}$$

The constants hidden in $\mathcal{O}(\,.\,)$ are independent of $\tilde{N}$ and ρ. For $\tilde{N} \in \mathbb{N}$, set the accuracy $\rho \coloneqq \tilde{N}^{-\min\{r-\frac{1}{2},t\}}$. Then it holds

$$\begin{aligned}
\operatorname{depth}(f_{\Lambda_{\tilde{N}},\rho}) &= \mathcal{O}\left(\log(\tilde{N})^2\log(\log(\tilde{N}))^2\right),\\
\operatorname{size}(f_{\Lambda_{\tilde{N}},\rho}) &= \mathcal{O}\left(\tilde{N}\log(\tilde{N})^2\log(\log(\tilde{N}))\right).
\end{aligned}$$

Proposition 51 shows that the ReLU bounds also hold for the RePU-case.

By Lemma 57 the assumptions of Lemma 56 are satisfied. Applying Lemma 56 (i) with $\alpha \coloneqq r - \tilde{\tau}/2$ and $\beta \coloneqq t - \tilde{\tau}/2$ gives a sequence $(m_i)_{i\in\mathbb{N}} \subset \mathbb{N}_0^{\mathbb{N}}$ such that $\sum_{i\in\mathbb{N}} m_i \leq \tilde{N}$ and

$$\sum_{i\in\mathbb{N}} \omega_{\boldsymbol{\nu_i}}^{\frac{1}{2}}\left(\sum_{j>m_i} c_{\boldsymbol{\nu_i},j}^2\right)^{\frac{1}{2}} \leq C\tilde{N}^{-\min\{r-1,t\}+\tilde{\tau}}. \tag{D.26}$$

We now define

$$\left(\tilde{\gamma}_{\tilde{N},j}\right) \coloneqq \sum_{\{i\in\mathbb{N}:m_i\geq j\}} \tilde{L}_{\boldsymbol{\nu_i},\rho}(\boldsymbol{y})c_{\boldsymbol{\nu_i},j} \tag{D.27}$$

for $j \in \mathbb{N}$, where empty sums are set to zero. Recall the uniform distribution π on $U = [-1,1]^{\mathbb{N}}$, see Example 3. With $\tilde{\gamma}_{\tilde{N}} = (\tilde{\gamma}_{\tilde{N},j})_{j\in\mathbb{N}}$ it holds

$$\begin{aligned}
\|G_0 \circ \sigma_R^r(\boldsymbol{y}) - \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}}(\boldsymbol{y})\|_{\mathcal{Y}} &= \left\| \sum_{i,j\in\mathbb{N}} c_{\boldsymbol{\nu_i},j} L_{\boldsymbol{\nu_i}}(\boldsymbol{y})\eta_j - \sum_{i\in\mathbb{N}} \sum_{j\le m_i} c_{\boldsymbol{\nu_i},j} \tilde{L}_{\boldsymbol{\nu_i},\rho}(\boldsymbol{y})\eta_j \right\|_{\mathcal{Y}} \\
&\le \left\| \sum_{i\in\mathbb{N}} L_{\boldsymbol{\nu_i}}(\boldsymbol{y}) \sum_{j>m_i} c_{\boldsymbol{\nu_i},j}\eta_j \right\|_{\mathcal{Y}} + \left\| \sum_{i\in\mathbb{N}} (L_{\boldsymbol{\nu_i}}(\boldsymbol{y}) - \tilde{L}_{\boldsymbol{\nu_i},\rho}(\boldsymbol{y})) \sum_{j\le m_i} c_{\boldsymbol{\nu_i},j}\eta_j \right\|_{\mathcal{Y}} \\
&\le \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \sum_{i\in\mathbb{N}} \underbrace{\|L_{\boldsymbol{\nu_i}}\|_{\infty,\pi}}_{\le \omega_{\boldsymbol{\nu_i}}^{\frac{1}{2}}} \left(\sum_{j>m_i} c_{\boldsymbol{\nu_i},j}^2 \right)^{\frac{1}{2}} + \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \rho \sum_{i\in\mathbb{N}} \left(\sum_{j\le m_i} c_{\boldsymbol{\nu_i},j}^2 \right)^{\frac{1}{2}} \\
&\le \tilde{C} \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \tilde{N}^{-\min\{r-1,t\}+\tilde{\tau}} + \tilde{C} \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \rho \le \tilde{C} \tilde{N}^{-\min\{r-1,t\}+\tilde{\tau}}
\end{aligned} \tag{D.28}$$

for all $\boldsymbol{y} \in U$. In (D.28) we used the definition of $\mathcal{D}_{\mathcal{Y}}$, (D.27) and (D.21) at the first equality. Furthermore, we used (D.26), the definition of ρ and

$$\begin{aligned}
\sum_{i\in\mathbb{N}} \left(\sum_{j\le m_i} c_{\boldsymbol{\nu_i},j}^2 \right)^{\frac{1}{2}} &\le \sum_{i\in\mathbb{N}} \left(\sum_{j\in\mathbb{N}} c_{\boldsymbol{\nu_i},j}^2 \right)^{\frac{1}{2}} \\
&\le \tilde{C} \sum_{i\in\mathbb{N}} \omega_{\boldsymbol{\nu_i}}^{\frac{1}{2}} \left(\sum_{j\in\mathbb{N}} \theta_j^{-2t'} c_{\boldsymbol{\nu_i},j}^2 \right)^{\frac{1}{2}} \\
&\le \tilde{C} \sum_{i\in\mathbb{N}} a_{\boldsymbol{\nu_i}} \le \tilde{C} \sum_{i\in\mathbb{N}} i^{-r+\tilde{\tau}/2} \le \tilde{C}
\end{aligned} \tag{D.29}$$

at the second-to-last inequality. We changed the constants $\tilde{C}$ from line to line in (D.28) and (D.29). The last line of (D.28) shows why the RePU-case does not improve the approximation property qualitatively. In the RePU-case, Proposition 51 gives a σ_q-NN f_Λ exactly realizing the tensorized Legendre polynomials, i.e. the case $\rho = 0$ from above. Therefore the second summand in the last line of (D.28) vanishes. This does not improve the approximation rate due to the first summand. This part depends on the summability properties of the Legendre coefficients $c_{\boldsymbol{\nu_i},j}$ following Assumption 2 and is therefore independent of the activation function σ.

Now we argue similarly to Herrmann et al. (2024, Proof of Theorem 1). Consider the scaling S_r from (3.5). It holds

$$S_r \circ \mathcal{E}_{\mathcal{X}}(a) \in U \qquad \forall a \in C_R^r(\mathcal{X}), \tag{D.30}$$

because of (3.8) and (3.5). We define $\tilde{\Gamma}_{\tilde{N}} \coloneqq \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} \circ S_r \circ \mathcal{E}_{\mathcal{X}}$ and calculate

$$\begin{aligned}
&\sup_{a \in C_R^r(\mathcal{X})} \left\| G_0(a) - \tilde{\Gamma}_{\tilde{N}}(a) \right\|_{\mathcal{Y}} = \sup_{a \in C_R^r(\mathcal{X})} \left\| G_0(a) - \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} \circ S_r \circ \mathcal{E}_{\mathcal{X}}(a) \right\|_{\mathcal{Y}} \\
&= \sup_{\{\boldsymbol{y} \in U \colon \sigma_R^r(\boldsymbol{y}) \in C_R^r(\mathcal{X})\}} \left\| G_0 \circ \sigma_R^r(\boldsymbol{y}) - \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} \circ S_r \circ \mathcal{E}_{\mathcal{X}} \circ \sigma_R^r(\boldsymbol{y}) \right\|_{\mathcal{Y}} \\
&\leq \sup_{\boldsymbol{y} \in U} \left\| G_0 \circ \sigma_R^r(\boldsymbol{y}) - \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}}(\boldsymbol{y}) \right\|_{\mathcal{Y}} \leq \tilde{C} \tilde{N}^{-\min\{r-1,t\}+\tilde{\tau}}, \qquad \text{(D.31)}
\end{aligned}$$

where we used (D.30) and (D.28).

In order to finish the proof of Theorem 52, we relate $\tilde{N}$ to N and show $\Gamma_N \coloneqq \tilde{\Gamma}_{\tilde{N}} \in \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$, i.e. we show that the approximation networks we constructed have the desired sparse structure. We simultaneously prove the ReLU- and RePU-case.

In order to analyze the NNs $\tilde{\gamma}_{\tilde{N}}$ from (D.27), we specify its structure. We set $n_j = |\{m_i \geq j\}|$ and define

$$\tilde{\gamma}_{\tilde{N}} = \left(\left\{ \Sigma_{n_j} \left(\left\{ \mathrm{Id}_{\mathbb{R}} \circ SM_{c_{\boldsymbol{\nu_i}},j} \circ \tilde{L}_{\boldsymbol{\nu_i},\rho} \right\}_{m_i \geq j} \right) \right\}_{j \in \mathbb{N}} \right). \qquad \text{(D.32)}$$

