Approximation spaces of deep neural networks

Rémi Gribonval, Gitta Kutyniok, Morten Nielsen, Felix Voigtlaender

APPROXIMATION SPACES OF DEEP NEURAL NETWORKS

RÉMI GRIBONVAL, GITTA KUTYNIOK, MORTEN NIELSEN, AND FELIX VOIGTLAENDER

Abstract. We study the expressivity of deep neural networks. Measuring a network's complexity by its number of connections or by its number of neurons, we consider the class of functions for which the error of best approximation with networks of a given complexity decays at a certain rate when increasing the complexity budget. Using results from classical approximation theory, we show that this class can be endowed with a (quasi)-norm that makes it a linear function space, called approximation space. We establish that allowing the networks to have certain types of "skip connections" does not change the resulting approximation spaces. We also discuss the role of the network's nonlinearity (also known as activation function) on the resulting spaces, as well as the role of depth. For the popular ReLU nonlinearity and its powers, we relate the newly constructed spaces to classical Besov spaces. The established embeddings highlight that some functions of very low Besov smoothness can nevertheless be well approximated by neural networks, if these networks are sufficiently deep.

1. Introduction

Today, we witness a worldwide triumphant march of deep neural networks, impacting not only various application fields, but also areas in mathematics such as inverse problems. Originally, neural networks were developed by McCulloch and Pitts [48] in 1943 to introduce a theoretical framework for artificial intelligence. At that time, however, the limited amount of data and the lack of sufficient computational power only allowed the training of shallow networks, that is, networks with only few layers of neurons, which did not lead to the anticipated results. The current age of big data and the significantly increased computer performance now make the application of deep learning algorithms feasible, leading to the successful training of very deep neural networks. For this reason, neural networks have seen an impressive comeback. The list of important applications in public life ranges from speech recognition systems on cell phones over self-driving cars to automatic diagnoses in healthcare. For applications in science, one can witness a similarly strong impact of deep learning methods in research areas such as quantum chemistry [61] and molecular dynamics [47], often allowing to resolve problems which were deemed unreachable before. This phenomenon is manifested similarly in certain fields of mathematics, foremost in inverse problems [2, 10], but lately also, for instance, in numerical analysis of partial differential equations [8].

Yet, most of the existing research related to deep learning is empirically driven and a profound and comprehensive mathematical foundation is still missing, in particular for the previously mentioned applications. This poses a significant challenge not only for mathematics itself, but in general for the "safe" applicability of deep neural networks [22].

A *deep neural network* in mathematical terms is a tuple

$$\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big) \tag{1.1}$$

consisting of affine-linear maps $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ (hence $T_\ell(x) = A_\ell\, x + b_\ell$ for appropriate matrices A_ℓ and vectors b_ℓ, often with a convolutional or Toeplitz structure) and of nonlinearities $\alpha_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$ that typically encompass componentwise rectification, possibly followed by a pooling operation.

The tuple in (1.1) encodes the architectural components of the neural network, where L denotes the *number of layers* of the network, while $L-1$ is the number of *hidden layers*. The highly structured function $\mathtt{R}(\Phi)$ implemented by such a network Φ is then defined by applying the different maps in an iterative (layer-wise) manner; precisely,

$$\mathtt{R}(\Phi) : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}, \quad \text{with} \quad \mathtt{R}(\Phi) := \alpha_L \circ T_L \circ \cdots \circ \alpha_1 \circ T_1\,.$$

We call this function the *realization* of the deep neural network Φ. It is worth pointing out that most of the literature calls this function itself the neural network; one can however—depending on the choice of the

2010 *Mathematics Subject Classification.* Primary 82C32, 41A65. Secondary 68T05, 41A46, 42C40.

Key words and phrases. Deep neural networks; sparsely connected networks; Approximation spaces; Besov spaces; direct estimates; inverse estimates; piecewise polynomials; ReLU activation function;

G.K. acknowledges partial support by the Bundesministerium fur Bildung und Forschung (BMBF) through the Berliner Zentrum for Machine Learning (BZML), Project AP4, RTG DAEDALUS (RTG 2433), Projects P1 and P3, RTG BIOQIC (RTG 2260), Projects P4 and P9, and by the Berlin Mathematics Research Center MATH+, Projects EF1-1 and EF1-4.

G.K. and F.V. acknowledge support by the European Commission-Project DEDALE (contract no. 665044) within the H2020 Framework.

activation functions—imagine the same function being realized by different architectural components, so that it would not make sense, for instance, to speak of the number of layers of $\mathtt{R}(\Phi)$; this is only well-defined when we talk about Φ itself. The *complexity* of a neural network can be captured by various numbers such as the *depth* L, the *number of hidden neurons* $N(\Phi) = \sum_{\ell=1}^{L-1} N_\ell$, or the *number of connections* (also called the *connectivity*, or the *number of weights*) given by $W(\Phi) = \sum_{\ell=1}^{L} \|A_\ell\|_{\ell^0}$, where $\|A_\ell\|_{\ell^0}$ denotes the number of non-zero entries of the matrix A_ℓ.

