Foundations of Computational Mathematics (2024) 24:1085–1143
https://doi.org/10.1007/s10208-023-09607-w

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Proof of the Theory-to-Practice Gap in Deep Learning via Sampling Complexity bounds for Neural Network Approximation Spaces

Philipp Grohs[1,2,3] · **Felix Voigtlaender**[1,4,5]

Received: 29 October 2021 / Revised: 15 September 2022 / Accepted: 26 September 2022 /
Published online: 12 July 2023
© The Author(s) 2023

Abstract

We study the computational complexity of (deterministic or randomized) algorithms based on point samples for approximating or integrating functions that can be well approximated by neural networks. Such algorithms (most prominently stochastic gradient descent and its variants) are used extensively in the field of deep learning. One of the most important problems in this field concerns the question of whether it is possible to realize theoretically provable neural network approximation rates by such algorithms. We answer this question in the negative by proving hardness results for the problems of approximation and integration on a novel class of neural network approximation spaces. In particular, our results confirm a conjectured and empirically observed theory-to-practice gap in deep learning. We complement our hardness

Communicated by Teresa Krick and Hans Munthe-Kaas.

Philipp Grohs and Felix Voigtlaender contributed equally to this work.

Invited paper associated to the FoCM 2021 Online Seminar lecture Deep Learning in Numerical Analysis presented by Philipp Grohs in May 2021.

F. Voigtlaender acknowledges support by the German Research Foundation (DFG) in the context of the Emmy Noether junior research group VO 2594/1–1.

✉ Philipp Grohs
philipp.grohs@univie.ac.at

1 Faculty of Mathematics, University of Vienna, Oskar-Morgenstern-Platz 1, 1090 Vienna, Austria

2 Research Platform Data Science @ Uni Vienna, Währinger Straße 29/S6, 1090 Vienna, Austria

3 Johann Radon Institute, Altenberger Straße 69, 4040 Linz, Austria

4 Department of Mathematics, Technical University of Munich, Boltzmannstr. 3, 85748 Garching, Germany

5 Mathematical Institute for Machine Learning and Data Science (MIDS), Catholic University Eichstätt-Ingolstadt (KU), Auf der Schanz 49, 85049 Ingolstadt, Germany

results by showing that error bounds of a comparable order of convergence are (at least theoretically) achievable.

Keywords Deep neural networks · Approximation spaces · Information based complexity · Gelfand numbers · Theory-to-computational gaps · Randomized approximation

Mathematics Subject Classification Primary 41A46 · 68T07; Secondary 41A65 · 41A25 · 68T05 · 65Y20

1 Introduction

The use of data-driven classification and regression algorithms based on deep neural networks—coined *deep learning*—has made a big impact in the areas of artificial intelligence, machine learning, and data analysis and has led to a number of breakthroughs in diverse areas of artificial intelligence, including image classification [24, 29, 32, 47], natural language processing [53], game playing [34, 45, 46, 51], and symbolic mathematics [31, 42].

More recently, these methods have been applied to problems from the natural sciences where data driven approaches are combined with physical models. Example applications in this field—called *scientific machine learning*—include the development of drugs [33], molecular dynamics [18], high-energy physics [5], protein folding [43], or numerically solving inverse problems and partial differential equations (PDEs) [4, 17, 26, 37, 40].

For this wide variety of different application areas, one can summarize the underlying computational problem as approximating an unknown function f (or a quantity of interest depending on f) based on possibly noisy and random samples $(f(x_i))_{i=1}^m$. In deep learning this is being done by fitting a neural network to these samples using stochastic optimization algorithms. While there is still no convincingly comprehensive explanation for the empirically observed success (or failure) of this methodology, its success critically hinges on the properties

A. that f can be well approximated by neural networks, and
B. that f (or a quantity of interest depending on f) can be efficiently and accurately reconstructed from a relatively small number of samples $(f(x_i))_{i=1}^m$.

In other words, the validity of both A and B constitutes a *necessary* condition for a deep learning approach to be efficient. This is especially true in applications related to scientific machine learning where often a guaranteed high accuracy is required and where obtaining samples is computationally expensive.

To date most theoretical contributions focused on Property A, namely studying which functions can be well approximated by neural networks. It is now well understood that neural networks are superior approximators compared to virtually all classical approximation methods, including polynomials, finite elements, wavelets, or low rank representations; see [15, 22] for two recent surveys. Beyond that it was recently shown that neural networks can approximate solutions of high dimensional

PDEs without suffering from the curse of dimensionality [21, 27, 30]. In light of these results it becomes clear that neural networks are a highly expressive and versatile function class whose theoretical approximation capabilities vastly outperform classical numerical function representations.

On the other hand, the question of whether property B holds, namely to which extent these superior approximation properties can be harnessed by an efficient algorithm based on point samples, remains one of the most relevant open questions in the field of deep learning. At present, almost no theoretical results exist in this direction. On the empirical side, Adcock and Dexter [1] recently performed a careful study finding that the theoretical approximation rates are in general not attained by common algorithms, meaning that the convergence rate of these algorithms does not match the theoretically postulated approximation rates. In [1] this empirically observed phenomenon is coined the *theory-to-practice gap* of deep learning. In this paper we prove the existence of this gap.

1.1 Description of Results

To provide an appropriate mathematical framework for understanding Properties A and B we introduce neural network spaces which classify functions $f : [0, 1]^d \to \mathbb{R}$ according to how rapidly the error of approximation by neural networks with n weights decays as $n \to \infty$. Specifically we consider neural networks using the rectified linear unit (ReLU) activation function, i.e., functions of the form

$$g = T_L \circ (\varrho \circ T_{L-1}) \circ \cdots \circ (\varrho \circ T_1), \tag{1.1}$$

where

$$T_\ell\, x = A_\ell\, x + b_\ell \tag{1.2}$$

are affine mappings and $\varrho\big((x_1, \dots, x_n)\big) = \big(\max\{x_1, 0\}, \dots, \max\{x_n, 0\}\big)$. Referring to L as the depth of the neural network (1.1) and to total number of nonzero coefficients of the matrix-vector pairs $(A_\ell, b_\ell)_{\ell=1}^L$ in (1.2) as number of weights of the neural network, we can formalize the property of being well approximable by neural networks as follows.

For $\alpha > 0$ let

$$\begin{aligned} U^\alpha := \big\{ f : [0,1]^d \to \mathbb{R} \ : \ &\text{for every } n \in \mathbb{N} \\ &\text{there is a ReLU neural network } g \text{ with depth } L \text{ and} \\ &n \text{ weights of magnitude at most 1 such that } \|f - g\|_\infty \le n^{-\alpha} \big\} \end{aligned} \tag{1.3}$$

In words, the sets U^α consist of all functions that are approximable by neural networks with depth L and at most n uniformly bounded coefficients to within uniform accuracy $\lesssim n^{-\alpha}$. For the remainder of the introduction we will say that f *can be approximated at rate* α *by depth* L *neural networks* if $f \in U^\alpha$.

We emphasize that all our results apply to much more general approximation spaces than the sets U^α (which is in fact the unit ball of some approximation space),

incorporating more complex constraints on the approximating neural network while considering approximation with respect to arbitrary L^p norms; see Sect. 2.2 for more details. In any case, for the current discussion it is sufficient to note that membership of f in such a space for large α simply means that Property A is satisfied.

For the mathematical formalization of Property B we employ the formalism of *Information Based Complexity* (more precisely we will study *sampling numbers of neural network approximation spaces*), as for example presented in [25]. This theory provides a general framework for studying the complexity of approximating a given solution mapping $S : U \to Y$, with $U \subset C([0,1]^d)$ bounded, and Y a Banach space, under the constraint that the approximating algorithm is only allowed to access *point samples* of the functions $f \in U$. Formally, a (deterministic) algorithm using m point samples is determined by a set of sample points $\boldsymbol{x} = (x_1, \dots, x_m) \in ([0,1]^d)^m$ and a map $Q : \mathbb{R}^m \to Y$ such that

$$A(f) = Q\big(f(x_1), \dots, f(x_m)\big) \qquad \forall\, f \in U.$$

The set of all such algorithms is denoted $\mathrm{Alg}_m(U, Y)$ and we define the optimal order for (deterministically) approximating $S : U \to Y$ using point samples as the best possible convergence rate with respect to the number of samples:

$$\beta_*^{\mathrm{det}}(U, S) := \sup\Big\{\beta \geq 0 : \ \exists C > 0\ \forall m \in \mathbb{N} : \ \inf_{A \in \mathrm{Alg}_m(U,Y)} \sup_{f \in U} \|A(f) - S(f)\|_Y \leq C \cdot m^{-\beta}\Big\}.$$

In a similar way one can define randomized algorithms and consider the optimal order $\beta_*^{\mathrm{ran}}(U, S)$ for approximating S using randomized algorithms based on point samples; see Sect. 2.4.2 below. We emphasize that *all currently used deep learning algorithms, such as stochastic gradient descent (SGD)* [44] *and its variants (such as ADAM* [28]*) are of this form.*

In this paper we derive bounds for the optimal orders $\beta_*^{\mathrm{det}}(U, S)$ and $\beta_*^{\mathrm{ran}}(U, S)$ for the unit ball $U = U^\alpha$ and the following solution mappings:

1. The embedding into $C([0,1]^d)$, i.e., $S = \iota_\infty$ for $\iota_\infty : U \to C([0,1]^d)$, $f \mapsto f$,
2. The embedding into $L^2([0,1]^d)$, i.e., $S = \iota_2$ for $\iota_2 : U \to L^2([0,1]^d)$, $f \mapsto f$, and
3. The definite integral, i.e., $S = T_\int$ for $T_\int : U \to \mathbb{R}$, $f \mapsto \int_{[0,1]^d} f(x)\,dx$.

1.1.1 Approximation with Respect to the Uniform Norm

We first consider the solution mapping $S = \iota_\infty$ operating on $U = U^\alpha$, i.e., the problem of approximation with respect to the uniform norm. Then the property $\beta_*^{\mathrm{ran}}(U, \iota_\infty) = \alpha$ would imply that the theoretical approximation rate α with respect to the uniform norm can in principle be realized by a (randomized) algorithm such as SGD and its variants. On the other hand, if $\beta_*^{\mathrm{ran}}(U, \iota_\infty) < \alpha$, then there cannot exist any (randomized) algorithm based on point samples that realizes the theoretical approximation rate α with respect to the uniform norm—that is, there exists a theory-to-practice gap. We now present (a slightly simplified version of) our first main result establishing such a gap for ι_∞.

Theorem 1.1 *(special case of Theorems 4.2 and 5.1) We have*

$$\beta_*^{\mathrm{ran}}\big(U^\alpha, \iota_\infty\big) = \beta_*^{\mathrm{det}}\big(U^\alpha, \iota_\infty\big) = \frac{1}{d} \cdot \frac{\alpha}{\lfloor L/2 \rfloor + \alpha} \in \big[0, \tfrac{1}{d}\big].$$

Theorem 1.1 states that for every $\beta < \frac{1}{d} \cdot \frac{\alpha}{\lfloor L/2 \rfloor + \alpha}$ and for every $m \in \mathbb{N}$ there exists an algorithm using m point samples such that every function $f \in U^\alpha$ (i.e., f can be approximated at rate α by depth L neural networks) can be reconstructed to within L^∞ error $\lesssim m^{-\beta}$. Conversely, this rate is the maximally achievable rate. Note the big discrepancy between the approximation rate α and the maximally achievable reconstruction rate $\frac{1}{d} \cdot \frac{\alpha}{\lfloor L/2 \rfloor + \alpha}$, especially for large input dimensions d. Probably the term "gap" is a vast understatement for the difference between the theoretical approximation rate α and the rate $\beta_* \leq \min\{\frac{1}{d}, \frac{\alpha}{d}\}$ that can actually be realized by a numerical algorithm. A particular consequence of Theorem 1.1 is that if all one knows is that a function f is well approximated by neural networks— no matter how rapidly the approximation error decays—any conceivable numerical algorithm based on function samples (such as SGD and its variants) requires at least $\Theta(\varepsilon^{-d})$ many samples to guarantee an error $\varepsilon > 0$ with respect to the uniform norm. Since evaluating f takes a certain minimum amount of time, *any conceivable numerical algorithm based on function samples (such as SGD and its variants) must have a worst-case runtime scaling at least as* $\Theta(\varepsilon^{-d})$ *to guarantee an error* $\varepsilon > 0$ *with respect to the uniform norm—irrespective of how well* f *can be theoretically approximated by neural networks.* In particular:

- Any conceivable numerical algorithm based on function samples (such as SGD and its variants) suffers from the curse of dimensionality—even if neural network approximations exist that do not.
- On the class of all functions well approximable by neural networks it is impossible to realize these high convergence rates for uniform approximation with any conceivable numerical algorithm based on function samples (such as SGD and its variants).
- If the number of layers is unbounded it is impossible to realize *any* positive convergence rate on the class of all functions well approximable by neural networks for the problem of uniform approximation with any conceivable numerical algorithm based on function samples (such as SGD and its variants).

Our findings disqualify deep learning-based methods for problems where high uniform accuracy is desired, at least if the only available information is that the function of interest is well approximated by neural networks.

1.1.2 Approximation with Respect to the L^2 Norm

Next we consider the solution mapping $S = \iota_2$ operating on $U = U^\alpha$, i.e., the problem of approximation with respect to the L^2 norm. Also in this case we establish a considerable theory-to-practice gap, albeit not as severe as in the case of $S = \iota_\infty$. A slightly simplified version of our main result is as follows.

Theorem 1.2 *(special case of Theorems 6.3 and 7.1) We have*

$$\beta_*^{\mathrm{ran}}\big(U^\alpha, \iota_2\big), \beta_*^{\det}\big(U^\alpha, \iota_2\big) \in \left[\frac{1}{2+2/\alpha}, \frac{1}{2} + \frac{\alpha}{\lfloor L/2 \rfloor + \alpha}\right].$$

We see again that it is impossible to realize a high convergence rate with any conceivable algorithm based on point samples, no matter how high the theoretically possible approximation rate α may be. Indeed, the theorem easily implies $\beta_*^{\mathrm{ran}}\big(U^\alpha, \iota_2\big), \beta_*^{\det}\big(U^\alpha, \iota_2\big) \leq \frac{3}{2}$, irrespective of α. This means that *any conceivable (possibly randomized) numerical algorithm based on function samples (such as SGD and its variants) must have a worst-case runtime scaling at least as* $\Theta(\varepsilon^{-2/3})$ *to guarantee an* L^2 *error* $\varepsilon > 0$*—irrespective of how well the function of interest can be theoretically approximated by neural networks.* On the positive side, there is a uniform lower bound of $\frac{1}{2+\frac{2}{\alpha}}$ for the optimal rate, which means that there exist algorithms (in the sense defined above) that almost realize an error bound of $\mathcal{O}(m^{-1/2})$, given m point samples, for α sufficiently large. Note however that the existence of such an algorithm by no means implies the existence of an *efficient* algorithm, say, with runtime scaling linearly or even polynomially in m.

Our findings disqualify deep learning-based methods for problems where a high convergence rate of the L^2 error is desired, at least if the only available information is that the function of interest is well approximated by neural networks. On the other hand, deep learning based methods may be a viable option for problems where a low—but dimension independent—convergence rate of the L^2 error is sufficient.

1.1.3 Integration

Finally we consider the solution mapping $S = T_\int$ operating on $U = U^\alpha$. The question of estimating $\beta_*^{\mathrm{ran}}\big(U^\alpha, T_\int\big)$ and $\beta_*^{\det}\big(U^\alpha, T_\int\big)$ can be equivalently stated as the question of determining the optimal order of (Monte Carlo or deterministic) quadrature on neural network approximation spaces. Again we exhibit a significant theory-to-practice gap that we summarize in the following simplified version of our main result.

Theorem 1.3 *(special case of Theorems 9.1, 9.4, 8.1 and 8.4) We have*

$$\beta_*^{\det}\big(U^\alpha, T_\int\big) \in \left[\frac{1}{2+1/\alpha}, 1 + \frac{\alpha}{\lfloor L/2 \rfloor + \alpha}\right].$$
$$\beta_*^{\mathrm{ran}}\big(U^\alpha, T_\int\big) \in \left[\frac{1}{2} + \frac{1}{2+2/\alpha}, 1 + \frac{\alpha}{\lfloor L/2 \rfloor + \alpha}\right].$$

We see in particular that *there are no (deterministic or Monte Carlo) quadrature schemes achieving a convergence order greater than 2. Further, if the number of layers is unbounded, there are no (deterministic or Monte Carlo) quadrature schemes achieving a convergence order greater than 1.* On the other hand there exist Monte Carlo algorithms that almost realize a rate 1 for α sufficiently large. This again does not

imply the existence of an *efficient* algorithm with this convergence rate; but it is well-known that the error bound $\mathcal{O}(m^{-1/2})$ can be *efficiently* realized by standard Monte Carlo integration, Theorem 1.3 implies that there is not much room for improvement.

1.1.4 General Comments

We close the overview of our results with the following general comments.

- Our results for the first time shed light on the question of which problem classes can be efficiently tackled by deep learning methods and which problem classes might be better handled using classical methods such as finite elements. These findings enable informed choices regarding the use of these methods. Concretely, we find that it is not advisable to use deep learning methods for problems where a high convergence rate and/or uniform accuracy is needed. In particular, *no high order (approximation or quadrature) algorithms exist*, provided that the only available information is that the function of interest is well approximated by neural networks.
- As another contribution, we exhibit the exact impact of the choice of the architecture, such as the number of layers, and magnitude of the coefficients. Particularly, we show that allowing the number of layers to be unbounded adversely affects the optimal rate β_*.
- Our hardness results hold universally across virtually all choices of network architectures. Concretely, all hardness results of Theorems 1.1, 1.2 and 1.3 hold true whenever at least 3 layers are used. This means that *limiting the number of layers will not help*. In this context we also note that it is known that at least $\lfloor \alpha/2d \rfloor$ layers are needed for ReLU neural networks to achieve the (essentially) optimal approximation rate $\frac{\alpha}{d}$ for all $f \in C^\alpha([0,1]^d)$; see [36, Theorem C.6].
- Our hardness results hold universally across all size constraints on the magnitudes of the approximating network weights. Furthermore, a careful analysis of our proofs reveals that our hardness results qualitatively remain true if analogous constraints are put on the ℓ^2 norms of the weights of the approximating networks. Such constraints constitute a common regularization strategy, termed *weight decay* [23]. This means that *applying standard regularization strategies—such as weight decay—will not help*.
- In many machine learning problems one assumes that one only has access to inexact (noisy) samples of a given function. Since this noise can be incorporated into the stochasticity of a randomized algorithm, our hardness results also hold for the case of noisy samples.

1.2 Related Work

To put our results in perspective we discuss related work.

1.2.1 Information-Based Complexity and Classical Function Spaces

The study of optimal rates β_* for approximating a given solution map based on point samples or general linear samples has a long tradition in approximation theory, function

space theory, spectral theory and information based complexity. It is closely related to so-called *Gelfand numbers* of linear operators—a classical and well-studied concept in function space theory and spectral theory [38, 39]. It is instructive to compare our findings to these classical results, for example for U the unit ball in a Sobolev spaces $W_\infty^\alpha([0,1]^d)$ and $S = \iota_\infty$. These Sobolev spaces can be (not quite but almost, see for example [49, Theorem 5.3.2] and [16, Theorem 12.1.1]) characterized by the property that its elements can be approximated by polynomials of degree $\leq n$ to within L^∞ accuracy $\mathcal{O}(n^{-\alpha})$. Since the set of polynomials of degree $\leq n$ in dimension d possesses $\asymp n^d$ degrees of freedom, this approximation rate can be fully harnessed by a deterministic, resp. randomized algorithm based on point samples if $\beta_*^{\det}(U, S) = \alpha/d$, resp. $\beta_*^{\mathrm{ran}}(U, S) = \alpha/d$. It is a classical result that this is indeed the case, see [25, Theorem 6.1]. This fact implies that there is no theory-to-practice gap in polynomial approximation and can be considered the basis of any high order (approximation or quadrature) algorithm in numerical analysis.

In the case of classical function spaces it is the generic behavior that the optimal rate β_* increases (linearly) with the underlying smoothness α, at least for fixed dimension d. On the other hand, our results show that neural network approximation spaces have the peculiar property that the optimal rate β_* is always uniformly bounded, regardless of the underlying "smoothness" α.

To put our results in a somewhat more abstract context we can compare the optimal rate β_* to other complexity measures of a function space. A well studied example is the metric entropy related to the covering numbers $\mathrm{Cov}(V, \varepsilon)$ of sets $V \subset C[0,1]^d$. The associated entropy exponent is

$$s_*(U) := \sup\left\{\lambda \geq 0 \,:\, \exists\, C > 0\ \forall\, \varepsilon \in (0,1) :\quad \mathrm{Cov}(U,\varepsilon) \leq \exp\left(C \cdot \varepsilon^{-1/\lambda}\right)\right\},$$

which, roughly speaking, determines the theoretically optimal rate $\mathcal{O}(m^{-s_*})$ at which an arbitrary element of U can be approximated from a representation using at most m bits. On the other hand, β_* determines the optimal rate $\mathcal{O}(m^{-\beta_*})$ that can actually be realized by an algorithm using m point samples of the input function $f \in U$. For a solution mapping S to be efficiently computable from point samples, one would therefore expect that $\beta_* = s_*$ or at least that β_* grows linearly with s_*. For example, for U the unit ball in a Sobolev spaces $W_\infty^\alpha([0,1]^d)$ and $S = \iota_\infty$ we have $s_*(U) = \beta_*^{\det}(U, \iota_\infty) = \beta_*^{\mathrm{ran}}(U, \iota_\infty) = \frac{\alpha}{d}$. In contrast, $U^\alpha = U^\alpha([0,1]^d)$ satisfies $s_*(U^\alpha) \geq \alpha$ according to Lemma 6.2, while Theorem 1.1 shows $\beta_*^{\det}(U^\alpha, \iota_\infty), \beta_*^{\mathrm{ran}}(U^\alpha, \iota_\infty) \leq \frac{1}{d}$ independent of α, and even $\beta_*^{\det}(U^\alpha, \iota_\infty) = \beta_*^{\mathrm{ran}}(U^\alpha, \iota_\infty) = 0$ if the number of layers is unbounded. This is yet another manifestation of the wide theory-to-practice gap in neural network approximation.

1.2.2 Other Hardness Results for Deep Learning

While we are not aware of any work addressing the optimal sampling complexity on neural network spaces, there exist a number of different approaches to establishing various "hardness" results for deep learning. We comment on some of them.

A prominent and classical research direction considers the computational complexity of fitting a neural network of a fixed architecture to given (training) samples. It

is known that this can be an NP complete problem for certain specific architectures and samples; see [9] for the first result in this direction that has inspired a large body of follow-up work. This line of work does however not consider the full scope of the problem, namely the relation between theoretically possible approximation rates and algorithmically realizable rates. In our results we do not take into account the computational efficiency of algorithms at all. Our results are stronger in the sense that they show that *even if there was an efficient algorithm for fitting a neural network to samples, one would need to access too many samples to achieve efficient runtimes.*

Another research direction considers the existence of convergent algorithms that only have access to inexact information about the samples, as is commonly the case when computing in floating point arithmetic. Specifically, [3] identifies various problems in sparse approximation that cannot be algorithmically solved based on inputs with finite precision using neural networks. The deeper underlying reason is that these problems cannot be solved by *any* algorithm based on inexact measurements. Thus, the results of [3] are not really specific to neural networks. In contrast, *our hardness results are highly specific to the structure of neural networks and do not occur for most other computational approaches.*

A different kind of hardness results appears in the neural network approximation theory literature. There, typically lower bounds are provided for the number of network weights and/or number of layers that a neural network needs to have in order to reach a desired accuracy in the approximation of functions from various classical smoothness spaces [10, 36, 48, 52]. Yet, these bounds exclusively concern theoretical approximation rates for classical smoothness spaces while *our results provide bounds for the realizability of these rates based on point samples*

1.2.3 Other Work on Neural Network Approximation Spaces

Our definition of neural network approximation spaces is inspired by [20] where such spaces were first introduced and some structural properties, such as embedding theorems into classical function spaces, are investigated. The neural network spaces $A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ introduced in the present work differ from those spaces in the sense that we also allow to take the size of the network weights into account. This is important, as such bounds on the weights are often enforced in applications through regularization procedures. Another construction of neural network approximation spaces can be found in [7] for the purpose of providing a calculus on functions that can be approximated by neural networks without curse of dimensionality. While all these works focus on aspects related to theoretical approximability of functions, our main focus concerns the algorithmic realization of such approximations.

1.3 Notation

For $n \in \mathbb{N}$, we write $\underline{n} := \{1, 2, \dots, n\}$. For any finite set $I \neq \varnothing$ and any sequence $(a_i)_{i\in I} \subset \mathbb{R}$, we define $\sum_{i\in I} a_i := \frac{1}{|I|}\sum_{i\in I} a_i$. The expectation of a random variable X will be denoted by $\mathbb{E}[X]$.

For a subset $M \subset X$ of a metric space X, we write $\overline{M}$ for the closure of M and M° for the interior of M. In particular, this notation applies to subsets of $\mathbb{R}^d$. We write $\boldsymbol{\lambda}(M)$ for the Lebesgue measure of a (measurable) set $M \subset \mathbb{R}^d$.

1.4 Structure of the paper

Section 2 formally introduces the neural network approximation spaces $A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}$ and furthermore provides a review of the most important notions and definitions from information based complexity. The basis for all our hardness results is developed in Sect. 3, where we show that the unit ball $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ in the approximation space $A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ contains a large family of "hat functions", depending on the precise properties of the functions $\boldsymbol{\ell}$, $\boldsymbol{c}$ and on $\alpha > 0$.

The remaining sections develop error bounds and hardness results for the problems of uniform approximation (Sects. 4 and 5), approximation in L^2 (Sects. 6 and 7), and numerical integration (Sects. 8 and 9). Several technical proofs and results are deferred to Sect. A.

2 The Notion of Sampling Complexity on Neural Network Approximation Spaces

In this section, we first formally introduce the neural network approximation spaces $A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}$ and then review the framework of information based complexity, including the notion of randomized algorithms and the concept of the optimal order of convergence based on point samples.

2.1 The Mathematical Formalization of Neural Networks

In our analysis, it will be helpful to distinguish between a neural network Φ as a set of weights and the associated function $R_\varrho\Phi$ computed by the network. Thus, we say that a *neural network* is a tuple $\Phi = \big((A_1, b_1), \dots, (A_L, b_L)\big)$, with $A_\ell \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ and $b_\ell \in \mathbb{R}^{N_\ell}$. We then say that $\boldsymbol{a}(\Phi) := (N_0, \dots, N_L) \in \mathbb{N}^{L+1}$ is the *architecture* of Φ, $L(\Phi) := L$ is the *number of layers* [1] of Φ, and $W(\Phi) := \sum_{j=1}^{L} (\|A_j\|_{\ell^0} + \|b_j\|_{\ell^0})$ denotes the *number of (nonzero) weights* of Φ. The notation $\|A\|_{\ell^0}$ used here denotes the number of nonzero entries of a matrix (or vector) A. Finally, we write $d_{\text{in}}(\Phi) := N_0$ and $d_{\text{out}}(\Phi) := N_L$ for the *input and output dimension* of Φ, and we set $\|\Phi\|_{\mathcal{NN}} := \max_{j=1,\dots,L} \max\{\|A_j\|_\infty, \|b_j\|_\infty\}$, where $\|A\|_\infty := \max_{i,j} |A_{i,j}|$.

To define the function $R_\varrho\Phi$ computed by Φ, we need to specify an *activation function*. In this paper, we will only consider the so-called *rectified linear unit (ReLU)* $\varrho : \mathbb{R} \to \mathbb{R}, x \mapsto \max\{0, x\}$, which we understand to act componentwise on $\mathbb{R}^n$, i.e., $\varrho\big((x_1, \dots, x_n)\big) = \big(\varrho(x_1), \dots, \varrho(x_n)\big)$. The function $R_\varrho\Phi : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$ computed by the network Φ (its *realization*) is then given by

$$R_\varrho\Phi := T_L \circ (\varrho \circ T_{L-1}) \circ \cdots \circ (\varrho \circ T_1) \quad \text{where} \quad T_\ell\, x = A_\ell\, x + b_\ell.$$

[1] Note that the number of *hidden* layers is given by $H = L - 1$.

2.2 Neural Network Approximation Spaces

Approximation spaces [14] classify functions according to how well they can be approximated by a family $\boldsymbol{\Sigma} = (\Sigma_n)_{n\in\mathbb{N}}$ of certain "simple functions" of increasing complexity n, as $n \to \infty$. Common examples consider the case where Σ_n is the set of polynomials of degree n, or the set of all linear combinations of n wavelets. The notion of *neural network approximation spaces* was originally introduced in [20], where Σ_n was taken to be a family of neural networks of increasing complexity. However, [20] *does not impose any restrictions on the size of the individual network weights*, which plays an important role in practice and—as we shall see—also influences the possible performance of algorithms based on point samples.

For this reason, we introduce a modified notion of neural network approximation spaces that also takes the size of the individual network weights into account. Precisely, given an input dimension $d \in \mathbb{N}$ (which we will keep fixed throughout this paper) and non-decreasing functions $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ (called the **depth-growth function** and the **coefficient growth function**, respectively), we define

$$\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} := \left\{ R_\varrho \Phi \ : \ \begin{array}{l} \Phi \text{ NN with } d_{\text{in}}(\Phi) = d, d_{\text{out}}(\Phi) = 1, \\ W(\Phi) \leq n, L(\Phi) \leq \boldsymbol{\ell}(n), \|\Phi\|_{\mathcal{NN}} \leq \boldsymbol{c}(n) \end{array} \right\}.$$

Then, given a measurable subset $\Omega \subset \mathbb{R}^d$, $p \in [1,\infty]$, and $\alpha \in (0,\infty)$, for each measurable $f : \Omega \to \mathbb{R}$, we define

$$\Gamma_{\alpha,p}(f) := \max\left\{ \|f\|_{L^p(\Omega)}, \quad \sup_{n\in\mathbb{N}} \left[n^\alpha \cdot d_p\big(f, \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}\big) \right] \right\} \in [0,\infty],$$

where $d_p(f,\Sigma) := \inf_{g\in\Sigma} \|f - g\|_{L^p(\Omega)}$.

The remaining issue is that since the set $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ is in general neither closed under addition nor under multiplication with scalars, $\Gamma_{\alpha,p}$ is *not a (quasi)-norm*. To resolve this issue, taking inspiration from the theory of Orlicz spaces (see e.g. [41, Theorem 3 in Section 3.2]), we define the *neural network approximation space quasi-norm* $\|\cdot\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}$ as

$$\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} := \inf\big\{\theta > 0 \ : \ \Gamma_{\alpha,p}(f/\theta) \leq 1\big\} \in [0,\infty],$$

giving rise to the *approximation space*

$$A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}} := A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}(\Omega) := \big\{ f \in L^p(\Omega) \ : \ \|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} < \infty \big\}.$$

The following lemma summarizes the main elementary properties of these spaces.

Lemma 2.1 *Let $\varnothing \neq \Omega \subset \mathbb{R}^d$ be measurable, let $p \in [1,\infty]$ and $\alpha \in (0,\infty)$. Then, $A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}} := A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}(\Omega)$ satisfies the following properties:*

1. $(A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}, \|\cdot\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}})$ *is a quasi-normed space. Precisely, given arbitrary measurable functions* $f, g : \Omega \to \mathbb{R}$, *it holds that* $\|f+g\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} \leq C \cdot (\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} + \|g\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}})$ *for* $C := 17^{\alpha}$.
2. *We have* $\Gamma_{\alpha,p}(cf) \leq |c|\,\Gamma_{\alpha,p}(f)$ *for* $c \in [-1,1]$.
3. $\Gamma_{\alpha,p}(f) \leq 1$ *if and only if* $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} \leq 1$.
4. $\Gamma_{\alpha,p}(f) < \infty$ *if and only if* $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} < \infty$.
5. $A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}(\Omega) \hookrightarrow L^p(\Omega)$. *Furthermore, if* $\Omega \subset \overline{\Omega^{\circ}}$, *then* $A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}(\Omega) \hookrightarrow C_b(\Omega)$, *where* $C_b(\Omega)$ *denotes the Banach space of continuous functions that are bounded and extend continuously to the closure* $\overline{\Omega}$ *of* Ω.