The round brackets in (D.32) denote a parallelization. The networks $SM_{c_{\boldsymbol{\nu_i}},j}$ denote the scalar multiplication networks from Definition 42. Furthermore, Σ_{n_j} denotes the summation network from Definition 41 and we use the identity networks $\mathrm{Id}_{\mathbb{R}}$ from Lemma 37 or 38 to synchronize the depth. Using the respective bounds for the summation and scalar multiplication networks and the NN calculus for parallelization and sparse concatenation we get

$$\begin{aligned}
\operatorname{depth}(\tilde{\gamma}_{\tilde{N}}) &\leq 2 + \max_{i,j \in \mathbb{N}^2,\, m_i \geq j} \left(\operatorname{depth}\left(\tilde{L}_{\boldsymbol{\nu_i},\rho} \right) + \operatorname{depth}\left(SM_{c_{\boldsymbol{\nu_i}},j} \right) \right) + \max_{j \in \mathbb{N}} \operatorname{depth}\left(\Sigma_{n_j} \right) \\
&\leq 3 + \mathcal{O}(\log(\tilde{N})^2 \log(\log(\tilde{N}))) + \max_{i,j \in \mathbb{N}^2,\, m_i \geq j} C_q \log\left(|c_{\boldsymbol{\nu_i},j}| \right) + 0 \\
&= \mathcal{O}(\log(\tilde{N})^2 \log(\log(\tilde{N}))), \qquad \text{(D.33)} \\
\operatorname{width}(\tilde{\gamma}_{\tilde{N}}) &\leq \max\Bigg\{ \sum_{j \in \mathbb{N}} \sum_{m_i \geq j} \operatorname{width}\left(\tilde{L}_{\boldsymbol{\nu_i},\rho} \right), \sum_{j \in \mathbb{N}} \sum_{m_i \geq j} \operatorname{width}\left(SM_{c_{\boldsymbol{\nu_i}},j} \right), \\
&\qquad \sum_{j \in \mathbb{N}} \sum_{m_i \geq j} \operatorname{width}\left(\mathrm{Id}_{\mathbb{R}} \right), \sum_{j \in \mathbb{N}} \operatorname{width}\left(\Sigma_{n_j} \right) \Bigg\} \\
&\leq C_q \max\left\{ O(\tilde{N}\log(\tilde{N})), \sum_{j \in \mathbb{N}} \sum_{m_i \geq j} C_q, \sum_{j \in \mathbb{N}} n_j \right\} = O(\tilde{N}\log(\tilde{N})),
\end{aligned}$$

as well as

$$\begin{aligned}
\operatorname{size}(\tilde{\gamma}_{\tilde{N}}) &\leq C_q \sum_{j\in\mathbb{N}} \sum_{m_i\geq j} \left(\operatorname{size}\left(\tilde{L}_{\boldsymbol{\nu_i},\rho}\right) + \operatorname{size}\left(SM_{c_{\boldsymbol{\nu_i},j}}\right) + \operatorname{size}\left(\mathrm{Id}_{\mathbb{R}}\right)\right) + C_q \sum_{j\in\mathbb{N}} \operatorname{size}\left(\Sigma_{n_j}\right) \\
&= \mathcal{O}\left(\tilde{N}\log(\tilde{N})^2\log(\log(\tilde{N}))\right) + \sum_{j\in\mathbb{N}} \left[\left(\sum_{m_i\geq j} C_q \log\left(|c_{\boldsymbol{\nu_i},j}|\right)\right) + C_q n_j\right] \\
&\leq \mathcal{O}\left(\tilde{N}\log(\tilde{N})^2\log(\log(\tilde{N}))\right) + C_q \sum_{j\in\mathbb{N}} \sum_{m_i\geq j} 1 \\
&= \mathcal{O}\left(\tilde{N}\log(\tilde{N})^2\log(\log(\tilde{N}))\right), \qquad \text{(D.34)} \\
\operatorname{mpar}(\tilde{\gamma}_{\tilde{N}}) &\leq C_q.
\end{aligned}$$

In (D.33)–(D.34) we used $\log\left(|c_{\boldsymbol{\nu_i},j}|\right) \leq C_q$ for all $i,j \in \mathbb{N}$ independent of n. Furthermore, we used

$$\sum_{j\in\mathbb{N}} n_j = \sum_{j\in\mathbb{N}} \sum_{m_i\geq j} 1 = \sum_{i\in\mathbb{N}} \sum_{j\leq m_i} 1 = \sum_{i\in\mathbb{N}} m_i \leq \tilde{N}.$$

To get rid of the logarithmic terms, we define $N = N(\tilde{N}) \coloneqq \max\{1, \tilde{N}\log(\tilde{N})^3\}$ and obtain a NN $\gamma_N = \tilde{\gamma}_{\tilde{N}}$ with

$$\begin{aligned}
\operatorname{depth}(\gamma_N) &= \mathcal{O}(\log(N)), \\
\operatorname{width}(\gamma_N) &= \mathcal{O}(N), \\
\operatorname{size}(\gamma_N) &= N, \\
\operatorname{mpar}(\gamma_N) &\leq C_q
\end{aligned}$$

and error less than

$$\tilde{C}\tilde{N}^{-\min\{r-1,t\}+\tilde{\tau}} = \tilde{C}\tilde{N}^{-\kappa} \leq \tilde{C}(3\kappa/\tilde{\tau})^{3\kappa} N^{-\kappa+\tilde{\tau}} \coloneqq C N^{-\min\{r-1,t\}+\tau}. \qquad \text{(D.35)}$$

Per definition $\operatorname{depth}(\gamma_N) = \mathcal{O}(\log(N))$ and $\operatorname{width}(\gamma_N) = \mathcal{O}(N)$ yields constants $\tilde{C}_L$, $\tilde{C}_p$ and $N_1, N_2 \in \mathbb{N}$ s.t.

$$\begin{aligned}
\operatorname{depth}(\gamma_N) &\leq \tilde{C}_L \max\{1, \log(N)\}, \qquad N \geq N_1, \\
\operatorname{width}(\gamma_N) &\leq \tilde{C}_p N, \qquad N \geq N_2.
\end{aligned}$$

Setting $C_L = \max\{\tilde{C}_L, \max_{N=2,\dots,N_1-1} \operatorname{depth}(\gamma_N)/\log(2)\}$ and
$C_p = \max\{\tilde{C}_p, \max_{N=1,\dots,N_2-1} \operatorname{width}(\gamma_N)\}$ shows

$$\begin{aligned}
\operatorname{depth}(\gamma_N) &\leq C_L \log(N), \qquad N \in \mathbb{N}, \\
\operatorname{width}(\gamma_N) &\leq C_p N, \qquad N \in \mathbb{N}.
\end{aligned}$$

In order to show $\Gamma_N \coloneqq \mathcal{D}_{\mathcal{Y}} \circ \gamma_N \circ S_r \circ \mathcal{E}_{\mathcal{X}} \in \boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_q, N)$, we are left to show that the maximum Euclidean norm $\|\cdot\|_2$ of γ_N in U is independent of N. It holds for all $\boldsymbol{y} \in U$ that

$\mathcal{D}_{\mathcal{X}} \circ S_r^{-1}(\boldsymbol{y}) \in C_R^r(\mathcal{X})$. We get

$$\begin{aligned}
\sup_{\boldsymbol{y}\in U} \|\gamma_N(\boldsymbol{y})\|_2 &= \sup_{\boldsymbol{y}\in U} \|\mathcal{E}_{\mathcal{Y}} \circ \Gamma_N \circ \mathcal{D}_{\mathcal{X}} \circ S_r^{-1}(\boldsymbol{y})\|_2 \\
&\leq \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \sup_{a\in C_R^r(\mathcal{X})} \|\Gamma_N(a)\|_{\mathcal{Y}} \\
&= \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \sup_{a\in C_R^r(\mathcal{X})} \|\Gamma_N(a) - G_0(a) + G_0(a)\|_{\mathcal{Y}} \\
&\leq \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \sup_{a\in C_R^r(\mathcal{X})} (\|\Gamma_N(a) - G_0(a)\|_{\mathcal{Y}} + c\|G_0(a)\|_{\mathcal{Y}^t}) \\
&\leq \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}}(C + cC_{G_0}) =: B, \qquad \text{(D.36)}
\end{aligned}$$

where $\Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}}$ denotes the upper frame bound of $\boldsymbol{\Psi}_{\mathcal{Y}}$ and $c = \theta_0^t$, see Definition 14. In (D.36) we used Assumption 2 and the approximation error from (D.35). Thus $\Gamma_N \in \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$ for all $N \in \mathbb{N}$, where we set $C_s = 1$. Using $\mathrm{supp}(\gamma) \subset C_R^r(\mathcal{X})$, see Assumption 2, in (D.31) finalizes the proof of Theorem 52. ■

Proof [Proof of Theorem 53] By Lemma 57 the assumptions of Lemma 56 are satisfied. Applying Lemma 56 (ii) with $\alpha \coloneqq r - \tilde{\tau}/2$ and $\beta \coloneqq t - \tilde{\tau}/2$ gives a sequence $(m_i)_{i\in\mathbb{N}} \subset \mathbb{N}_0^{\mathbb{N}}$ such that $\sum_{i\in\mathbb{N}} m_i \leq \tilde{N}$ and

$$\left(\sum_{i\in\mathbb{N}} \omega_{\boldsymbol{\nu}_i} \sum_{j>m_i} c_{\boldsymbol{\nu_i},j}^2\right)^{\frac{1}{2}} \leq \tilde{C}\tilde{N}^{-\min\left\{r-\frac{1}{2},t\right\}+\tilde{\tau}}. \qquad \text{(D.37)}$$

Define $\tilde{\gamma}_{\tilde{N}} = (\tilde{\gamma}_{\tilde{N},j})_{j\in\mathbb{N}}$ for all $y \in U$ with $\tilde{\gamma}_{\tilde{N},j}$ as in (D.27). Then it holds

$$\begin{aligned}
\|G_0 \circ \sigma_R^r - \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}}\|_{L^2(U,\pi;\mathcal{Y})} &\leq \left\|\sum_{i\in\mathbb{N}} \sum_{j>m_i} c_{\boldsymbol{\nu_i},j} L_{\boldsymbol{\nu_i}} \eta_j\right\|_{L^2(U,\pi;\mathcal{Y})} \\
&\quad + \left\|\sum_{i\in\mathbb{N}} \sum_{j\leq m_i} c_{\boldsymbol{\nu_i},j}\eta_j \left(L_{\boldsymbol{\nu_i}} - \tilde{L}_{\boldsymbol{\nu_i},\rho}\right)\right\|_{L^2(U,\pi;\mathcal{Y})} \\
&\leq \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \left(\sum_{i\in\mathbb{N}} \underbrace{\|L_{\boldsymbol{\nu_i}}\|_{\infty,\pi}^2}_{\leq \omega_{\boldsymbol{\nu_i}}} \sum_{j>m_i} c_{\boldsymbol{\nu_i},j}^2\right)^{\frac{1}{2}} + \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}}\rho \left(\sum_{i\in\mathbb{N}} \sum_{j\leq m_i} c_{\boldsymbol{\nu_i},j}^2\right)^{\frac{1}{2}} \\
&\leq \tilde{C}\Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}}\tilde{N}^{-\min\{r-\frac{1}{2},t\}+\tilde{\tau}} + \tilde{C}\Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}}\rho \leq \tilde{C}\tilde{N}^{-\min\{r-\frac{1}{2},t\}+\tilde{\tau}}. \qquad \text{(D.38)}
\end{aligned}$$

In (D.38) we used the definition of $\mathcal{D}_{\mathcal{Y}}$, (D.27) and (D.21) at the first inequality. Additionally we used that $(L_{\boldsymbol{\nu}}\eta_j)_{\boldsymbol{\nu},j}$ is a frame of $L^2(U,\pi;\mathcal{Y})$ at the second inequality. Finally we used (D.37), the definition of ρ and an argument similar to (D.29) at the second-to-last inequality. Note that again we changed the constants $\tilde{C}$ from line to line in (D.38).

Since $\boldsymbol{\Psi}_{\mathcal{X}}$ is a Riesz basis, we have (see Section 3.1.2 and 3.5)

$$C_R^r(\mathcal{X}) = \{\sigma_R^r(\boldsymbol{y}), \boldsymbol{y} \in U\} \quad \text{and} \quad \mathcal{E}_{\mathcal{X}} \circ \sigma_R^r(\boldsymbol{y}) = S_r^{-1}(\boldsymbol{y}). \qquad \text{(D.39)}$$

With $\tilde{\Gamma}_{\tilde{N}} \coloneqq \mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} \circ S_r \circ \mathcal{E}_{\mathcal{X}}$ we calculate

$$\begin{aligned}\left\|\tilde{\Gamma}_{\tilde{N}} - G_0\right\|_{L^2(C_R^r(\mathcal{X}),(\sigma_R^r)_\#\pi;\mathcal{Y})} &= \left\|\mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} \circ S_r \circ \mathcal{E}_{\mathcal{X}} - G_0\right\|_{L^2(C_R^r(\mathcal{X}),(\sigma_R^r)_\#\pi;\mathcal{Y})} \\ &= \left\|\mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} \circ S_r \circ \mathcal{E}_{\mathcal{X}} \circ \sigma_R^r - G_0 \circ \sigma_R^r\right\|_{L^2(U,\pi;\mathcal{Y})} \\ &= \left\|\mathcal{D}_{\mathcal{Y}} \circ \tilde{\gamma}_{\tilde{N}} - G_0 \circ \sigma_R^r\right\|_{L^2(U,\pi;\mathcal{Y})} \leq \tilde{C}\tilde{N}^{-\min\left\{r-\frac{1}{2},t\right\}+\tau},\end{aligned}$$

where we used (D.39) and (D.38). Defining $N = N(\tilde{N}) \coloneqq \max\{1, \tilde{N}\log(\tilde{N})^3\}$ we can proceed similar to the proof of Theorem 52 from (D.31) on. The reason this works is that the NNs $\tilde{\gamma}_{\tilde{N}}$ are defined in the same way in the L^2- and the L^∞-case (only the sequence $\boldsymbol{m}$ changes, but not its properties). This shows $\Gamma_N \coloneqq \tilde{\Gamma}_{\tilde{N}} \in \boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$ for all $N \in \mathbb{N}$ and thus finishes the proof of Theorem 53. ∎

D.4 Proof of Lemma 20

The arguments in the following proof are based on entropy bounds for feedforward neural network classes, first established in Schmidt-Hieber (2020a, Proof of Lemma 5).

Define the supremum norm $\|\cdot\|_{\infty,\infty}$ on $\boldsymbol{g}_{\mathrm{FN}}$ as

$$\|g\|_{\infty,\infty} \coloneqq \sup_{\boldsymbol{y}\in\mathbb{R}^{p_0}} \|g(\boldsymbol{y})\|_\infty, \quad g \in \boldsymbol{g}_{\mathrm{FN}}, \tag{D.40}$$

where $\|\cdot\|_\infty$ denotes the maximum norm in $\mathbb{R}^n$. Then Petersen et al. (2021, Proposition 3.5) shows that $(\boldsymbol{g}_{\mathrm{FN}}, \|\cdot\|_{\infty,\infty})$ is compact. Since the map $i : \boldsymbol{g}_{\mathrm{FN}} \to \boldsymbol{G}_{\mathrm{FN}}$, $g \to G = \mathcal{D}_{\mathcal{Y}} \circ g \circ \mathcal{E}_{\mathcal{X}}$ is linear, also $(\boldsymbol{G}_{\mathrm{FN}}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)})$ and hence $(\boldsymbol{G}_{\mathrm{FN}}, \|\cdot\|_n)$ is compact. We now show the entropy bounds for $\boldsymbol{G}_{\mathrm{FN}}$.

Step 1. Recall $\mathrm{depth}(g) \leq L$, $\mathrm{depth}(g) \leq p$, $\mathrm{size}(g) \leq s$ and $\mathrm{mpar}(g) \leq M$ for $g \in \boldsymbol{g}_{\mathrm{FN}}$. We first estimate the entropy $H(\boldsymbol{G}_{\mathrm{FN}}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)}, \delta)$ against the respective entropy of $\boldsymbol{g}_{\mathrm{FN}}$. For $G, G' \in \boldsymbol{G}_{\mathrm{FN}}$ and $g, g' \in \boldsymbol{G}_{\mathrm{FN}}$ with $G = \mathcal{D}_{\mathcal{Y}} \circ g \circ S_r \circ \mathcal{E}_{\mathcal{X}}$, $G' = \mathcal{D}_{\mathcal{Y}} \circ g' \circ S_r \circ \mathcal{E}_{\mathcal{X}}$, it holds

$$\begin{aligned}\|G - G'\|_{\infty,\sigma_R^r(U)} &= \sup_{x\in\sigma_R^r(U)} \|\mathcal{D}_{\mathcal{Y}} \circ g \circ S_r \circ \mathcal{E}_{\mathcal{X}}(x) - \mathcal{D}_{\mathcal{Y}} \circ g' \circ S_r \circ \mathcal{E}_{\mathcal{X}}(x)\|_{\mathcal{Y}} \\ &\leq \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \sup_{\boldsymbol{y}\in U} \|g(\boldsymbol{y}) - g'(\boldsymbol{y})\|_2 \leq \Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}} \sqrt{p} \|g - g'\|_{\infty,\infty},\end{aligned} \tag{D.41}$$

where we used $\sigma_R^r = \mathcal{D}_{\mathcal{X}} \circ S_r^{-1}$ and $\|\cdot\|_{\infty,\infty}$ from (D.40). Furthermore, we used $\|g(u)\|_2 \leq \sqrt{p}\|g\|_\infty$ for all $g \in \boldsymbol{g}_{\mathrm{FN}}$, since NNs $g \in \boldsymbol{g}_{\mathrm{FN}}$ have $\mathrm{width}(g) \leq p$. Then (D.41) yields

$$H(\boldsymbol{G}_{\mathrm{FN}}, \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta) \leq H\left(\boldsymbol{g}_{\mathrm{FN}}, \|\cdot\|_{\infty,\infty}, \frac{\delta}{\Lambda_{\boldsymbol{\Psi}_{\mathcal{Y}}}\sqrt{p}}\right). \tag{D.42}$$

Step 2. It remains to bound $H(\boldsymbol{g}_{\mathrm{FN}}, \|\cdot\|_{\infty,\infty}, \delta) = \log(N(\boldsymbol{g}_{\mathrm{FN}}, \|\cdot\|_{\infty,\infty}, \delta))$. To this end we follow the proof and notation of Schmidt-Hieber (2020a, Lemma 5). For $l = 1, \dots L+1$, define the matrices $W_l = (w_{i,j}^l)_{i,j} \in \mathbb{R}^{p_{l-1}\times p_l}$ and the vectors $B_l = (b_j^l)_j \in \mathbb{R}^{p_l}$. Furthermore, define

$$\begin{aligned}\sigma^{B_l} &: \mathbb{R}^{p_l} \to \mathbb{R}^{p_l}, \quad \sigma^{B_l}(x) = \sigma_1(x + B_l) = \max\{0, x + B_l\}, \quad l = 1, \dots L, \\ \sigma^{B_{L+1}} &: \mathbb{R}^{p_{L+1}} \to \mathbb{R}^{p_{L+1}}, \quad \sigma^{B_{L+1}}(x) = x + B_{L+1}.\end{aligned} \tag{D.43}$$

Then we can write a NN $g \in \boldsymbol{g}_{\mathrm{FN}}$ as a functional composition of σ^{B_l} and W_l, i.e.

$$g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_{L+1}}, \quad g(x) = \sigma^{B_{L+1}} W_{L+1} \sigma^{B_l} \dots W_2 \sigma^{B_1} W_1 x.$$

For $k \in \{1, \dots, L+1\}$ we define the functions

$$\begin{aligned} A_k^+ g : \mathbb{R}^{p_0} \to \mathbb{R}^{p_k}, \quad A_k^+ g(x) &= \sigma^{B_k} W_k \dots \sigma^{B_1} W_1 x, \\ A_k^- g : \mathbb{R}^{p_{k-1}} \to \mathbb{R}^{p_{L+1}}, \quad A_k^- g(x) &= \sigma^{B_{L+1}} W_{L+1} \dots \sigma^{B_k} W_k x. \end{aligned} \tag{D.44}$$

Furthermore, set $A_0^+ g = \mathrm{Id}_{\mathbb{R}^{p_0}}$ and $A_{L+2}^- g = \mathrm{Id}_{\mathbb{R}^{p_{L+1}}}$. For all $1 \le l \le L+1$ holds

$$\begin{aligned} \|\sigma^{B_l}(x)\|_\infty &\le \|x\|_\infty + M \\ \|W^l(x)\|_\infty &\le \|W^l\|_\infty \|x\|_\infty \le Mp \|x\|_\infty. \end{aligned}$$

We claim that for $k \in \{1, \dots, L+1\}$

$$\sup_{x \in [-1,1]^{p_0}} \|A_k^+ g(x)\|_\infty \le (M(p+1))^k$$

and proceed by induction. The case $k = 0$ is trivial. To go from $k-1$ to k we compute

$$\begin{aligned} \sup_{x \in [-1,1]^{p_0}} \|A_k^+ x\|_\infty &= \sup_{x \in [-1,1]^{p_0}} \|\sigma^{B_k} W_k (\sigma^{B_{k-1}} W_{k-1} \cdots \sigma^{B_1} W_1 x)\|_\infty \\ &\le \sup_{x \in [-(M(p+1))^{k-1}, (M(p+1))^{k-1}]^{p_{k-1}}} \|\sigma^{B_k} W^k x\|_\infty \\ &\le (Mp(M(p+1))^{k-1} + M) \le (M(p+1))^k, \end{aligned} \tag{D.45}$$

as claimed.