Proof See Sect. A.1. □

2.3 Quantities Characterizing the Complexity of the Network Architecture

To conveniently summarize those aspects of the growth behavior of the functions $\boldsymbol{\ell}$ and $\boldsymbol{c}$ most relevant to us, we introduce three quantities that will turn out to be crucial for characterizing the sample complexity of the neural network approximation spaces. First, we set

$$\boldsymbol{\ell}^* := \sup_{n\in\mathbb{N}} \boldsymbol{\ell}(n) \in \mathbb{N} \cup \{\infty\}. \tag{2.1}$$

Furthermore, we define

$$\begin{aligned} \gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) &:= \sup\left\{\gamma \in [0,\infty) \colon \exists\, L \in \mathbb{N}_{\leq \boldsymbol{\ell}^*} \text{ and } C > 0 \quad \forall\, n \in \mathbb{N} : n^{\gamma} \leq C \cdot (\boldsymbol{c}(n))^L \cdot n^{\lfloor L/2 \rfloor}\right\}, \\ \gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c}) &:= \inf\left\{\gamma \in [0,\infty) \colon \exists\, C > 0 \quad \forall\, n \in \mathbb{N}, L \in \mathbb{N}_{\leq \boldsymbol{\ell}^*} : (\boldsymbol{c}(n))^L \cdot n^{\lfloor L/2 \rfloor} \leq C \cdot n^{\gamma}\right\}. \end{aligned} \tag{2.2}$$

Remark 2.2 Clearly, $\gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) \leq \gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c})$. Furthermore, since we will only consider settings in which $\boldsymbol{\ell}^* \geq 2$, we always have $\gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c}) \geq \gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) \geq 1$. Next, note that if $\boldsymbol{\ell}^* = \infty$ (i.e., if $\boldsymbol{\ell}$ is unbounded), then $\gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) = \gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c}) = \infty$. Finally, we remark that if $\boldsymbol{\ell}^* < \infty$ and if $\boldsymbol{c}$ satisfies the natural growth condition $\boldsymbol{c}(n) \asymp n^{\theta} \cdot (\ln(2n))^{\kappa}$ for certain $\theta \geq 0$ and $\kappa \in \mathbb{R}$, then $\gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) = \gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c}) = \theta \cdot \boldsymbol{\ell}^* + \lfloor \boldsymbol{\ell}^*/2 \rfloor$. Thus, in most natural cases—but not always—$\gamma^{\flat}$ and $\gamma^{\sharp}$ agree.

An explicit example where $\gamma^{\flat}$ is not identical to $\gamma^{\sharp}$ is as follows: Define $c_1 := c_2 := c_3 := 1$ and for $n, m \in \mathbb{N}$ with $2^{2^m} \leq n < 2^{2^{m+1}}$, define $c_n := 2^{2^m}$. Then, assume that $\gamma_1, \gamma_2 \in [0,\infty)$ and $\kappa_1, \kappa_2 > 0$ satisfy $\kappa_1\, n^{\gamma_1} \leq c_n \leq \kappa_2\, n^{\gamma_2}$ for all $n \in \mathbb{N}$. Applying the upper estimate for arbitrary $m \in \mathbb{N}$ and $n = n_m = 2^{2^m}$, we see $n = c_n \leq \kappa_2\, n^{\gamma_2}$; since $n_m = 2^{2^m} \to \infty$ as $m \to \infty$, this is only possible if $\gamma_2 \geq 1$. On the other hand, if we apply the lower estimate for arbitrary $m \in \mathbb{N}$ and $n = n_m = 2^{2^{m+1}} - 1$, we see because of $c_n = 2^{2^m} = 2^{2^{m+1}/2} = \sqrt{2^{2^{m+1}}} = \sqrt{n+1}$ that $\kappa_1\, n^{\gamma_1} \leq c_n = \sqrt{n+1}$. Again, since $n_m = 2^{2^{m+1}} - 1 \to \infty$ as $m \to \infty$, this is only possible if $\gamma_1 \leq \frac{1}{2}$.

Given these considerations, it is easy to see for $\boldsymbol{\ell} \equiv L \in \mathbb{N}_{\geq 2}$ that $\gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) \leq \frac{L}{2} + \lfloor \frac{L}{2} \rfloor$, while $\gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c}) \geq L + \lfloor \frac{L}{2} \rfloor$. In particular, $\gamma^{\flat}(\boldsymbol{\ell},\boldsymbol{c}) < \gamma^{\sharp}(\boldsymbol{\ell},\boldsymbol{c})$. △

2.4 The Framework of Sampling Complexity

Let $d \in \mathbb{N}$, let $\varnothing \neq U \subset C([0,1]^d)$ be bounded, and let Y be a Banach space. We are interested in numerically approximating a given **solution mapping** $S : U \to Y$, where the numerical procedure is only allowed to access *point samples* of the functions $f \in U$. The procedure can be either deterministic or probabilistic (Monte Carlo). In this short section, we discuss the mathematical formalization of this problem, based on the setup of *numerical complexity theory*, as for instance outlined in [25, Section 2].

The reader should keep in mind that we are mostly interested in the setting where U is the unit ball in the neural network approximation space $A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$, i.e.,

$$U = U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) := \big\{ f \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \,:\, \|f\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}} \leq 1 \big\}, \tag{2.3}$$

and where the solution mapping is one of the following:

1. The embedding into $C([0,1]^d)$, i.e., $S = \iota_\infty$ for $\iota_\infty : U \to C([0,1]^d)$, $f \mapsto f$,
2. The embedding into $L^2([0,1]^d)$, i.e., $S = \iota_2$ for $\iota_2 : U \to L^2([0,1]^d)$, $f \mapsto f$, or
3. The definite integral, i.e., $S = T_\int$ for $T_\int : U \to \mathbb{R}$, $f \mapsto \int_{[0,1]^d} f(x)\,dx$.

2.4.1 The Deterministic Setting

A (potentially nonlinear) map $A : U \to Y$ is called a **deterministic method using $\mathbf{m} \in \mathbb{N}$ point measurements** (written $A \in \mathrm{Alg}_m(U,Y)$) if there exists $\boldsymbol{x} = (x_1, \ldots, x_m) \in ([0,1]^d)^m$ and a map $Q : \mathbb{R}^m \to Y$ such that

$$A(f) = Q\big(f(x_1), \ldots, f(x_m)\big) \qquad \forall\, f \in U.$$

Given a (solution) mapping $S : U \to Y$, we define the error of A in approximating S as

$$e(A, U, S) := \sup_{f \in U} \|A(f) - S(f)\|_Y.$$

The **optimal error for (deterministically) approximating $S : U \to Y$ using m point samples** is then

$$e^{\mathrm{det}}_m(U, S) := \inf_{A \in \mathrm{Alg}_m(U,Y)} e(A, U, S).$$

Finally, the **optimal order for (deterministically) approximating $S : U \to Y$ using point samples** is

$$\beta^{\mathrm{det}}_*(U, S) := \sup \big\{ \beta \geq 0 \,:\, \exists\, C > 0 \;\forall\, m \in \mathbb{N} : \quad e^{\mathrm{det}}_m(U, S) \leq C \cdot m^{-\beta} \big\}. \tag{2.4}$$

2.4.2 The Randomized Setting

A **randomized method using** $m \in \mathbb{N}$ **point measurements (in expectation)** is a tuple $(\boldsymbol{A}, \boldsymbol{m})$ consisting of a family $\boldsymbol{A} = (A_\omega)_{\omega\in\Omega}$ of (potentially nonlinear) maps $A_\omega : U \to Y$ indexed by a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable function $\boldsymbol{m} : \Omega \to \mathbb{N}$ with the following properties:

1. for each $f \in U$, the map $\Omega \to Y$, $\omega \mapsto A_\omega(f)$ is measurable (with respect to the Borel σ-algebra on Y),
2. for each $\omega \in \Omega$, we have $A_\omega \in \mathrm{Alg}_{\boldsymbol{m}(\omega)}(U, Y)$,
3. $\mathbb{E}_\omega[\boldsymbol{m}(\omega)] \le m$.

We write $(\boldsymbol{A}, \boldsymbol{m}) \in \mathrm{Alg}_m^{\mathrm{ran}}(U, Y)$ if these conditions are satisfied. We say that $(\boldsymbol{A}, \boldsymbol{m})$ is *strongly measurable* if the map $\Omega \times U \to Y$, $(\omega, f) \mapsto A_\omega(f)$ is measurable, where $U \subset C([0,1]^d)$ is equipped with the Borel σ-algebra induced by $C([0,1]^d)$.

Remark In most of the literature (see e.g. [25, Section 2]), randomized algorithms are always assumed to be strongly measurable. All randomized algorithms that we construct will have this property. On the other hand, all our hardness results apply to arbitrary randomized algorithms satisfying Properties 1–3 from above. Thus, using the terminology just introduced we obtain stronger results than we would get using the usual definition.

The **expected error** of a randomized algorithm $(\boldsymbol{A}, \boldsymbol{m})$ for approximating a (solution) mapping $S : U \to Y$ is defined as

$$e\big((\boldsymbol{A}, \boldsymbol{m}), U, S\big) := \sup_{f\in U} \mathbb{E}_\omega\big[\|S(f) - A_\omega(f)\|_Y\big].$$

The **optimal randomized error for approximating** $S : U \to Y$ **using** m **point samples (in expectation)** is

$$e_m^{\mathrm{ran}}(U, S) := \inf_{(\boldsymbol{A},\boldsymbol{m})\in \mathrm{Alg}_m^{\mathrm{ran}}(U,Y)} e\big((\boldsymbol{A}, \boldsymbol{m}), U, S\big).$$

Finally, the **optimal randomized order for approximating** $S : U \to Y$ **using point samples** is

$$\beta_*^{\mathrm{ran}}(U, S) := \sup\big\{\beta \ge 0 : \ \exists\, C > 0 \ \forall\, m \in \mathbb{N} : \quad e_m^{\mathrm{ran}}(U, S) \le C \cdot m^{-\beta}\big\}.$$

The remainder of this paper is concerned with deriving upper and lower bounds for the exponents $\beta_*^{\mathrm{det}}(U, S)$ and $\beta_*^{\mathrm{ran}}(U, S)$, where $U = U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$ is the unit ball in $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$, and S is either the embedding of $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$ into $C([0,1]^d)$, the embedding into $L^2([0,1]^d)$, or the definite integral $Sf = \int_{[0,1]^d} f(t)\,dt$.

For deriving upper bounds (i.e., hardness bounds) for randomized algorithms, we will frequently use the following lemma, which is a slight adaptation of [25, Proposition 4.1]. In a nutshell, the lemma shows that if one can establish a hardness result that holds for deterministic algorithms *in the average case*, then this implies a hardness result for randomized algorithms.

Lemma 2.3 *Let $\varnothing \neq U \subset C([0,1]^d)$ be bounded, let Y be a Banach space, and let $S : U \to Y$. Assume that there exist $\lambda \in [0,\infty)$, $\kappa > 0$, and $m_0 \in \mathbb{N}$ such that for every $m \in \mathbb{N}_{\geq m_0}$ there exists a finite set $\Gamma_m \neq \varnothing$ and a family of functions $(f_\gamma)_{\gamma\in\Gamma_m} \subset U$ satisfying*

$$\sum_{\gamma\in\Gamma_m} \|S(f_\gamma) - A(f_\gamma)\|_Y \geq \kappa \cdot m^{-\lambda} \qquad \forall\, A \in \mathrm{Alg}_m(U,Y). \tag{2.5}$$

Then $\beta^{\mathrm{det}}_*(U,S), \beta^{\mathrm{ran}}_*(U,S) \leq \lambda$.

Proof **Step 1** (**proving** $\beta^{\mathrm{det}}_*(U,S) \leq \lambda$): For every $A \in \mathrm{Alg}_m(U,Y)$, Eq. (2.5) implies because of $f_\gamma \in U$ that

$$e(A,U,S) = \sup_{f\in U} \|A(f) - S(f)\|_Y \geq \sum_{\gamma\in\Gamma_m} \|S(f_\gamma) - A(f_\gamma)\|_Y \geq \kappa\, m^{-\lambda}.$$

Since this holds for every $m \in \mathbb{N}_{\geq m_0}$ and every $A \in \mathrm{Alg}_m(U,Y)$, with κ independent of A, m, this easily implies $e^{\mathrm{det}}_m(U,S) \geq \kappa\, m^{-\lambda}$ for all $m \in \mathbb{N}_{\geq m_0}$, and then $\beta^{\mathrm{det}}_*(U,S) \leq \lambda$.

Step 2 (**proving** $\beta^{\mathrm{ran}}_*(U,S) \leq \lambda$): Let $m \in \mathbb{N}_{\geq m_0}$ and let $(\boldsymbol{A},\boldsymbol{m}) \in \mathrm{Alg}^{\mathrm{ran}}_m(U,Y)$ be arbitrary, with $\boldsymbol{A} = (A_\omega)_{\omega\in\Omega}$ for a probability space $(\Omega,\mathcal{F},\mathbb{P})$. Define $\Omega_0 := \{\omega \in \Omega : \boldsymbol{m}(\omega) \leq 2m\}$ and note $m \geq \mathbb{E}_\omega[\boldsymbol{m}(\omega)] \geq 2m \cdot \mathbb{P}(\Omega_0^c)$, which shows $\mathbb{P}(\Omega_0^c) \leq \frac{1}{2}$ and hence $\mathbb{P}(\Omega_0) \geq \frac{1}{2}$.

Note that $A_\omega \in \mathrm{Alg}_{2m}(U,Y)$ for each $\omega \in \Omega_0$, so that Eq. (2.5) (with $2m$ instead of m) shows $\sum_{\gamma\in\Gamma_{2m}} \big\|S(f_\gamma) - A_\omega(f_\gamma)\big\|_Y \geq \kappa \cdot (2m)^{-\lambda} \geq \widetilde{\kappa} \cdot m^{-\lambda}$ for a constant $\widetilde{\kappa} = \widetilde{\kappa}(\kappa,\lambda) > 0$. Therefore,

$$\begin{aligned} e\big((\boldsymbol{A},\boldsymbol{m}),U,S\big) = \sup_{f\in U} \mathbb{E}_\omega \|S(f) - A_\omega(f)\|_Y &\geq \sum_{\gamma\in\Gamma_{2m}} \mathbb{E}_\omega \big\|S(f_\gamma) - A_\omega(f_\gamma)\big\|_Y \\ &\geq \mathbb{E}_\omega\left[\mathbb{1}_{\Omega_0}(\omega) \sum_{\gamma\in\Gamma_{2m}} \big\|S(f_\gamma) - A_\omega(f_\gamma)\big\|_Y\right] \\ &\geq \mathbb{P}(\Omega_0)\cdot\widetilde{\kappa}\cdot m^{-\lambda} \geq \frac{\widetilde{\kappa}}{2}\cdot m^{-\lambda}, \end{aligned} \tag{2.6}$$

and hence $e^{\mathrm{ran}}_m\big(U,S\big) \geq \frac{\widetilde{\kappa}}{2}\cdot m^{-\lambda}$, since Eq. (2.6) holds for any randomized algorithm $(\boldsymbol{A},\boldsymbol{m}) \in \mathrm{Alg}^{\mathrm{ran}}_m(U,Y)$. Finally, since $m \in \mathbb{N}_{\geq m_0}$ can be chosen arbitrarily, we see as claimed that $\beta^{\mathrm{ran}}_*(U,S) \leq \lambda$. □

3 Richness of the Unit Ball in the Spaces $A^{\alpha,\infty}_{\ell,c}$

In this section, we show that ReLU networks with a limited number of neurons and bounded weights can well approximate several different functions of "hat-function type," as shown in Fig. 1. The fact that this is possible implies that the unit ball $U^{\alpha,\infty}_{\ell,c} \subset A^{\alpha,\infty}_{\ell,c}$ is quite rich; this will be the basis of all of our hardness results.

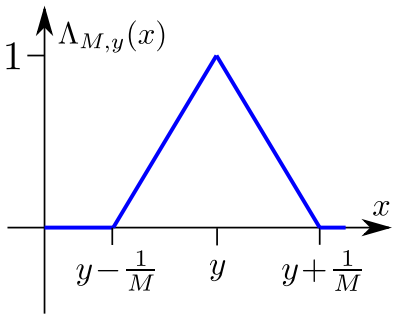

Fig. 1 A plot of the "hat-function" $\Lambda_{M,y}$ formally defined in Eq. (3.1)

We begin by considering the most basic "hat function" $\Lambda_{M,y} : \mathbb{R} \to [0, 1]$, defined for $M > 0$ and $y \in \mathbb{R}$ by

$$\Lambda_{M,y}(x) = \begin{cases} 0, & \text{if } x \le y - M^{-1}, \\ M \cdot (x - y + M^{-1}), & \text{if } y - M^{-1} \le x \le y, \\ -M \cdot (x - y - M^{-1}), & \text{if } y \le x \le y + M^{-1}, \\ 0, & \text{if } y + M^{-1} \le x. \end{cases} \tag{3.1}$$

For later use, we note that $\int_{\mathbb{R}} \Lambda_{M,y}(x)\, dx = M^{-1}$. Furthermore, we "lift" $\Lambda_{M,y}$ to a function on $\mathbb{R}^d$ by setting $\Lambda^*_{M,y} : \mathbb{R}^d \to \mathbb{R}, x = (x_1, \dots, x_d) \mapsto \Lambda_{M,y}(x_1)$. The following lemma gives a bound on how economically sums of the functions $\Lambda_{M,y}$ can be implemented by ReLU networks.

Lemma 3.1 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\ge 2} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ be non-decreasing. Let $M \ge 1$, $n \in \mathbb{N}$, and $0 < C \le \boldsymbol{c}(n)$, as well as $L \in \mathbb{N}_{\ge 2}$ with $L \le \boldsymbol{\ell}(n)$.*

Then

$$\frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4Mn} \sum_{i=1}^{n} \varepsilon_i \Lambda^*_{M,y_i} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{(2L+8)n}. \qquad \forall\, \varepsilon_1, \dots, \varepsilon_n \in [-1, 1] \text{ and } y_1, \dots, y_n \in [0, 1].$$

Proof Let $\varepsilon_1, \dots, \varepsilon_n \in [-1, 1]$ and $y_1, \dots, y_n \in [0, 1]$. Let $e_1 := (1, 0, \dots, 0) \in \mathbb{R}^{1\times d}$ and define

$$A_1 := \frac{C}{2} \begin{pmatrix} e_1 \\ \vdots \\ e_1 \end{pmatrix} \in \mathbb{R}^{3n\times d}, \qquad A_2^{(0)} := \frac{C}{2} \cdot \begin{pmatrix} \varepsilon_1 & -2\varepsilon_1 & \varepsilon_1 & \cdots & \varepsilon_n & -2\varepsilon_n & \varepsilon_n \end{pmatrix} \in \mathbb{R}^{1\times 3n},$$
$$A_2 := \begin{pmatrix} A_2^{(0)} \\ -A_2^{(0)} \end{pmatrix} \in \mathbb{R}^{2\times 3n},$$

as well as

$$b_1 := \frac{C}{2} \cdot \left(-y_1 + M^{-1} \middle| -y_1 \middle| -y_1 - M^{-1} \middle| \cdots \middle| -y_n + M^{-1} \middle| -y_n \middle| -y_n - M^{-1} \right)^T \in \mathbb{R}^{3n}.$$

Finally, set $E := (C \mid -C) \in \mathbb{R}^{1\times 2}$ and

$$A := C \cdot \begin{pmatrix} 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \\ \hline 0 & 1 \\ \vdots & \vdots \\ 0 & 1 \end{pmatrix} \begin{matrix} \Big\} n \\ \\ \Big\} n \end{matrix} \in \mathbb{R}^{2n \times 2}, \qquad B := C \cdot \left(\begin{array}{c|c} \overbrace{1 \cdots 1}^{n} & \overbrace{0 \cdots 0}^{n} \\ 0 \cdots 0 & 1 \cdots 1 \end{array}\right) \in \mathbb{R}^{2 \times 2n},$$

$$D := C \cdot \left(1 \ldots 1 \middle| -1 \ldots -1\right) \in \mathbb{R}^{1 \times 2n}.$$

Note that $\|A\|_\infty, \|B\|_\infty, \|D\|_\infty, \|E\|_\infty, \|A_1\|_\infty, \|A_2\|_\infty, \|A_2^{(0)}\|_\infty \leq C$. Furthermore, since $y_j \in [0,1]$ and $M \geq 1$, we also see $\|b_1\|_\infty \leq C$. Next, note that $\|A_1\|_{\ell^0}, \|A_2^{(0)}\|_{\ell^0}, \|b_1\|_{\ell^0} \leq 3n$, $\|A_2\|_{\ell^0} \leq 6n$, $\|A\|_{\ell^0}, \|B\|_{\ell^0}, \|D\|_{\ell^0} \leq 2n$, and $\|E\|_{\ell^0} \leq 2 \leq 2n$.

For brevity, set $\gamma := \frac{C^L n^{\lfloor L/2 \rfloor}}{4nM}$ and $\Xi := \sum_{i=1}^n \varepsilon_i \Lambda^*_{M,y_i}$, so that $\Xi : \mathbb{R}^d \to \mathbb{R}$. Before we describe how to construct a network Φ implementing $\gamma \cdot \Xi$, we collect a few auxiliary observations. First, a direct computation shows that

$$\tfrac{C}{2M} \Lambda_{M,y}(x) = \varrho\big(\tfrac{C}{2}(x - y + \tfrac{1}{M})\big) - 2\varrho\big(\tfrac{C}{2}(x-y)\big) + \varrho\big(\tfrac{C}{2}(x - y - \tfrac{1}{M})\big).$$

Based on this, it is easy to see

$$\begin{aligned} A_2^{(0)}\big[\varrho(A_1 x + b_1)\big] &= \frac{C}{2} \sum_{j=1}^n \Big[\varepsilon_j \cdot \Big(\varrho\big(\tfrac{C}{2}(x_1 - y_j + \tfrac{1}{M})\big) \\ &\qquad - 2\varrho\big(\tfrac{C}{2}(x_1 - y_j)\big) + \varrho\big(\tfrac{C}{2}(x_1 - y_j - \tfrac{1}{M})\big)\Big)\Big] \\ &= \frac{C}{2}\frac{C}{2M} \sum_{j=1}^n \varepsilon_j \, \Lambda^*_{M,y_j}(x) = \frac{C^2}{4M} \Xi(x). \end{aligned} \tag{3.2}$$

By definition of A_2, this shows $F(x) = \frac{C^2}{4M}\big(\varrho(\Xi(x)), \varrho(-\Xi(x))\big)^T$ for all $x \in \mathbb{R}^d$, for the function $F := \varrho \circ A_2 \circ \varrho \circ (A_1 \bullet + b_1) : \mathbb{R}^d \to \mathbb{R}^2$.

A further direct computation shows for $x, y \in \mathbb{R}$ that

$$\begin{aligned} &\Big[B\varrho\big(A\big(\tfrac{x}{y}\big)\big)\Big]_1 = C \sum_{j=1}^n \varrho\Big(\big(A\big(\tfrac{x}{y}\big)\big)_j\Big) = C \sum_{j=1}^n \varrho(Cx) = C^2 n \, \varrho(x) \\ &\text{and similarly} \quad \Big[B\varrho\big(A\big(\tfrac{x}{y}\big)\big)\Big]_2 = C^2 n \, \varrho(y). \end{aligned} \tag{3.3}$$

Thus, setting $G := B \circ \varrho \circ A : \mathbb{R}^2 \to \mathbb{R}^2$, we see $G(x,y) = C^2 n\big(\varrho(x), \varrho(y)\big)^T$. Therefore, denoting by $G^j := G \circ \cdots \circ G$ the j-fold composition of G with itself, we see $G^j(x,y) = (C^2 n)^j \cdot \big(\varrho(x), \varrho(y)\big)^T$ for $j \in \mathbb{N}$, and hence

$$G^j\big(F(x)\big) = \frac{C^{2j+2} n^j}{4M} \cdot \big(\varrho(\Xi(x)), \quad \varrho(-\Xi(x))\big)^T \quad \forall\, j \in \mathbb{N}_0 \text{ and } x \in \mathbb{R}^d, \tag{3.4}$$

where the case $j = 0$ (in which it is understood that $G^j = \mathrm{id}_{\mathbb{R}^2}$) is easy to verify separately.

In a similar way, we see for $H := D \circ \varrho \circ A : \mathbb{R}^2 \to \mathbb{R}$ that

$$\begin{aligned} H(x,y) &= D\Big[\varrho\Big(A\binom{x}{y}\Big)\Big] = C \cdot \Big(\sum_{j=1}^{n} \varrho(Cx) - \sum_{j=1}^{n} \varrho(Cy)\Big) \\ &= C^2 n\big(\varrho(x) - \varrho(y)\big) \qquad \forall\, x, y \in \mathbb{R}. \end{aligned} \tag{3.5}$$

Now, we prove the claim of the lemma, distinguishing three cases regarding $L \in \mathbb{N}_{\geq 2}$.

Case 1 ($L = 2$): Define $\Phi := \big((A_1, b_1), (A_2^{(0)}, 0)\big)$. Then Eq. (3.2) shows $R_\varrho \Phi = \frac{C^2}{4M}\,\Xi$. Because of $\frac{C^L\, n^{\lfloor L/2 \rfloor}}{4nM} = \frac{C^2}{4M}$ for $L = 2$, this implies the claim, once we note that

$$L(\Phi) = L \leq \boldsymbol{\ell}(n) \leq \boldsymbol{\ell}((2L+8)n) \quad \text{and} \quad \|\Phi\|_{\mathcal{NN}} \leq C \leq \boldsymbol{c}(n) \leq \boldsymbol{c}((2L+8)n),$$

as well as $W(\Phi) \leq 9n \leq (2L+8)n$, since $L = 2$.

Case 2 ($L \geq 4$ is even): In this case, define

$$\Phi := \Big((A_1, b_1), (A_2, 0), \underbrace{(A,0), (B,0), \ldots, (A,0), (B,0)}_{\frac{L-4}{2} \text{ copies of } (A,0),(B,0)}, (A,0), (D,0)\Big)$$

and note for $j := \frac{L-4}{2}$ that $j + 1 = \frac{L-2}{2} = \lfloor L/2 \rfloor - 1$, so that a combination of Eqs. (3.5) and (3.4) shows

$$\begin{aligned} R_\varrho \Phi(x) &= (H \circ G^j \circ F)(x) = C^2 n \cdot \frac{C^{2j+2}\, n^j}{4M} \cdot \big(\varrho(\Xi(x)) - \varrho(-\Xi(x))\big) \\ &= \frac{C^L\, n^{\lfloor L/2 \rfloor}}{4Mn} \cdot \Xi(x), \end{aligned}$$

since $\varrho(\varrho(z)) = \varrho(z)$ and $\varrho(z) - \varrho(-z) = z$ for all $z \in \mathbb{R}$. Finally, we note as in the previous case that $L(\Phi) = L \leq \boldsymbol{\ell}((2L+8)n)$ and $\|\Phi\|_{\mathcal{NN}} \leq C \leq \boldsymbol{c}((2L+8)n)$, and furthermore that

$$W(\Phi) \leq 3n + 3n + 6n + \frac{L-4}{2}\big(2n + 2n\big) + 4n = 16n + (2L-8)n = (8+2L)n.$$

Overall, we see also in this case that $\gamma \cdot \Xi \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{(2L+8)n}$, as claimed.

Case 3 ($L \geq 3$ is odd): In this case, define

$$\Phi := \Big((A_1, b_1), (A_2, 0), \underbrace{(A,0), (B,0), \ldots, (A,0), (B,0)}_{\frac{L-3}{2} \text{ copies of } (A,0),(B,0)}, (E,0)\Big).$$

Then, setting $j := \frac{L-3}{2}$ and noting $j = \lfloor L/2 \rfloor - 1$, we see thanks to Eq. (3.4) and because of $E = (C \mid -C)$ that

$$R_\varrho \Phi(x) = E\Big(G^j\big(F(x)\big)\Big) = C \cdot \frac{C^{2j+2}\, n^j}{4M} \cdot \big(\varrho(\Xi(x)) - \varrho(-\Xi(x))\big)$$
$$= \frac{C^L\, n^{\lfloor L/2 \rfloor}}{4Mn} \cdot \Xi(x).$$

It remains to note as before that $L(\Phi) = L \leq \boldsymbol{\ell}((2L+8)n)$ and $\|\Phi\|_{\mathcal{NN}} \leq C \leq \boldsymbol{c}((2L+8)n)$, and finally that $W(\Phi) \leq 3n + 3n + 6n + \frac{L-3}{2}(2n+2n) + 2 = 2 + 6n + 2Ln \leq (8+2L)n$, so that indeed $\gamma \cdot \Xi \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{(8+2L)n}$ also in this case. □

As an application of Lemma 3.1, we now describe a large class of functions contained in the unit ball of the approximation space $A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$.

Lemma 3.2 *Let $\alpha > 0$ and let $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ be non-decreasing. Let $\sigma \geq 2$, $0 < \gamma < \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$, $\theta \in (0,\infty)$ and $\lambda \in [0,1]$ with $\theta\lambda \leq 1$ be arbitrary and define*

$$\omega := \min\big\{-\theta\alpha,\quad \theta \cdot (\gamma - \lambda) - 1\big\} \in (-\infty, 0).$$

Then there exists a constant $\kappa = \kappa(\alpha, \theta, \lambda, \gamma, \sigma, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ such that for every $m \in \mathbb{N}$, the following holds:

Setting $M := 4m$ and $z_j := \frac{1}{4m} + \frac{j-1}{2m}$ for $j \in \underline{2m}$, the functions $\big(\Lambda^*_{M,z_j}\big)_{j\in\underline{2m}}$ are supported in $[0,1]^d$ and have disjoint supports, up to a null-set. Furthermore, for any $\boldsymbol{\nu} = (\nu_j)_{j\in\underline{2m}} \in [-1,1]^{2m}$ and $J \subset \underline{2m}$ satisfying $|J| \leq \sigma \cdot m^{\theta\lambda}$, we have

$$f_{\boldsymbol{\nu},J} := \kappa \cdot m^\omega \cdot \sum_{j\in J} \nu_j\, \Lambda^*_{M,z_j} \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \qquad \text{and} \qquad \big\|f_{\boldsymbol{\nu},J}\big\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)} \leq 1.$$

Proof Since $\gamma < \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$, we see by definition of $\gamma^\flat$ that there exist $L = L(\gamma, \boldsymbol{\ell}, \boldsymbol{c}) \in \mathbb{N}_{\leq \boldsymbol{\ell}^*}$ and $C_1 = C_1(\gamma, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ such that $n^\gamma \leq C_1 \cdot (\boldsymbol{c}(n))^L \cdot n^{\lfloor L/2 \rfloor}$ for all $n \in \mathbb{N}$. Because of $\boldsymbol{\ell} \geq 2$, we can assume without loss of generality that $L \geq 2$. Furthermore, since $L \leq \boldsymbol{\ell}^*$, we can choose $n_0 = n_0(\gamma, \boldsymbol{\ell}, \boldsymbol{c}) \in \mathbb{N}$ satisfying $L \leq \boldsymbol{\ell}(n_0)$.