Moreover, for $l = 1, \dots, L+1$, $W_l : (\mathbb{R}^{p_{l-1}}, \|\cdot\|_\infty) \to (\mathbb{R}^{p_l}, \|\cdot\|_\infty)$ is Lipschitz with constant Mp and $\sigma^{B_l} : (\mathbb{R}^{p_l}, \|\cdot\|_\infty) \to (\mathbb{R}^{p_l}, \|\cdot\|_\infty)$ is Lipschitz with constant 1. Thus we can estimate the Lipschitz constant of $A_k^- g$ for $k = 1, \dots, L+1$. It holds

$$\begin{aligned} \left\|A_k^- g(x) - A_k^- g(y)\right\|_\infty &= \left\|\sigma^{B_{L+1}} W_{L+1} \dots \sigma^{B_k} W_k x - \sigma^{B_{L+1}} W_{L+1} \dots \sigma^{B_k} W_k y\right\|_\infty \\ &\le Mp \left\|\sigma^{B_L} W_L \dots \sigma^{B_k} W_k x - \sigma^{B_L} W_L \dots \sigma^{B_k} W_k y\right\|_\infty \\ &\le \dots \le (Mp)^{L+2-k} \|x - y\|_\infty \quad \text{for } x, y \in \mathbb{R}^{p_{k-1}}. \end{aligned} \tag{D.46}$$

Now let $g, g^* \in \boldsymbol{g}_{\mathrm{FN}}$ be two NN such that $|w_{i,j}^l - w_{i,j}^{l,*}| < \varepsilon$ and $|b_i^l - b_i^{l,*}| < \varepsilon$ for all $i \le p_{l+1}$, $j \le p_l$, $l \le L+1$. Then

$$\begin{aligned} \|g - g^*\|_{\infty,\infty} &\le \sum_{k=1}^{L+1} \left\| A_{k+1}^- g \sigma^{B_k} W_k A_{k-1}^+ g^* - A_{k+1}^- g \sigma^{B_k^*} W_k^* A_{k-1}^+ g^* \right\|_{\infty,\infty} \\ &\le \sum_{k=1}^{L+1} (Mp)^{L+1-k} \left\| \sigma^{B_k} W_k A_{k-1}^+ g^* - \sigma^{B_k^*} W_k^* A_{k-1}^+ g^* \right\|_{\infty,\infty} \\ &\le \sum_{k=1}^{L+1} (Mp)^{L+1-k} \left(\left\| (W_k - W_k^*) A_{k-1}^+ g^* \right\|_{\infty,\infty} + \|B_k - B_k^*\|_\infty \right) \\ &\le \varepsilon \sum_{k=1}^{L+1} (Mp)^{L+1-k} \left(pM^{k-1}(p+1)^{k-1} + 1 \right) \\ &< \varepsilon (L+1) M^L (p+1)^{L+1}, \end{aligned} \tag{D.47}$$

where we used (D.45), (D.46) and $M \geq 1$. The total number of weight and biases is less than $(L+1)(p^2+p)$. Therefore there are at most

$$\binom{(L+1)(p^2+p)}{s} \leq ((L+1)(p^2+p))^s$$

combinations to pick s nonzero parameters. Since all parameters are bounded by M, we choose $\varepsilon = \delta/((L+1)M^L(p+1)^{L+1})$ and obtain the covering bound for all $\delta > 0$

$$\begin{aligned}
N(\boldsymbol{g}_{\mathrm{FN}}, \|\cdot\|_{\infty,\infty}, \delta) &\leq \max\left\{1, \sum_{s^*=1}^{s} \left(2M\epsilon^{-1}(L+1)(p^2+p)\right)^{s^*}\right\} \\
&\leq \max\left\{1, \sum_{s^*=1}^{s} \left(2\delta^{-1}(L+1)M^{L+1}(p+1)^{L+1}(L+1)(p^2+p)\right)^{s^*}\right\} \\
&\leq \max\left\{1, \sum_{s^*=1}^{s} \left(2\delta^{-1}(L+1)^2M^{L+1}(p+1)^{L+3}\right)^{s^*}\right\} \\
&\leq \left(2^{L+6}L^2M^{L+1}p^{L+3}\max\left\{1,\delta^{-1}\right\}\right)^{s+1}, \qquad \text{(D.48)}
\end{aligned}$$

where we used $L \geq 1$ and $p \geq 1$ at the last inequality. Eqs. (D.48) and (D.42) show (3.16).

Applying (3.16) to the sparse FrameNet class $\boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_1, N)$ gives

$$\begin{aligned}
&H\left(\boldsymbol{G}^{\mathrm{sp}}_{\mathrm{FN}}(\sigma_1,N), \|\cdot\|_{\infty,\sigma^r_R(U)}, \delta\right) \\
&\leq (\mathrm{size}_N + 1)\log\left(2^{\mathrm{depth}_N+6}\Lambda_{\boldsymbol{\Psi}_y}\mathrm{depth}_N^2 M^{\mathrm{depth}_N+1}\mathrm{width}_N^{\mathrm{depth}_N+4}\max\left\{1,\delta^{-1}\right\}\right) \\
&\leq (C_sN+1) \\
&\quad\times\log\left(2^{C_L\log(N)+6}\Lambda_{\boldsymbol{\Psi}_y}(C_L\log(N))^2 M^{C_L\log(N)+1}(C_pN)^{C_L\log(N)+4}\max\left\{1,\delta^{-1}\right\}\right) \\
&\leq C^{\mathrm{SP}}_H N\left(1+\log(N)^2+\log\left(\max\left\{1,\delta^{-1}\right\}\right)\right), \quad N\in\mathbb{N}, \quad \delta>0, \qquad \text{(D.49)}
\end{aligned}$$

where we defined

$$\begin{aligned}
C^{\mathrm{SP}}_H = 2C_s\Big((C_L+6)\log(2) + \log(\Lambda_{\boldsymbol{\Psi}_y}) + C_L^2 + \\
+ (C_L+1)\log(M) + (C_L+4)(\log(C_p)+1)\Big).
\end{aligned}$$

Applying (3.16) to the fully connected FrameNet class $\boldsymbol{G}^{\mathrm{full}}_{\mathrm{FN}}(\sigma_1, N)$ gives

$$\begin{aligned}
&H\left(\boldsymbol{G}^{\mathrm{full}}_{\mathrm{FN}}(\sigma_1,N), \|\cdot\|_{\infty,\sigma^r_R(U)}, \delta\right) \\
&\leq (s^{\mathrm{FC}}(N)+1)\log\left(2^{\mathrm{depth}_N+6}\Lambda_{\boldsymbol{\Psi}_y}\mathrm{depth}_N^2 M^{\mathrm{depth}_N+1}\mathrm{width}_N^{\mathrm{depth}_N+4}\max\left\{1,\delta^{-1}\right\}\right) \\
&\leq \left((\mathrm{depth}_N+1)\left(\mathrm{width}_N^2+\mathrm{width}_N\right)+1\right) \\
&\quad\times\log\left(2^{C_L\log(N)+6}\Lambda_{\boldsymbol{\Psi}_y}(C_L\log(N))^2 M^{C_L\log(N)+1}(C_pN)^{C_L\log(N)+4}\max\left\{1,\delta^{-1}\right\}\right) \\
&\leq \left((C_L\log(N)+1)\left(C_p^2N^2+C_pN\right)+1\right) \\
&\quad\times\log\left(2^{C_L\log(N)+6}\Lambda_{\boldsymbol{\Psi}_y}(C_L\log(N))^2 M^{C_L\log(N)+1}(C_pN)^{C_L\log(N)+4}\max\left\{1,\delta^{-1}\right\}\right) \\
&\leq C^{\mathrm{FC}}_H N^2\left(1+\log(N)^3+\log\left(\max\left\{1,\delta^{-1}\right\}\right)\right), \quad N\in\mathbb{N}, \qquad \text{(D.50)}
\end{aligned}$$

where we defined

$$C_H^{\mathrm{FC}} = 8C_L C_p^2 \Big((C_L+6)\log(2) + \log(\Lambda_{\Psi_y}) + C_L^2 + \\ + (C_L+1)\log(M) + (C_L+4)(\log(C_p)+1) \Big).$$

Equations (D.49) and (D.50) finish the proof of Lemma 20.

D.5 Proof of Lemma 22

The following proof is a modification of Schmidt-Hieber (2020a, Proof of Lemma 5) to the case where the activation function is not globally, but only locally Lipschitz continuous. The compactness of $(\boldsymbol{G}, \|\cdot\|_{\infty,\mathrm{supp}(\gamma)})$ follows similarly to the ReLU case since Petersen et al. (2021, Proposition 3.5) holds for any continuous activation function.