Let $m \in \mathbb{N}$ and let $\boldsymbol{\nu}$ and J be as in the statement of the lemma. For brevity, define $f^{(0)}_{\boldsymbol{\nu},J} := \sum_{j\in J} \nu_j \Lambda^*_{M,z_j}$. We note that Λ^*_{M,z_j} is continuous with $0 \leq \Lambda^*_{M,z_j} \leq 1$ and

$$\operatorname{supp} \Lambda^*_{M,z_j} \subset \big\{x \in \mathbb{R}^d : x_1 \in z_j + [-\tfrac{1}{M}, \tfrac{1}{M}]\big\} \subset \big\{x \in \mathbb{R}^d : x_1 \in \tfrac{j-1}{2m} + [0, \tfrac{1}{2m}]\big\}.$$

This shows that the supports of the functions Λ^*_{M,z_j} are contained in $[0,1]^d$ and are pairwise disjoint (up to null-sets), which then implies $\big\|f^{(0)}_{\boldsymbol{\nu},J}\big\|_{L^\infty} \leq 1$.

Next, since $\theta\lambda \leq 1$, we have $\lceil m^{\theta\lambda} \rceil \leq \lceil m \rceil = m \leq 2m$. Thus, by possibly enlarging the set $J \subset \underline{2m}$ and setting $\nu_j := 0$ for the added elements, we can without loss of generality assume that $|J| \geq \lceil m^{\theta\lambda} \rceil \geq 1$. Note that the extended set still satisfies $|J| \leq \sigma \cdot m^{\theta\lambda}$ since $\lceil m^{\theta\lambda} \rceil \leq 2m^{\theta\lambda}$ and $\sigma \geq 2$.

Now, define $N := n_0 \cdot \lceil m^{(1-\lambda)\theta} \rceil$ and $n := N \cdot |J|$, noting that $n \geq n_0$. Furthermore, writing $J = \{i_1, \dots, i_{|J|}\}$, define

$$(\varepsilon_1, \dots, \varepsilon_n) := \Big(\underbrace{\nu_{i_1}, \dots, \nu_{i_1}}_{N \text{ times}}, \dots, \underbrace{\nu_{i_{|J|}}, \dots, \nu_{i_{|J|}}}_{N \text{ times}} \Big) \quad \text{and}$$
$$(y_1, \dots, y_n) := \Big(\underbrace{z_{i_1}, \dots, z_{i_1}}_{N \text{ times}}, \dots, \underbrace{z_{i_{|J|}}, \dots, z_{i_{|J|}}}_{N \text{ times}} \Big).$$

By choice of C_1, we have $n^\gamma \leq C_1 \cdot (\boldsymbol{c}(n))^L \cdot n^{\lfloor L/2 \rfloor}$, so that we can choose $0 < C \leq \boldsymbol{c}(n)$ satisfying $n^\gamma \leq C_1 \cdot C^L \cdot n^{\lfloor L/2 \rfloor}$. Since we also have $L \geq 2$ and $L \leq \boldsymbol{\ell}(n_0) \leq \boldsymbol{\ell}(n)$, Lemma 3.1 shows that

$$\Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{(2L+8)n} \ni \frac{C^L\, n^{\lfloor L/2 \rfloor}}{4Mn} \sum_{i=1}^{n} \varepsilon_i \Lambda^*_{M,y_i} = \frac{C^L\, n^{\lfloor L/2 \rfloor} N}{4Mn} \cdot f^{(0)}_{\boldsymbol{\nu},J};$$

here the final equality comes from our choice of $\varepsilon_1, \dots, \varepsilon_n$ and $y_1, \dots, y_n$.

To complete the proof, we first collect a few auxiliary estimates. First, we see because of $|J| \geq m^{\theta\lambda}$ that $n \geq n_0\, m^{(1-\lambda)\theta}\, m^{\theta\lambda} \geq m^\theta$.

Thus, setting $C_2 := 16\sigma C_1$ and recalling that $\omega \leq \theta \cdot (\gamma - \lambda) - 1$ by choice of ω, we see for any $0 < \kappa \leq C_2^{-1}$ that

$$\kappa \cdot m^\omega \leq \frac{m^{\theta\gamma - \theta\lambda - 1}}{16\sigma C_1} \leq \frac{C_1^{-1} n^\gamma \cdot \sigma^{-1} m^{-\theta\lambda}}{4 \cdot 4m} \leq \frac{C^L n^{\lfloor L/2 \rfloor} \cdot \sigma^{-1} m^{-\theta\lambda}}{4M} \leq \frac{C^L n^{\lfloor L/2 \rfloor} N}{4Mn}.$$

Here, we used in the last step that $|J| \leq \sigma\, m^{\theta\lambda}$, which implies $\frac{N}{n} = |J|^{-1} \geq \sigma^{-1} m^{-\theta\lambda}$. Thus, noting that $c\Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_t \subset \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_t$ for $c \in [-1, 1]$, we see $\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{(2L+8)n}$ as long as $0 < \kappa \leq C_2^{-1}$.

Finally, set $C_3 := \max\{1,\ C_2,\ (2L+8)^\alpha\, (2n_0\sigma)^\alpha\}$. We claim that $\Gamma_{\alpha,\infty}\big(\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J}\big) \leq 1$ for $\kappa := C_3^{-1}$. Once this is shown, Lemma 2.1 will show that $\big\|\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J}\big\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}} \leq 1$ as well. To see $\Gamma_{\alpha,\infty}\big(\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J}\big) \leq 1$, first note that $\big\|\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J}\big\|_{L^\infty} \leq \|f^{(0)}_{\boldsymbol{\nu},J}\|_{L^\infty} \leq 1$ since $\omega < 0$ and $\kappa = C_3^{-1} \leq 1$. Furthermore, for $t \in \mathbb{N}$ there are two cases: For $t \geq (2L+8)n$ we have shown above that $\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{(2L+8)n} \subset \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_t$ and hence $t^\alpha\, d_\infty(\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J}; \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_t) = 0 \leq 1$. On the other hand, if $t \leq (2L+8)n$ then we see because of $\lceil m^{(1-\lambda)\theta} \rceil \leq 1 + m^{(1-\lambda)\theta} \leq 2 \cdot m^{(1-\lambda)\theta}$ and $|J| \leq \sigma\, m^{\theta\lambda}$ that $n \leq 2n_0\sigma\, m^\theta$. Since we also have $\omega \leq -\theta\alpha$, this implies

$$\begin{aligned} t^\alpha\, d_\infty(\kappa\, m^\omega\, f^{(0)}_{\vec{\nu},J}; \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_t) &\leq (2L+8)^\alpha\, n^\alpha\, \kappa\, m^\omega \left\| f^{(0)}_{\boldsymbol{\nu},J} \right\|_{L^\infty} \\ &\leq (2L+8)^\alpha (2n_0\sigma)^\alpha\, \kappa\, m^{\theta\alpha} m^{-\theta\alpha} \leq 1. \end{aligned}$$

All in all, this shows $\Gamma_{\alpha,\infty}\big(\kappa\, m^\omega\, f^{(0)}_{\boldsymbol{\nu},J}\big) \leq 1$. As seen above, this completes the proof. □

For later use, we also collect the following technical result which shows how to select a large number of "hat functions" as in Lemma 3.2 that are annihilated by a given set of sampling points.

Lemma 3.3 *Let $m \in \mathbb{N}$ and let $M = 4m$ and $z_j = \frac{1}{4m} + \frac{j-1}{2m}$ as in Lemma 3.2. Given arbitrary points $\boldsymbol{x} = (x_1, \dots, x_m) \in ([0,1]^d)^m$, define*

$$I_{\boldsymbol{x}} := \big\{ i \in \underline{2m} \ : \ \forall\, n \in \underline{m} : \Lambda^*_{M,z_i}(x_n) = 0 \big\}.$$

Then $|I_{\boldsymbol{x}}| \geq m$.

Proof Let $I^c_{\boldsymbol{x}} := \underline{2m} \setminus I_{\boldsymbol{x}}$. For each $i \in I^c_{\boldsymbol{x}}$, there exists $n_i \in \underline{m}$ satisfying $\Lambda^*_{M,z_i}(x_{n_i}) \neq 0$. The map $I^c_{\boldsymbol{x}} \to \underline{m}, i \mapsto n_i$ is injective, since $\Lambda^*_{M,z_i}\Lambda^*_{M,z_\ell} \equiv 0$ for $i \neq \ell$ (see Lemma 3.2). Therefore, $|I^c_{\boldsymbol{x}}| \leq m$ and hence $|I_{\boldsymbol{x}}| = 2m - |I^c_{\boldsymbol{x}}| \geq m$.

The function $\Lambda^*_{M,y} : \mathbb{R}^d \to \mathbb{R}$ has a controlled support with respect to the first coordinate of x, but unbounded support with respect to the remaining variables. For proving more refined hardness bounds, we shall therefore use the following modified construction of a function of "hat-type" with controlled support. As we will see in Lemma 3.5 below, this function can also be well implemented by ReLU networks, provided one can use networks *with at least two hidden layers*. □

Lemma 3.4 *Given $d \in \mathbb{N}$, $M > 0$ and $y \in \mathbb{R}^d$, define*

$$\begin{aligned} \theta : \ & \mathbb{R} \to [0,1], && x \mapsto \varrho(x) - \varrho(x-1), \\ \Delta_{M,y} : \ & \mathbb{R}^d \to \mathbb{R}, && x \mapsto \left[\sum_{j=1}^{d} \Lambda_{M,y_j}(x_j) \right] - (d-1), \\ \text{and} \quad \vartheta_{M,y} : \ & \mathbb{R}^d \to [0,1], && x \mapsto \theta\big(\Delta_{M,y}(x)\big). \end{aligned}$$

Then the function $\vartheta_{M,y}$ has the following properties:

a) $\vartheta_{M,y}(x) = 0$ *for all* $x \in \mathbb{R}^d \setminus \big(y + M^{-1}(-1,1)^d\big)$;
b) $\|\vartheta_{M,y}\|_{L^p(\mathbb{R}^d)} \leq (2/M)^{d/p}$ *for arbitrary* $p \in (0,\infty]$;
c) *For any* $p \in (0,\infty]$ *there is a constant* $C = C(d,p) > 0$ *satisfying*

$$\|\vartheta_{M,y}\|_{L^p([0,1]^d)} \geq C \cdot M^{-d/p}, \qquad \forall\, y \in [0,1]^d \text{ and } M \geq \tfrac{1}{2d}.$$

Proof of Lemma 3.4 **Ad a)** For $x \in \mathbb{R}^d \setminus \big(y + M^{-1}(-1,1)^d\big)$, there exists $\ell \in \underline{d}$ with $|x_\ell - y_\ell| \geq M^{-1}$ and hence $\Lambda_{M,y_\ell}(x_\ell) = 0$; see Fig. 1. Because of $0 \leq \Lambda_{M,y_j} \leq 1$, this implies

$$\Delta_{M,y}(x) = \sum_{j \in \underline{d} \setminus \{\ell\}} \Lambda_{M,y_j}(x_j) - (d-1) \leq d - 1 - (d-1) = 0.$$

By elementary properties of the function θ (see Fig. 2), this shows $\vartheta_{M,y}(x) = \theta(\Delta_{M,y}(x)) = 0$.

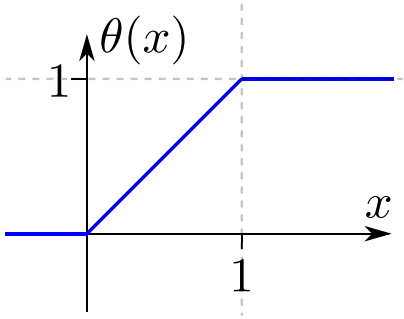

Fig. 2 A plot of the function θ appearing in Lemma 3.4. Note that θ is non-decreasing and satisfies $\theta(x) = 0$ for $x \leq 0$ as well as $\theta(x) = 1$ for $x \geq 1$

Ad b) Since $0 \leq \theta \leq 1$, we also have $0 \leq \vartheta_{M,y} \leq 1$. Combined with Part a), this implies $\|\vartheta_{M,y}\|_{L^p} \leq \left[\lambda(y + M^{-1}(-1,1)^d)\right]^{1/p} = (2/M)^{d/p}$, as claimed.

Ad c) Set $T := \frac{1}{2dM} \in (0,1]$ and $P := y + [-T,T]^d$. For $x \in P$ and arbitrary $j \in \underline{d}$, we have $|x_j - y_j| \leq \frac{1}{2dM}$. Since Λ_{M,y_j} is Lipschitz with $\mathrm{Lip}(\Lambda_{M,y_j}) \leq M$ (see Fig. 1) and $\Lambda_{M,y_j}(y_j) = 1$, this implies

$$\Lambda_{M,y_j}(x_j) \geq \Lambda_{M,y_j}(y_j) - \left|\Lambda_{M,y_j}(y_j) - \Lambda_{M,y_j}(x_j)\right| \geq 1 - M \cdot \frac{1}{2dM} = 1 - \frac{1}{2d}.$$

Since this holds for all $j \in \underline{d}$, we see $\Delta_{M,y}(x) = \sum_{j=1}^d \Lambda_{M,y_j}(x_j) - (d-1) \geq d \cdot (1 - \frac{1}{2d}) - (d-1) = \frac{1}{2}$, and hence $\vartheta_{M,y}(x) = \theta(\Delta_{M,y}(x)) \geq \theta(\frac{1}{2}) = \frac{1}{2}$, since θ is non-decreasing.

Finally, Lemma A.2 shows for $Q = [0,1]^d$ that $\lambda(Q \cap P) \geq 2^{-d} T^d \geq C_1 \cdot M^{-d}$ with $C_1 = C_1(d) > 0$. Hence, $\|\vartheta_{M,y}\|_{L^p([0,1]^d)} \geq \frac{1}{2}[\lambda(Q \cap P)]^{1/p} \geq C_1^{1/p} M^{-d/p}$, which easily yields the claim.

The next lemma shows how well the function $\vartheta_{M,y}$ can be implemented by ReLU networks. We emphasize that the lemma requires using networks with $L \geq 3$, i.e., with at least two hidden layers. □

Lemma 3.5 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ be non-decreasing. Let $M \geq 1$, $n \in \mathbb{N}$ and $0 < C \leq \boldsymbol{c}(n)$, as well as $L \in \mathbb{N}_{\geq 3}$ with $L \leq \boldsymbol{\ell}(n)$. Then*

$$\frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4M} \cdot \vartheta_{M,y} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{15(d+L)n} \qquad \forall\, y \in [0,1]^d.$$

Proof Let $y \in [0,1]^d$ be fixed. For $j \in \underline{d}$, denote by $e_j \in \mathbb{R}^{d \times 1}$ the j-th standard basis vector. Define $A_1 \in \mathbb{R}^{4nd \times d}$ and $b_1 \in \mathbb{R}^{4nd}$ by

$$A_1^T := \frac{C}{2} \cdot \Big(\underbrace{e_1|\ldots|e_1}_{3n \text{ times}}, \underbrace{0|\ldots|0}_{n \text{ times}}, \underbrace{e_2|\ldots|e_2}_{3n \text{ times}}, \underbrace{0|\ldots|0}_{n \text{ times}}, \quad \ldots, \quad \underbrace{e_d|\ldots|e_d}_{3n \text{ times}}, \underbrace{0|\ldots|0}_{n \text{ times}} \Big),$$

$$b_1 := -\frac{C}{2} \cdot \Big(\underbrace{y_1 - \tfrac{1}{M}, \ldots, y_1 - \tfrac{1}{M}}_{n \text{ times}}, \underbrace{y_1, \ldots, y_1}_{n \text{ times}}, \underbrace{y_1 + \tfrac{1}{M}, \ldots, y_1 + \tfrac{1}{M}}_{n \text{ times}}, \underbrace{-1, \ldots, -1}_{n \text{ times}},$$

$$\underbrace{y_2 - \tfrac{1}{M}, \dots, y_2 - \tfrac{1}{M}}_{n \text{ times}}, \underbrace{y_2, \dots, y_2}_{n \text{ times}}, \underbrace{y_2 + \tfrac{1}{M}, \dots, y_2 + \tfrac{1}{M}}_{n \text{ times}}, \underbrace{-1, \dots, -1}_{n \text{ times}},$$

$$\dots,$$

$$\underbrace{y_d - \tfrac{1}{M}, \dots, y_d - \tfrac{1}{M}}_{n \text{ times}}, \underbrace{y_d, \dots, y_d}_{n \text{ times}}, \underbrace{y_d + \tfrac{1}{M}, \dots, y_d + \tfrac{1}{M}}_{n \text{ times}} \underbrace{-1, \dots, -1}_{n \text{ times}} \Big)^T$$

Furthermore, set $b_2 := 0 \in \mathbb{R}^2$ and $b_3 := 0 \in \mathbb{R}^n$, let $\zeta := -\frac{1}{M}\frac{d-1}{d}$ and $\xi := -\frac{1}{M}$, and define $A_2 \in \mathbb{R}^{2\times 4nd}$ and $A_3 \in \mathbb{R}^{n\times 2}$ by

$$A_2 := \frac{C}{2}\left(\begin{matrix} \overbrace{1,\dots,1}^{n \text{ times}}, \overbrace{-2,\dots,-2}^{n \text{ times}}, \overbrace{1,\dots,1}^{n \text{ times}}, \overbrace{\zeta,\dots,\zeta}^{n \text{ times}}, \dots, \overbrace{1,\dots,1}^{n \text{ times}}, \overbrace{-2,\dots,-2}^{n \text{ times}}, \overbrace{1,\dots,1}^{n \text{ times}}, \overbrace{\zeta,\dots,\zeta}^{n \text{ times}} \\ \underbrace{1,\dots,1}_{n \text{ times}}, \underbrace{-2,\dots,-2}_{n \text{ times}}, \underbrace{1,\dots,1}_{n \text{ times}}, \underbrace{\xi,\dots,\xi}_{n \text{ times}}, \dots, \underbrace{1,\dots,1}_{n \text{ times}}, \underbrace{-2,\dots,-2}_{n \text{ times}}, \underbrace{1,\dots,1}_{n \text{ times}}, \underbrace{\xi,\dots,\xi}_{n \text{ times}} \end{matrix}\right),$$

$$A_3^T := C\begin{pmatrix} 1,\dots,1 \\ -1,\dots,-1 \end{pmatrix} \in \mathbb{R}^{2\times n}.$$

Finally, set $A := C \cdot (1,\dots,1) \in \mathbb{R}^{1\times n}$, $B := C \cdot (1,\dots,1)^T \in \mathbb{R}^{n\times 1}$, and $D := C \cdot (1,-1) \in \mathbb{R}^{1\times 2}$, as well as $E := (C) \in \mathbb{R}^{1\times 1}$. Note that $\|A_1\|_\infty, \|A_2\|_\infty, \|A_3\|_\infty, \|A\|_\infty, \|B\|_\infty, \|D\|_\infty, \|E\|_\infty \le C$ and $\|b_1\|_\infty, \|b_2\|_\infty \le C$, since $M \ge 1$ and $y \in [0,1]^d$. Furthermore, note $\|A_1\|_{\ell^0} \le 3dn$, $\|A_2\|_{\ell^0} \le 8dn$, $\|A_3\|_{\ell^0} \le 2n$, $\|A\|_{\ell^0}, \|B\|_{\ell^0} \le n$, $\|D\|_{\ell^0} \le 2$, and finally $\|b_1\|_{\ell^0} \le 4dn$ and $\|b_2\|_{\ell^0} = 0$. Furthermore, note $C \le \boldsymbol{c}(n) \le \boldsymbol{c}(15(d+L)n)$ and likewise $L \le \boldsymbol{\ell}(n) \le \boldsymbol{\ell}(15(d+L)n)$ thanks to the monotonicity of $\boldsymbol{c}, \boldsymbol{\ell}$.

A direct computation shows that

$$\tfrac{C/2}{M}\Lambda_{M,y}(x) = \varrho\big(\tfrac{C}{2}(x - y + \tfrac{1}{M})\big) - 2\varrho\big(\tfrac{C}{2}(x-y)\big) + \varrho\big(\tfrac{C}{2}(x - y - \tfrac{1}{M})\big).$$

Combined with the positive homogeneity of the ReLU (i.e., $\varrho(tx) = t\varrho(x)$ for $t \ge 0$), this shows

$$\begin{aligned}
&\big(A_2\,\varrho(A_1 x + b_1) + b_2\big)_1 \\
&= \frac{C}{2}\sum_{j=1}^{d}\sum_{\ell=1}^{n}\Big[\varrho\big(\tfrac{C}{2}(\langle x, e_j\rangle - (y_j - \tfrac{1}{M}))\big) - 2\varrho\big(\tfrac{C}{2}(\langle x, e_j\rangle - y_j)\big) \\
&\quad + \varrho\big(\tfrac{C}{2}(\langle x, e_j\rangle - (y_j + \tfrac{1}{M}))\big) + \zeta\,\varrho(\tfrac{C}{2})\Big] \\
&= \frac{C^2 n}{4M}\sum_{j=1}^{d}\Big[\Lambda_{M,y_j}(x_j) - \frac{d-1}{d}\Big] = \frac{C^2 n}{4M}\Delta_{M,y}(x).
\end{aligned}$$

In the same way, it follows that $\big(A_2\,\varrho(A_1 x + b_1) + b_2\big)_2 = \frac{C^2 n}{4M} \cdot (\Delta_{M,y}(x) - 1)$. We now distinguish three cases:

Case 1: $L = 3$. In this case, set $\Phi := \big((A_1, b_1), (A_2, b_2), (D, 0)\big)$. Then the calculation from above, combined with the positive homogeneity of the ReLU shows

$$\begin{aligned} R_\varrho \Phi(x) &= C \cdot \Big(\varrho\big(\tfrac{C^2 n}{4M} \Delta_{M,y}(x)\big) - \varrho\big(\tfrac{C^2 n}{4M}(\Delta_{M,y}(x) - 1)\big)\Big) \\ &= \tfrac{C^3 n}{4M} \theta(\Delta_{M,y}(x)) = \tfrac{C^3 n}{4M} \vartheta_{M,y}(x). \end{aligned}$$

Furthermore, it is straightforward to see $W(\Phi) \leq 3dn + 4dn + 8dn + 2 \leq 2 + 15dn \leq 15(L+d)n$. Combined with our observations from above, and noting $\lfloor \frac{L}{2} \rfloor = 1$, we thus see as claimed that $\frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4M} \vartheta_{M,y} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{15(L+d)n}$.

Case 2: $L \geq 4$ is even. In this case, define

$$\Phi = \Big((A_1, b_1), (A_2, b_2), (A_3, b_3), (A, 0), \underbrace{(B, 0), (A, 0), \ldots, (B, 0), (A, 0)}_{(L-4)/2 \text{ copies of } \text{“}(B,0),(A,0)\text{”}} \Big)$$

Similar arguments as in Case 1 show that $\big(A_3\, \varrho\big(A_2\, \varrho(A_1 x + b_1) + b_2\big) + b_3\big)_j = \frac{C^3 n}{4M} \vartheta_{M,y}(x)$ for all $j \in \underline{n}$, and hence $A \circ \varrho \circ A_3 \circ \varrho \circ A_2 \circ \varrho \circ (A_1 \bullet + b_1)$ Furthermore, using similar arguments as in Eq. (3.3), we see for $z \in [0, \infty)$ that $A(\varrho(Bz)) = C^2 n z$. Combining all these observations, we see

$$R_\varrho \Phi(x) = (C^2 n)^{(L-4)/2} \cdot \frac{C^4 n^2}{4M} \vartheta_{M,y}(x) = \frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4M} \cdot \vartheta_{M,y}(x).$$

Since also $W(\Phi) \leq 3dn + 4dn + 8dn + 2n + n + \frac{L-4}{2} \cdot 2n \leq 15(d+L)n$, we see overall as claimed that $\frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4M} \vartheta_{M,y} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{15(d+L)n}$.

Case 3: $L \geq 5$ is odd. In this case, define

$$\Phi := \Big((A_1, b_1), (A_2, b_2), (A_3, b_3), (A, 0), \underbrace{(B, 0), (A, 0), \ldots, (B, 0), (A, 0)}_{(L-5)/2 \text{ copies of } \text{“}(B,0),(A,0)\text{”}}, (E, 0) \Big).$$

A variant of the arguments in Case 2 shows that $R_\varrho \Phi = C \cdot (C^2 n)^{(L-5)/2} \cdot \frac{C^4 n^2}{4M} \vartheta_{M,y} = \frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4M} \vartheta_{M,y}$ and $W(\Phi) \leq 15dn + 2n + \frac{L-5}{2} \cdot 2n + 1 \leq 15(d+L)n$, and hence $\frac{C^L \cdot n^{\lfloor L/2 \rfloor}}{4M} \vartheta_{M,y} \in \Sigma^{\boldsymbol{\ell},\boldsymbol{c}}_{15(d+L)n}$ also in this last case. □

Lemma 3.6 *Let $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ be non-decreasing with $\boldsymbol{\ell}^* \geq 3$. Let $d \in \mathbb{N}$, $\alpha \in (0, \infty)$, and $0 < \gamma < \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$. Then there exists a constant $\kappa = \kappa(\gamma, \alpha, d, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ such that for any $M \in [1, \infty)$ and $y \in [0,1]^d$, we have*

$$g_{M,y} := \kappa \cdot M^{-\alpha/(\alpha+\gamma)}\, \vartheta_{M,y} \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \qquad \text{with} \qquad \big\| g_{M,y} \big\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}} \leq 1.$$

Proof Since $\gamma < \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$, there exist $L = L(\gamma, \boldsymbol{\ell}, \boldsymbol{c}) \in \mathbb{N}_{\geq \ell^*}$ and $C_1 = C_1(\gamma, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ satisfying $n^\gamma \leq C_1 \cdot (\boldsymbol{c}(n))^L \cdot n^{\lfloor L/2 \rfloor}$ for all $n \in \mathbb{N}$. Since $\boldsymbol{\ell}^* \geq 3$, we can assume

without loss of generality that $L \geq 3$. Furthermore, since $L \leq \ell^*$, there exists $n_0 = n_0(\gamma, \boldsymbol{\ell}, \boldsymbol{c}) \in \mathbb{N}$ satisfying $L \leq \boldsymbol{\ell}(n_0)$.

Given $M \in [1, \infty)$ and $y \in [0,1]^d$, set $n := n_0 \cdot \lceil M^{1/(\alpha+\gamma)} \rceil$, noting that $n \geq n_0$. Since $n^\gamma \leq C_1 \cdot (\boldsymbol{c}(n))^L \cdot n^{\lfloor L/2 \rfloor}$, there exists $0 < C \leq \boldsymbol{c}(n)$ satisfying $n^\gamma \leq C_1 \cdot C^L n^{\lfloor L/2 \rfloor}$.

Set $\kappa := \min\{(15(d+L))^{-\alpha}(2n_0)^{-\alpha}, (4\,C_1)^{-1}\} > 0$ and note $\kappa = \kappa(d, \alpha, \gamma, \boldsymbol{\ell}, \boldsymbol{c})$. Furthermore, note that $n \geq M^{1/(\alpha+\gamma)}$ and hence $\kappa\, M^{-\frac{\alpha}{\alpha+\gamma}} = \frac{\kappa}{M} M^{\frac{\gamma}{\alpha+\gamma}} \leq \kappa \frac{n^\gamma}{M} \leq 4C_1\,\kappa \frac{C^L n^{\lfloor L/2 \rfloor}}{4M} \leq \frac{C^L n^{\lfloor L/2 \rfloor}}{4M}$. Combining this with the inclusion $c\Sigma_t^{\boldsymbol{\ell},\boldsymbol{c}} \subset \Sigma_t^{\boldsymbol{\ell},\boldsymbol{c}}$ for $c \in [-1,1]$, we see from Lemma 3.5 and because of $3 \leq L \leq \boldsymbol{\ell}(n_0) \leq \boldsymbol{\ell}(n)$ that $g_{M,y} = \kappa\, M^{-\alpha/(\alpha+\gamma)}\, \vartheta_{M,y} \in \Sigma_{15(d+L)n}^{\boldsymbol{\ell},\boldsymbol{c}}$.

We claim that $\Gamma_{\alpha,\infty}(g_{M,y}) \leq 1$. To see this, first note $\|g_{M,y}\|_{L^\infty} \leq \|\vartheta_{M,y}\|_{L^\infty} \leq 1$. Furthermore, for $t \in \mathbb{N}$, there are two cases: For $t \geq 15(d+L)n$, we have $g_{M,y} \in \Sigma_t^{\boldsymbol{\ell},\boldsymbol{c}}$, and hence $t^\alpha\, d_\infty(g_{M,y}, \Sigma_t^{\boldsymbol{\ell},\boldsymbol{c}}) = 0 \leq 1$. On the other hand, if $t \leq 15(d+L)n$, then we see because of $n \leq 1n_0 + n_0\, M^{1/(\alpha+\gamma)} \leq 2n_0\, M^{1/(\alpha+\gamma)}$ that

$$\begin{aligned} t^\alpha\, d_\infty(g_{M,y}, \Sigma_t^{\boldsymbol{\ell},\boldsymbol{c}}) &\leq \big(15(d+L)\big)^\alpha\, n^\alpha\, \|g_{M,y}\|_{L^\infty} \leq \big(15(d+L)\big)^\alpha\, \kappa\, n^\alpha\, M^{-\alpha/(\alpha+\gamma)} \\ &\leq \big(15(d+L)\big)^\alpha\, (2n_0)^\alpha\, \kappa\, M^{\alpha/(\alpha+\gamma)}\, M^{-\alpha/(\alpha+\gamma)} \leq 1. \end{aligned}$$

Overall, this shows $\Gamma_{\alpha,p}(g_{M,y}) \leq 1$, so that Lemma 2.1 shows as claimed that $\|g_{M,y}\|_{A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}} \leq 1$. □

4 Error Bounds for Uniform Approximation

In this section, we derive an upper bound on how many point samples of a function $f \in A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$ are needed in order to uniformly approximate f up to error $\varepsilon \in (0,1)$. The crucial ingredient will be the following estimate of the Lipschitz constant of functions $F \in \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$. The bound in the lemma is one of the reasons for our choice of the quantities $\gamma^\flat$ and $\gamma^\sharp$ introduced in Eq. (2.2).

Lemma 4.1 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to [1, \infty]$ be non-decreasing. Let $n \in \mathbb{N}$ and assume that $L := \boldsymbol{\ell}(n)$ and $C := \boldsymbol{c}(n)$ are finite. Then each $F \in \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ satisfies*

$$\mathrm{Lip}_{(\mathbb{R}^d, \|\cdot\|_{\ell^1}) \to \mathbb{R}}(F) \leq C^L \cdot n^{\lfloor L/2 \rfloor} \quad \text{and} \quad \mathrm{Lip}_{(\mathbb{R}^d, \|\cdot\|_{\ell^\infty}) \to \mathbb{R}}(F) \leq d \cdot C^L \cdot n^{\lfloor L/2 \rfloor}.$$

Proof **Step 1:** For any matrix $A \in \mathbb{R}^{k \times m}$, define $\|A\|_\infty := \max_{i,j} |A_{i,j}|$ and denote by $\|A\|_{\ell^0}$ the number of nonzero entries of A. In this step, we show that

$$\|A\|_{\ell^1 \to \ell^\infty} \leq \|A\|_\infty \quad \text{and} \quad \|A\|_{\ell^\infty \to \ell^1} \leq \|A\|_\infty\, \|A\|_{\ell^0}. \tag{4.1}$$

To prove the first part, note for arbitrary $x \in \mathbb{R}^m$ and any $i \in \underline{k}$ that

$$\big|(Ax)_i\big| \leq \sum_{j=1}^m |A_{i,j}|\, |x_j| \leq \|A\|_\infty \sum_{j=1}^m |x_j| = \|A\|_\infty\, \|x\|_{\ell^1},$$

showing that $\|Ax\|_{\ell^\infty} \leq \|A\|_\infty \|x\|_{\ell^1}$. To prove the second part, note for arbitrary $x \in \mathbb{R}^m$ that

$$\|Ax\|_{\ell^1} = \sum_{i=1}^{k} \big|(Ax)_i\big| \leq \sum_{i,j} |A_{i,j}|\,|x_j| \leq$$
$$\|x\|_{\ell^\infty} \|A\|_\infty \sum_{i,j} \mathbb{1}_{A_{i,j}\neq 0} = \|A\|_\infty \|A\|_{\ell^0} \|x\|_{\ell^\infty}.$$

Step 2 (Completing the proof): Let $F \in \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ be arbitrary, so that $F = R_\varrho \Phi$ for a network $\Phi = \big((A_1, b_1), \dots, (A_{\widetilde{L}}, b_{\widetilde{L}})\big)$ satisfying $\widetilde{L} \leq \boldsymbol{\ell}(n) = L$ and $\|A_j\|_\infty \leq \|\Phi\|_{\mathcal{NN}} \leq \boldsymbol{c}(n) = C$, as well as $\|A_j\|_{\ell^0} \leq W(\Phi) \leq n$ for all $j \in \underline{\widetilde{L}}$.