Let $q \in \mathbb{N}$, $q \geq 2$ and let $\sigma_q : \mathbb{R} \to \mathbb{R}$, $\sigma_q(x) = \max\{0,x\}^q$ denote the RePU activation function. Recall $\mathrm{depth}(g) \leq L$, $\mathrm{width}(g) \leq p$, $\mathrm{size}(g) \leq s$ and $\mathrm{mpar}(g) \leq M$ for $g \in \boldsymbol{g}_{\mathrm{FN}}$. We argue analogously to the ReLU-case in Lemma 20 and bound the entropy of the NN class $\boldsymbol{g}_{\mathrm{FN}}(\sigma_q, L, p, s, M, B)$. Recall the definitions of σ^{B_l}, $A_k^+ g$ and $A_k^- g$ from (D.43)-(D.44). Similar to (D.45) it holds that

$$\begin{aligned}
\|A_k^+ g\|_{\infty,\infty} &= \sup_{x\in[-1,1]^{p_0}} \|A_k^+ g(x)\|_\infty \\
&\leq \sup_{x\in[-M(p+1),M(p+1)]^{p_1}} \left\|\sigma^{B_k} W_k \sigma^{B_{k-1}} \dots W_2 \sigma_q x\right\|_\infty \\
&\leq \sup_{x\in[-M^q(p+1)^q,M^q(p+1)^q]^{p_1}} \left\|\sigma^{B_k} W_k \sigma^{B_{k-1}} \dots W_2 x\right\|_\infty \\
&\leq \dots \leq (M(p+1))^{\sum_{j=1}^k q^j} \leq (M(p+1))^{q^{k+1}},
\end{aligned}$$

where we used $M \geq 1$.

In the RePU-case, $A_k^- g$ is only locally Lipschitz: Since $|\sigma_q(x)'| \leq q|x|^{q-1}$ it holds

$$|\sigma_q(x) - \sigma_q(y)| \leq q \max\{|x|,|y|\}^{q-1}|x-y| \qquad \forall x,y \in \mathbb{R}.$$

Therefore for $k = 1,\dots,L+1$ and $x, y \in \mathbb{R}^{p_{k-1}}$, $\|x\|_\infty, \|y\|_\infty \leq C$, we get

$$\begin{aligned}
&\left\|A_k^- g(x) - A_k^- g(y)\right\|_\infty \\
&= \left\|\sigma^{B_{L+1}} W_{L+1} \sigma^{B_L} \dots W_k x - \sigma^{B_{L+1}} W_{L+1} \sigma^{B_L} \dots W_k y\right\|_\infty \\
&\leq Mp \left\|\sigma^{B_L} W_L \sigma^{B_{L-1}} \dots W_k x - \sigma^{B_L} W_L \sigma^{B_{L-1}} \dots W_k y\right\|_\infty \\
&\leq Mpq \left(\sup_{\|x\|_\infty \leq C} \left\|W_L \sigma^{B_{L-1}} \dots W_k x\right\|_\infty \right)^{q-1} \\
&\qquad \times \left\|W_L \sigma^{B_{L-1}} \dots W_k x - W_L \sigma^{B_{L-1}} \dots W_k y\right\|_\infty \\
&\leq (Mpq)^{L+2-k} \left(\sup_{\|x\|_\infty \leq C} \left\|W_L \sigma^{B_{L-1}} \dots W_k x\right\|_\infty \right)^{q-1} \\
&\qquad \times \left(\sup_{\|x\|_\infty \leq C} \left\|W_{L-1} \sigma^{B_{L-2}} \dots W_k x\right\|_\infty \right)^{q-1} \times \dots \times \left(\sup_{\|x\|_\infty \leq C} \|W_k x\|_\infty \right)^{q-1} \|x-y\|_\infty.
\end{aligned}$$

Using

$$\begin{aligned}\sup_{\|x\|_\infty\le C}\left\|W_j\sigma^{B_{j-1}}\dots W_k x\right\|_\infty &\le \sup_{\|x\|_\infty\le (M(p+1)C)^q}\left\|W_j\sigma^{B_{j-1}}\dots W_{k+1}x\right\|_\infty\\ &\le \sup_{\|x\|_\infty\le (M(p+1))^{q+q^2}C^{q^2}}\left\|W_j\sigma^{B_{j-1}}\dots W_{k+2}x\right\|_\infty\\ &\le\dots\le \sup_{\|x\|_\infty\le (M(p+1))^{\sum_{l=1}^{j-k}q^l}C^{q^{j-k}}}\|W_j x\|_\infty\\ &\le (M(p+1))^{\sum_{l=0}^{j-k}q^l}\,C^{q^{j-k}}\le (M(p+1)C)^{q^{j-k+1}}\end{aligned}$$

for $j=k,\dots,L$ and $C,M\ge 1$, we get

$$\begin{aligned}\left\|A_k^- g(x)-A_k^- g(y)\right\|_\infty &\le (Mpq)^{L+2-k}\prod_{j=k}^{L}\left(M(p+1)\hat{C}\right)^{q^{j-k+1}}\|x-y\|_\infty \qquad\text{(D.51)}\\ &\le (Mpq)^{L+2-k}\left(M(p+1)\hat{C}\right)^{q^{L+2-k}}\|x-y\|_\infty,\quad x,y\in\mathbb{R}^{p_{k-1}}.\end{aligned}$$

Now we proceed similarly to (D.47). Let $g,g^*\in\boldsymbol{g}_{\mathrm{FN}}$ be two NN such that $|w_{i,j}^l-w_{i,j}^{l,*}|<\varepsilon$ and $|b_i^l-b_i^{l,*}|<\varepsilon$ for all $i\le p_{l+1}$, $j\le p_l$, $l\le L+1$. Then with $A_0^+ g=\mathrm{Id}_{\mathbb{R}^{p_0}}$ and $A_{L+2}^- g=\mathrm{Id}_{\mathbb{R}^{p_{L+1}}}$ we estimate

$$\begin{aligned}&\|g-g^*\|_{\infty,\infty}\\ &\le \sum_{k=1}^{L+1}\left\|A_{k+1}^- g\sigma^{B_k}W_kA_{k-1}^+g^*-A_{k+1}^- g\sigma^{B_k^*}W_k^*A_{k-1}^+g^*\right\|_{\infty,\infty}\\ &\le \sum_{k=1}^{L+1}(Mpq)^{L+1-k}\left(M(p+1)\left(M(p+1)\right)^{q^{k+1}}\right)^{q^{L+1-k}}\\ &\qquad\left\|\sigma^{B_k}W_kA_{k-1}^+g^*-\sigma^{B_k^*}W_k^*A_{k-1}^+g^*\right\|_{\infty,\infty}\\ &\le \sum_{k=1}^{L+1}(Mpq)^{L+1-k}\left(M(p+1)\left(M(p+1)\right)^{q^{k+1}}\right)^{q^{L+1-k}}\\ &\qquad\left(\left\|(W_k-W_k^*)A_{k-1}^+g^*\right\|_{\infty,\infty}+\|B_k-B_k^*\|_{\infty,\infty}\right)q\left(M(p+1)\left\|A_{k-1}^+g^*\right\|_{\infty,\infty}\right)^{q-1}\\ &\le 2\varepsilon\sum_{k=1}^{L+1}(Mpq)^{L+2-k}\left(M(p+1)\left(M(p+1)\right)^{q^{k+1}}\right)^{q^{L+1-k}}\left\|A_{k-1}^+g^*\right\|_{\infty,\infty}\\ &\qquad\left(M(p+1)\left\|A_{k-1}^+g^*\right\|_{\infty,\infty}\right)^{q-1}\\ &\le 2\varepsilon(L+1)\left(M(p+1)q\right)^{L+q}\left(M(p+1)\left(M(p+1)\right)^{q^{L+2}}\right)^{q^L}\left(\left(M(p+1)\right)^{q^{L+1}}\right)^q\\ &< \varepsilon L q^{L+q}\left(2pM\right)^{4q^{2L+2}}\left(2M\sqrt{p}(p^2+p)(L+1)\right)^{-1}.\end{aligned}\qquad\text{(D.52)}$$

In (D.52) we used the Lipschitz bound (D.51) with

$$C=\max\left\{1,\|\sigma^{B_k}W_kA_{k-1}^+g^*\|_{\infty,\infty},\|\sigma^{B_k^*}W_k^*A_{k-1}^+g^*\|_{\infty,\infty}\right\}\le\left(M(p+1)\right)^{q^{k+1}},$$

and $p \geq 1$, $L \geq 1$, $q \geq 2$ at the last inequality.

As in the proof of Lemma 20, there are

$$\binom{(L+1)(p^2+p)}{s} \leq ((L+1)(p^2+p))^s$$

combinations to pick s nonzero weights and biases. Since all parameters are bounded by M, we choose

$$\varepsilon = \frac{2M\sqrt{p}(p^2+p)(L+1)\delta}{Lq^{L+q}\left(2pM\right)^{4q^{2L+2}}}$$

and obtain the covering bound

$$\begin{aligned} N(\boldsymbol{g}_{\mathrm{FN}}, \|\cdot\|_{\infty,\infty}, \delta) &\leq \max\left\{1, \sum_{s^*=1}^{s} \left(2M\epsilon^{-1}(L+1)(p^2+p)\right)^{s^*}\right\} \\ &\leq \max\left\{1, \sum_{s^*=1}^{s} \left(Lq^{L+q}\left(2pM\right)^{4q^{2L+2}}(\sqrt{p}\delta)^{-1}\right)^{s^*}\right\} \\ &\leq \left(Lq^{L+q}\left(2pM\right)^{4q^{2L+2}}\sqrt{p}^{-1}\max\left\{1,\delta^{-1}\right\}\right)^{s+1}. \end{aligned} \tag{D.53}$$

Eqs. (D.53) and (D.42) show (3.17).

Applying (3.17) to the sparse FrameNet class $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N)$ gives

$$\begin{aligned} &H\left(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{sp}}(\sigma_q, N), \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta\right) \\ &\leq (s^{SP}(N)+1)\log\left(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}\mathrm{depth}_N q^{\mathrm{depth}_N+q}\left(2\mathrm{width}_N M\right)^{4q^{2\mathrm{depth}_N+2}}\max\left\{1,\delta^{-1}\right\}\right) \\ &\leq (C_s N+1)\log\left(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}C_L\log(N)q^{C_L\log(N)+q}\left(2C_pNM\right)^{4q^{2C_L\log(N)+2}}\max\left\{1,\delta^{-1}\right\}\right) \\ &\leq C_H^{\mathrm{SP}}N^{1+2C_L\log(q)}\left(1+\log(N)+\log\left(\max\left\{1,\delta^{-1}\right\}\right)\right), \end{aligned} \tag{D.54}$$

where we set

$$C_H^{\mathrm{SP}} = 2C_s\left(\log(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}) + C_L + (C_L+q)\log(q) + 4q^2\left(\log(2C_pM)+1\right)\right).$$

Applying (3.17) to the fully connected FrameNet class $\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N)$ gives the entropy bound

$$\begin{aligned} &H\left(\boldsymbol{G}_{\mathrm{FN}}^{\mathrm{full}}(\sigma_q, N), \|\cdot\|_{\infty,\sigma_R^r(U)}, \delta\right) \\ &\quad\leq (s^{FC}(N)+1)\log\left(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}\mathrm{depth}_N q^{\mathrm{depth}_N+q}\left(2\mathrm{width}_N M\right)^{4q^{2\mathrm{depth}_N+2}}\max\left\{1,\delta^{-1}\right\}\right) \\ &\quad\leq \left((\mathrm{depth}_N+1)\left(\mathrm{width}_N^2+\mathrm{width}_N\right)+1\right) \\ &\qquad\times\log\left(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}\mathrm{depth}_N q^{\mathrm{depth}_N+q}\left(2\mathrm{width}_N M\right)^{4q^{2\mathrm{depth}_N+2}}\max\left\{1,\delta^{-1}\right\}\right) \\ &\quad\leq \left((C_L\log(N)+1)\left(C_p^2N^2+C_pN\right)+1\right) \\ &\qquad\times\log\left(\Lambda_{\boldsymbol{\Psi}_{\mathrm{y}}}C_L\log(N)q^{C_L\log(N)+q}\left(2C_pNM\right)^{4q^{2C_L\log(N)+2}}\max\left\{1,\delta^{-1}\right\}\right) \\ &\quad\leq C_H^{\mathrm{FC}}N^{2+2C_L\log(q)}\left(1+\log(N)^2+\log\left(\max\left\{1,\delta^{-1}\right\}\right)\right), \end{aligned} \tag{D.55}$$

where we set

$$C_H^{\mathrm{FC}} = 8C_sC_p^2\bigg(\log(\Lambda_{\mathbf{\Psi}_y}) + C_L + (C_L+q)\log(q) + 4q^2\left(\log(2C_pM)+1\right)\bigg).$$

Equations (D.54) and (D.55) finish the proof of Lemma 22.

Appendix E. Proofs of Section 4

E.1 Proof of Theorem 27

In Herrmann et al. (2024, Proof of Proposition 3, Step 1), the holomorphy in Assumption 2 is verified for $\mathcal{X}$, $\mathcal{Y}$ in (4.5) with $r_0 > d/2$ and $t \in [0, (1 + r_0 - d/2 - t_0)/d)$. Moreover, $\gamma = (\sigma_R^r)_\# \pi$ in particular shows $\mathrm{supp}(\gamma) \subseteq C_R^r(\mathcal{X})$ and hence verifies the second part of Assumption 2. Substituting $\mathfrak{s} = r_0 + rd$, i.e. $r = \frac{\mathfrak{s}-r_0}{d}$, and taking $t = (1+r_0-d/2-t_0)/d - \tau$ with some small τ, Theorem 23 (i) then gives

$$\mathbb{E}_{G_0}[\|\hat{G}_n - G_0\|^2_{L^2(\gamma)}] \leq Cn^{-\frac{\kappa}{\kappa+1}+\tau},$$

where

$$\kappa = 2\min\left\{\frac{\mathfrak{s}-r_0}{d} - \frac{1}{2}, \frac{1+r_0-\frac{d}{2}-t_0}{d}\right\} - \tau$$

for all $r_0 > d/2$ and $t_0 \in [0,1]$.

From here on the proof is essentially the same as Herrmann et al. (2024, Proof of Proposition 3, Step 2); the only difference is that while Herrmann et al. (2024) uses the uniform bound in Theorem 18 (i), we require the L^2-bound in Theorem 18 (ii). For completeness, we repeat the argument. The constraint $r > 1$ implies $\mathfrak{s} > r_0 + d$ on $\mathfrak{s}$. We now choose $r_0 > \frac{d}{2}$ in order to maximize the convergence rate. Solving

$$\frac{\mathfrak{s}-r_0}{d} - \frac{1}{2} = \frac{1+r_0-\frac{d}{2}-t_0}{d}$$

for r_0 gives

$$r_0 = \frac{\mathfrak{s}+t_0-1}{2}. \tag{E.1}$$

The constraint $r_0 > \frac{d}{2}$ implies the constraint $\mathfrak{s} > d + 1 - t_0$.

We look at two cases separately. First, if $\mathfrak{s} \in (\frac{3d}{2}, 2d+1-t_0]$, we set $r_0 := \frac{d}{2} + \tau_2$, where we choose $\tau_2 > 0$ s.t. $\tau_2 < \mathfrak{s} - 3d/2$ which guarantees $\mathfrak{s} > r_0 + d$. For $\tau < \tau_2/d$, we obtain the convergence rate

$$\kappa = 2\min\left\{\frac{\mathfrak{s}-\frac{d}{2}-\tau_2}{d} - \frac{1}{2}, \frac{1+\frac{d}{2}+\tau_2-\frac{d}{2}-t_0}{d}\right\} - \tau \geq 2\min\left\{\frac{\mathfrak{s}}{d} - 1 - \frac{2\tau_2}{d}, \frac{1-t_0}{d}\right\}.$$

In the case $\mathfrak{s} > 2d + 1 - t_0$, define r_0 as in (E.1). The constraint $\mathfrak{s} > r_0 + d$ amounts to

$$\mathfrak{s} > \frac{\mathfrak{s}+t_0-1}{2} + d \qquad \Leftrightarrow \qquad \mathfrak{s} > 2d + t_0 - 1,$$

which holds since $\mathfrak{s} > 2d + 1 - t_0 \geq 2d + t_0 - 1$ for all $t_0 \in [0,1]$. In this case we get the convergence rate

$$\kappa = 2\frac{\mathfrak{s}-r_0}{d} - 1 - \tau = \frac{\mathfrak{s}+1-t_0}{d} - 1 - \tau.$$

Choosing $\tau_1 > 8\tau_2/d > 8\tau$ shows (4.6) and finishes the proof of Theorem 27.

References

Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55. US Government printing office, 1948.

Ben Adcock, Nick Dexter, and Sebastian Moraga. Optimal deep learning of holomorphic operators between banach spaces. *Advances in Neural Information Processing Systems*, 37:27725–27789, 2024.

Ben Adcock, Simone Brugiapaglia, Nick Dexter, and Sebastian Moraga. Near-optimal learning of banach-valued, high-dimensional functions via deep neural networks. *Neural Networks*, 181:106761, 2025. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2024.106761. URL https://www.sciencedirect.com/science/article/pii/S0893608024006853.

Sergios Agapiou and Sven Wang. Laplace priors and spatial inhomogeneity in bayesian inverse problems. *Bernoulli*, 30(2):878–910, 2024.

Anima Anandkumar, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Nikola Kovachki, Zongyi Li, Burigede Liu, and Andrew Stuart. Neural operator: Graph kernel network for partial differential equations. In *ICLR 2020 Workshop on Integration of Deep Neural Models and Differential Equations*, 2019. URL https://openreview.net/forum?id=fg2ZFmXFO3.

Ivo Babuška, Fabio Nobile, and Raúl Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. *SIAM Rev.*, 52(2):317–355, 2010. ISSN 0036-1445,1095-7200. doi: 10.1137/100786356. URL https://doi.org/10.1137/100786356.

Markus Bachmayr, Albert Cohen, Ronald DeVore, and Giovanni Migliorati. Sparse polynomial approximation of parametric elliptic PDEs. Part II: Lognormal coefficients. *ESAIM Math. Model. Numer. Anal.*, 51(1):341–363, 2017. ISSN 2822-7840,2804-7214. doi: 10.1051/m2an/2016051. URL https://doi.org/10.1051/m2an/2016051.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability theory and related fields*, 113:301–413, 1999.

Sebastian Becker, Arnulf Jentzen, Marvin S Müller, and Philippe von Wurstemberger. Learning the random variables in monte carlo simulations with stochastic gradient descent: Machine learning for parametric pdes and financial derivative pricing. *Mathematical Finance*, 34(1):90–150, 2024.

Kaushik Bhattacharya, Bamdad Hosseini, Nikola B. Kovachki, and Andrew M. Stuart. Model reduction and neural networks for parametric PDEs. *SMAI J. Comput. Math.*, 7:121–157, 2021.

Lucien Birgé and Pascal Massart. Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields*, 97:113–150, 1993.

Ismaël Castillo and Richard Nickl. Nonparametric bernstein–von mises theorems in gaussian white noise. *The Annals of Statistics*, 41(4), 2013. ISSN 0090-5364. doi: 10.1214/13-AOS1133.

Gaëlle Chagny, Anouar Meynaoui, and Angelina Roche. Adaptive nonparametric estimation in the functional linear model with functional output. *Electronic Journal of Statistics*, 19 (1):2990–3039, 2025.

Abdellah Chkifa, Albert Cohen, Ronald DeVore, and Christoph Schwab. Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. *ESAIM Math. Model. Numer. Anal.*, 47(1):253–280, 2013. ISSN 0764-583X. doi: 10.1051/m2an/2012027. URL https://doi.org/10.1051/m2an/2012027.

Abdellah Chkifa, Albert Cohen, Giovanni Migliorati, Fabio Nobile, and Raul Tempone. Discrete least squares polynomial approximation with random evaluations—application to parametric and stochastic elliptic PDEs. *ESAIM Math. Model. Numer. Anal.*, 49(3): 815–837, 2015a. ISSN 0764-583X. doi: 10.1051/m2an/2014050. URL https://doi.org/10.1051/m2an/2014050.