Set $p_j := 1$ if j is even and $p_j := \infty$ otherwise. Choose N_j such that $A_j \in \mathbb{R}^{N_j \times N_{j-1}}$, and define $T_j\, x := A_j\, x + b_j$. By Step 1, we then see that $T_j : \big(\mathbb{R}^{N_{j-1}}, \|\cdot\|_{\ell^{p_{j-1}}}\big) \to \big(\mathbb{R}^{N_j}, \|\cdot\|_{\ell^{p_j}}\big)$ is Lipschitz with

$$\mathrm{Lip}(T_j) = \|A_j\|_{\ell^{p_{j-1}} \to \ell^{p_j}} \leq \begin{cases} \|A_j\|_\infty \|A_j\|_{\ell^0} \leq Cn, & \text{if } j \text{ is even,} \\ \|A_j\|_\infty \leq C, & \text{if } j \text{ is odd.} \end{cases}$$

Next, a straightforward computation shows that the "vector-valued ReLU" is 1-Lipschitz as a map $\varrho : (\mathbb{R}^k, \|\cdot\|_{\ell^p}) \to (\mathbb{R}^k, \|\cdot\|_{\ell^p})$, for arbitrary $p \in [1, \infty]$ and any $k \in \mathbb{N}$. As a consequence, we see that

$$F = R_\varrho \Phi = T_{\widetilde{L}} \circ (\varrho \circ T_{\widetilde{L}-1}) \circ \cdots \circ (\varrho \circ T_1) :$$
$$(\mathbb{R}^d, \|\cdot\|_{\ell^1}) \to (\mathbb{R}, \|\cdot\|_{\ell^{p_{\widetilde{L}}}}) = (\mathbb{R}, |\cdot|)$$

is Lipschitz continuous as a composition of Lipschitz maps, with overall Lipschitz constant

$$\mathrm{Lip}(R_\varrho \Phi) \leq \prod_{j=1}^{\widetilde{L}} \big(C \cdot n_j\big) = C^{\widetilde{L}} \cdot n^{\lfloor \widetilde{L}/2 \rfloor} \leq C^L \cdot n^{\lfloor L/2 \rfloor}.$$

where we used the notation $n_j := n$ if j is even and $n_j := 1$ otherwise. Furthermore, we used in the last step that $C \geq 1$. The final claim of the lemma follows from the elementary estimate $\|x\|_{\ell^1} \leq d \cdot \|x\|_{\ell^\infty}$ for $x \in \mathbb{R}^d$. □

Based on the preceding lemma, we can now prove an error bound for the computational problem of uniform approximation on the neural network approximation space $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}([0,1]^d)$.

Theorem 4.2 *Let $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ be non-decreasing and suppose that $\gamma^\sharp(\boldsymbol{\ell}, \boldsymbol{c}) < \infty$. Let $d \in \mathbb{N}$ and $\alpha \in (0, \infty)$ be arbitrary, and let $U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}([0,1]^d)$ as in Eq. (2.3). Furthermore, let $\iota_\infty : A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}([0,1]^d) \to C([0,1]^d)$, $f \mapsto f$.*

Then, we have

$$\beta_*^{\det}\big(U_{\ell,c}^{\alpha,\infty}([0,1]^d), \iota_\infty\big) \geq \frac{1}{d} \cdot \frac{\alpha}{\gamma^\sharp(\ell, c) + \alpha}.$$

Remark *a)* The proof shows that choosing the uniform grid $\{0, \frac{1}{N}, \dots, \frac{N-1}{N}\}^d$ as the set of sampling points (with $N \sim m^{1/d}$) yields an essentially optimal sampling scheme.

b) It is well-known (see [25, Proposition 3.3]) that the error of an optimal randomized algorithm is at most two times the error of an optimal deterministic algorithm. Therefore, the theorem also implies that

$$\beta_*^{\mathrm{ran}}\big(U_{\ell,c}^{\alpha,\infty}([0,1]^d), \iota_\infty\big) \geq \frac{1}{d} \cdot \frac{\alpha}{\gamma^\sharp(\ell, c) + \alpha}.$$

Proof Since $\gamma^\sharp(\ell, c) < \infty$, Remark 2.2 shows that $L := \ell^* < \infty$. Let $\gamma > \gamma^\sharp(\ell, c) \geq 1$ be arbitrary. By definition of $\gamma^\sharp(\ell, c)$, it follows that there exists some $\gamma' \in \big(\gamma^\sharp(\ell, c), \gamma\big)$ and a constant $C_0 = C_0(\gamma', \ell, c) = C_0(\gamma, \ell, c) > 0$ satisfying $(c(n))^L \cdot n^{\lfloor L/2 \rfloor} \leq C_0 \cdot n^{\gamma'} \leq C_0 \cdot n^\gamma$ for all $n \in \mathbb{N}$. Let $m \in \mathbb{N}$ be arbitrary and choose

$$N := \big\lfloor m^{1/d} \big\rfloor \geq 1 \qquad \text{and} \qquad n := \big\lceil m^{1/(d\cdot(\gamma+\alpha))} \big\rceil \in \mathbb{N}.$$

Furthermore, let $I := \big\{0, \frac{1}{N}, \dots, \frac{N-1}{N}\big\}^d \subset [0,1]^d$ and set $C := c(n)$ and $\mu := d \cdot C^L \cdot n^{\lfloor L/2 \rfloor}$, noting that $\mu \leq d\, C_0\, n^\gamma =: C_1\, n^\gamma$ and $|I| = N^d \leq m$.

Next, set $B := U := U_{\ell,c}^{\alpha,\infty}([0,1]^d) = \big\{f \in A_{\ell,c}^{\alpha,\infty}([0,1]^d) \colon \|f\|_{A_{\ell,c}^{\alpha,\infty}} \leq 1\big\}$ and define $S := \Omega(B)$ for

$$\Omega : \quad C([0,1]^d) \to \mathbb{R}^I, \quad f \mapsto \big(f(i)\big)_{i\in I}.$$

For each $y = (y_i)_{i\in I} \in S$, choose some $f_y \in B$ satisfying $y = \Omega(f_y)$. Note by Lemma 2.1 that $\Gamma_{\alpha,\infty}(f_y) \leq 1$; by definition of $\Gamma_{\alpha,\infty}$, we can thus choose $F_y \in \Sigma_n^{\ell,c}$ satisfying $\|f_y - F_y\|_{L^\infty} \leq 2 \cdot n^{-\alpha}$. Given this choice, define

$$Q : \quad \mathbb{R}^I \to C([0,1]^d), \quad y \mapsto \begin{cases} F_y, & \text{if } y \in S, \\ 0, & \text{otherwise.} \end{cases}$$

We claim that $\|f - Q(\Omega(f))\|_{L^\infty} \leq C_2 \cdot m^{-\alpha/(d\cdot(\gamma+\alpha))}$ for all $f \in B$, for a suitable constant $C_2 = C_2(d, \gamma, \ell, c)$. Once this is shown, it follows that $\beta_*^{\det}(U, \iota_\infty) \geq \frac{1}{d}\frac{\alpha}{\gamma+\alpha}$, which then implies the claim of the theorem, since $\gamma > \gamma^\sharp(\ell, c)$ was arbitrary.

Thus, let $f \in B$ be arbitrary and set $y := \Omega(f) \in S$. By the same arguments as above, there exists $F \in \Sigma_n^{\ell,c}$ satisfying $\|f - F\|_{L^\infty} \leq 2 \cdot n^{-\alpha}$. Now, we see for each $i \in I$ because of $f(i) = (\Omega(f))_i = y_i = (\Omega(f_y))_i = f_y(i)$ that

$$\begin{aligned}|F(i) - F_y(i)| &\leq |F(i) - f(i)| + |f_y(i) - F_y(i)| \\ &\leq \|F - f\|_{L^\infty} + \|f_y - F_y\|_{L^\infty} \\ &\leq 4 \cdot n^{-\alpha}.\end{aligned}$$

Furthermore, Lemma 4.1 shows that $F - F_y : (\mathbb{R}^d, \|\cdot\|_{\ell^\infty}) \to (\mathbb{R}, |\cdot|)$ is Lipschitz continuous with Lipschitz constant at most 2μ. Now, given any $x \in [0,1]^d$, we can choose $i = i(x) \in I$ satisfying $\|x - i\|_{\ell^\infty} \leq N^{-1}$. Therefore, $|(F - F_y)(x)| \leq \frac{2\mu}{N} + |(F - F_y)(i)| \leq \frac{2\mu}{N} + 4\,n^{-\alpha}$. Overall, we have thus shown $\|F - F_y\|_{L^\infty} \leq \frac{2\mu}{N} + 4\,n^{-\alpha}$, which finally implies because of $Q(\Omega(f)) = Q(y) = F_y$ that

$$\big\|f - Q(\Omega(f))\big\|_{L^\infty} \leq \|f - F\|_{L^\infty} + \|F - F_y\|_{L^\infty} \leq 6\,n^{-\alpha} + \frac{2\mu}{N}.$$

It remains to note that our choice of N and n implies $m^{1/d} \leq 1 + N \leq 2N$ and hence $\frac{1}{N} \leq 2m^{-1/d}$ and furthermore $n \leq 1 + m^{1/(d\cdot(\gamma+\alpha))} \leq 2\,m^{1/(d\cdot(\gamma+\alpha))}$. Hence, recalling that $\mu \leq C_1\,n^\gamma$, we see

$$\frac{\mu}{N} \leq 2C_1 m^{-1/d} n^\gamma \leq 2^{1+\gamma} C_1 m^{\frac{1}{d}\left(\frac{\gamma}{\gamma+\alpha} - 1\right)} = 2^{1+\gamma} C_1 m^{-\frac{\alpha}{d\cdot(\gamma+\alpha)}}.$$

Furthermore, since $n \geq m^{1/(d\cdot(\gamma+\alpha))}$, we also have $n^{-\alpha} \leq m^{-\frac{\alpha}{d\cdot(\gamma+\alpha)}}$. Combining all these observations, it is easy to see that $\|f - Q(\Omega(f))\|_{L^\infty} \leq C_2 \cdot m^{-\frac{\alpha}{d\cdot(\gamma+\alpha)}}$, for a suitable constant $C_2 = C_2(d, \gamma, \boldsymbol{\ell}, \boldsymbol{c}) > 0$. Since $f \in B$ was arbitrary, this completes the proof. □

5 Hardness of Uniform Approximation

In this section, we show that the error bound for uniform approximation provided by Theorem 4.2 is optimal, at least in the common case where $\gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c}) = \gamma^\sharp(\boldsymbol{\ell}, \boldsymbol{c})$ and $\boldsymbol{\ell}^* \geq 3$. This latter condition means that the approximation for defining the approximation space $A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ is performed using networks with at least *two hidden layers*. We leave it as an interesting question for future work whether a similar result even holds for approximation spaces associated to shallow networks.

Theorem 5.1 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ be non-decreasing with $\boldsymbol{\ell}^* \geq 3$. Given $d \in \mathbb{N}$ and $\alpha \in (0, \infty)$, let $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} = U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ as in Eq. (2.3) and consider the embedding $\iota_\infty : A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \hookrightarrow C([0,1]^d)$. Then*

$$\beta^{\mathrm{det}}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_\infty), \beta^{\mathrm{ran}}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_\infty) \leq \frac{1}{d}\frac{\alpha}{\alpha + \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})}.$$

Proof Set $K := [0,1]^d$ and $U := U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$.

Step 1: Let $0 < \gamma < \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$. Let $m \in \mathbb{N}$ be arbitrary and $\Gamma_m := \underline{2k}^d \times \{\pm 1\}$, where $k := \lceil m^{1/d} \rceil$. In this step, we show that there is a constant $\kappa = \kappa(d, \alpha, \gamma, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ (independent of m) and a family of functions $(f_{\ell,\nu})_{(\ell,\nu)\in\Gamma_m} \subset U$ which satisfies

$$\sum_{(\ell,\nu)\in\Gamma_m} \big\| f_{\ell,\nu} - A(f_{\ell,\nu}) \big\|_{L^\infty} \geq \kappa \cdot m^{-\frac{1}{d}\frac{\alpha}{\alpha+\gamma}} \qquad \forall\, A \in \mathrm{Alg}_m\big(U, C([0,1]^d)\big). \tag{5.1}$$

To see this, set $M := 4k$, and for $\ell \in \underline{2k}^d$ define $y^{(\ell)} := \frac{(1,\dots,1)}{4k} + \frac{\ell - (1,\dots,1)}{2k} \in \mathbb{R}^d$. Then, we have

$$\begin{aligned} y^{(\ell)} + (-M^{-1}, M^{-1})^d &= \frac{2}{M}\big(\ell - (1,\dots,1)\big) + \frac{(1,\dots,1)}{M} + (-M^{-1}, M^{-1})^d \\ &= \frac{2}{M}\Big(\ell - (1,\dots,1) + (0,1)^d\Big) \subset (0,1)^d, \end{aligned}$$

which shows that the functions $\vartheta_{M,y^{(\ell)}}$, $\ell \in \underline{2k}^d$, (with $\vartheta_{M,y}$ as defined in Lemma 3.4), have disjoint supports contained in $[0,1]^d$. Furthermore, Lemma 3.6 yields a constant $\kappa_1 = \kappa_1(\gamma, \alpha, d, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ such that $f_{\ell,\nu} := \kappa_1 \cdot M^{-\alpha/(\alpha+\gamma)} \cdot \nu \cdot \vartheta_{M,y^{(\ell)}} \in U$ for arbitrary $(\ell, \nu) \in \Gamma_m$.

To prove Eq. (5.1), let $A \in \mathrm{Alg}_m(U, C([0,1]^d))$ be arbitrary. By definition, there exist $\boldsymbol{x} = (x_1, \dots, x_m) \in K^m$ and a function $Q : \mathbb{R}^m \to \mathbb{R}$ satisfying $A(f) = Q(f(x_1), \dots, f(x_m))$ for all $f \in U$. Choose $I := I_{\boldsymbol{x}} := \big\{\ell \in \underline{2k}^d : \forall\, n \in \underline{m} : \vartheta_{M,y^{(\ell)}}(x_n) = 0\big\}$. Then for each $\ell \in I^c = \underline{2k}^d \setminus I$, there exists $n_\ell \in \underline{m}$ such that $\vartheta_{M,y^{(\ell)}}(x_{n_\ell}) \neq 0$. Then the map $I^c \to \underline{m}$, $\ell \mapsto n_\ell$ is injective, since $\vartheta_{M,y^{(\ell)}}\, \vartheta_{M,y^{(t)}} = 0$ for $t, \ell \in \underline{2k}^d$ with $t \neq \ell$. Therefore, $|I^c| \leq m$ and hence $|I| \geq (2k)^d - m \geq m$, because of $k \geq m^{1/d}$.

Define $h := Q(0, \dots, 0)$. Then for each $\ell \in I_{\boldsymbol{x}}$ and $\nu \in \{\pm 1\}$, we have $f_{\ell,\nu}(x_n) = 0$ for all $n \in \underline{m}$ and hence $A(f_{\ell,\nu}) = Q(0, \dots, 0) = h$. Therefore,

$$\begin{aligned} &\|f_{\ell,1} - A(f_{\ell,1})\|_{L^\infty} + \|f_{\ell,-1} - A(f_{\ell,-1})\|_{L^\infty} \\ &= \|f_{\ell,1} - h\|_{L^\infty} + \| - f_{\ell,1} - h\|_{L^\infty} = \|f_{\ell,1} - h\|_{L^\infty} + \|h + f_{\ell,1}\|_{L^\infty} \\ &\geq \|f_{\ell,1} - h + h + f_{\ell,1}\|_{L^\infty} = 2\,\|f_{\ell,1}\|_{L^\infty} = 2\kappa_1 \cdot M^{-\alpha/(\alpha+\gamma)} \qquad \forall\, \ell \in I_{\boldsymbol{x}}. \end{aligned} \tag{5.2}$$

Furthermore, since $k \leq 1 + m^{1/d} \leq 2m^{1/d}$, we see $k^d \leq 2^d m$ and $M = 4k \leq 8\, m^{1/d}$ and hence $M^{\frac{\alpha}{\alpha+\gamma}} \leq 8^{\frac{\alpha}{\alpha+\gamma}} m^{\frac{1}{d}\frac{\alpha}{\alpha+\gamma}}$. Combining these estimates with Eq. (5.2) and recalling that $|I| \geq m$, we finally see

$$\begin{aligned} \sum_{(\ell,\nu)\in\Gamma_m} \|f_{\ell,\nu} - A(f_{\ell,\nu})\|_{L^\infty} &\geq (2k)^{-d} \sum_{\ell\in I_{\boldsymbol{x}}} \sum_{\nu\in\{\pm 1\}} \|f_{\ell,\nu} - A(f_{\ell,\nu})\|_{L^\infty} \\ &\geq (2k)^{-d} \cdot |I| \cdot \kappa_1 \cdot M^{-\frac{\alpha}{\alpha+\gamma}} \geq \frac{\kappa_1}{4^d} \cdot m^{-1}\, |I| \cdot M^{-\frac{\alpha}{\alpha+\gamma}} \\ &\geq \frac{\kappa_1/8}{4^d} \cdot m^{-\frac{1}{d}\frac{\alpha}{\alpha+\gamma}}, \end{aligned}$$

which establishes Eq. (5.1) for $\kappa := \frac{\kappa_1/8}{4^d}$.

Step 2 (Completing the proof): Given Eq. (5.1), a direct application of Lemma 2.3 shows that $\beta_*^{\det}(U, \iota_\infty), \beta_*^{\mathrm{ran}}(U, \iota_\infty) \leq \frac{1}{d}\frac{\alpha}{\alpha+\gamma}$. Since this holds for arbitrary $0 < \gamma < \gamma^{\flat}(\boldsymbol{\ell}, \boldsymbol{c})$, we easily obtain the claim of the theorem. □

6 Error Bounds for Approximation in L^2

This section provides error bounds for the approximation of functions in $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}([0,1]^d)$ based on point samples, with error measured in L^2. In a nutshell, the argument is based on combining bounds from statistical learning theory (specifically from [13]) with bounds for the *covering numbers* of the neural network sets $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$.

For completeness, we mention that the *ε-covering number* $\mathrm{Cov}(\Sigma, \varepsilon)$ (with $\varepsilon > 0$) of a (non-empty) subset Σ of a metric space (X, d) is the minimal number $N \in \mathbb{N}$ for which there exist $f_1, \dots, f_N \in \Sigma$ satisfying $\Sigma \subset \bigcup_{j=1}^N \overline{B}_\varepsilon(f_j)$. Here, $\overline{B}_\varepsilon(f) := \{g \in X : d(f, g) \leq \varepsilon\}$. If no such $N \in \mathbb{N}$ exists, then $\mathrm{Cov}(\Sigma, \varepsilon) = \infty$. If we want to emphasize the metric space X, we also write $\mathrm{Cov}_X(\Sigma, \varepsilon)$.

For the case where one considers networks *of a given architecture*, bounds for the covering numbers of network sets have been obtained for instance in [8, Proposition 2.8]. Here, however, we are interested in *sparsely connected networks with unspecified architecture*. For this case, the following lemma provides covering bounds.

Lemma 6.1 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N}$ be non-decreasing. The covering numbers of the neural network set $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ (considered as a subset of the metric space $C([0,1]^d)$) can be estimated by*

$$\mathrm{Cov}_{C([0,1]^d)}(\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon) \leq \left(\frac{44}{\varepsilon} \cdot \big(\boldsymbol{\ell}(n)\big)^4 \cdot \big(\boldsymbol{c}(n) \max\{d, n\}\big)^{1+\boldsymbol{\ell}(n)}\right)^n$$

for arbitrary $\varepsilon \in (0, 1]$ and $n \in \mathbb{N}$.

Proof Define $L := \boldsymbol{\ell}(n)$ and $R := \boldsymbol{c}(n)$. We will use some results and notation from [8]. Precisely, given a network architecture $\boldsymbol{a} = (a_0, \dots, a_K) \in \mathbb{N}^{K+1}$, we denote by

$$\mathcal{NN}(\boldsymbol{a}) := \prod_{j=1}^{K} \left([-R, R]^{a_j \times a_{j-1}} \times [-R, R]^{a_j}\right)$$

the set of all network weights with architecture $\boldsymbol{a}$ and all weights bounded (in magnitude) by R. Let us also define the index set $I(\boldsymbol{a}) := \biguplus_{j=1}^{K} \big(\{j\} \times \{1, ..., a_j\} \times \{1, ..., 1 + a_{j-1}\}\big)$, noting that $\mathcal{NN}(\boldsymbol{a}) \cong [-R, R]^{I(\boldsymbol{a})}$. In the following, we will equip $\mathcal{NN}(\boldsymbol{a})$ with the ℓ^∞-norm. Then, [8, Theorem 2.6] shows that the realization map $R_\varrho : \mathcal{NN}(\boldsymbol{a}) \to C([0,1]^d), \Phi \mapsto R_\varrho \Phi$ is Lipschitz continuous on $\mathcal{NN}(\boldsymbol{a})$, with Lipschitz constant bounded by $2K^2 R^{K-1} \|\boldsymbol{a}\|_\infty^K$, a fact that we will use below.

For $\ell \in \{1, \dots, L\}$, define $\boldsymbol{a}^{(\ell)} := (d, n, \dots, n, 1) \in \mathbb{N}^{\ell+1}$ and $I_\ell := I(\boldsymbol{a}^{(\ell)})$, as well as

$$\Sigma_\ell := \left\{ R_\varrho \Phi \,:\, \begin{array}{l} \Phi \text{ NN with } d_{\text{in}}(\Phi) = d, d_{\text{out}}(\Phi) = 1, \\ W(\Phi) \leq n, L(\Phi) = \ell, \|\Phi\|_{\mathcal{NN}} \leq R \end{array} \right\}.$$

By dropping "dead neurons," it is easy to see that each $f \in \Sigma_\ell$ is of the form $f = R_\varrho \Phi$ for some $\Phi \in \mathcal{NN}(\boldsymbol{a}^{(\ell)})$ satisfying $W(\Phi) \leq n$. Thus, keeping the identification $\mathcal{NN}(\boldsymbol{a}) \cong [-R, R]^{I(\boldsymbol{a})}$, given a subset $S \subset I_\ell$, let us write $\mathcal{NN}_{S,\ell} := \left\{\Phi \in \mathcal{NN}(\boldsymbol{a}^{(\ell)}) \colon \operatorname{supp} \Phi \subset S\right\}$; then we have $\Sigma_\ell = \bigcup_{S \subset I_\ell, |S| = \min\{n, |I_\ell|\}} R_\varrho(\mathcal{NN}_{S,\ell})$. Moreover, it is easy to see that $|I_\ell| \leq 2d$ if $\ell = 1$ while if $\ell \geq 2$ then $|I_\ell| = 1 + n(d+2) + (\ell - 2)(n^2 + n)$. This implies in all cases that $|I_\ell| \leq 2n(Ln + d)$.

Now we collect several observations which in combination will imply the claimed bound. First, directly from the definition of covering numbers, we see that if Θ is Lipschitz continuous, then $\operatorname{Cov}(\Theta(\Omega), \varepsilon) \leq \operatorname{Cov}(\Omega, \frac{\varepsilon}{\operatorname{Lip}(\Theta)})$, and furthermore $\operatorname{Cov}(\bigcup_{j=1}^K \Omega_j, \varepsilon) \leq \sum_{j=1}^K \operatorname{Cov}(\Omega_j, \varepsilon)$. Moreover, since $\mathcal{NN}_{S,\ell} \cong [-R, R]^{|S|}$, we see by [8, Lemma 2.7] that $\operatorname{Cov}_{\ell^\infty}(\mathcal{NN}_{S,\ell}, \varepsilon) \leq \lceil R/\varepsilon \rceil^n \leq (2R/\varepsilon)^n$. Finally, [50, Exercise 0.0.5] provides the bound $\binom{N}{n} \leq (eN/n)^n$ for $n \leq N$.

Recall that the realization map $R_\varrho : \mathcal{NN}(\boldsymbol{a}^{(\ell)}) \to C([0,1]^d)$ is Lipschitz continuous with $\operatorname{Lip}(R_\varrho) \leq C := 2L^2 R^{L-1} \max\{d, n\}^L$. Combining this with the observations from the preceding paragraph and recalling that $|I_\ell| \leq 2n(Ln + d)$, we see

$$\begin{aligned}
\operatorname{Cov}_{C([0,1]^d)}(\Sigma_\ell, \varepsilon) &\leq \sum_{S \subset I_\ell, |S| = \min\{n, |I_\ell|\}} \operatorname{Cov}_{C([0,1]^d)}\big(R_\varrho(\mathcal{NN}_{S,\ell}), \varepsilon\big) \\
&\leq \sum_{S \subset I_\ell, |S| = \min\{n, |I_\ell|\}} \operatorname{Cov}_{\ell^\infty}(\mathcal{NN}_{S,\ell}, \tfrac{\varepsilon}{C}) \\
&\leq \sum_{S \subset I_\ell, |S| = \min\{n, |I_\ell|\}} \Big(\frac{2CR}{\varepsilon}\Big)^{|S|} \leq \binom{|I_\ell|}{\min\{n, |I_\ell|\}} \cdot \Big(\frac{2CR}{\varepsilon}\Big)^n \\
&\leq \Big(\frac{e|I_\ell|}{\min\{n, |I_\ell|\}}\Big)^n \cdot \Big(\frac{2CR}{\varepsilon}\Big)^n \leq \big(2e(Ln + d)\big)^n \cdot \Big(\frac{2CR}{\varepsilon}\Big)^n.
\end{aligned}$$

Finally, noting that $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} = \bigcup_{\ell=1}^L \Sigma_\ell$ and setting $\eta := \max\{d, n\}$, we see via elementary estimates that

$$\begin{aligned}
&\operatorname{Cov}_{C([0,1]^d)}(\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon) \leq L \cdot \big(4e(Ln + d)RC/\varepsilon\big)^n \leq L \cdot \big(16e\, L^3 \eta^{L+1} R^L/\varepsilon\big)^n \\
&\leq \big(44\, L^4\, \eta^{L+1} R^L/\varepsilon\big)^n,
\end{aligned}$$

which implies the claim of the lemma. □

Using the preceding bounds for the covering numbers of the network sets $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$, we now derive covering number bounds for the (closure of the) unit ball $U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$ of the approximation space $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$.

Lemma 6.2 *Let $d \in \mathbb{N}$, $C_1, C_2, \alpha \in (0,\infty)$, and $\theta, \nu \in [0,\infty)$. Assume that $\boldsymbol{c}(n) \leq C_1 \cdot n^\theta$ and $\boldsymbol{\ell}(n) \leq C_2 \cdot \ln^\nu(2n)$ for all $n \in \mathbb{N}$.*

Then there exists $C = C(d,\alpha,\theta,\nu,C_1,C_2) > 0$ such that for any $\varepsilon \in (0,1]$, the unit ball

$$U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} := \big\{ f \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \colon \|f\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}} \leq 1 \big\}$$

satisfies

$$\mathrm{Cov}_{C([0,1]^d)}\big(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon\big) \leq \exp\big(C \cdot \varepsilon^{-1/\alpha} \cdot \ln^{\nu+1}(2/\varepsilon)\big).$$

Here, we denote by $\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ the closure of $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ in $C([0,1]^d)$.

Proof Let $n := \lceil (8/\varepsilon)^{1/\alpha} \rceil \in \mathbb{N}_{\geq 2}$, noting $n^{-\alpha} \leq \varepsilon/8$. Set $C := \boldsymbol{c}(n)$ and $L := \boldsymbol{\ell}(n)$. Lemma 6.1 provides an absolute constant $C_3 > 0$ and $N \in \mathbb{N}$ such that $N \leq \big(\frac{C_3}{\varepsilon} L^4 \cdot (C \max\{d,n\})^{1+L}\big)^n$ and functions $h_1, \dots, h_N \in \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ satisfying $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} \subset \bigcup_{j=1}^N \overline{B}_{\varepsilon/4}(h_j)$; here, $\overline{B}_\varepsilon(h)$ is the closed ball in $C([0,1]^d)$ of radius ε around h. For each $j \in \underline{N}$ choose $g_j \in U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \cap \overline{B}_{\varepsilon/2}(h_j)$, provided that the intersection is non-empty; otherwise choose $g_j := 0 \in U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$.

We claim that $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset \bigcup_{j=1}^N \overline{B}_\varepsilon(g_j)$. To see this, let $f \in U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ be arbitrary; then Lemma 2.1 shows that $\Gamma_{\alpha,\infty}(f) \leq 1$. Directly from the definition of $\Gamma_{\alpha,\infty}$ we see that we can choose $h \in \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ satisfying $n^\alpha \|f - h\|_{L^\infty} \leq 2$ and hence $\|f - h\|_{L^\infty} \leq \frac{\varepsilon}{4}$. By choice of $h_1, \dots, h_N$, there exists $j \in \underline{N}$ satisfying $\|h - h_j\|_{L^\infty} \leq \frac{\varepsilon}{4}$. This implies $\|f - h_j\|_{L^\infty} \leq \frac{\varepsilon}{2}$ and therefore $f \in \overline{B}_{\varepsilon/2}(h_j) \cap U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \neq \varnothing$. By our choice of g_j, we thus have $g_j \in U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \cap \overline{B}_{\varepsilon/2}(h_j)$ and hence $\|f - g_j\|_{L^\infty} \leq \varepsilon$. All in all, we have thus shown $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset \bigcup_{j=1}^N \overline{B}_{\varepsilon/2}(g_j)$ and hence also $\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset \bigcup_{j=1}^N \overline{B}_{\varepsilon/2}(g_j)$. This implies $\mathrm{Cov}_{C([0,1]^d)}(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon) \leq N$, so that it remains to estimate N sufficiently well.

To estimate N, first note that

$$\begin{aligned} n &\leq 1 + (\tfrac{8}{\varepsilon})^{1/\alpha} \leq 2 \cdot 8^{1/\alpha}\, \varepsilon^{-1/\alpha} \quad \text{and} \\ \ln(n) &\leq \ln(2n) \leq \ln(4 \cdot 8^{1/\alpha}) + \tfrac{1}{\alpha} \ln(\tfrac{1}{\varepsilon}) \leq C_4 \cdot \ln(\tfrac{2}{\varepsilon}) \end{aligned} \tag{6.1}$$

for a suitable constant $C_4 = C_4(\alpha) > 0$. This implies

$$L \leq 1 + L \leq 2L \leq 2C_2 \ln^\nu(2n) \leq 2C_2 C_4^\nu \cdot \ln^\nu(\tfrac{2}{\varepsilon}) \leq C_5 \cdot \ln^\nu(\tfrac{2}{\varepsilon})$$

with a constant $C_5 = C_5(C_2,\nu,\alpha) \geq 1$.