Abdellah Chkifa, Albert Cohen, and Christoph Schwab. Breaking the curse of dimensionality in sparse polynomial approximation of parametric PDEs. *J. Math. Pures Appl. (9)*, 103(2):400–428, 2015b. ISSN 0021-7824,1776-3371. doi: 10.1016/j.matpur.2014.04.009. URL https://doi.org/10.1016/j.matpur.2014.04.009.

Ole Christensen. *An Introduction to Frames and Riesz Bases [recurso electrónico]*. Applied and Numerical Harmonic Analysis. Birkhauser, Cham, second edition, 2016. ISBN 978-3-319-25613-9.

Ludovica Cicci, Stefania Fresca, and Andrea Manzoni. Deep-hyromnet: A deep learning-based operator approximation for hyper-reduction of nonlinear parametrized pdes. *Journal of Scientific Computing*, 93(2):57, 2022.

K. A. Cliffe, M. B. Giles, R. Scheichl, and A. L. Teckentrup. Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Comput. Vis. Sci.*, 14(1):3–15, 2011. ISSN 1432-9360,1433-0369. doi: 10.1007/s00791-011-0160-x. URL https://doi.org/10.1007/s00791-011-0160-x.

Albert Cohen and Ronald DeVore. Approximation of high-dimensional parametric PDEs. *Acta Numer.*, 24:1–159, 2015. ISSN 0962-4929. doi: 10.1017/S0962492915000033. URL https://doi.org/10.1017/S0962492915000033.

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. ISSN 1615-3375. doi: 10.1007/s10208-010-9072-2. URL http://dx.doi.org/10.1007/s10208-010-9072-2.

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. ISSN 0219-5305. doi: 10.1142/S0219530511001728. URL http://dx.doi.org/10.1142/S0219530511001728.

Albert Cohen, Giovanni Migliorati, and Fabio Nobile. Discrete least-squares approximations over optimized downward closed polynomial spaces in arbitrary dimension. *Constr. Approx.*, 45(3):497–519, 2017. ISSN 0176-4276. doi: 10.1007/s00365-017-9364-8. URL https://doi.org/10.1007/s00365-017-9364-8.

Albert Cohen, Christoph Schwab, and Jakob Zech. Shape holomorphy of the stationary navier-stokes equations. *SIAM J. Math. Analysis*, 50(2):1720–1752, 2018. doi: https://doi.org/10.1137/16M1099406.

Niccolò Dal Santo, Simone Deparis, and Luca Pegolotti. Data driven approximation of parametrized pdes by reduced basis and neural networks. *Journal of Computational Physics*, 416:109550, 2020.

Oleg Davydov and Rob Stevenson. Hierarchical riesz bases for $h^s(\omega)$, $1 < s < 5/2$. *Constructive Approximation*, 22(3):365–394, 2005. ISSN 1432-0940. doi: 10.1007/s00365-004-0593-2.

Maarten V. de Hoop, Nikola B. Kovachki, Nicholas H. Nelsen, and Andrew M. Stuart. Convergence rates for learning linear operators from noisy data. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):480–513, 2023. doi: 10.1137/21M1442942. URL https://doi.org/10.1137/21M1442942.

Tim De Ryck, Samuel Lanthaler, and Siddhartha Mishra. On the approximation of functions by tanh neural networks. *Neural Networks*, 143:732–750, 2021.

R. DeVore, R. Howard, and C. Micchelli. Optimal nonlinear approximation. *Manuscripta mathematica*, 63(4):469–478, 1989. URL http://eudml.org/doc/155392.

R.A. DeVore and G.G. Lorentz. *Constructive Approximation*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 1993. ISBN 9783540506270. URL https://books.google.de/books?id=cDqNW6k7_ZwC.

Sjoerd Dirksen. Tail bounds via generic chaining. *Electronic Journal of Probability*, 20, 2015. ISSN 1083-6489. doi: 10.1214/EJP.v20-3760.

Alireza Doostan and Houman Owhadi. A non-adapted sparse approximation of pdes with stochastic inputs. *Journal of Computational Physics*, 230(8):3015–3034, 2011. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2011.01.002. URL https://www.sciencedirect.com/science/article/pii/S0021999111000106.

Dinh Dung, Van Kien Nguyen, Christoph Schwab, and Jakob Zech. *Analyticity and sparsity in uncertainty quantification for PDEs with Gaussian random field inputs*, volume 2334 of *Lecture Notes in Mathematics*. Springer, Cham, 2023. ISBN 978-3-031-38383-0. doi: 10.1007/978-3-031-38384-7. URL https://doi.org/10.1007/978-3-031-38384-7.

D. E. Edmunds and H. Triebel. *Function spaces, entropy numbers, differential operators*, volume 120 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 1996. ISBN 0-521-56036-5. doi: 10.1017/CBO9780511662201.

Dennis Elbrachter, Dmytro Perekrestenko, Philipp Grohs, and Helmut Bolcskei. Deep neural network approximation theory. *IEEE Transactions on Information Theory*, 67(5):2581–2623, 2021. ISSN 0018-9448. doi: 10.1109/TIT.2021.3062161.

Utku Evci, Fabian Pedregosa, Aidan Gomez, and Erich Elsen. The difficulty of training sparse neural networks. *arXiv preprint arXiv:1906.10732*, 2019.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, New York, 2016.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, New York (NY), 2016. ISBN 1107043166.

Sonja Greven and Fabian Scheipl. A general framework for functional regression modelling. *Statistical Modelling*, 17(1-2):1–35, 2017.

Helmut Harbrecht, Michael Peters, and Markus Siebenmorgen. Analysis of the domain mapping method for elliptic diffusion problems on random domains. *Numerische Mathematik*, 134(4):823–856, 2016.

Fernando Henríquez and Christoph Schwab. Shape holomorphy of the calderón projector for the laplacian in r 2. *Integral Equations and Operator Theory*, 93(4):43, 2021.

Lukas Herrmann, Christoph Schwab, and Jakob Zech. Neural and spectral operator surrogates: unified construction and expression rate bounds. *Advances in Computational Mathematics*, 50(4):1–43, 2024. ISSN 1019-7168. doi: 10.1007/s10444-024-10171-2. URL `https://link.springer.com/article/10.1007/s10444-024-10171-2`.

Jan S. Hesthaven, Gianluigi Rozza, and Benjamin Stamm. *Certified reduced basis methods for parametrized partial differential equations*. SpringerBriefs in Mathematics. Springer, Cham; BCAM Basque Center for Applied Mathematics, Bilbao, 2016. ISBN 978-3-319-22469-5. doi: 10.1007/978-3-319-22470-1. URL `https://doi.org/10.1007/978-3-319-22470-1`. BCAM SpringerBriefs.

J.S. Hesthaven and S. Ubbiali. Non-intrusive reduced order modeling of nonlinear problems using neural networks. *Journal of Computational Physics*, 363:55–78, 2018. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.02.037. URL `https://www.sciencedirect.com/science/article/pii/S0021999118301190`.

R. Hiptmair, L. Scarabosio, C. Schillings, and Ch. Schwab. Large deformation shape uncertainty quantification in acoustic scattering. *Advances in Computational Mathematics*, 44(5):1475–1518, 10 2018. ISSN 1572-9044. doi: 10.1007/s10444-018-9594-8. URL `https://doi.org/10.1007/s10444-018-9594-8`.

Viet Ha Hoang and Christoph Schwab. N-term wiener chaos approximation rates for elliptic pdes with lognormal gaussian random inputs. *Mathematical Models and Methods in Applied Sciences*, 24(04):797–826, 2014. doi: 10.1142/S0218202513500681. URL `https://doi.org/10.1142/S0218202513500681`.

Carlos Jerez-Hanckes, Christoph Schwab, and Jakob Zech. Electromagnetic wave scattering by random surfaces: shape holomorphy. *Math. Models Methods Appl. Sci.*, 27(12):2229–2259, 2017. ISSN 0218-2025. doi: 10.1142/S0218202517500439. URL `https://doi.org/10.1142/S0218202517500439`.

Padraig Kirwan. *Complexifications of multilinear and polynomial mappings*. PhD thesis, University College Dublin, 1997. Ph.D. thesis, National University of Ireland, Galway.

Nikola B Kovachki, Samuel Lanthaler, and Hrushikesh Mhaskar. Data complexity estimates for operator learning. *arXiv preprint arXiv:2405.15992*, 2024a.

Nikola B Kovachki, Samuel Lanthaler, and Andrew M Stuart. Operator learning: Algorithms and analysis. *Handbook of Numerical Analysis*, 25:419–467, 2024b.

Fabian Kröpfl, Roland Maier, and Daniel Peterseim. Operator compression with deep neural networks. *Advances in Continuous and Discrete Models*, 2022(1):29, 2022.

Gitta Kutyniok, Philipp Petersen, Mones Raslan, and Reinhold Schneider. A theoretical analysis of deep neural networks and parametric PDEs. *Constr. Approx.*, 55(1):73–125, 2022. ISSN 0176-4276,1432-0940. doi: 10.1007/s00365-021-09551-4. URL `https://doi.org/10.1007/s00365-021-09551-4`.

Samuel Lanthaler. Operator learning with pca-net: upper and lower complexity bounds. *Journal of Machine Learning Research*, 24(318):1–67, 2023. URL `http://jmlr.org/papers/v24/23-0478.html`.

Samuel Lanthaler, Siddhartha Mishra, and George E. Karniadakis. Error estimates for deeponets: a deep learning framework in infinite dimensions. *Transactions of Mathematics and Its Applications*, 6(1), 2022. doi: 10.1093/imatrm/tnac001.

Bo Li. Better approximations of high dimensional smooth functions by deep neural networks with rectified power units. *Communications in Computational Physics*, 27(2):379–411, 2020. ISSN 1815-2406. doi: 10.4208/cicp.OA-2019-0168.

Bo Li, Shanshan Tang, and Haijun Yu. Powernet: Efficient representations of polynomials and smooth functions by deep neural networks with rectified power units. *arXiv preprint arXiv:1909.05136*, 2019.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

Hao Liu, Haizhao Yang, Minshuo Chen, Tuo Zhao, and Wenjing Liao. Deep nonparametric estimation of operators between infinite dimensional spaces. *Journal of Machine Learning Research*, 25(24):1–67, 2024.