Now, using Eq. (6.1) and noting $\max\{d,n\} \le d\,n$, we obtain $C_6 = C_6(d,\alpha,C_1) > 0$ and $C_7 = C_7(d,\alpha,\theta,\nu,C_1,C_2) > 0$ satisfying

$$\begin{aligned}\big(C\,\max\{d,n\}\big)^{1+L} &\le \big(C_1\,d\cdot n^{\theta+1}\big)^{1+L} \le \big(C_6\cdot n^{1+\theta}\big)^{1+L} \le \big(C_6\cdot n^{1+\theta}\big)^{C_5\,\ln^\nu(2/\varepsilon)}\\ &= \exp\Big(\big(\ln(C_6)+(1+\theta)\,\ln(n)\big)\cdot C_5\,\ln^\nu(2/\varepsilon)\Big)\\ &\le \exp\Big(\big(\ln(C_6)+(1+\theta)\,C_4\,\ln(2/\varepsilon)\big)\cdot C_5\,\ln^\nu(2/\varepsilon)\Big)\\ &\le \exp\Big(C_7\cdot\ln^{\nu+1}(2/\varepsilon)\Big).\end{aligned} \tag{6.2}$$

Furthermore, using the elementary estimate $\ln x \le x$ for $x > 0$, we see

$$\begin{aligned}\frac{C_3}{\varepsilon}\,L^4 &\le C_3C_5^4\cdot\ln^{4\nu}(2/\varepsilon)\cdot\varepsilon^{-1} \le 2^{4\nu}C_3C_5^4\cdot\varepsilon^{-(1+4\nu)}\\ &= \exp\big(C_8+(1+4\nu)\cdot\ln(1/\varepsilon)\big) \le \exp\big(C_9\,\ln(2/\varepsilon)\big) \le \exp\big(C_{10}\,\ln^{\nu+1}(2/\varepsilon)\big)\end{aligned} \tag{6.3}$$

for suitable constants C_8, C_9, C_10 all only depending on ν,α,C_2.

Overall, recalling the estimate for N from the beginning of the proof and using Eqs. (6.1), (6.2) and (6.3), we finally see

$$\begin{aligned}N &\le \left(\frac{C_3}{\varepsilon}\,L^4\cdot\big(C\,\max\{d,n\}\big)^{1+L}\right)^n \le \exp\Big((C_{10}+C_7)\cdot n\cdot\ln^{\nu+1}(2/\varepsilon)\Big)\\ &\le \exp\Big(2\cdot 8^{1/\alpha}\cdot(C_{10}+C_7)\cdot\varepsilon^{-1/\alpha}\cdot\ln^{\nu+1}(2/\varepsilon)\Big),\end{aligned}$$

which easily implies the claim of the lemma. □

Combining the preceding covering number bounds with bounds from statistical learning theory, we now prove the following error bound for approximating functions $f \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ from point samples, with error measured in L^2.

Theorem 6.3 *Let $d \in \mathbb{N}$, $C_1, C_2, \alpha \in (0,\infty)$, and $\theta, \nu \in [0,\infty)$. Let $\boldsymbol{c} : \mathbb{N} \to \mathbb{N}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\ge 2}$ be non-decreasing and such that $\boldsymbol{c}(n) \le C_1\cdot n^\theta$ and $\boldsymbol{\ell}(n) \le C_2\cdot\ln^\nu(2n)$ for all $n \in \mathbb{N}$. Let $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ as in Eq. (2.3), and denote by $\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ the closure of $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ in $C([0,1]^d)$.*

Then there exists a constant $C = C(\alpha,\theta,\nu,d,C_1,C_2) > 0$ such that for each $m \in \mathbb{N}$, there are points $x_1,\dots,x_m \in [0,1]^d$ with the following property:

$$\begin{aligned}&\forall\, f,g \in \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \text{ with } f(x_i) = g(x_i) \text{ for all } i \in \underline{m}:\\ &\quad \|f-g\|_{L^2([0,1]^d)} \le C\cdot\big(\ln^{1+\nu}(2m)/m\big)^{\frac{\alpha/2}{1+\alpha}}.\end{aligned} \tag{6.4}$$

In particular, this implies for the embedding $\iota_2 : A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \hookrightarrow L^2([0,1]^d)$ that

$$\beta^{\det}_*\big(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}},\iota_2\big) \ge \frac{\alpha/2}{1+\alpha}.$$

Remark The proof shows that the points $x_1, \dots, x_m$ can be obtained with positive probability by uniformly and independently sampling $x_1, \dots, x_m$ from $[0,1]^d$. In fact, an inspection of the proof shows for each $m \in \mathbb{N}$ that this sampling procedure will result in "good" points with probability at least

$$1 - \exp\Big(- \big[m \cdot \ln^{\alpha\cdot(1+\nu)}(2m)\big]^{1/(1+\alpha)}\Big).$$

Proof **Step 1:** An essential ingredient for our proof is [13, Proposition 7]. In this step, we briefly recall the general setup from [13] and describe how it applies to our setting.

Let us fix a function $f_0 \in \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ for the moment. In [13], one starts with a probability measure ρ on $Z = X \times Y$, where X is a compact domain and $Y = \mathbb{R}$. In our case we take $X = [0,1]^d$ and we define $\rho(M) := \rho_{f_0}(M) := \boldsymbol{\lambda}(\{x \in [0,1]^d : (x, f_0(x)) \in M\})$ for any Borel set $M \subset X \times Y$. In other words, ρ is the distribution of the random variable $\xi = (\eta, f_0(\eta))$, where η is uniformly distributed in $X = [0,1]^d$. Then, in the notation of [13], the measure ρ_X on X is simply the Lebesgue measure on $[0,1]^d$ and the conditional probability measure $\rho(\bullet \mid x)$ on Y is $\rho(\bullet \mid x) = \delta_{f_0(x)}$. Furthermore, the *regression function* f_ρ considered in [13] is simply $f_\rho = f_0$, and the *(least squares) error* $\mathcal{E}(f)$ of $f : X \to Y$ is $\mathcal{E}(f) = \int_{[0,1]^d} |f(x) - f_0(x)|^2 \, d\boldsymbol{\lambda}(x) = \|f - f_0\|^2_{L^2}$; to emphasize the role of f_0, we shall write $\mathcal{E}(f; f_0) = \|f - f_0\|^2_{L^2}$ instead. The *empirical error* of $f : X \to Y$ with respect to a sample $\boldsymbol{z} \in Z^m$ is

$$\mathcal{E}_{\boldsymbol{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} \big(f(x_i) - y_i\big)^2 \quad \text{where} \quad \boldsymbol{z} = \big((x_1, y_1), \dots, (x_m, y_m)\big).$$

We shall also use the notation

$$\mathcal{E}_{\boldsymbol{x}}(f; f_0) := \mathcal{E}_{\boldsymbol{z}}(f) = \frac{1}{m} \sum_{i=1}^{m} \big(f(x_i) - f_0(x_i)\big)^2 \quad \text{where} \quad y_i = f_0(x_i) \text{ for } i \in \underline{m}.$$

Furthermore, as the *hypothesis space* $\mathcal{H}$ we choose $\mathcal{H} := \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$. As required in [13], this is a compact subset of $C(X)$; indeed $\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset C([0,1]^d)$ is closed and has finite covering numbers $\mathrm{Cov}(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon)$ for arbitrarily small $\varepsilon > 0$ (see Lemma 6.2). Thus, $\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset C([0,1]^d)$ is compact; see for instance [2, Theorem 3.28].

Moreover, since every $(x, y) \in Z$ satisfies $y = f_0(x)$ almost surely (with respect to $\rho = \rho_{f_0}$), and since all $f \in \mathcal{H} = \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ satisfy $\|f\|_{C([0,1]^d)} \le 1$, we see that ρ_{f_0}-almost surely, the estimate $|f(x) - y| = |f(x) - f_0(x)| \le 2 =: M$ holds for all $f \in \mathcal{H}$. Furthermore, in [13], the function $f_{\mathcal{H}} \in \mathcal{H}$ is a minimizer of $\mathcal{E}$ over $\mathcal{H}$; in our case, since $f_0 \in \mathcal{H}$, we easily see that $f_{\mathcal{H}} = f_0$ and $\mathcal{E}(f_{\mathcal{H}}) = 0$. Therefore, the *error in* $\mathcal{H}$ of $f \in \mathcal{H}$ as considered in [13] is simply $\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\mathcal{H}}) = \mathcal{E}(f)$. Finally, the *empirical error in* $\mathcal{H}$ of $f \in \mathcal{H}$ is given by $\mathcal{E}_{\mathcal{H},\boldsymbol{z}}(f) = \mathcal{E}_{\boldsymbol{z}}(f) - \mathcal{E}_{\boldsymbol{z}}(f_{\mathcal{H}})$. Hence, if $\boldsymbol{z} = \big((x_1, y_1), \dots, (x_m, y_m)\big)$ satisfies $y_i = f_0(x_i)$ for all $i \in \underline{m}$, then $\mathcal{E}_{\mathcal{H},\boldsymbol{z}}(f) = \mathcal{E}_{\boldsymbol{z}}(f) = \mathcal{E}_{\boldsymbol{x}}(f; f_0)$, because of $f_{\mathcal{H}} = f_0$.

Now, let $\boldsymbol{x} = (x_1, \dots, x_m)$ be i.i.d. uniformly distributed in $[0,1]^d$ and set $y_i = f_0(x_i)$ for $i \in \underline{m}$ and $\boldsymbol{z} = (z_1, \dots, z_m) = \big((x_1, y_1), \dots, (x_m, y_m)\big)$. Then

$z_1, \dots, z_m \overset{iid}{\sim} \rho_{f_0}$. Therefore, [13, Proposition 7] (applied with $\alpha = \frac{1}{6}$) shows for arbitrary $\varepsilon > 0$ and $m \in \mathbb{N}$ that there is a measurable set

$$\begin{aligned} &E = E(m, \varepsilon, f_0) \subset ([0,1]^d)^m \cong [0,1]^{dm} \\ \text{with}\quad &\lambda(E) \leq \mathrm{Cov}\big(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \tfrac{\varepsilon}{48}\big) \cdot e^{-m\varepsilon/288} \text{satisfying} \\ &\sup_{f \in \mathcal{H}} \frac{\mathcal{E}(f; f_0) - \mathcal{E}_{\boldsymbol{x}}(f; f_0)}{\mathcal{E}(f; f_0) + \varepsilon} = \sup_{f \in \mathcal{H}} \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},z}(f)}{\mathcal{E}_{\mathcal{H}}(f) + \varepsilon} \leq \frac{1}{2} \qquad \forall \boldsymbol{x} \in \big([0,1]^d\big)^m \setminus E. \end{aligned} \tag{6.5}$$

Here, we remark that [13, Proposition 7] requires the hypothesis space $\mathcal{H}$ to be convex, which is not in general satisfied in our case. However, as shown in [13, Remark 13], the assumption of convexity can be dropped provided that $f_\rho \in \mathcal{H}$, which is satisfied in our case.

Step 2: In this step, we prove the first claim of the theorem. To this end, we first apply Lemma 6.2 to obtain a constant $C_3 = C_3(\alpha, \nu, \theta, d, C_1, C_2) > 0$ satisfying

$$\mathrm{Cov}\big(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon\big) \leq N_\varepsilon := \mathrm{Cov}\big(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \tfrac{\varepsilon}{48}\big) \leq \exp\big(C_3 \cdot \varepsilon^{-1/\alpha} \cdot \ln^{1+\nu}(2/\varepsilon)\big) \quad \forall\, \varepsilon \in (0,1]. \tag{6.6}$$

Next, define $C_4 := 1 + \frac{\alpha}{1+\alpha}$ and $C_5 := C_4^{1+\nu}$, and choose $C_6 = C_6(\alpha, \nu, \theta, d, C_1, C_2) \geq 1$ such that $2C_3C_5 - \frac{C_6}{288} \leq -1 < 0$.

Let $m \in \mathbb{N}$ be arbitrary with $m \geq m_0 = m_0(\alpha, \nu, \theta, d, C_1, C_2) \geq 2$, where m_0 is chosen such that $\varepsilon := C_6 \cdot \big(\ln^{1+\nu}(2m)/m\big)^{\alpha/(1+\alpha)}$ satisfies $\varepsilon \in (0,1]$; the case $m \leq m_0$ will be considered below. Let $N := N_\varepsilon$ as in Eq. (6.6). Since $\mathrm{Cov}(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \varepsilon) \leq N$, we can choose $f_1, \dots, f_N \in \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ such that $\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset \bigcup_{j=1}^N \overline{B}_\varepsilon(f_j)$, where $\overline{B}_\varepsilon(f) := \big\{g \in C([0,1]^d) \colon \|f - g\|_{L^\infty} \leq \varepsilon\big\}$. Now, for each $j \in \underline{N}$, choose $E_j := E(m, \varepsilon, f_j) \subset ([0,1]^d)^m$ as in Eq. (6.5), and define $E^* := \bigcup_{j=1}^N E_j$.

Note because of $C_6 \geq 1$ and $\ln(2m) \geq \ln(4) \geq 1$ that $\varepsilon \geq \big(\ln^{1+\nu}(2m)/m\big)^{\alpha/(1+\alpha)} \geq m^{-\alpha/(1+\alpha)}$ and hence

$$\ln(2/\varepsilon) \leq \ln(2) + \tfrac{\alpha}{1+\alpha} \ln(m) \leq C_4 \ln(2m) \quad \text{and thus} \quad \ln^{1+\nu}(2/\varepsilon) \leq C_5 \ln^{1+\nu}(2m).$$

Using the estimate for $N = N_\varepsilon$ from Eq. (6.6) and the bound for the measure of E_j from Eq. (6.5), we thus see

$$\begin{aligned} \lambda(E^*) &\leq N \cdot \mathrm{Cov}\big(\overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \tfrac{\varepsilon}{48}\big) \cdot e^{-m\varepsilon/288} \leq \exp\big(2C_3 \cdot \varepsilon^{-1/\alpha} \cdot \ln^{1+\nu}(2/\varepsilon) - m\varepsilon/288\big) \\ &\leq \exp\Big(2C_3C_5 \cdot \big(m / \ln^{1+\nu}(2m)\big)^{1/(1+\alpha)} \cdot \ln^{1+\nu}(2m) - \tfrac{C_6}{288} \cdot m^{1-\frac{\alpha}{1+\alpha}} \cdot \big(\ln(2m)\big)^{(1+\nu)\frac{\alpha}{1+\alpha}}\Big) \\ &\leq \exp\Big(m^{\frac{1}{1+\alpha}} \cdot \big(\ln(2m)\big)^{(1+\nu)\frac{\alpha}{1+\alpha}} \cdot \big(2C_3C_5 - \tfrac{C_6}{288}\big)\Big) \\ &\leq \exp\Big(-m^{\frac{1}{1+\alpha}} \cdot \big(\ln(2m)\big)^{(1+\nu)\frac{\alpha}{1+\alpha}}\Big) < 1. \end{aligned}$$

Thus, we can choose $\boldsymbol{x} = (x_1, \dots, x_m) \in ([0,1]^d)^m \setminus E^*$. We claim that every such choice satisfies the property stated in the first part of the theorem.

To see this, let $f, g \in \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ be arbitrary with $f(x_i) = g(x_i)$ for all $i \in \underline{m}$. By choice of $f_1, \dots, f_N$, there exists some $j \in \underline{N}$ satisfying $\|f - f_j\|_{L^\infty} \leq \varepsilon$.

Since $\boldsymbol{x} \notin E^*$, we have $\boldsymbol{x} \notin E_j = E(m, \varepsilon, f_j)$. In view of Eq. (6.5), this implies $\mathcal{E}(g; f_j) - \mathcal{E}_{\boldsymbol{x}}(g; f_j) \leq \frac{1}{2}(\mathcal{E}(g; f_j) + \varepsilon)$, and after rearranging, this yields $\mathcal{E}(g; f_j) \leq 2\,\mathcal{E}_{\boldsymbol{x}}(g; f_j) + \varepsilon$. Because of $\|g - f_j\|_{L^2} \leq \|g\|_{L^\infty} + \|f_j\|_{L^\infty} \leq 2$ and thanks to the elementary estimate $(a+\varepsilon)^2 = a^2 + 2a\varepsilon + \varepsilon^2 \leq a^2 + 5\varepsilon$ for $0 \leq a \leq 2$, we thus see

$$\begin{aligned}\|g - f\|_{L^2}^2 &\leq \big(\|g - f_j\|_{L^2} + \|f_j - f\|_{L^2}\big)^2 \\ &\leq \|g - f_j\|_{L^2}^2 + 5\varepsilon = \mathcal{E}(g; f_j) + 5\varepsilon \leq 2\,\mathcal{E}_{\boldsymbol{x}}(g; f_j) + 6\varepsilon.\end{aligned}$$

But directly from the definition and because of $g(x_i) = f(x_i)$ and $\|f - f_j\|_{L^\infty} \leq \varepsilon$, we see $\mathcal{E}_{\boldsymbol{x}}(g; f_j) = \frac{1}{m}\sum_{i=1}^m \big(g(x_i) - f_j(x_i)\big)^2 \leq \varepsilon^2 \leq \varepsilon$. Overall, we thus see that

$$\begin{aligned}&\|g - f\|_{L^2}^2 \leq 8\varepsilon = 8C_6\left(\ln^{1+\nu}(2m)/m\right)^{\frac{\alpha}{1+\alpha}} \\ &\quad \forall\, f, g \in \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty} \text{ satisfying } f(x_i) = g(x_i) \text{ for all } i \in \underline{m}.\end{aligned}$$

We have thus proved the claim for $m \geq m_0$. Since $\|g - f\|_{L^2} \leq \|f\|_{L^\infty} + \|g\|_{L^\infty} \leq 2$ for *arbitrary* $f, g \in \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$, it is easy to see that this proves the claim for all $m \in \mathbb{N}$, possibly after enlarging C.

Step 3: To complete the proof of the theorem, for each $\boldsymbol{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$, choose a fixed $f_{\boldsymbol{y}} \in \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$ satisfying

$$f_{\boldsymbol{y}} \in \operatorname*{argmin}_{f \in \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}} \sum_{i=1}^m \big(f(x_i) - y_i\big)^2;$$

existence of $f_{\boldsymbol{y}}$ is an easy consequence of the compactness of $\overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty} \subset C([0,1]^d)$. Define

$$\begin{aligned}\Phi:&\quad \mathbb{R}^m \to \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}, \quad \boldsymbol{y} \mapsto f_{\boldsymbol{y}} \quad \text{and} \\ A:&\quad \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty} \to \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}, \quad f \mapsto \Phi\big((f(x_1), \dots, f(x_m))\big).\end{aligned}$$

Then given any $f \in \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$, the function $g := Af \in \overline{U}_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$ satisfies $f(x_i) = g(x_i)$ for all $i \in \underline{m}$, and hence $\|f - Af\|_{L^2} \leq C \cdot \big(\ln^{1+\nu}(2m)/m\big)^{\frac{\alpha/2}{1+\alpha}}$, as shown in the previous step. By definition of $\beta_*^{\mathrm{det}}(\overline{U}_{\vec{\ell},\vec{c}}^{\alpha,\infty}, \iota_2)$, this easily entails $\beta_*^{\mathrm{det}}(\overline{U}_{\vec{\ell},\vec{c}}^{\alpha,\infty}, \iota_2) \geq \frac{\alpha/2}{1+\alpha}$. □

7 Hardness of Approximation in L^2

This section presents hardness results for approximating the embedding $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}([0,1]^d) \hookrightarrow L^2([0,1]^d)$ using point samples.

Theorem 7.1 *Let $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ be non-decreasing with $\boldsymbol{\ell}^* \geq 2$. Let $d \in \mathbb{N}$ and $\alpha \in (0, \infty)$. Set $\gamma^\flat := \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$ as in Eq. (2.2) and let $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ as in Eq. (2.3). For the embedding $\iota_2 : U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \to L^2([0,1]^d)$, $f \mapsto f$, we then have*

$$\begin{aligned}\beta^{\mathrm{det}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big),\ \beta^{\mathrm{ran}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big) &\leq \begin{cases} \min\left\{\frac{1}{2} + \frac{\alpha}{\alpha+\gamma^\flat},\ \frac{2\alpha}{\alpha+\gamma^\flat}\right\}, & \text{if } \alpha + \gamma^\flat < 2, \\ \min\left\{\frac{1}{2} + \frac{\alpha}{\alpha+\gamma^\flat},\ \alpha,\ \frac{1}{2} + \frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}\right\}, & \text{if } \alpha + \gamma^\flat \geq 2 \end{cases} \\ &= \begin{cases} \frac{2\alpha}{\alpha+\gamma^\flat}, & \text{if } \alpha + \gamma^\flat < 2, \\ \alpha, & \text{if } \alpha + \gamma^\flat \geq 2 \text{ and } \alpha \leq \frac{1}{2}, \\ \frac{1}{2} + \frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}, & \text{if } \alpha + \gamma^\flat \geq 2 \text{ and } \frac{1}{2} \leq \alpha \leq \gamma^\flat, \\ \frac{1}{2} + \frac{\alpha}{\alpha+\gamma^\flat}, & \text{if } \alpha + \gamma^\flat \geq 2 \text{ and } \alpha \geq \gamma^\flat. \end{cases} \end{aligned} \tag{7.1}$$

Remark The bound from above might seem intimidating at first sight, so we point out two important consequences: First, we always have $\beta^{\mathrm{det}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big),\ \beta^{\mathrm{ran}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big) \leq \frac{1}{2} + \frac{\alpha}{\alpha+\gamma^\flat} \leq \frac{3}{2}$, which shows that *no matter how large the approximation rate α is*, one can never get a better convergence rate than $m^{-3/2}$. Furthermore, in the important case where $\gamma^\flat = \infty$ (for instance if the depth-growth function $\boldsymbol{\ell}$ is unbounded), then $\beta^{\mathrm{det}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big),\ \beta^{\mathrm{ran}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big) \leq \frac{1}{2} + \frac{\alpha}{\alpha+\gamma^\flat} = \frac{1}{2}$. These two bounds are the interesting bounds for the regime of large α.

For small $\alpha > 0$, the theorem shows

$$\beta^{\mathrm{det}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big),\ \beta^{\mathrm{ran}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, \iota_2\big) \leq \max\left\{\tfrac{2\alpha}{\alpha+\gamma^\flat}, \alpha\right\} \leq \max\left\{\tfrac{2}{\gamma^\flat}, 1\right\} \cdot \alpha \leq 2\alpha,$$

since $\gamma^\flat \geq 1$. This shows that one cannot get a good rate of convergence for small exponents $\alpha > 0$.

Proof **Step 1 (preparation):** Let $0 < \gamma < \gamma^\flat$ be arbitrary and let $\theta \in (0, \infty)$ and $\lambda \in [0, 1]$ with $\theta\lambda \leq 1$ and set $\omega := \min\{-\theta\alpha,\ \theta \cdot (\gamma - \lambda) - 1\} \in (-\infty, 0)$.

Let $m \in \mathbb{N}$ be arbitrary and set $M := 4m$ and $z_j := \frac{1}{4m} + \frac{j-1}{2m}$ for $j \in \underline{2m}$. Then, Lemma 3.2 yields a constant $\kappa = \kappa(\gamma, \alpha, \lambda, \theta, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ (independent of m) such that

$$f_{\boldsymbol{v},J} := \kappa \cdot m^\omega \cdot \sum_{j \in J} v_j\, \Lambda^*_{M,z_j} \in U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \quad \forall\, J \subset \underline{2m} \text{ with } |J| \leq 2 \cdot m^{\theta\lambda}$$
$$\text{and } \boldsymbol{v} = (v_j)_{j \in \underline{2m}} \in [-1, 1]^{2m}.$$

Furthermore, Lemma 3.2 shows that the functions $(\Lambda^*_{M,z_i})_{i \in \underline{2m}}$ have supports contained in $[0, 1]^d$ which are pairwise disjoint (up to null-sets). By continuity, this implies $\Lambda^*_{M,z_i} \Lambda^*_{M,z_\ell} \equiv 0$ for $i \neq \ell$.

Let $k := \lceil m^{\theta\lambda} \rceil$, noting because of $\theta\lambda \leq 1$ that $k \leq \lceil m \rceil = m$ and $k \leq 1 + m^{\theta\lambda} \leq 2 \cdot m^{\theta\lambda}$. Set $\mathcal{P}_k(\underline{2m}) := \left\{J \subset \underline{2m} \colon |J| = k\right\}$ and $\Gamma_m := \{\pm 1\}^{2m} \times \mathcal{P}_k(\underline{2m})$. The idea of the proof is to show that Lemma 2.3 is applicable to the family $(f_{\boldsymbol{v},J})_{(\boldsymbol{v},J) \in \Gamma_m}$.

Step 2: In this step, we prove

$$\sum_{(\boldsymbol{\nu},J)\in\Gamma_m} \big\| f_{\boldsymbol{\nu},J} - A(f_{\boldsymbol{\nu},J}) \big\|_{L^2([0,1]^d)} \geq \frac{\kappa}{32}\cdot m^{\omega+\frac{1}{2}(\theta\lambda-1)} \qquad \forall\, A\in \mathrm{Alg}_m\big(U, L^2([0,1]^d)\big). \tag{7.2}$$

To see this, let $\boldsymbol{x} = (x_1,\dots,x_m)\in([0,1]^d)^m$ and $Q:\mathbb{R}^m\to L^2([0,1]^d)$ be arbitrary. Define $I := I_{\boldsymbol{x}} := \big\{ i\in \underline{2m} \,:\, \forall\, n\in\underline{m} : \Lambda^*_{M,z_i}(x_n) = 0\big\}$ as in Lemma 3.3 and recall the estimate $|I|\geq m$ from that lemma.

Now, given $\boldsymbol{\nu}^{(1)}\in\{\pm 1\}^I$ and $\boldsymbol{\nu}^{(2)}\in\{\pm 1\}^{I^c}$ as well as $J\in\mathcal{P}_k(\underline{2m})$, define

$$F_{\boldsymbol{\nu}^{(1)},J} := \kappa\cdot m^\omega\cdot \sum_{j\in I\cap J} \nu_j^{(1)}\Lambda^*_{M,z_j} \qquad \text{and} \qquad g_{\boldsymbol{\nu}^{(2)},J} := \kappa\cdot m^\omega\cdot \sum_{j\in I^c\cap J} \nu_j^{(2)}\Lambda^*_{M,z_j}$$

and finally $h_{\boldsymbol{\nu}^{(2)},J} := g_{\boldsymbol{\nu}^{(2)},J} - Q\big(g_{\boldsymbol{\nu}^{(2)},J}(x_1),\dots,g_{\boldsymbol{\nu}^{(2)},J}(x_m)\big)$. Note by choice of $I = I_{\boldsymbol{x}}$ that $f_{\boldsymbol{\nu},J}(x_n) = g_{\boldsymbol{\nu}^{(2)},J}(x_n)$ for all $n\in\underline{m}$, if we identify $\boldsymbol{\nu}$ with $(\boldsymbol{\nu}^{(1)},\boldsymbol{\nu}^{(2)})$, as we will continue to do for the remainder of the proof. Thus, we see for fixed but arbitrary $\boldsymbol{\nu}^{(2)}\in\{\pm1\}^{I^c}$ and $J\in\mathcal{P}_k(\underline{2m})$ that

$$\begin{aligned}
&\sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \big\| f_{\boldsymbol{\nu},J} - Q\big(f_{\boldsymbol{\nu},J}(x_1),\dots,f_{\boldsymbol{\nu},J}(x_m)\big)\big\|_{L^2([0,1]^d)} \\
&= \sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \big\| F_{\boldsymbol{\nu}^{(1)},J} + h_{\boldsymbol{\nu}^{(2)},J}\big\|_{L^2([0,1]^d)} \\
&= \frac{1}{2}\sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \Big(\big\| F_{\boldsymbol{\nu}^{(1)},J} + h_{\boldsymbol{\nu}^{(2)},J}\big\|_{L^2([0,1]^d)} + \big\| F_{-\boldsymbol{\nu}^{(1)},J} + h_{\boldsymbol{\nu}^{(2)},J}\big\|_{L^2([0,1]^d)}\Big) \qquad (7.3)\\
&\overset{(*)}{\geq} \sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \|F_{\boldsymbol{\nu}^{(1)},J}\|_{L^2([0,1]^d)} \\
&\overset{(\blacklozenge)}{\geq} 2^{|I|}\cdot\frac{\kappa}{8}\cdot m^\omega\cdot\left(\frac{|I\cap J|}{m}\right)^{1/2}.
\end{aligned}$$

Here, the step marked with $(*)$ used the identity $F_{-\boldsymbol{\nu}^{(1)},J} = -F_{\boldsymbol{\nu}^{(1)},J}$ and the elementary estimate $\|f+g\|_{L^2} + \|-f+g\|_{L^2} = \|f+g\|_{L^2} + \|f-g\|_{L^2} \geq \|f+g+f-g\|_{L^2} = 2\|f\|_{L^2}$. Finally, the step marked with $(*)$ used that the functions $\big(\Lambda^*_{M,z_i}\big)_{i\in\underline{2m}}$ have disjoint supports (up to null-sets) contained in $[0,1]^d$ and that $\Lambda^*_{M,z_j}(x)\geq\frac{1}{2}$ for all $x\in[0,1]^d$ satisfying $|x_1-z_j|\leq\frac{1}{2M}$; since $M=4m$, this easily implies $\|\Lambda^*_{M,z_i}\|_{L^2([0,1]^d)}\geq\frac{1}{2}\big(\frac{1}{2M}\big)^{1/2}\geq\frac{m^{-1/2}}{8}$ and hence

$$\begin{aligned}
\|F_{\boldsymbol{\nu}^{(1)},J}\|_{L^2([0,1]^d)} &= \kappa\cdot m^\omega\cdot\Big\|\sum_{j\in I\cap J}\nu_j^{(1)}\,\Lambda^*_{M,z_j}\Big\|_{L^2([0,1]^d)} \\
&= \kappa\cdot m^\omega\cdot\Big(\sum_{j\in I\cap J}|\nu_j^{(1)}|^2\,\|\Lambda^*_{M,z_j}\|^2_{L^2([0,1]^d)}\Big)^{1/2} \geq \frac{\kappa}{8}\cdot m^\omega\cdot\Big(|I\cap J|\,\Big/\,m\Big)^{1/2}.
\end{aligned}$$

Combining Eq. (7.3) with Lemma A.4 and recalling that $k \geq m^{\theta\lambda}$, we finally see

$$\begin{aligned}
&\sum_{(\boldsymbol{\nu},J)\in\Gamma_m} \left\| f_{\boldsymbol{\nu},J} - Q\big(f_{\boldsymbol{\nu},J}(x_1),\dots,f_{\boldsymbol{\nu},J}(x_m)\big)\right\|_{L^2([0,1]^d)} \\
&\geq \sum_{J\in\mathcal{P}_k(\underline{2m})} \sum_{\boldsymbol{\nu}^{(2)}\in\{\pm 1\}^{I^c}} \sum_{\boldsymbol{\nu}^{(1)}\in\{\pm 1\}^{I}} \left\| f_{\boldsymbol{\nu},J} - Q\big(f_{\boldsymbol{\nu},J}(x_1),\dots,f_{\boldsymbol{\nu},J}(x_m)\big)\right\|_{L^2([0,1]^d)} \\
&\geq \frac{\kappa}{8}\cdot m^{\omega} \sum_{J\in\mathcal{P}_k(\underline{2m})} \left(\frac{|I_{\boldsymbol{x}}\cap J|}{m}\right)^{1/2} \geq \frac{\kappa}{32}\cdot m^{\omega+\frac{1}{2}(\theta\lambda-1)}.
\end{aligned}$$

Recall that this holds for any $m \in \mathbb{N}$, arbitrary $\boldsymbol{x} = (x_1,\dots,x_m) \in ([0,1]^d)^m$ and any map $Q : \mathbb{R}^m \to L^2([0,1]^d)$. Thus, we have established Eq. (7.2).

Step 3: In view of Eq. (7.2), an application of Lemma 2.3 shows that

$$\beta_*^{\mathrm{det}}(U,\iota_2), \beta_*^{\mathrm{ran}}(U,\iota_2) \leq \tfrac{1}{2} - \omega - \tfrac{\theta\lambda}{2} = \tfrac{1}{2} + \max\left\{\theta\cdot(\alpha-\tfrac{\lambda}{2}),\ 1+\theta\cdot(\tfrac{\lambda}{2}-\gamma)\right\} \quad (7.4)$$

for arbitrary $0 < \gamma < \gamma^{\flat}$, $\theta \in (0,\infty)$ and $\lambda \in [0,1]$ with $\theta\lambda \leq 1$; here, we note that $\frac{1}{2} - \frac{\theta\lambda}{2} \geq 0$ and $-\omega \geq 0$.