Lu Lu, Pengzhan Jin, Guofei Pang, Zhongqiang Zhang, and George Em Karniadakis. Learning nonlinear operators via deeponet based on the universal approximation theorem of operators. *Nature Machine Intelligence*, 3(3):218–229, 3 2021. ISSN 2522-5839. doi: 10.1038/s42256-021-00302-5. URL https://doi.org/10.1038/s42256-021-00302-5.

H. N. Mhaskar and N. Hahm. Neural networks for functional approximation and system identification. *Neural computation*, 9(1):143–159, 1997. ISSN 0899-7667. doi: 10.1162/neco.1997.9.1.143.

Hrushikesh N Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural computation*, 8(1):164–177, 1996.

Jeffrey S Morris and Raymond J Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(2):179–199, 2006.

Gustavo A. Muñoz, Yannis Sarantopoulos, and Andrew Tonge. Complexifications of real Banach spaces, polynomials and multilinear maps. *Studia Math.*, 134(1):1–33, 1999. ISSN 0039-3223.

Nicholas H Nelsen and Andrew M Stuart. Operator learning using random features: A tool for scientific computing. *SIAM Review*, 66(3):535–571, 2024.

R. Nickl, S. van de Geer, and S. Wang. Convergence rates for penalised least squares estimators in PDE-constrained regression problems. *SIAM J. Uncert. Quant.*, 8, 2020.

Richard Nickl. Donsker-type theorems for nonparametric maximum likelihood estimators. *Probability Theory and Related Fields*, 138(3-4), 2007. ISSN 0178-8051. doi: 10.1007/s00440-006-0031-4.

Richard Nickl. Bernstein–von mises theorems for statistical inverse problems i: Schrödinger equation. *Journal of the European Mathematical Society*, 22(8):2697–2750, 2020. ISSN 1435-9855. doi: 10.4171/JEMS/975.

Richard Nickl. *Bayesian Non-linear Statistical Inverse Problems*. EMS press, 2023.

Richard Nickl and Sven Wang. On polynomial-time computation of high-dimensional posterior measures by Langevin-type algorithms. *Journal of the European Mathematical Society*, 2024.

F. Nobile, R. Tempone, and C. G. Webster. A sparse grid stochastic collocation method for partial differential equations with random input data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008. doi: 10.1137/060663660. URL https://doi.org/10.1137/060663660.

Thomas O'Leary-Roseberry, Peng Chen, Umberto Villa, and Omar Ghattas. Derivative-informed neural operator: An efficient framework for high-dimensional parametric derivative learning. *Journal of Computational Physics*, 496:112555, 2024. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2023.112555. URL https://www.sciencedirect.com/science/article/pii/S0021999123006502.

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, 2022a. ISSN 1432-0940. doi: 10.1007/s00365-021-09542-5. URL https://link.springer.com/article/10.1007/s00365-021-09542-5#citeas.

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. ISSN 0219-5305. doi: gjh43n.

Joost A. A. Opschoor, Christoph Schwab, and Jakob Zech. Deep learning in high dimension: ReLU neural network expression for Bayesian PDE inversion. In *Optimization and control for partial differential equations—uncertainty quantification, open and closed-loop control, and shape optimization*, volume 29 of *Radon Ser. Comput. Appl. Math.*, pages 419–462. De Gruyter, Berlin, 2022b. ISBN 978-3-11-069596-0. doi: 10.1515/9783110695984-015. URL https://doi.org/10.1515/9783110695984-015.

Houman Owhadi and Gene Ryan Yoo. Kernel flows: From learning kernels from data into the abyss. *Journal of Computational Physics*, 389:22–47, 2019.

Thomas O'Leary-Roseberry, Umberto Villa, Peng Chen, and Omar Ghattas. Derivative-informed projected neural networks for high-dimensional parametric maps governed by pdes. *Computer Methods in Applied Mechanics and Engineering*, 388:114199, 2022. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2021.114199. URL https://www.sciencedirect.com/science/article/pii/S0045782521005302.

Philipp Petersen and Felix Voigtlaender. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural networks : the official journal of the International Neural Network Society*, 108:296–330, 2018. doi: 10.1016/j.neunet.2018.08.019.

Philipp Petersen, Mones Raslan, and Felix Voigtlaender. Topological properties of the set of functions generated by neural networks of fixed size. *Foundations of Computational Mathematics*, 21(2):375–444, 2021. ISSN 1615-3375. doi: 10.1007/s10208-020-09461-0. URL https://link.springer.com/article/10.1007/s10208-020-09461-0.

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. ISBN 0-521-77088-2. doi: 10.1017/S0962492900002919. URL https://doi.org/10.1017/S0962492900002919.

Tomaso Poggio, Hrushikesh Mhaskar, Lorenzo Rosasco, Brando Miranda, and Qianli Liao. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. *International Journal of Automation and Computing*, 14(5):503–519, 2017.

David Pollard. *Convergence of Stochastic Processes*. Springer New York, New York, NY, 1984. ISBN 978-1-4612-9758-1. doi: 10.1007/978-1-4612-5254-2.

Alfio Quarteroni, Andrea Manzoni, and Federico Negri. *Reduced basis methods for partial differential equations*, volume 92 of *Unitext*. Springer, Cham, 2016. ISBN 978-3-319-15430-5. doi: 10.1007/978-3-319-15431-2. URL https://doi.org/10.1007/978-3-319-15431-2. An introduction, La Matematica per il 3+2.

Bogdan Raonic, Roberto Molinaro, Tobias Rohner, Siddhartha Mishra, and Emmanuel de Bezenac. Convolutional neural operators. In *ICLR 2023 Workshop on Physics for Machine Learning*, 2023.

Holger Rauhut and Christoph Schwab. Compressive sensing Petrov-Galerkin approximation of high-dimensional parametric operator equations. *Math. Comp.*, 86(304):661–700, 2017. ISSN 0025-5718. doi: 10.1090/mcom/3113. URL http://dx.doi.org/10.1090/mcom/3113. Report 2014-14, Seminar for Applied Mathematics, ETH Zürich.

Markus Reiß. Asymptotic equivalence for nonparametric regression with multivariate and random design. *Ann. Statist.*, 36(4):1957–1982, 2008. ISSN 0090-5364. doi: 10.1214/07-AOS525. URL http://dx.doi.org/10.1214/07-AOS525.

Johannes Schmidt-Hieber. Supplement to "nonparametric regression using deep neural networks with relu activation function". *The Annals of Statistics*, 48(4), 2020a. ISSN 0090-5364. doi: 10.1214/19-AOS1875SUPP.

Johannes Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *The Annals of Statistics*, 48(4), 2020b. doi: 10.1214/19-aos1875. URL https://doi.org/10.1214%2F19-aos1875.

C. Schwab and A. M. Stuart. Sparse deterministic approximation of Bayesian inverse problems. *Inverse Problems*, 28(4):045003, 32, 2012. ISSN 0266-5611,1361-6420. doi: 10.1088/0266-5611/28/4/045003. URL https://doi.org/10.1088/0266-5611/28/4/045003.

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. ISSN 0219-5305,1793-6861. doi: 10.1142/S0219530518500203. URL https://doi.org/10.1142/S0219530518500203.

Christoph Schwab and Jakob Zech. Deep learning in high dimension: neural network expression rates for analytic functions in $L^2(\mathbb{R}^d, \gamma_d)$. *SIAM/ASA J. Uncertain. Quantif.*, 11(1):199–234, 2023. ISSN 2166-2525. doi: 10.1137/21M1462738. URL https://doi.org/10.1137/21M1462738.

Christoph Schwab, Andreas Stein, and Jakob Zech. Deep operator network approximation rates for lipschitz operators. *Analysis and Applications*, pages 1–41, 2025. doi: 10.1142/S0219530525500307. URL https://doi.org/10.1142/S0219530525500307.

Euan A Spence and Jared Wunsch. Wavenumber-explicit parametric holomorphy of helmholtz solutions in the context of uncertainty quantification. *SIAM/ASA Journal on Uncertainty Quantification*, 11(2):567–590, 2023.

Andrew M. Stuart. Inverse problems: a Bayesian perspective. *Acta Numer.*, 19:451–559, 2010. ISSN 0962-4929. doi: 10.1017/S0962492910000061.

Taiji Suzuki. Adaptivity of deep relu network for learning in besov and mixed smooth besov spaces: optimal rate and curse of dimensionality. In *he 7th International Conference on Learning Representations (ICLR2019)*, volume 7, 2019.

Michel Talagrand. *The generic chaining: Upper and lower bounds for stochastic processes*. Springer, Berlin and New York, 2005. ISBN 3-540-24518-9. doi: 10.1007/3-540-27499-5.

Michel Talagrand. *Upper and lower bounds for stochastic processes*, volume 60. Springer, 2014.

Hans Triebel. *Interpolation theory, function spaces, differential operators*. Barth, Heidelberg and Leipzig, 2., rev. and enl. ed. edition, 1995. ISBN 3335004205.

Alexandre B Tsybakov. *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, Dordrecht, 2009. doi: 10.1007/b13794. URL `https://cds.cern.ch/record/1315296`.

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge U. Press, 2000.

Sara van de Geer. Least squares estimation with complexity penalties. *Mathematical Methods of statistics*, 10(3):355, 2001.

Sara A. van de Geer. *Applications of empirical process theory*. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, digitally printed version edition, 2000. ISBN 052165002X.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47 of *Cambridge series in statistical and probabilistic mathematics*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, 2018. ISBN 1108415199.

Dongbin Xiu and George Em Karniadakis. The wiener–askey polynomial chaos for stochastic differential equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002. doi: 10.1137/S1064827501387826. URL `https://doi.org/10.1137/S1064827501387826`.

Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural networks : the official journal of the International Neural Network Society*, 94:103–114, 2017. doi: 10.1016/j.neunet.2017.07.002.

Jakob Zech. *Sparse-Grid Approximation of High-Dimensional Parametric PDEs*. PhD thesis, ETH Zurich, 2018.