From Eq. (7.4), it is easy (but slightly tedious) to deduce the first line of Eq. (7.1); the details are given in Lemma A.5. Finally, the second line of Eq. (7.1) follows by a straightforward case distinction. □

8 Error Bounds for Numerical Integration

In this section, we derive error bounds for the numerical integration of functions $f \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ based on point samples. We first consider (in Theorem 8.1) deterministic algorithms, which surprisingly provide a strictly positive rate of convergence, even for neural network approximation spaces *without restrictions on the size of the network weights*. Then, in Theorem 8.4, we consider the case of randomized (Monte Carlo) algorithms. As usual for such algorithms, they improve on the deterministic rate of convergence (essentially) by a factor of $m^{-1/2}$, at the cost of having a non-deterministic algorithm and (in our case) of requiring a non-trivial (albeit mild) condition on the growth function $\boldsymbol{c}$ used to define the space $A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$.

Theorem 8.1 *Let $d \in \mathrm{Nand}C, \sigma, \alpha \in (0,\infty)$. Let $\boldsymbol{c} : \mathbb{N} \to \mathbb{N}\cup\{\infty\}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2}\cup\{\infty\}$ be non-decreasing and assume that $\boldsymbol{\ell}(n) \leq C\cdot(\ln(en))^{\sigma}$ for all $n \in \mathbb{N}$. Then, with $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ as in Eq. (2.3) and with $T_{\int} : A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \to \mathbb{R}$, $f \mapsto \int_{[0,1]^d} f(x)\,dx$, we have*

$$\beta_*^{\mathrm{det}}\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_{\int}\big) \geq \frac{\alpha}{1+2\alpha} \in \Big(0,\frac{1}{2}\Big).$$

The proof relies on VC-dimension-based bounds for empirical processes. For the convenience of the reader, we briefly review the notion of VC dimension. Let $\Omega \neq \varnothing$

be a set, and let $\varnothing \neq \mathcal{H} \subset \{0,1\}^{\Omega}$ be arbitrary. In the terminology of machine learning, $\mathcal{H}$ is called a *hypothesis class*. The *growth function* of $\mathcal{H}$ is defined as

$$\tau_{\mathcal{H}}: \quad \mathbb{N} \to \mathbb{N}, \quad m \mapsto \sup_{x_1,\dots,x_m \in \Omega} \left| \left\{ \big(f(x_1), \dots, f(x_m)\big) : f \in \mathcal{H} \right\} \right|,$$

see [35, Definition 3.6]. That is, $\tau_{\mathcal{H}}(m)$ describes the maximal number of different ways in which the hypothesis class $\mathcal{H}$ can partition points $x_1, \dots, x_m \in \Omega$. Clearly, $\tau_{\mathcal{H}}(m) \leq 2^m$ for each $m \in \mathbb{N}$. This motivates the definition of the *VC-dimension* $\mathrm{VC}(\mathcal{H}) \in \mathbb{N}_0 \cup \{\infty\}$ of $\mathcal{H}$ as

$$\mathrm{VC}(\mathcal{H}) := \begin{cases} 0, & \text{if } \tau_{\mathcal{H}}(1) < 2^1, \\ \sup\left\{ m \in \mathbb{N} : \tau_{\mathcal{H}}(m) = 2^m \right\} \in \mathbb{N} \cup \{\infty\}, & \text{otherwise.} \end{cases}$$

For applying existing learning bounds based on the VC dimension in our setting, the following lemma will be essential.

Lemma 8.2 *Let $C_1 \geq 1$ and $C_2, \sigma_1, \sigma_2 > 0$. Then there exist constants $n_0 = n_0(C_1, C_2, \sigma_1, \sigma_2) \in \mathbb{N}$ and $C = C(C_1) > 0$ such that for every $n \in \mathbb{N}_{\geq n_0}$ and every $L \in \mathbb{N}$ with $L \leq C_2 \cdot (\ln(en))^{\sigma_2}$, the following holds:*

For any set $\Omega \neq \varnothing$ and any hypothesis classes $\varnothing \neq \mathcal{H}_1, \dots, \mathcal{H}_N \subset \{0,1\}^{\Omega}$ satisfying

$$N \leq L \cdot \binom{Ln^2}{n} \quad \text{and} \quad \mathrm{VC}(\mathcal{H}_j) \leq C_1 \cdot n \cdot (\ln(en))^{\sigma_1} \text{ for all } j \in \underline{N},$$

we have

$$\mathrm{VC}(\mathcal{H}_1 \cup \dots \cup \mathcal{H}_N) \leq C \cdot n \cdot (\ln(en))^{1+\sigma_1}.$$

Proof Choose $C_0 = 10\, C_1$ so that $\ln 2 - \frac{C_1}{C_0} \geq \frac{1}{2}$; here we used that $\ln 2 \approx 0.693 \geq \frac{6}{10}$. Set $C_3 := 1 + 2\ln(C_2) + 2\sigma_2$ and choose $n_0 = n_0(C_1, C_2, \sigma_1, \sigma_2) \in \mathbb{N}$ so large that for every $n \geq n_0$, we have $C_3 \cdot (\ln(en))^{-\sigma_1} \leq \frac{1}{6}$ and $C_1 \ln(20e) \cdot (\ln(en))^{-1} \leq \frac{1}{6}$.

For any subset $\varnothing \neq \mathcal{H} \subset \{0,1\}^{\Omega}$, Sauer's lemma shows that if $d_{\mathcal{H}} := \mathrm{VC}(\mathcal{H}) \in \mathbb{N}$, then $\tau_{\mathcal{H}}(m) \leq (em/d_{\mathcal{H}})^{d_{\mathcal{H}}}$ for all $m \geq d_{\mathcal{H}}$; see [35, Corollary 3.18]. An elementary calculation shows that the function $(0,\infty) \to \mathbb{R}$, $x \mapsto (em/x)^x$ is non-decreasing on $(0, m]$; thus, we see

$$\tau_{\mathcal{H}}(m) \leq (em/d)^d \qquad \forall\, m \in \mathbb{N} \text{ and } d \in [d_{\mathcal{H}}, m] \cap [1,\infty); \tag{8.1}$$

this trivially remains true if $d_{\mathcal{H}} = 0$.

Let $n \in \mathbb{N}_{\geq n_0}$, L, and $\mathcal{H}_1, \dots, \mathcal{H}_N$ as in the statement of the lemma. Set $\mathcal{H} := \mathcal{H}_1 \cup \dots \cup \mathcal{H}_N$ and $m := \lceil C_0 \cdot n \cdot (\ln(en))^{\sigma_1+1} \rceil$; we want to show that $\mathrm{VC}(\mathcal{H}) \leq m$. By definition of the VC dimension, it is sufficient to show that

$\tau_{\mathcal{H}}(m) < 2^m$. To this end, first note by a standard estimate for binomial coefficients (see [50, Exercise 0.0.5]) that

$$N \leq L \cdot \binom{Ln^2}{n} \leq L \cdot \left(eLn^2/n\right)^n \leq (eL^2 n)^n = \exp\big(n \cdot \ln(eL^2 n)\big) \leq \exp\big(C_3 n \ln(en)\big),$$

thanks to the elementary estimate $\ln x \leq x$, since $\ln(en) \geq 1$ and $L \leq C_2 \cdot (\ln(en))^{\sigma_2}$, and by our choice of C_3 at the beginning of the proof.

Next, recall that $C_0 = 10\, C_1$ and note $d_{\mathcal{H}_j} \leq d := C_1 \cdot n \cdot (\ln(en))^{\sigma_1} \in [1, m]$, so that Eq. (8.1) shows because of $m \leq 2C_0 \cdot n \cdot (\ln(en))^{\sigma_1+1}$ that

$$\tau_{\mathcal{H}_j}(m) \leq \left(\frac{em}{C_1 \cdot n \cdot (\ln(en))^{\sigma_1}}\right)^{C_1\, n\, (\ln(en))^{\sigma_1}} \leq \big(20e \ln(en)\big)^{C_1\, n\, (\ln(en))^{\sigma_1}}.$$

Combining all these observations and using the subadditivity property $\tau_{\mathcal{H}_1 \cup \mathcal{H}_2} \leq \tau_{\mathcal{H}_1} + \tau_{\mathcal{H}_2}$ and the bounds $m \geq C_0\, n\, (\ln(en))^{\sigma_1+1}$ and $\ln(2) - \frac{C_1}{C_0} \geq \frac{1}{2}$ as well as $C_0 \geq 1$, we see with $\theta := C_0\, n\, (\ln(en))^{\sigma_1+1}$ that

$$\begin{aligned}
\frac{\tau_{\mathcal{H}}(m)}{2^m} &\leq \frac{N}{2^m} \cdot \big(20e \ln(en)\big)^{C_1\, n\, (\ln(en))^{\sigma_1}} \\
&\leq \exp\big(C_3 n\, \ln(en) + C_1 n\, (\ln(en))^{\sigma_1}\, \ln(20e \ln(en)) - m \ln(2)\big) \\
&\leq \exp\Big(-\theta \cdot \Big[\ln(2) - \frac{C_1}{C_0} - \frac{C_1 \ln(20e)}{\ln(en)} - \frac{C_3}{(\ln(en))^{\sigma_1}}\Big]\Big) \\
&\leq \exp\Big(-\theta \cdot \Big[\frac{1}{2} - \frac{1}{6} - \frac{1}{6}\Big]\Big) = \exp\big(-\theta/6\big) < 1,
\end{aligned}$$

since $n \geq n_0$ and thanks to our choice of n_0 from the beginning of the proof.

Overall, we have thus shown $\tau_{\mathcal{H}}(m) < 2^m$ and hence $\mathrm{VC}(\mathcal{H}) \leq m \leq 2C_0 \cdot n \cdot (\ln(en))^{\sigma_1+1}$, which completes the proof, for $C := 2C_0 = 20\, C_1$. □

As a consequence, we get the following VC-dimension bounds for the network classes $\Sigma_n^{\ell,\infty}$.

Lemma 8.3 *Let $d \in \mathbb{N}$ and $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2}$ such that $\boldsymbol{\ell}(n) \leq C \cdot (\ln(en))^{\sigma}$ for all $n \in \mathbb{N}$ and certain $C, \sigma > 0$. Then there exist $n_0 = n_0(C, \sigma, d) \in \mathbb{N}$ and $C' = C'(C) > 0$ such that for all $\lambda \in \mathbb{R}$ and $n \geq n_0$, we have*

$$\mathrm{VC}\big(\big\{\mathbb{1}_{g>\lambda} : g \in \Sigma_n^{\ell,\infty}\big\}\big) \leq C' \cdot n \cdot (\ln(en))^{\sigma+2}.$$

Proof Given a network architecture $\boldsymbol{a} = (a_0, \ldots, a_K) \in \mathbb{N}^{K+1}$, we denote the set of all networks with architecture $\boldsymbol{a}$ by

$$\mathcal{NN}(\boldsymbol{a}) := \prod_{j=1}^{K} \big(\mathbb{R}^{a_j \times a_{j-1}} \times \mathbb{R}^{a_j}\big),$$

and by $I(\boldsymbol{a}) := \biguplus_{j=1}^{K} \big(\{j\} \times \{1, ..., a_j\} \times \{1, ..., 1 + a_{j-1}\}\big)$ the corresponding index set, so that $\mathcal{NN}(\boldsymbol{a}) \cong \mathbb{R}^{I(\boldsymbol{a})}$.

Define $L := \boldsymbol{\ell}(n)$. For $\ell \in \{1, \dots, L\}$, define $I_\ell := I(\boldsymbol{a}^{(\ell)})$ and $\boldsymbol{a}^{(\ell)} := (d, n, \dots, n, 1) \in \mathbb{N}^{\ell+1}$, as well as

$$\Sigma_\ell := \left\{ R_\varrho \Phi \,:\, \begin{array}{l} \Phi \text{ NN with } d_{\text{in}}(\Phi) = d, d_{\text{out}}(\Phi) = 1, \\ W(\Phi) \leq n, L(\Phi) = \ell, \end{array} \right\}.$$

By dropping "dead neurons," it is easy to see that each $f \in \Sigma_\ell$ is of the form $f = R_\varrho \Phi$ for some $\Phi \in \mathcal{NN}(\boldsymbol{a}^{(\ell)})$ satisfying $W(\Phi) \leq n$. In other words, keeping the identification $\mathcal{NN}(\boldsymbol{a}) \cong \mathbb{R}^{I(\boldsymbol{a})}$, given a subset $S \subset I_\ell$, let us write

$$\mathcal{NN}_{S,\ell} := \left\{ R_\varrho \Phi \in \mathcal{NN}(\boldsymbol{a}^{(\ell)}) \,:\, \operatorname{supp} \Phi \subset S \right\};$$

then $\Sigma_\ell = \bigcup_{S \subset I_\ell, |S| = \min\{n, |I_\ell|\}} \mathcal{NN}_{S,\ell}$. Moreover, $|I_\ell| \leq 2d$ if $\ell = 1$ while $|I_\ell| = 1 + n(d+2) + (\ell - 2)(n^2 + n)$ for $\ell \geq 2$, and this implies in all cases that $|I_\ell| \leq 2n(Ln + d) \leq L' \cdot n^2$ for $L' := 4d\,L$.

Overall, given a class $\mathcal{F} \subset \{f : \mathbb{R}^d \to \mathbb{R}\}$ and $\lambda \in \mathbb{R}$, let us write $\mathcal{F}(\lambda) := \{\mathbb{1}_{f > \lambda} \colon f \in \mathcal{F}\}$. Then the considerations from the preceding paragraph show that

$$\Sigma_n^{\boldsymbol{\ell},\infty}(\lambda) \subset \bigcup_{\ell=1}^{L} \bigcup_{S \subset I_\ell, |S| = \min\{n, |I_\ell|} \mathcal{NN}_{S,\ell}(\lambda). \tag{8.2}$$

Now, the set $\mathcal{NN}_{S,\ell}$ can be seen as all functions obtained by a fixed ReLU network (architecture) with at most n nonzero weights and ℓ layers, in which the weights are allowed to vary. Therefore, [6, Eq. (2)] shows for a suitable absolute constant $C^{(0)} > 0$ that

$$\operatorname{VC}(\mathcal{NN}_{S,\ell}(\lambda)) \leq C^{(0)} \cdot n\ell \ln(en) \leq C^{(0)} C \cdot n \cdot (\ln(en))^{\sigma+1}.$$

Finally, noting that the number of sets over which the union is taken in Eq. (8.2) is bounded by $\sum_{\ell=1}^{L} \binom{|I_\ell|}{n \min\{n, |I_\ell|\}} \leq \sum_{\ell=1}^{L} \binom{L' n^2}{n} \leq L \cdot \binom{L' n^2}{n} \leq L' \cdot \binom{L' n^2}{n}$, we can apply Lemma 8.2 (with $\sigma_1 = \sigma + 1$, $\sigma_2 = \sigma$, $C_1 = \max\{1, C^{(0)} C\}$, and $C_2 = 4dC$) to obtain $n_0 = n_0(d, C, \sigma) \in \mathbb{N}$ and $C' = C'(C) > 0$ satisfying $\operatorname{VC}(\Sigma_n^{\boldsymbol{\ell},\infty}(\lambda)) \leq C' \cdot n \cdot (\ln(en))^{\sigma+2}$ for all $n \geq n_0$. □

Proof of Theorem 8.1 Define $\theta := \frac{1}{1+2\alpha}$ and $\gamma := -\frac{\sigma+2}{1+2\alpha}$. Let $m \geq m_0$ with m_0 chosen such that $n := \lfloor m^\theta \cdot (\ln(em))^\gamma \rfloor$ satisfies $n \geq n_0$ for $n_0 = n_0(\sigma, C, d) \in \mathbb{N}$ provided by Lemma 8.3. Let $\mathcal{G} := \{g \in \Sigma_n^{\boldsymbol{\ell},\infty} \colon \|g\|_{L^\infty} \leq 3\}$ and note that Lemma 8.3 shows for every $\lambda \in \mathbb{R}$ that $\operatorname{VC}(\{\mathbb{1}_{g > \lambda} \colon g \in \mathcal{G}\}) \leq C' \cdot n \cdot (\ln(en))^{\sigma+2}$ for a suitable constant $C' = C'(C) > 0$. Therefore, [11, Proposition A.1] yields a universal constant $\kappa > 0$

such that if $X_1, \dots, X_m \overset{\text{iid}}{\sim} U([0,1]^d)$, then

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}\left|\int_{[0,1]^d} g(x)dx - \frac{1}{m}\sum_{j=1}^{m} g(X_j)\right|\right] \le 6\kappa\sqrt{\frac{C'\, n\,(\ln(en))^{\sigma+2}}{m}}.$$

In particular, there exists $\boldsymbol{x} = (X_1, \dots, X_m) \in ([0,1]^d)^m$ such that

$$\left|\int_{[0,1]^d} g(x)dx - \frac{1}{m}\sum_{j=1}^{m} g(X_j)\right| \le 6\kappa\sqrt{\frac{C'\, n\,(\ln(en))^{\sigma+2}}{m}} =: \varepsilon_1 \qquad \forall\, g \in \mathcal{G}.$$

Next, note because of $\gamma < 0$ that $n \le m^\theta\,(\ln(em))^\gamma \le m^\theta$ and hence $\ln(en) \lesssim \ln(em)$. Therefore,

$$\varepsilon_1 \lesssim \sqrt{\frac{n\cdot(\ln(en))^{\sigma+2}}{m}} \lesssim m^{\frac{\theta-1}{2}}\cdot(\ln(em))^{\frac{\sigma+2+\gamma}{2}} = m^{-\frac{\alpha}{1+2\alpha}}\cdot(\ln(em))^{-\alpha\gamma} =: \varepsilon_2,$$

where the implied constant only depends on α. Similarly, we have $n^{-\alpha} \lesssim m^{-\alpha\theta}(\ln(em))^{-\alpha\gamma} = \varepsilon_2$, because of $m^\theta\cdot(\ln(em))^\gamma \le n+1 \le 2n$.

Finally, set $Q : \mathbb{R}^m \to \mathbb{R}$, $(y_1,\dots,y_m) \mapsto \frac{1}{m}\sum_{j=1}^m y_j$ and let $f \in A^{\alpha,\infty}_{\boldsymbol{\ell},\infty}$ with $\|f\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\infty}} \le 1$ be arbitrary. By Lemma 2.1, we have $\Gamma_{\alpha,\infty}(f) \le 1$, which implies that $\|f\|_{L^\infty} \le 1$, and furthermore that there is some $g \in \Sigma_n^{\boldsymbol{\ell},\infty}$ satisfying $\|f-g\|_{L^\infty} \le 2n^{-\alpha} \le 2$, which in particular implies that $g \in \mathcal{G}$. Therefore,

$$\begin{aligned}
&\left|\int_{[0,1]^d} f(x)dx - Q\big(f(X_1),\dots,f(X_m)\big)\right| \\
&\le \left|\int_{[0,1]^d} f(x)-g(x)\,dx\right| + \left|\int_{[0,1]^d} g(x)\,dx - \frac{1}{m}\sum_{j=1}^m g(X_j)\right| + \left|\frac{1}{m}\sum_{j=1}^m (g-f)(X_j)\right| \\
&\le 2\|f-g\|_{L^\infty} + \varepsilon_1 \lesssim \varepsilon_2.
\end{aligned}$$

Since this holds for all $f \in U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$, with an implied constant independent of f and m, and since $\varepsilon_2 = m^{-\frac{\alpha}{1+2\alpha}}\cdot(\ln(em))^{-\alpha\gamma}$, this easily implies $\beta^{\det}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_{\int}) \ge \frac{\alpha}{1+2\alpha}$. □

Our next result shows that randomized (Monte Carlo) algorithms can improve the rate of convergence of the deterministic algorithm from Theorem 8.1 by (essentially) a factor $m^{-1/2}$. The proof is based on our error bounds for L^2 approximation from Theorem 6.3.

Theorem 8.4 *Let $d \in \mathbb{N}$, $C_1, C_2, \alpha \in (0,\infty)$, and $\theta, \nu \in [0,\infty)$. Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\ge 2}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N}$ be non-decreasing and such that $\boldsymbol{c}(n) \le C_1\cdot n^\theta$ and $\boldsymbol{\ell}(n) \le C_2\cdot\ln^\nu(2n)$ for all $n \in \mathbb{N}$. Let $U := U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} = \big\{f \in A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} : \|f\|_{A^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}} \le 1\big\}$.*

There exists $C = C(\alpha, \theta, \nu, d, C_1, C_2) > 0$ such that for every $m \in \mathbb{N}$, there exists a strongly measurable randomized (Monte Carlo) algorithm $(\boldsymbol{A}, \boldsymbol{m})$ with $\boldsymbol{m} \equiv m$ and $\boldsymbol{A} = (A_\omega)_{\omega\in\Omega}$ that satisfies

$$\left(\mathbb{E}\left|A_\omega(f) - \int_{[0,1]^d} f(t)\,dt\right|\right)^2 \le \mathbb{E}\left[\left|A_\omega(f) - \int_{[0,1]^d} f(t)\,dt\right|^2\right] \le C \cdot \frac{1}{m} \cdot \left(\ln^{1+\nu}(2m)/m\right)^{\frac{\alpha}{1+\alpha}} \tag{8.3}$$

for all $f \in U$. In particular, this implies

$$\beta^{\mathrm{ran}}_*\left(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_f\right) \ge \frac{1}{2} + \frac{\alpha/2}{1+\alpha}. \tag{8.4}$$

Proof Set $Q := [0,1]^d$. Let $m \in \mathbb{N}_{\ge 2}$ and $m' := \lfloor \frac{m}{2} \rfloor \in \mathbb{N}$ and note that $\frac{m}{2} \le m' + 1 \le 2m'$ and hence $\frac{m}{4} \le m' \le \frac{m}{2}$. Let $C = C(\alpha, \theta, \nu, d, C_1, C_2) > 0$ and $\boldsymbol{x} = (x_1, \dots, x_{m'}) \in Q^{m'}$ as provided by Theorem 6.3 (applied with m' instead of m). Note that $\mathcal{H} := \overline{U}^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}} \subset C(Q)$ is closed and nonempty, with finite covering numbers $\mathrm{Cov}_{C(Q)}(\mathcal{H}, \varepsilon)$, for arbitrary $\varepsilon > 0$; see Lemma 6.2. Hence, $\mathcal{H} \subset C(Q)$ is compact, see for instance [2, Theorem 3.28]. Let us equip $\mathcal{H}$ with the Borel σ-algebra induced by $C(Q)$. Then, it is easy to see from Lemma A.3 that the map $M : \mathcal{H} \to \mathbb{R}^{m'}$, $f \mapsto \big(f(x_1), \dots, f(x_{m'})\big)$ is measurable and that there is a measurable map $B : \mathbb{R}^{m'} \to \mathcal{H}$ satisfying $B(\boldsymbol{y}) \in \mathrm{argmin}_{g\in\mathcal{H}} \sum_{i=1}^{m'} \big(g(x_i) - y_i\big)^2$ for all $\boldsymbol{y} \in \mathbb{R}^{m'}$.

Given $f \in \mathcal{H}$, note that $g := B(M(f)) \in \mathcal{H}$ satisfies $g(x_i) = f(x_i)$ for all $i \in \underline{m'}$, so that Theorem 6.3 shows

$$\big\|f - B(M(f))\big\|_{L^2} \le C \cdot \left(\ln^{1+\nu}(2m')/m'\right)^{\frac{\alpha/2}{1+\alpha}} \le C' \cdot \left(\ln^{1+\nu}(2m)/m\right)^{\frac{\alpha/2}{1+\alpha}}, \tag{8.5}$$

for a suitable constant $C' = C'(\alpha, \theta, \nu, d, C_1, C_2) > 0$.

Now, consider the probability space $\Omega = Q^{m'} \cong [0,1]^{m'd}$, equipped with the Lebesgue measure $\boldsymbol{\lambda}$. For $\boldsymbol{z} \in \Omega$, write $\Omega \ni \boldsymbol{z} = (z_1, \dots, z_{m'})$ and define

$$\Psi : \quad \Omega \times C(Q) \to \mathbb{R}, \quad (\boldsymbol{z}, g) \mapsto \frac{1}{m'} \sum_{j=1}^{m'} g(z_j).$$

It is easy to see that Ψ is continuous and hence measurable; see Eq. (A.2) for more details.

Note that for $\boldsymbol{z} = (z_1, \dots, z_{m'}) \in \Omega$, the random vectors $z_1, \dots, z_{m'} \in Q$ are stochastically independent. Furthermore, for arbitrary $g \in C(Q)$, we have $\mathbb{E}_{\boldsymbol{z}}[g(z_j)] = \int_{[0,1]^d} g(t)\,dt = T_f(g)$. Using the additivity of the variance for

independent random variables, this entails

$$\begin{aligned}\mathbb{E}_z\Big[\big(\Psi(z,g)-T_\int(g)\big)^2\Big] &= \mathrm{Var}\Psi(z,g) = \big(1/m'\big)^2 \sum_{j=1}^{m'} \mathrm{Var}\big(g(z_j)\big) \\ &\le \big(1/m'\big)^2 \sum_{j=1}^{m'} \int_{[0,1]^d} |g(x)|^2\,dx = \frac{\|g\|_{L^2}^2}{m'}.\end{aligned} \tag{8.6}$$

Finally, for each $z \in \Omega$ define

$$A_z : \quad \mathcal{H} \to \mathbb{R}, \quad f \mapsto \Psi\big(z, f - B(M(f))\big) + T_\int\big(B(M(f))\big)$$

Since the map $T_\int : C([0,1]^d) \to \mathbb{R}$ is continuous and hence measurable, it is easy to verify that $\Omega \times U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty} \ni (z,f) \mapsto A_z(f)$ is measurable. Furthermore, explicitly writing out the definition of A_z shows that

$$A_z(f) = \frac{1}{m'}\sum_{j=1}^{m'} f(z_j) - \frac{1}{m'}\sum_{j=1}^{m'} B\big(f(x_1),\dots,f(x_{m'})\big)(z_j) + T_\int\big(B(f(x_1),\dots,f(x_{m'}))\big)$$

only depends on $m' + m' \le m$ point samples of f. Thus, if we set $\boldsymbol{m} \equiv m$, then $(\boldsymbol{A},\boldsymbol{m})$ is a strongly measurable randomized (Monte Carlo) algorithm $(\boldsymbol{A},\boldsymbol{m}) \in \mathrm{Alg}_m^{\mathrm{ran}}(U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty},\mathbb{R})$.

To complete the proof, note that a combination of Eqs. (8.5) and (8.6) shows

$$\begin{aligned}\mathbb{E}_z\Big[\big(A_z(f)-T_\int(f)\big)^2\Big] &= \mathbb{E}_z\Big[\big(\Psi(z,f-B(M(f)))-T_\int(f-B(M(f)))\big)^2\Big] \\ &\le \frac{1}{m'}\big\|f-B(M(f))\big\|_{L^2}^2 \le 4\,(C')^2\cdot m^{-1}\cdot\big(\ln^{1+\nu}(2m)/m\big)^{\frac{\alpha}{1+\alpha}}\end{aligned}$$

for all $f \in U$. Combined with Jensen's inequality, this proves Eq. (8.3) for the case $m \in \mathbb{N}_{\ge 2}$. The case $m = 1$ can be handled by taking $A_\omega \equiv 0$ and possibly enlarging the constant C in Eq. (8.3). Directly from the definition of $\beta_*^{\mathrm{ran}}(U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}, T_\int)$, we see that Eq. (8.3) implies Eq. (8.4). □

9 Hardness of Numerical Integration

Our goal in this section is to prove upper bounds for the optimal order $\beta_*(U_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}, T_\int)$ of quadrature on the neural network approximation spaces, both for deterministic and randomized algorithms. Our bounds for the deterministic setting in particular show that *regardless of the "approximation exponent"* α, the quadrature error given m point samples can never decay faster than $\mathcal{O}\big(m^{-\min\{2,2\alpha\}}\big)$. In fact, if the depth growth function $\boldsymbol{\ell}$ is unbounded, or if the weight growth function $\boldsymbol{c}$ grows sufficiently fast (so that $\gamma^\flat(\boldsymbol{\ell},\boldsymbol{c}) = \infty$), then no better rate than $\mathcal{O}\big(m^{-\min\{1,\alpha\}}\big)$ is possible.

For the case of randomized (Monte Carlo) algorithms, the bound that we derive shows that the expected quadrature error given at most m point samples (in expectation) can never decay faster than $\mathcal{O}\big(m^{-\min\{2,\frac{1}{2}+2\alpha\}}\big)$. In fact, if $\gamma^\flat = \infty$ then the error cannot decay faster than $\mathcal{O}\big(m^{-\min\{1,\frac{1}{2}+\alpha\}}\big)$.

Our precise bound for the deterministic setting reads as follows:

Theorem 9.1 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ be non-decreasing, and let $d \in \mathbb{N}$ and $\alpha > 0$. Let $\gamma^\flat := \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$ as in Eq. (2.2) and $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ as in Eq. (2.3). For the operator $T_\int : U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \to \mathbb{R}$, $f \mapsto \int_{[0,1]^d} f(x)\,dx$, we then have*

$$\beta^{\det}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \leq \begin{cases} \frac{2\alpha}{\alpha+\gamma^\flat}, & \text{if } \alpha+\gamma^\flat < 2, \\ \min\left\{\alpha, 1+\frac{\alpha-1}{\alpha+\gamma^\flat-1}\right\} & \text{if } \alpha+\gamma^\flat \geq 2 \end{cases} \tag{9.1}$$

$$= \begin{cases} \frac{2\alpha}{\alpha+\gamma^\flat}, & \text{if } \alpha+\gamma^\flat \leq 2 \\ \alpha, & \text{if } \alpha+\gamma^\flat > 2 \text{ and } \alpha \leq 1, \\ 1+\frac{\alpha-1}{\alpha+\gamma^\flat-1} & \text{if } \alpha+\gamma^\flat \geq 2 \text{ and } \alpha > 1. \end{cases} \tag{9.2}$$

Remark Since the bound above might seem intimidating at first sight, we discuss a few specific consequences. First, the theorem implies $\beta^{\det}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \leq \max\left\{\alpha, \frac{2\alpha}{\alpha+\gamma^\flat}\right\} \leq \max\{1, \frac{2}{\gamma^\flat}\}\alpha \leq 2\alpha$ and hence $\beta^{\det}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \to 0$ as $\alpha \downarrow 0$. Furthermore, the theorem shows that $\beta^{\det}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \leq 2$, and if $\gamma^\flat = \infty$, then in fact $\beta^{\det}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \leq \min\{\alpha, 1\}$.

Proof For brevity, set $U := U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$.

Step 1: Let $0 < \gamma < \gamma^\flat$, $\theta \in (0,\infty)$, and $\lambda \in [0,1]$ with $\theta\lambda \leq 1$ be arbitrary and define $\omega := \min\{-\theta\alpha, \theta\cdot(\gamma-\lambda)-1\}$. In this step, we show that

$$e(A, U, T_\int) \geq \kappa_2 \cdot m^{-(1-\omega-\theta\lambda)} \qquad \forall\, m \in \mathbb{N} \text{ and } A \in \mathrm{Alg}_m(U, \mathbb{R}), \tag{9.3}$$

for a suitable constant $\kappa_2 = \kappa_2(\alpha, \gamma, \theta, \lambda, \boldsymbol{\ell}, \boldsymbol{c}) > 0$.

To see this, let $m \in \mathbb{N}$ and $A \in \mathrm{Alg}_m(U, \mathbb{R})$ be arbitrary. By definition, this means that there exist $Q : \mathbb{R}^m \to \mathbb{R}$ and $\boldsymbol{x} = (x_1, \dots, x_m) \in ([0,1]^d)^m$ satisfying $A(f) = Q\big(f(x_1), \dots, f(x_m)\big)$ for all $f \in U$. Set $M := 4m$ and let $z_j := \frac{1}{4m} + \frac{j-1}{2m}$ for $j \in \underline{2m}$ as in Lemma 3.2. Furthermore, choose $I := I_{\boldsymbol{x}} := \big\{i \in \underline{2m} \;:\; \forall\, n \in \underline{m} : \Lambda^*_{M,z_i}(x_n) = 0\big\}$ and recall from Lemma 3.3 that $|I| \geq m$. Define $k := \lceil m^{\theta\lambda} \rceil$ and note $k \leq 1 + m^{\theta\lambda} \leq 2\, m^{\theta\lambda}$. Since $\theta\lambda \leq 1$, we also have $k \leq \lceil m \rceil = m \leq |I|$. Hence, there is a subset $J \subset I$ satisfying $|J| = k$.

Now, an application of Lemma 3.2 yields a constant $\kappa_1 = \kappa_1(\alpha, \gamma, \theta, \lambda, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ (independent of m and A) such that $f := \kappa_1\, m^\omega \sum_{j\in J} \Lambda^*_{M,z_j}$ satisfies $\pm f \in U$. Since $J \subset I$, we see by definition of $I = I_{\boldsymbol{x}}$ that $f(x_n) = 0$ for all $n \in \underline{m}$ and hence $A(\pm f) = Q(0, \dots, 0) =: \mu$. Using the elementary estimate $\max\{|x-\mu|, |-x-\mu|\} \geq$

$\frac{1}{2}\big(|x-\mu|+|x+\mu|\big) \geq \frac{1}{2}|x-\mu+x+\mu| = |x|$, we thus see

$$\begin{aligned} e(A, U, T_f) &\geq \max\Big\{\big|T_f(f) - Q\big(f(x_1), \dots, f(x_m)\big)\big|, \\ &\qquad \big|T_f(-f) - Q\big(-f(x_1), \dots, -f(x_m)\big)\big|\Big\} \\ &\geq \max\Big\{\big|T_f(f)-\mu\big|, \quad \big|-T_f(f)-\mu\big|\Big\} \\ &\geq |T_f(f)| = \kappa_1 \cdot m^{\omega} \cdot \frac{|J|}{M} \overset{(*)}{\geq} \frac{\kappa_1}{4} \cdot m^{\omega-1+\theta\lambda} =: \kappa_2 \cdot m^{-(1-\omega-\theta\lambda)}, \end{aligned}$$

as claimed in Eq. (9.3). Here, the step marked with $(*)$ used that $|J| = k \geq m^{\theta\lambda}$ and that $M = 4m$.

Step 2 (Completing the proof): Eq. (9.3) shows that $e_m^{\mathrm{det}}(U, T_f) \geq \kappa_2 \cdot m^{-(1-\omega-\theta\lambda)}$ for all $m \in \mathbb{N}$, with $\kappa_2 > 0$ independent of m. Directly from the definition of $\beta_*^{\mathrm{det}}(U, T_f)$ and ω, this shows

$$\begin{aligned} \beta_*^{\mathrm{det}}(U, T_f) &\leq 1 - \omega - \theta\lambda \\ &= 1 + \max\big\{\theta\cdot(\alpha-\lambda), \quad 1+\theta\cdot(\lambda-\gamma)-\theta\lambda\big\} \\ &= 1 + \max\big\{\theta\cdot(\alpha-\lambda), \quad 1-\theta\gamma\big\}, \end{aligned}$$

and this holds for arbitrary $0 < \gamma < \gamma^\flat, \theta \in (0,\infty)$, and $\lambda \in [0,1]$ satisfying $\theta\lambda \leq 1$. It is easy (but somewhat tedious) to shows that this implies Eq. (9.1); see Lemma A.6 for the details. Finally, Eq. (9.2) follows from Eq. (9.1) via an easy case distinction. □

As our next results, we derive a hardness results for randomized (Monte Carlo) algorithms for integration on the neural network approximation space $A_{\boldsymbol{\ell},\boldsymbol{c}}^{\alpha,\infty}$. The proof hinges on *Khintchine's inequality*, which states the following:

Proposition 9.2 *([12, Theorem 1 in Section 10.3]) Let $n \in \mathbb{N}$ and let $(X_i)_{i=1,\dots,n}$ be independent random variables (on some probability space $(\Omega, \mathcal{F}, \mathbb{P})$) that are Rademacher distributed (i.e., $\mathbb{P}(X_i = 1) = \frac{1}{2} = \mathbb{P}(X_i = -1)$ for each $i \in \underline{n}$). Then for each $p \in (0,\infty)$ there exist constants $A_p, B_p \in (0,\infty)$ (only depending on p) such that for arbitrary $c = (c_i)_{i=1,\dots,n} \subset \mathbb{R}$, the following holds:*

$$A_p \cdot \left(\sum_{i=1}^n c_i^2\right)^{1/2} \leq \left\| \sum_{i=1}^n c_i\, X_i \right\|_{L^p(\mathbb{P})} = \left(\mathbb{E}\left|\sum_{i=1}^n c_i\, X_i\right|^p\right)^{1/p} \leq B_p \cdot \left(\sum_{i=1}^n c_i^2\right)^{1/2}$$

Remark 9.3 Applying Khintchine's inequality for $p = 1$ and $c_i = 1$, we see

$$\sum_{\nu\in\{\pm 1\}^n} \left| \sum_{i=1}^n \nu_i \right| \geq A_1 \cdot n^{1/2}, \tag{9.4}$$

which is what we will actually use below.

Our precise hardness result for integration using randomized (Monte Carlo) algorithms reads as follows.

Theorem 9.4 *Let $\boldsymbol{\ell} : \mathbb{N} \to \mathbb{N}_{\geq 2} \cup \{\infty\}$ and $\boldsymbol{c} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$ be non-decreasing. Let $d \in \mathbb{N}$ and $\alpha \in (0, \infty)$. Let $\gamma^\flat := \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$ as in Eq. (2.2) and $U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d)$ as in Eq. (2.3). For the operator $T_\int : U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}([0,1]^d) \to \mathbb{R}$, $f \mapsto \int_{[0,1]^d} f(x)\,dx$, we then have*

$$\beta^{\mathrm{ran}}_*\big(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int\big) \leq \begin{cases} \min\left\{1 + \frac{\alpha}{\alpha+\gamma^\flat},\ \frac{1}{2} + \frac{2\alpha}{\alpha+\gamma^\flat}\right\}, & \text{if } \alpha + \gamma^\flat < 2, \\ \min\left\{1 + \frac{\alpha}{\alpha+\gamma^\flat},\ \frac{1}{2} + \alpha,\ 1 + \frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}\right\}, & \text{if } \alpha + \gamma^\flat \geq 2. \end{cases}$$

$$= \begin{cases} \frac{1}{2} + \frac{2\alpha}{\alpha+\gamma^\flat}, & \text{if } \alpha + \gamma^\flat < 2, \\ \frac{1}{2} + \alpha, & \text{if } \alpha + \gamma^\flat \geq 2 \text{ and } \alpha \leq \frac{1}{2}, \\ 1 + \frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}, & \text{if } \alpha + \gamma^\flat \geq 2 \text{ and } \frac{1}{2} \leq \alpha \leq \gamma^\flat, \\ 1 + \frac{\alpha}{\alpha+\gamma^\flat}, & \text{if } \alpha + \gamma^\flat \geq 2 \text{ and } \alpha \geq \gamma^\flat. \end{cases} \tag{9.5}$$

Remark We discuss a few special cases. First, we always have $\beta^{\mathrm{ran}}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \leq 1 + \frac{\alpha}{\alpha+\gamma^\flat} \leq 2$, which shows that *no matter how large the approximation rate α is*, one can never get an (asymptotically) better error bound than m^{-2}. Furthermore, if $\gamma^\flat = \infty$ (for instance if $\boldsymbol{\ell}$ is unbounded), then $\beta^{\mathrm{ran}}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}, T_\int) \leq 1 + \frac{\alpha}{\alpha+\gamma^\flat} = 1$.

The previous bounds are informative for (somewhat) large α. For small $\alpha > 0$, it is more useful to note that the theorem shows $\beta^{\mathrm{ran}}_*(U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}) \leq \frac{1}{2} + \max\left\{\frac{2\alpha}{\alpha+\gamma^\flat},\ \alpha\right\} \leq \frac{1}{2} + \max\left\{\frac{2}{\gamma^\flat},\ 1\right\}\alpha \leq \frac{1}{2} + 2\alpha$.

Proof For brevity, set $U := U^{\alpha,\infty}_{\boldsymbol{\ell},\boldsymbol{c}}$ and $\gamma^\flat := \gamma^\flat(\boldsymbol{\ell}, \boldsymbol{c})$.

The main idea of the proof is to apply Lemma 2.3 for a suitable choice of the family of functions $(f_{\boldsymbol{\nu},J})_{(\boldsymbol{\nu},J)\in\Gamma_m} \subset U$.

Step 1 (Preparation): Let $0 < \gamma < \gamma^\flat, \theta \in (0, \infty)$, and $\lambda \in [0, 1]$ with $\theta\lambda \leq 1$ be arbitrary and define $\omega := \min\{-\theta\alpha, \theta\cdot(\gamma-\lambda)-1\}$. Given a fixed but arbitrary $m \in \mathbb{N}$, set $M := 4m$ and $z_j := \frac{1}{4m} + \frac{j-1}{2m}$ as in Lemma 3.2. Furthermore, let $k := \lceil m^{\theta\lambda} \rceil$ and note because of $\theta\lambda \leq 1$ that $k \leq \lceil m \rceil = m$ and $k \leq 1 + m^{\theta\lambda} \leq 2\,m^{\theta\lambda}$.

Define $\mathcal{P}_k(\underline{2m}) := \{J \subset \underline{2m} \colon |J| = k\}$ and $\Gamma_m := \{\pm 1\}^{2m} \times \mathcal{P}_k(\underline{2m})$. Then, Lemma 3.2 yields a constant $\kappa_1 = \kappa_1(\gamma, \theta, \lambda, \alpha, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ such that for any $(\boldsymbol{\nu}, J) \in \Gamma_m$, the function

$$f_{\boldsymbol{\nu},J} := \kappa_1\, m^\omega \sum_{j\in J} \nu_j\, \Lambda^*_{M,z_j} \quad \text{satisfies} \quad f_{\boldsymbol{\nu},J} \in U.$$

Step 2: We show for $\gamma, \theta, \lambda, \omega$ as in Step 2 that there exists $\kappa_3 = \kappa_3(\gamma, \theta, \lambda, \alpha, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ (independent of $m \in \mathbb{N}$) such that

$$\sum_{(\boldsymbol{\nu},J)\in\Gamma_m} \left|T_\int(f_{\boldsymbol{\nu},J}) - A(f_{\boldsymbol{\nu},J})\right| \geq \kappa_3 \cdot m^{-(1-\frac{\theta\lambda}{2}-\omega)} \qquad \forall\, m \in \mathbb{N} \text{ and } A \in \mathrm{Alg}_m(U, \mathbb{R}). \tag{9.6}$$

To see this, let $A \in \mathrm{Alg}_m(U, \mathbb{R})$ be arbitrary. By definition, we have $A(f) = Q\big(f(x_1), \dots, f(x_m)\big)$ for all $f \in U$, for suitable $\boldsymbol{x} = (x_1, \dots, x_m) \in ([0,1]^d)^m$ and $Q : \mathbb{R}^m \to \mathbb{R}$. Now, define $I := I_{\boldsymbol{x}} := \{j \in \underline{2m} : \forall\, n \in \underline{m} : \Lambda^*_{M,z_j}(x_n) = 0\}$ and recall from Lemma 3.3 that $|I| \geq m$.

Set $I^c := \underline{2m} \setminus I$. For $\boldsymbol{\nu}^{(1)} = (\nu_j)_{j\in I} \in \{\pm 1\}^I$ and $\boldsymbol{\nu}^{(2)} := (\nu_j)_{j\in I^c} \in \{\pm 1\}^{I^c}$ and $J \in \mathcal{P}_k(\underline{2m})$, define

$$g_{\boldsymbol{\nu}^{(1)},J} := \kappa_1\, m^\omega \sum_{j\in J\cap I} \nu_j^{(1)}\, \Lambda^*_{M,z_j} \qquad \text{and} \qquad h_{\boldsymbol{\nu}^{(2)},J} := \kappa_1\, m^\omega \sum_{j\in J\cap I^c} \nu_j^{(2)}\, \Lambda^*_{M,z_j}.$$

Furthermore, define $\mu_{\boldsymbol{\nu}^{(2)},J} := T_\int(h_{\boldsymbol{\nu}^{(2)},J}) - Q\big(h_{\boldsymbol{\nu}^{(2)},J}(x_1), \dots, h_{\boldsymbol{\nu}^{(2)},J}(x_m)\big)$. By choice of I, we have $g_{\boldsymbol{\nu}^{(1)},J}(x_n) = 0$ for all $n \in \underline{m}$, and hence $f_{\boldsymbol{\nu},J}(x_n) = h_{\boldsymbol{\nu}^{(2)},J}(x_n)$, if we identify $\boldsymbol{\nu}$ with $(\boldsymbol{\nu}^{(1)}, \boldsymbol{\nu}^{(2)})$, as we will do for the remainder of this step.

Finally, recall from Lemma 3.2 that $\operatorname{supp} \Lambda^*_{M,z_j} \subset [0,1]^d$ and hence $T_\int(\Lambda^*_{M,z_j}) = M^{-1} = \frac{1}{4m}$. Overall, we thus see for arbitrary $J \in \mathcal{P}_k(\underline{2m})$ and $\boldsymbol{\nu}^{(2)} \in \{\pm 1\}^{I^c}$ that

$$\begin{aligned}
&\sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \Big| T_\int(f_{\boldsymbol{\nu},J}) - Q\big(f_{\boldsymbol{\nu},J}(x_1), \dots, f_{\boldsymbol{\nu},J}(x_m)\big)\Big| \\
&= \sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \Big| T_\int(g_{\boldsymbol{\nu}^{(1)},J}) + \mu_{\boldsymbol{\nu}^{(2)},J}\Big| = \sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \Big| \frac{\kappa_1\, m^\omega}{M} \sum_{j\in J\cap I} \nu_j^{(1)} + \mu_{\boldsymbol{\nu}^{(2)},J}\Big| \\
&\overset{(*)}{=} \frac{1}{2} \sum_{\boldsymbol{\nu}^{(3)}\in\{\pm1\}^{I\cap J}} \left(\Big| \frac{\kappa_1\, m^\omega}{M} \sum_{j\in J\cap I} \nu_j^{(3)} + \mu_{\boldsymbol{\nu}^{(2)},J}\Big| + \Big| \frac{\kappa_1\, m^\omega}{M} \sum_{j\in J\cap I} (-\nu_j^{(3)}) + \mu_{\boldsymbol{\nu}^{(2)},J}\Big| \right) \\
&\overset{(\blacklozenge)}{\geq} \sum_{\boldsymbol{\nu}^{(3)}\in\{\pm1\}^{I\cap J}} \Big| \frac{\kappa_1\, m^\omega}{M} \sum_{j\in J\cap I} \nu_j^{(3)}\Big| \geq \kappa_2\, m^{\omega-1} \cdot |J\cap I|^{1/2}
\end{aligned} \tag{9.7}$$

for a suitable constant $\kappa_2 = \kappa_2(\gamma, \theta, \lambda, \alpha, \boldsymbol{\ell}, \boldsymbol{c}) > 0$. Here, the very last step used Eq. (9.4) and the identity $M = 4m$. Furthermore, the step marked with $(*)$ used that

$$\sum_{\boldsymbol{\sigma}\in\{\pm1\}^K} a_{\boldsymbol{\sigma}} = \frac{1}{2}\Big(\sum_{\boldsymbol{\sigma}\in\{\pm1\}^K} a_{\boldsymbol{\sigma}} + \sum_{\boldsymbol{\sigma}\in\{\pm1\}^K} a_{\boldsymbol{\sigma}}\Big) = \frac{1}{2}\Big(\sum_{\boldsymbol{\sigma}\in\{\pm1\}^K} a_{\boldsymbol{\sigma}} + \sum_{\boldsymbol{\sigma}\in\{\pm1\}^K} a_{-\boldsymbol{\sigma}}\Big),$$

while the elementary estimate $|x+y|+|-x+y| = |x+y|+|x-y| \geq |x+y+x-y| = 2|x|$ was used at the step marked with $(\blacklozenge)$.

Combining Eq. (9.7) and Lemma A.4, we finally obtain $\kappa_3 = \kappa_3(\gamma, \theta, \lambda, \alpha, \boldsymbol{\ell}, \boldsymbol{c}) > 0$ satisfying

$$\begin{aligned}
\sum_{(\boldsymbol{\nu},J)\in\Gamma_m} \big|T_\int(f_{\boldsymbol{\nu},J}) - A(f_{\boldsymbol{\nu},J})\big| &= \sum_{J\in\mathcal{P}_k(m)} \sum_{\boldsymbol{\nu}^{(2)}\in\{\pm1\}^{I^c}} \sum_{\boldsymbol{\nu}^{(1)}\in\{\pm1\}^I} \big|T_\int(f_{\boldsymbol{\nu},J}) - A(f_{\boldsymbol{\nu},J})\big| \\
&\geq \kappa_2\, m^{\omega-1} \sum_{J\in\mathcal{P}_k(m)} |J\cap I|^{1/2} \geq \kappa_3\, m^{\omega-1} \cdot k^{1/2} \geq \kappa_3\, m^{\omega-1+\frac{\theta\lambda}{2}},
\end{aligned}$$

as claimed in Eq. (9.6). Since $m \in \mathbb{N}$ and $A \in \mathrm{Alg}_m(U;\mathbb{R})$ were arbitrary and κ_3 is independent of A and m, Step 2 is complete.

Step 3: In view of Eq. (9.6), a direct application of Lemma 2.3 shows that

$$\beta_*^{\mathrm{ran}}(U, T_\int) \leq 1 - \omega - \tfrac{\theta\lambda}{2} = 1 + \max\left\{\theta \cdot (\alpha - \tfrac{\lambda}{2}),\ 1 + \theta \cdot (\tfrac{\lambda}{2} - \gamma)\right\}$$

for arbitrary $0 < \gamma < \gamma^{\mathrm{b}}$, $\theta \in (0,\infty)$, and $\lambda \in [0,1]$ with $\theta\lambda \leq 1$. From this, the first part of Eq. (9.5) follows by a straightforward but technical computation; see Lemma A.5 for the details. The second part of Eq. (9.5) follows from the first one by a straightforward case distinction. □

Funding Open access funding provided by University of Vienna.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

A Postponed Technical Results and Proofs

A.1 Proof of Lemma 2.1

This section provides the proof of Lemma 2.1, which is based on the following lemma concerning closure properties of the sets $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$.

Lemma A.1 *With $\widetilde{\boldsymbol{\ell}}(n) := \min\{\boldsymbol{\ell}(n), n\}$, we have $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} = \Sigma_n^{\widetilde{\boldsymbol{\ell}},\boldsymbol{c}}$. Furthermore, for every $n \in \mathbb{N}$, we have $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} + \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} \subset \Sigma_{9n}^{\boldsymbol{\ell},\boldsymbol{c}}$.*

Proof We first prove $\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} = \Sigma_n^{\widetilde{\boldsymbol{\ell}},\boldsymbol{c}}$. To this end, we prove for fixed $n \in \mathbb{N}$ by induction on $\ell \in \mathbb{N}_{\geq n}$ that $\Sigma_n^{\ell,\boldsymbol{c}} \subset \Sigma_n^{n,\boldsymbol{c}}$. For $\ell = n$, this is trivial. Thus, suppose that $\Sigma_n^{\ell,\boldsymbol{c}} \subset \Sigma_n^{n,\boldsymbol{c}}$ for some $\ell \in \mathbb{N}_{\geq n}$ and let $f \in \Sigma_n^{\ell+1,\boldsymbol{c}}$, say $f = R_\varrho \Phi$ with $\|\Phi\|_{\mathcal{NN}} \leq \boldsymbol{c}(n)$ and $W(\Phi) \leq n$, as well as $L(\Phi) \leq \ell + 1$. If $L(\Phi) \leq \ell$, then $f \in \Sigma_n^{\ell,\boldsymbol{c}} \subset \Sigma_n^{n,\boldsymbol{c}}$ by induction. Hence, we can assume that $L(\Phi) = \ell + 1$.

Writing $\Phi = \big((A_1, b_1), \ldots, (A_{\ell+1}, b_{\ell+1})\big)$ with $b_m \in \mathbb{R}^{N_m}$ and $A_m \in \mathbb{R}^{N_m \times N_{m-1}}$, we have $A_j = b_j = 0$ for some $j \in \underline{\ell+1}$, since otherwise $n+1 \leq \ell+1 \leq \sum_{j=1}^{\ell+1} \big(\|A_j\|_{\ell^0} + \|b_j\|_{\ell^0}\big) = W(\Phi) \leq n$. If $j = \ell+1$, we trivially have $f \equiv 0 \in \Sigma_n^{n,\boldsymbol{c}}$; thus, let us assume $j \leq \ell$ and define

$$\widetilde{\Phi} := \big((0_{N_{j+1} \times d}, b_{j+1}), (A_{j+2}, b_{j+2}), \ldots, (A_{\ell+1}, b_{\ell+1})\big).$$

Since $A_j = b_j = 0$ and $\varrho(0) = 0$, it is straightforward to verify $R_\varrho \widetilde{\Phi} = R_\varrho \Phi = f$. Since furthermore $\|\widetilde{\Phi}\|_{\mathcal{NN}} \leq \|\Phi\|_{\mathcal{NN}} \leq \boldsymbol{c}(n)$ and $W(\widetilde{\Phi}) \leq W(\Phi) \leq n$, as well as

$L(\widetilde{\Phi}) \leq \ell - j + 1 \leq \ell$, this implies $f \in \Sigma_n^{\ell,\boldsymbol{c}} \subset \Sigma_n^{n,\boldsymbol{c}}$, where the last inclusion holds by induction. This completes the induction.

To prove $\Sigma_n^{\ell,\boldsymbol{c}} + \Sigma_n^{\ell,\boldsymbol{c}} \subset \Sigma_{5n}^{\ell,\boldsymbol{c}}$, let $f, g \in \Sigma_n^{\ell,\boldsymbol{c}}$, so that $f = R_\varrho \Phi$ and $g = R_\varrho \Psi$ for networks Φ, Ψ satisfying $W(\Phi), W(\Psi) \leq n$ and $\|\Phi\|_{\mathcal{NN}}, \|\Psi\|_{\mathcal{NN}} \leq \boldsymbol{c}(n)$, as well as $L(\Phi), L(\Psi) \leq \min\{n, \boldsymbol{\ell}(n)\}$; here we used the first part of the lemma. By possibly swapping Φ, Ψ and f, g, we can assume that $k := L(\Phi) \leq L(\Psi) =: \ell$. If $k = \ell$, define $\widetilde{\Phi} := \Phi$. If otherwise $k < \ell$, write $\Phi = \big((A_1, b_1), \dots, (A_k, b_k)\big)$ where $A_k \in \mathbb{R}^{1 \times N_{k-1}}$ and $b_k \in \mathbb{R}^1$, and define $\Gamma := \left(\begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right)$ and $\Lambda := \big((1, -1), 0\big)$ and finally

$$\widetilde{\Phi} := \left((A_1, b_1), \dots, (A_{k-1}, b_{k-1}), \left(\begin{pmatrix} A_k \\ -A_k \end{pmatrix}, \begin{pmatrix} b_k \\ -b_k \end{pmatrix} \right), \Gamma, \dots, \Gamma, \Lambda \right),$$

where Γ appears $\ell - k - 1$ times, so that $L(\widetilde{\Phi}) = \ell$. Using the identities $x = \varrho(x) - \varrho(-x)$ and $\varrho(\varrho(x)) = \varrho(x)$, it is easy to see $R_\varrho \widetilde{\Phi} = R_\varrho \Phi = f$. Moreover, $\|\widetilde{\Phi}\|_{\mathcal{NN}} \leq \max\{1, \boldsymbol{c}(n)\} = \boldsymbol{c}(n)$ and $W(\widetilde{\Phi}) \leq 2W(\Phi) + 2(\ell - k) \leq 4n$.

Finally, explicitly writing $\widetilde{\Phi} = \big((B_1, c_1), \dots, (B_\ell, c_\ell)\big)$ and $\Psi = \big((C_1, e_1), \dots, (C_\ell, e_\ell)\big)$ with $c_\ell, e_\ell \in \mathbb{R}^1$ and $B_\ell, C_\ell \in \mathbb{R}^{1 \times N_{\ell-1}}$, define

$$\Theta_1 := \left(\begin{pmatrix} B_1 \\ C_1 \\ 0_{4\times 1} \end{pmatrix}, \begin{pmatrix} c_1 \\ e_1 \\ c_\ell \\ -c_\ell \\ e_\ell \\ -e_\ell \end{pmatrix} \right) \quad \text{and}$$

$$\Theta_m := \left(\begin{pmatrix} B_m & 0 & 0 \\ 0 & C_m & 0 \\ 0 & 0 & I_{4\times 4} \end{pmatrix}, \begin{pmatrix} c_m \\ e_m \\ 0_{4\times 1} \end{pmatrix} \right) \quad \text{for } m \in \{2, \dots, \ell - 1\},$$

and set

$$\Xi := \Big(\Theta_1, \dots, \Theta_{\ell-1}, \big((B_\ell \mid C_\ell \mid 1 \mid -1 \mid 1 \mid -1), 0 \big) \Big).$$

Using the identities $\varrho(\varrho(x)) = \varrho(x)$ and $x = \varrho(x) - \varrho(-x)$, it is then straightforward to verify $R_\varrho \Xi = R_\varrho \widetilde{\Phi} + R_\varrho \Psi = f + g$. Moreover, $\|\Xi\|_{\mathcal{NN}} \leq \boldsymbol{c}(n) \leq \boldsymbol{c}(9n)$, $L(\Xi) = \ell \leq \boldsymbol{\ell}(n) \leq \boldsymbol{\ell}(9n)$, and $W(\Xi) \leq W(\widetilde{\Phi}) + W(\Psi) + 4\,\ell \leq 9n$. Here, we used that $\boldsymbol{\ell}$ and $\boldsymbol{c}$ are non-decreasing and that $\ell \leq n$. Overall, we have shown $f + g \in \Sigma_{9n}^{\ell,\boldsymbol{c}}$, as claimed. □

With Lemma A.1 at our disposal, we can now prove Lemma 2.1.

Proof of Lemma 2.1 **Step 1** *(Showing $\Gamma_{\alpha,p}(f+g) \leq C \cdot (\Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g))$)*: To see this, let $n \in \mathbb{N}_{\geq 9}$ and write $n = 9m + k$ with $m \in \mathbb{N}$ and $k \in \{0, \dots, 8\}$, noting that $n \leq 17m$. By Lemma A.1, we have $\Sigma_n^{\ell,\boldsymbol{c}} \supset \Sigma_{9m}^{\ell,\boldsymbol{c}} \supset \Sigma_m^{\ell,\boldsymbol{c}} + \Sigma_m^{\ell,\boldsymbol{c}}$ and hence

$$\begin{aligned} n^\alpha\, d_p(f+g, \Sigma_n^{\ell,\boldsymbol{c}}) &\leq n^\alpha\, d_p(f+g, \Sigma_m^{\ell,\boldsymbol{c}} + \Sigma_m^{\ell,\boldsymbol{c}}) \\ &\leq 17^\alpha m^\alpha \cdot \big(d_p(f, \Sigma_m^{\ell,\boldsymbol{c}}) + d_p(g, \Sigma_m^{\ell,\boldsymbol{c}}) \big) \\ &\leq 17^\alpha \cdot \big(\Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g) \big). \end{aligned}$$

Moreover, if $n \leq 8$, then we see because of $0 \in \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ that

$$n^\alpha \, d_p(f+g, \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}) \leq 8^\alpha \|f+g\|_{L^p} \leq 8^\alpha \cdot (\|f\|_{L^p} + \|g\|_{L^p}) \leq 8^\alpha \cdot (\Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g)).$$

Overall, we thus see for every $n \in \mathbb{N}$ that $n^\alpha \, d_p(f+g, \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}) \leq C \cdot (\Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g))$. Since also $\|f+g\|_{L^p} \leq \|f\|_{L^p} + \|g\|_{L^p} \leq \Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g) \leq C \cdot (\Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g))$, we see that $\Gamma_{\alpha,p}(f+g) \leq C \cdot (\Gamma_{\alpha,p}(f) + \Gamma_{\alpha,p}(g))$, as claimed in this step.

Step 2 (*Showing* $\Gamma_{\alpha,p}(cf) \leq |c| \, \Gamma_{\alpha,p}(f)$ *for* $|c| \leq 1$): Since $|c| \leq 1$, it is straightforward to see $c\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}} \subset \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}$ and hence $n^\alpha \, d_p(cf, \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}) \leq n^\alpha \, d_p(cf, c\Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}}) = |c| n^\alpha \, d_p(f, \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}})$. This implies $\Gamma_{\alpha,p}(cf) \leq \max\left\{\|cf\|_{L^p}, |c| \sup_{n\in\mathbb{N}} \left[n^\alpha \, d_p(f, \Sigma_n^{\boldsymbol{\ell},\boldsymbol{c}})\right]\right\} = |c| \, \Gamma_{\alpha,p}(f)$.

Step 3 (*Showing* $\Gamma_{\alpha,p}(f) < \infty \iff \|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} < \infty$):

"$\Rightarrow$:" For $\theta := 1 + \Gamma_{\alpha,p}(f) \in [1,\infty)$, Step 2 shows $\Gamma_{\alpha,p}(f/\theta) \leq \frac{1}{\theta}\Gamma_{\alpha,p}(f) \leq 1$, and hence $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} < \infty$.

"$\Leftarrow$:" Let $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} < \infty$. Hence, there exists $\theta > 0$ satisfying $\Gamma_{\alpha,p}(f/\theta) \leq 1 < \infty$. Step 1 shows $\Gamma_{\alpha,p}(2g) = \Gamma_{\alpha,p}(g+g) \leq 2C\Gamma_{\alpha,p}(g)$. Inductively, this implies $\Gamma_{\alpha,p}(2^m g) \leq (2C)^m \Gamma_{\alpha,p}(g)$ for every $m \in \mathbb{N}$. Now, choosing $m \in \mathbb{N}$ such that $\theta \leq 2^m$, Step 2 shows

$$\Gamma_{\alpha,p}(f) = \Gamma_{\alpha,p}\left(\tfrac{\theta}{2^m} 2^m \tfrac{f}{\theta}\right) \leq \Gamma_{\alpha,p}(2^m \tfrac{f}{\theta}) \leq (2C)^m \Gamma_{\alpha,p}(\tfrac{f}{\theta}) < \infty.$$

Step 4 (*Homogeneity of* $\|\cdot\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}$): It is easy to see $\|0\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} = 0$. Moreover, given $c \in \mathbb{R} \setminus \{0\}$, Step 2 shows that $\Gamma_{\alpha,p}(\pm f) = \Gamma_{\alpha,p}(f)$. Therefore,

$$\begin{aligned} \|cf\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} &= \inf\{\theta > 0 \colon \Gamma_{\alpha,p}(cf/\theta) \leq 1\} \\ &= |c| \cdot \inf\left\{\tfrac{\theta}{|c|} : \theta > 0 \text{ and } \Gamma_{\alpha,p}\left(\tfrac{f}{\theta/|c|}\right) \leq 1\right\} \\ &= |c| \, \|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}. \end{aligned}$$

Step 5 (*Definiteness of* $\|\cdot\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}$): If $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} = 0$, then for each $n \in \mathbb{N}$ there exists $\theta_n \in (0, \frac{1}{n})$ satisfying $\Gamma_{\alpha,p}(f/\theta_n) \leq 1$. By Step 2, this implies

$$\|f\|_{L^p} \leq \Gamma_{\alpha,p}(f) = \Gamma_{\alpha,p}\left(\theta_n \tfrac{f}{\theta_n}\right) \leq \theta_n \Gamma_{\alpha,p}\left(\tfrac{f}{\theta_n}\right) \leq \theta_n \xrightarrow[n\to\infty]{} 0,$$

and hence $f = 0$.

Step 6 (*If* $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}} \in (0,\infty)$, *then* $\Gamma_{\alpha,p}(f/\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}) \leq 1$): By definition of $\|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}$, there exists a sequence $(\theta_n)_{n\in\mathbb{N}} \subset (0,\infty)$ satisfying $\theta_n \to \theta := \|f\|_{A^{\alpha,p}_{\boldsymbol{\ell},\boldsymbol{c}}}$ and $\Gamma_{\alpha,p}(f/\theta_n) \leq 1$ for all $n \in \mathbb{N}$. Since $\frac{f}{\theta_n} \to \frac{f}{\theta}$ and since $d_p(\cdot, \Sigma_m^{\boldsymbol{\ell},\boldsymbol{c}})$ is continuous with respect to $\|\cdot\|_{L^p}$, this implies for each $m \in \mathbb{N}$ that

$$\max\left\{\left\|\tfrac{f}{\theta}\right\|_{L^p}, \; m^\alpha \, d_p\left(\tfrac{f}{\theta}, \Sigma_m^{\boldsymbol{\ell},\boldsymbol{c}}\right)\right\} = \lim_{n\to\infty} \max\left\{\left\|\tfrac{f}{\theta_n}\right\|_{L^p}, \; m^\alpha \, d_p\left(\tfrac{f}{\theta_n}, \Sigma_m^{\boldsymbol{\ell},\boldsymbol{c}}\right)\right\} \leq 1,$$

and hence $\Gamma_{\alpha,p}(f/\theta) \le 1$.

Step 7 *(Showing* $\|f+g\|_{A^{\alpha,p}_{\ell,c}} \le C\cdot(\|f\|_{A^{\alpha,p}_{\ell,c}} + \|g\|_{A^{\alpha,p}_{\ell,c}})$*)*: The claim is trivial if $\|f\|_{A^{\alpha,p}_{\ell,c}} \in \{0,\infty\}$ or $\|g\|_{A^{\alpha,p}_{\ell,c}} \in \{0,\infty\}$. Hence, we can assume that $A := \|f\|_{A^{\alpha,p}_{\ell,c}} \in (0,\infty)$ and $B := \|g\|_{A^{\alpha,p}_{\ell,c}} \in (0,\infty)$. By Steps 1, 2, and 6, this implies

$$\begin{aligned}\Gamma_{\alpha,p}\big(\tfrac{f+g}{C(A+B)}\big) &\le \tfrac{1}{C}\Gamma_{\alpha,p}\big(\tfrac{f}{A+B} + \tfrac{g}{A+B}\big)\\ &\le \Gamma_{\alpha,p}\big(\tfrac{A}{A+B}\,\tfrac{f}{A}\big) + \Gamma_{\alpha,p}\big(\tfrac{B}{A+B}\,\tfrac{g}{B}\big)\\ &\le \tfrac{A}{A+B}\Gamma_{\alpha,p}\big(\tfrac{f}{A}\big) + \tfrac{B}{A+B}\Gamma_{\alpha,p}\big(\tfrac{g}{B}\big) \le 1,\end{aligned}$$

and hence $\|f+g\|_{A^{\alpha,p}_{\ell,c}} \le C\cdot(A+B) = C\cdot(\|f\|_{A^{\alpha,p}_{\ell,c}} + \|g\|_{A^{\alpha,p}_{\ell,c}})$, as claimed.

Step 8 *(Showing* $\Gamma_{\alpha,p}(f) \le 1 \iff \|f\|_{A^{\alpha,p}_{\ell,c}} \le 1$*)*: "$\Rightarrow$" follows by definition of $\|\cdot\|_{A^{\alpha,p}_{\ell,c}}$.

"$\Leftarrow$" is trivial if $f=0$. Otherwise, Steps 6 and 2 show for $\theta := \|f\|_{A^{\alpha,p}_{\ell,c}} \in (0,1]$ that $\Gamma_{\alpha,p}(f) = \Gamma_{\alpha,p}(\theta\,\frac{f}{\theta}) \le \Gamma_{\alpha,p}(f/\theta) \le 1$.

Step 9: In this step, we prove the last part of Lemma 2.1. First, note that if $\|f\|_{A^{\alpha,p}_{\ell,c}(\Omega)} \le 1$, then $\|f\|_{L^p} \le \Gamma_{\alpha,p}(f) \le 1$ thanks to Step 8. This proves $A^{\alpha,p}_{\ell,c}(\Omega) \hookrightarrow L^p(\Omega)$.

Next, if $\Omega \subset \overline{\Omega^\circ}$, then it is easy to see for $f \in C_b(\Omega)$ that $\|f\|_{\sup,\Omega} := \sup_{x\in\Omega}|f(x)| = \|f\|_{L^\infty(\Omega)}$, and this implies that $C_b(\Omega) \subset L^\infty(\Omega)$ is closed. Therefore, it suffices to show $A^{\alpha,\infty}_{\ell,c}(\Omega) \subset \overline{C_b(\Omega)}$. To see this, let $f \in A^{\alpha,\infty}_{\ell,c}(\Omega)$; by Step 3, this implies $\theta := \Gamma_{\alpha,\infty}(f) < \infty$. Furthermore, $\|f\|_{L^\infty} < \infty$. By definition of $\Gamma_{\alpha,\infty}$, for each $n \in \mathbb{N}$ there exists $F_n \in \Sigma^{\ell,c}_n$ satisfying $\|F_n - f\|_{L^\infty} \le 2Cn^{-\alpha} \to 0$ as $n\to\infty$; in particular, $\|F_n\|_{\sup,\Omega} = \|F_n\|_{L^\infty} < \infty$. Finally, since F_n can be extended to a continuous function on all of $\mathbb{R}^d$, we see $F_n \in C_b(\Omega)$ and hence $f \in \overline{C_b(\Omega)} = C_b(\Omega)$. □

A.2 A Technical Result used in Sect. 3

Lemma A.2 *For each* $d\in\mathbb{N}$, $T\in(0,1]$, *and* $x\in[0,1]^d$, *we have*

$$\lambda\big([0,1]^d \cap (x+[-T,T]^d)\big) \ge 2^{-d}\,T^d.$$

Proof For brevity, set $Q := [0,1]^d$. Below, we show

$$\lambda\big(Q\cap(x+[-T,T]^d)\big) \ge T^d \qquad \forall x\in Q \text{ and } T\in(0,\tfrac12], \tag{A.1}$$

which clearly implies the claim for these T. Furthermore, for $T\in[\frac12,1]$, the above estimate shows $\lambda\big(Q\cap(x+[-T,T]^d)\big) \ge \lambda\big(Q\cap(x+[-\frac12,\frac12]^d)\big) \ge 2^{-d} \ge 2^{-d}T^d$, which proves the claim for general $T\in(0,1]$.

Thus, let $x\in Q$ and $T\in(0,\frac12]$. For each $j\in\underline{d}$, define $\varepsilon_j := -1$ if $x_j \ge \frac12$ and $\varepsilon_j := 1$ otherwise. Let $P := \prod_{j=1}^d \big(\varepsilon_j\,[0,T]\big) \subset [-T,T]^d$. We claim that $x+P\subset Q$.

Once this is shown, it follows that $\boldsymbol{\lambda}\big(Q \cap (x + [-T, T]^d)\big) \geq \boldsymbol{\lambda}(x + P) = T^d$, proving Eq. (A.1).

To see that indeed $x + P \subset Q$, let $y \in P$ be arbitrary. For each $j \in \underline{d}$, there are then two cases:

1. If $x_j \geq \frac{1}{2}$, then $\varepsilon_j = -1$ and $-\frac{1}{2} \leq -T \leq y_j \leq 0$. Thus, $0 \leq x_j - \frac{1}{2} \leq x_j + y_j \leq x_j \leq 1$, meaning $(x + y)_j \in [0, 1]$.
2. If $x_j < \frac{1}{2}$, then $\varepsilon_j = 1$ and $0 \leq y_j \leq T \leq \frac{1}{2}$. Thus, $0 \leq x_j \leq x_j + y_j \leq \frac{1}{2} + \frac{1}{2} = 1$, so that we see again $(x + y)_j \in [0, 1]$.

Overall, this shows in both cases that $x + y \in [0, 1]^d = Q$. □

A.3 A Technical Result Regarding Measurability

Lemma A.3 *Let $\varnothing \neq \Omega \subset \mathbb{R}^d$ be compact and let $\varnothing \neq \mathcal{H} \subset C(\Omega)$ be compact. Then, equipping $\mathcal{H}$ with the Borel σ-algebra induced from $C(\Omega)$, the following hold:*

1. *The map*

$$M : \Omega^m \times \mathcal{H} \to \Omega^m \times \mathbb{R}^m,$$
$$(\boldsymbol{x}, f) = \big((x_1, \ldots, x_m), f\big) \mapsto \Big(\boldsymbol{x}, \big(f(x_1), \ldots, f(x_m)\big)\Big)$$

   is continuous and hence measurable;
2. *there is a measurable map $B : \Omega^m \times \mathbb{R}^m \to \mathcal{H}$ satisfying*

$$B(\boldsymbol{x}, \boldsymbol{y}) \in \operatorname*{argmin}_{g \in \mathcal{H}} \sum_{i=1}^{m} \big(g(x_i) - y_i\big)^2$$
$$\forall\, \boldsymbol{x} = (x_1, \ldots, x_m) \in \Omega^m \text{ and } \boldsymbol{y} = (y_1, \ldots, y_m) \in \mathbb{R}^m.$$

Proof **Part 1:** It is enough to prove continuity of each of the components of M. For the component $(\boldsymbol{x}, f) \mapsto \boldsymbol{x}$ this is trivial. For the component $(\boldsymbol{x}, f) \mapsto f(x_j)$ note that if $\Omega^m \ni \boldsymbol{x}^{(n)} \to \boldsymbol{x} \in \Omega^m$ and $\mathcal{H} \ni f_n \to f \in \mathcal{H}$ (with convergence in $C(\Omega)$), then

$$\begin{aligned} \big|f(\boldsymbol{x}_j) - f_n\big(\boldsymbol{x}_j^{(n)}\big)\big| &\leq \big|f(\boldsymbol{x}_j) - f\big(\boldsymbol{x}_j^{(n)}\big)\big| + \big|f\big(\boldsymbol{x}_j^{(n)}\big) - f_n\big(\boldsymbol{x}_j^{(n)}\big)\big| \\ &\leq \big|f(\boldsymbol{x}_j) - f\big(\boldsymbol{x}_j^{(n)}\big)\big| + \|f - f_n\|_{C(\Omega)} \xrightarrow[n\to\infty]{} 0, \end{aligned} \tag{A.2}$$

since f is continuous. Thus, M is continuous. To see that this implies that M is measurable, note that both Ω^m and $\mathcal{H}$ are separable metric spaces (and hence second countable), so that the product σ-algebra on $\Omega^m \times \mathcal{H}$ coincides with the Borel σ-algebra on $\Omega^m \times \mathcal{H}$; see for instance [19, Theorem 7.20].

Part 2: For this part, we use the "Measurable Maximum Theorem," [2, Theorem 18.19]. Thanks to this theorem, setting $S := \Omega^m \times \mathbb{R}^m$, it is enough to show that

1. the set-valued map [2] $\varphi : S \twoheadrightarrow C(\Omega)$, $(\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathcal{H}$ is weakly measurable with nonempty, compact values;
2. the map $F : S \times C(\Omega) \to \mathbb{R}$, $\big((\boldsymbol{x}, \boldsymbol{y}), g\big) \mapsto -\sum_{i=1}^{m} \big(g(x_i) - y_i\big)^2$ is a Carathéodory function (see [2, Definition 4.50]).

By our assumptions on $\mathcal{H}$, it is clear that φ has nonempty, compact values. The weak measurability of φ follows directly from the definition, see [2, Definition 18.1]. For the second property, it is enough to show that F is continuous. This follows as in Eq. (A.2). □

A.4 A Technical Result Regarding Random Subsets of {1, . . . , m}

Lemma A.4 *Let $m \in \mathbb{N}$ and $1 \leq k \leq 2m$. Write $\mathcal{P}_k(\underline{2m}) := \{J \subset \underline{2m} : |J| = k\}$. Then, for each subset $I \subset \underline{2m}$ with $|I| \geq m$, we have*

$$\sum_{J \in \mathcal{P}_k(\underline{2m})} |J \cap I|^{1/2} \geq \frac{1}{4} \cdot k^{1/2}.$$

Proof Let $I^c := \underline{2m} \setminus I$. We note for any $T \subset \underline{2m}$ that the quantity $\psi(T) := \sum_{J \in \mathcal{P}_k(\underline{2m})} |J \cap T|^{1/2}$ only depends on the cardinality $|T|$ and that $\psi(T) \leq \psi(S)$ if $|T| \leq |S|$. Since $|I| \geq m \geq |I^c|$, this implies $\psi(I) \geq \psi(I^c)$. Combined with the estimate

$$\begin{aligned} |J \cap I|^{\frac{1}{2}} + |J \cap I^c|^{\frac{1}{2}} &\geq \Big[\max\big\{|J \cap I|, |J \cap I^c|\big\}\Big]^{\frac{1}{2}} \geq \Big[\tfrac{1}{2}\big(|J \cap I| + |J \cap I^c|\big)\Big]^{\frac{1}{2}} \\ &\geq \big(\tfrac{1}{2}|J|\big)^{\frac{1}{2}} \geq \tfrac{1}{2}|J|^{\frac{1}{2}} = \tfrac{1}{2}k^{\frac{1}{2}} \end{aligned}$$

which holds for all $J \in \mathcal{P}_k(\underline{2m})$, we finally see

$$\begin{aligned} \sum_{J \in \mathcal{P}_k(\underline{2m})} |J \cap I|^{1/2} = \psi(I) &\geq \frac{\psi(I) + \psi(I^c)}{2} \\ &= \frac{1}{2} \sum_{J \in \mathcal{P}_k(\underline{2m})} \big(|J \cap I|^{\frac{1}{2}} + |J \cap I^c|^{\frac{1}{2}}\big) \geq \frac{1}{4}k^{\frac{1}{2}}. \end{aligned}$$

□

A.5 Two Technical Optimization Results

Lemma A.5 *Let $\gamma^\flat \in [1, \infty]$ and $\alpha > 0$. Let*

$$\Psi := \big\{(\gamma, \theta, \lambda) \in (0, \infty) \times (0, \infty) \times [0, 1] : \gamma < \gamma^\flat \text{ and } \theta\lambda \leq 1\big\}. \tag{A.3}$$

[2] A set-valued map $f : X \twoheadrightarrow Y$ is a map $f : X \to 2^Y$, into the power set 2^Y of Y.

Then

$$\inf_{(\gamma,\theta,\lambda)\in\Psi} \max\left\{\theta\cdot(\alpha-\tfrac{\lambda}{2}),\ \ 1+\theta\cdot(\tfrac{\lambda}{2}-\gamma)\right\} \leq \begin{cases} \min\left\{\frac{\alpha}{\alpha+\gamma^\flat},\ \frac{2\alpha}{\alpha+\gamma^\flat}-\frac{1}{2}\right\}, & \text{if } \alpha+\gamma^\flat<2,\\ \min\left\{\frac{\alpha}{\alpha+\gamma^\flat},\ \alpha-\frac{1}{2},\ \frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}\right\}, & \text{if } \alpha+\gamma^\flat\geq 2. \end{cases}$$

Remark In fact, one has equality. But since we do not need this, we omit the proof (and the explicit statement) of this fact.

Proof **Step 1 (Preparations):** Define $f_1(\gamma,\theta,\lambda) := \theta\cdot(\alpha-\frac{\lambda}{2})$ and $f_2(\gamma,\theta,\lambda) := 1+\theta\cdot(\frac{\lambda}{2}-\gamma)$ as well as $f := \max\{f_1, f_2\}$ and $\beta_* := \inf_{(\gamma,\theta,\lambda)\in\Psi} f(\gamma,\theta,\lambda)$. For arbitrary $0<\gamma<\gamma^\flat$, we have $(\gamma,\frac{1}{\alpha+\gamma},0)\in\Psi$ and hence $\beta_*\leq f(\gamma,\frac{1}{\alpha+\gamma},0) = \max\left\{\frac{\alpha}{\alpha+\gamma}, 1-\frac{\gamma}{\alpha+\gamma}\right\} = \frac{\alpha}{\alpha+\gamma}$. Letting $\gamma\uparrow\gamma^\flat$, this implies

$$\beta_* \leq \frac{\alpha}{\alpha+\gamma^\flat}. \tag{A.4}$$

Step 2 (The case $\gamma^\flat=\infty$): Let us first consider the case $\gamma^\flat=\infty$. In this case, Eq. (A.4) shows $\beta_*\leq 0$. Furthermore, given $0<\gamma<\gamma^\flat=\infty$, we have $(\gamma,1,1)\in\Psi$, which shows that $\beta_*\leq f(\gamma,1,1) = \max\left\{\alpha-\frac{1}{2},\frac{3}{2}-\gamma\right\}$. Letting $\gamma\to\infty$, we thus see $\beta_*\leq\alpha-\frac{1}{2}$ and hence $\beta_*\leq\min\{0,\alpha-\frac{1}{2}\}$. It is easy to see that this implies the claim for $\gamma^\flat=\infty$.

Hence, we can assume from now on that $\gamma^\flat$ is finite. Then, we easily see for $g_1(\theta,\lambda) := \theta\cdot(\alpha-\frac{\lambda}{2})$ and $g_2(\theta,\lambda) := 1+\theta\cdot(\frac{\lambda}{2}-\gamma^\flat)$ as well as $g := \max\{g_1,g_2\}$ and $\Omega := \{(\theta,\lambda)\in(0,\infty)\times[0,1]\colon \theta\lambda\leq 1\}$ that $\beta_*\leq\inf_{(\theta,\lambda)\in\Omega} g(\theta,\lambda)$.
Step 3 (The case $\alpha+\gamma^\flat<2$): In this case, we have $\frac{2}{\alpha+\gamma^\flat}\in(1,\infty)$ and hence $(\frac{2}{\alpha+\gamma^\flat},\frac{\alpha+\gamma^\flat}{2})\in\Omega$. Furthermore, $g_1(\frac{2}{\alpha+\gamma^\flat},\frac{\alpha+\gamma^\flat}{2}) = g_2(\frac{2}{\alpha+\gamma^\flat},\frac{\alpha+\gamma^\flat}{2}) = \frac{2\alpha}{\alpha+\gamma^\flat}-\frac{1}{2}$ and hence $\beta_*\leq\frac{2\alpha}{\alpha+\gamma^\flat}-\frac{1}{2}$. Together with Eq. (A.4), this proves the claim for $\alpha+\gamma^\flat<2$.
Step 4 (The case $\alpha+\gamma^\flat\geq 2$): Note $g_1(1,1)=\alpha-\frac{1}{2}$ and $g_2(1,1)=\frac{3}{2}-\gamma^\flat\leq\frac{3}{2}-(2-\alpha)=\alpha-\frac{1}{2}$. Since $(1,1)\in\Omega$, this implies $\beta_*\leq g(1,1)=\alpha-\frac{1}{2}$. Furthermore, $\theta_0 := \frac{1}{\alpha+\gamma^\flat-1}\in(0,1]$ and hence $(\theta_0,1)\in\Omega$. It is easy to see $g_1(\theta_0,1)=g_2(\theta_0,1)=\frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}$ and hence $\beta_*\leq g(\theta_0,1)=\frac{\alpha-\frac{1}{2}}{\alpha+\gamma^\flat-1}$. Combining these two estimates with Eq. (A.4) completes the proof for the case $\alpha+\gamma^\flat\geq 2$. □

Lemma A.6 *Let $\gamma^\flat\in[1,\infty]$ and $\alpha>0$. Let Ψ be as in Eq. (A.3). Then*

$$\inf_{(\gamma,\theta,\lambda)\in\Psi} \max\left\{\theta\cdot(\alpha-\lambda),\quad 1-\theta\gamma\right\} \leq \begin{cases} \frac{2\alpha}{\alpha+\gamma^\flat}-1, & \text{if } \alpha+\gamma^\flat\leq 2,\\ \min\left\{\alpha-1,\frac{\alpha-1}{\alpha+\gamma^\flat-1}\right\}, & \text{if } \alpha+\gamma^\flat>2. \end{cases} \tag{A.5}$$

Proof For brevity, denote the left-hand side of Eq. (A.5) by β_*.

We first consider the special case $\gamma^\flat = \infty$. Define $g := \max\{g_1, g_2\}$, where $g_1(\gamma, \theta, \lambda) := \theta \cdot (\alpha - \lambda)$ and $g_2(\gamma, \theta, \lambda) := 1 - \theta\gamma$. For any $\gamma > 0$, we have $g_1(\gamma, 1, 1) = \alpha - 1$ and $g_2(\gamma, 1, 1) = 1 - \gamma$ and furthermore $(\gamma, 1, 1) \in \Psi$. Therefore, $\beta_* \leq g(\gamma, 1, 1) = \max\{\alpha - 1,\ 1 - \gamma\} \xrightarrow[\gamma\to\infty]{} \alpha - 1$. Furthermore, for arbitrary $\gamma > 0$ we have $(\gamma, \frac{1}{\gamma}, 0) \in \Psi$ and $g_1(\gamma, \frac{1}{\gamma}, 0) = \frac{\alpha}{\gamma}$ and $g_2(\gamma, \frac{1}{\gamma}, 0) = 0$, so that $\beta_* \leq \min\{0, \frac{\alpha}{\gamma}\} \xrightarrow[\gamma\to\infty]{} 0$. Overall, we have thus shown $\beta_* \leq \min\{\alpha - 1, 0\}$, which easily implies that Eq. (A.5) holds in case of $\gamma^\flat = \infty$.

Hence, we can assume that $\gamma^\flat < \infty$. Then, setting $\Omega := \{(\theta, \lambda) \in (0, \infty) \times [0, 1] \colon \theta\lambda \leq 1\}$ and furthermore $f := \max\{f_1, f_2\}$ for $f_1(\theta, \lambda) := \theta(\alpha - \lambda)$ and $f_2(\theta, \lambda) := 1 - \theta\gamma^\flat$, it is easy to see by continuity that $\beta_* \leq \inf_{(\theta,\lambda)\in\Omega} f(\theta, \lambda)$. We now distinguish two cases:

Case 1 ($\alpha + \gamma^\flat \leq 2$): In this case, $\theta_0 := \frac{2}{\alpha+\gamma^\flat} \in [1, \infty)$ and $\lambda_0 := \frac{1}{\theta_0} \in (0, 1]$ satisfy $(\theta_0, \lambda_0) \in \Omega$. Furthermore, it is easy to see $f_1(\theta_0, \lambda_0) = \frac{2\alpha}{\alpha+\gamma^\flat} - 1 = f_2(\theta_0, \lambda_0)$. Thus, $\beta_* \leq f(\theta_0, \lambda_0) = \frac{2\alpha}{\alpha+\gamma^\flat} - 1$, which proves Eq. (A.5) in this case.

Case 2 ($\alpha + \gamma^\flat > 2$): First note because of $\alpha + \gamma^\flat > 2$ that $f_1(1, 1) = \alpha - 1 > 1 - \gamma^\flat = f_2(1, 1)$ and hence $\beta_* \leq f(1, 1) = \alpha - 1$. Furthermore, we have $\theta^* := \frac{1}{\alpha+\gamma^\flat-1} \in (0, 1)$ and hence $(\theta^*, 1) \in \Omega$. Furthermore, it is easy to see $f_1(\theta^*, 1) = \frac{\alpha-1}{\alpha+\gamma^\flat-1} = f_2(\theta^*, 1)$ which implies $\beta_* \leq f(\theta^*, 1) = \frac{\alpha-1}{\alpha+\gamma^\flat-1}$. Overall, we see $\beta_* \leq \min\left\{\alpha - 1, \frac{\alpha-1}{\alpha+\gamma^\flat-1}\right\}$, which shows that Eq. (A.5) holds for $\alpha + \gamma^\flat > 2$. □

References

1. B. Adcock and N. Dexter. The gap between theory and practice in function approximation with deep neural networks. SIAM Journal on Mathematics of Data Science, 3(2):624–655, 2021.
2. C. D. Aliprantis and K. C. Border. Infinite dimensional analysis. Springer, Berlin, third edition, 2006.
3. V. Antun, M. J. Colbrook, and A. C. Hansen. The difficulty of computing stable and accurate neural networks: On the barriers of deep learning and Smale's 18th problem. Applied Mathematics, 119(12):e21071511, 2022.
4. S. Arridge, P. Maass, O. Öktem, and C.-B. Schönlieb. Solving inverse problems using data-driven models. Acta Numerica, 28:1–174, 2019.
5. P. Baldi, P. Sadowski, and D. Whiteson. Searching for exotic particles in high-energy physics with deep learning. Nature communications, 5(1):1–9, 2014.
6. P. L. Bartlett, N. Harvey, C. Liaw, and A. Mehrabian. Nearly-tight VC-dimension and Pseudodimension Bounds for Piecewise Linear Neural Networks. Journal of Machine Learning Research, 20(63):1–17, 2019.
7. P. Beneventano, P. Cheridito, A. Jentzen, and P. von Wurstemberger. High-dimensional approximation spaces of artificial neural networks and applications to partial differential equations. arXiv preprint 2012.04326, 2020.
8. J. Berner, P. Grohs, and A. Jentzen. Analysis of the Generalization Error: Empirical Risk Minimization over Deep Artificial Neural Networks Overcomes the Curse of Dimensionality in the Numerical Approximation of Black–Scholes Partial Differential Equations. SIAM Journal on Mathematics of Data Science, 2(3):631–657, 2020.
9. A. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. In Advances in neural information processing systems, pages 494–501, 1989.
10. H. Bölcskei, P. Grohs, G. Kutyniok, and P. C. Petersen. Optimal approximation with sparsely connected deep neural networks. SIAM J. Math. Data Sci., 1:8–45, 2019.

11. A. Caragea, P. Petersen, and F. Voigtlaender. Neural network approximation and estimation of classifiers with classification boundary in a Barron class. arXiv preprint 2011.09363, 2020.
12. Y. S. Chow and H. Teicher. Probability theory. Springer Texts in Statistics. Springer-Verlag, New York, third edition, 1997.
13. F. Cucker and S. Smale. On the mathematical foundations of learning. Bull. Amer. Math. Soc. (N.S.), 39(1):1–49, 2002.
14. R. A. DeVore and G. G. Lorentz. Constructive approximation, volume 303 of Grundlehren der Mathematischen Wissenschaften. Springer-Verlag, Berlin, 1993.
15. R. DeVore, B. Hanin, and G. Petrova. Neural network approximation. Acta Numerica, 30:327–444, 2021.
16. Z. Ditzian and V. Totik. Moduli of smoothness, volume 9. Springer Science & Business Media, 2012.
17. W. E and B. Yu. The deep ritz method: a deep learning-based numerical algorithm for solving variational problems. Communications in Mathematics and Statistics, 6(1):1–12, 2018.
18. F. A. Faber, L. Hutchison, B. Huang, J. Gilmer, S. S. Schoenholz, G. E. Dahl, O. Vinyals, S. Kearnes, P. F. Riley, and O. A. Von Lilienfeld. Prediction errors of molecular machine learning models lower than hybrid DFT error. Journal of chemical theory and computation, 13(11):5255–5264, 2017.
19. G. B. Folland. Real analysis. Pure and Applied Mathematics (New York). John Wiley & Sons, Inc., New York, second edition, 1999.
20. R. Gribonval, G. Kutyniok, M. Nielsen, and F. Voigtlaender. Approximation spaces of deep neural networks. Constructive Approximation, 55:259–367, 2022.
21. P. Grohs, F. Hornung, A. Jentzen, and P. Von Wurstemberger. A proof that artificial neural networks overcome the curse of dimensionality in the numerical approximation of Black-Scholes partial differential equations. Memoirs of the American Mathematical Society, 2020.
22. P. Grohs, D. Perekrestenko, D. Elbrächter, and H. Bölcskei. Deep neural network approximation theory. IEEE Transactions on Information Theory, 67(5):2581–2623, 2021.
23. A. Gupta and S. M. Lam. Weight decay backpropagation for noisy data. Neural Networks, 11(6):1127–1138, 1998.
24. K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.
25. S. Heinrich. Random approximation in numerical analysis. In Functional analysis (Essen, 1991), volume 150 of Lecture Notes in Pure and Appl. Math., pages 123–171. Dekker, New York, 1994.
26. J. Hermann, Z. Schätzle, and F. Noé. Deep-neural-network solution of the electronic Schrödinger equation. Nature Chemistry, 12(10):891–897, 2020.
27. M. Hutzenthaler, A. Jentzen, T. Kruse, and T. A. Nguyen. A proof that rectified deep neural networks overcome the curse of dimensionality in the numerical approximation of semilinear heat equations. SN Partial Differential Equations and Applications, 1(2):1–34, 2020.
28. D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. arXiv preprint 1412.6980, 2014.
29. A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 25, pages 1097–1105. Curran Associates, Inc., 2012.
30. G. Kutyniok, P. Petersen, M. Raslan, and R. Schneider. A theoretical analysis of deep neural networks and parametric PDEs. arXiv preprint 1904.00377, 2019.
31. G. Lample and F. Charton. Deep learning for symbolic mathematics. In International Conference on Learning Representations, 2019.
32. Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.
33. J. Ma, R. P. Sheridan, A. Liaw, G. E. Dahl, and V. Svetnik. Deep neural nets as a method for quantitative structure–activity relationships. Journal of chemical information and modeling, 55(2):263–274, 2015.
34. V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing Atari with deep reinforcement learning. arXiv preprint 1312.5602, 2013.
35. M. Mohri, A. Rostamizadeh, and A. Talwalkar. Foundations of machine learning. MIT Press, Cambridge, MA, 2018.
36. P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. Neural Networks, 108:296–330, 2018.
37. D. Pfau, J. S. Spencer, A. G. Matthews, and W. M. C. Foulkes. Ab initio solution of the many-electron Schrödinger equation with deep neural networks. Physical Review Research, 2(3):033429, 2020.
38. A. Pietsch. Eigenvalues and s-numbers. Cambridge University Press, 1986.

39. A. Pinkus. N-widths in Approximation Theory, volume 7. Springer Science & Business Media, 2012.
40. M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. Journal of Computational Physics, 378:686–707, 2019.
41. M. M. Rao and Z. D. Ren. Theory of Orlicz spaces, volume 146 of Monographs and Textbooks in Pure and Applied Mathematics. Marcel Dekker, Inc., New York, 1991.
42. D. Saxton, E. Grefenstette, F. Hill, and P. Kohli. Analysing mathematical reasoning abilities of neural models. In International Conference on Learning Representations, 2018.
43. A. W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. W. Nelson, and A. Bridgland. Improved protein structure prediction using potentials from deep learning. Nature, 577(7792):706–710, 2020.
44. S. Shalev-Shwartz and S. Ben-David. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.
45. D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, and M. Lanctot. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587):484–489, 2016.
46. D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, and A. Bolton. Mastering the game of Go without human knowledge. Nature, 550(7676):354–359, 2017.
47. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1–9, 2015.
48. M. Telgarsky. Benefits of depth in neural networks. In Conference on learning theory, pages 1517–1539. PMLR, 2016.
49. A. F. Timan. Theory of approximation of functions of a real variable. Elsevier, 2014.
50. R. Vershynin. High-dimensional probability, volume 47 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2018.
51. O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, and P. Georgiev. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature, 575(7782):350–354, 2019.
52. D. Yarotsky. Optimal approximation of continuous functions by very deep ReLU networks. In Conference on Learning Theory, pages 639–649. PMLR, 2018.
53. T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. IEEE Computational intelligence magazine, 13(3):55–75, 2018.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.