From a mathematical perspective, the central task of a deep neural network is to approximate a function $f : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}$, which for instance encodes a classification problem. Given a training data set $\big(x_i, f(x_i)\big)_{i=1}^m$ a loss function $\mathcal{L} : \mathbb{R}^{N_L} \times \mathbb{R}^{N_L} \to \mathbb{R}$, and a regularizer $\mathcal{P}$, which imposes, for instance, sparsity conditions on the weights of the neural network Φ, solving the optimization problem

$$\min_{\Phi} \sum_{i=1}^{m} \mathcal{L}\big(\mathtt{R}(\Phi)(x_i), f(x_i)\big) + \lambda \mathcal{P}(\Phi) \tag{1.2}$$

typically through a variant of stochastic gradient descent, yields a learned neural network $\widehat{\Phi}$. The objective is to achieve $\mathtt{R}(\widehat{\Phi}) \approx f$, which is only possible if the function f can indeed be well approximated by (the realization of) a network with the prescribed architecture. Various theoretical results have already been published to establish the ability of neural networks—often with specific architectural constraints—to approximate functions from certain function classes; this is referred to as analyzing the *expressivity* of neural networks. However, the fundamental question asking which function spaces are truly natural for deep neural networks has never been comprehensively addressed. Such an approach may open the door to a novel viewpoint and lead to a refined understanding of the expressive power of deep neural networks.

In this paper we introduce approximation spaces associated to neural networks. This leads to an extensive theoretical framework for studying the expressivity of deep neural networks, allowing us also to address questions such as the impact of the depth and of the activation function, or of so-called (and widely used) skip connections on the approximation power of deep neural networks.

1.1. **Expressivity of Deep Neural Networks.** The first theoretical results concerning the expressivity of neural networks date back to the early 90s, at that time focusing on shallow networks, mainly in the context of the *universal approximation theorem* [43, 36, 16, 35]. The breakthrough-result of the ImageNet competition in 2012 [38], and the ensuing worldwide success story of neural networks has brought renewed interest to the study of neural networks, now with an emphasis on *deep* networks. The surprising effectiveness of such networks in applications has motivated the study of the effect of depth on the expressivity of these networks. Questions related to the learning phase are of a different nature, focusing on aspects of statistical learning and optimization, and hence constitute a different research field.

Let us recall some of the key contributions in the area of expressivity, in order to put our results into perspective. The universal approximation theorems by Hornik [35] and Cybenko [16] can be counted as a first highlight, stating that neural networks with only one hidden layer can approximate continuous functions on compact sets arbitrarily well. Examples of further work in this early stage, hence focusing on networks with a single hidden layer, are approximation error bounds in terms of the number of neurons for functions with bounded first Fourier moments [5, 6], the failure of those networks to provide localized approximations [13], a fundamental lower bound on approximation rates [18, 12], and the approximation of smooth/analytic functions [50, 52]. Some of the early contributions already study networks with multiple hidden layers, such as [29] for approximating continuous functions, and [53] for approximating functions together with their derivatives. Also [13], which shows in certain instances that deep networks can perform better than single-hidden-layer networks can be counted towards this line of research. For a survey of those early results, we refer to [24, 57].

More recent work focuses predominantly on the analysis of the effect of depth. Some examples—again without any claim of completeness—are [23], in which a function is constructed which cannot be expressed by a small two-layer network, but which is implemented by a three-layer network of low complexity, or [51] which considers so-called compositional functions, showing that such functions can be approximated by neural networks without suffering from the curse of dimensionality. A still different viewpoint is taken in [14, 15], which focus on a similar problem as [51] but attacking it by utilizing results on tensor decompositions. Another line of research aims to study the approximation rate when approximating certain function classes by neural networks with growing complexity [62, 9, 55, 68, 49].

1.2. **The classical notion of approximation spaces.** In classical approximation theory, the notion of approximation spaces refers to (quasi)-normed spaces that are defined by their elements satisfying a specific decay of a certain approximation error; see for instance [21] In this introduction, we will merely sketch the key construction and properties; we refer to Section 3 for more details.

Let X be a quasi-Banach space equipped with the quasi-norm $\|\cdot\|_X$. Furthermore, here, as in the rest of the paper, let us denote by $\mathbb{N} = \{1, 2, \ldots\}$ the set of natural numbers, and write $\mathbb{N}_0 = \{0\} \cup \mathbb{N}$, $\mathbb{N}_{\geq m} = \{n \in \mathbb{N}, n \geq m\}$. For a prescribed family $\Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}$ of subsets $\Sigma_n \subset X$, one aims to classify functions $f \in X$ by the decay (as $n \to \infty$) of the error of best approximation by elements from Σ_n, given by $E(f, \Sigma_n)_X := \inf_{g\in\Sigma_n} \|f - g\|_X$. The desired rate of decay of this error is prescribed by a discrete weighted ℓ^q-norm, where the weight depends on the parameter $\alpha > 0$. For $q = \infty$, this leads to the class

$$A^\alpha_\infty(X, \Sigma) := \left\{ f \in X \,:\, \sup_{n\geq 1} [n^\alpha \cdot E(f, \Sigma_{n-1})_X] < \infty \right\}.$$

Thus, intuitively speaking, this class consists of those elements of X for which the error of best approximation by elements of Σ_n decays at least as $\mathcal{O}(n^{-\alpha})$ for $n \to \infty$. This general philosophy also holds for the more general classes $A^\alpha_q(X, \Sigma)$, $q > 0$.

If the initial family Σ of subsets of X satisfies some quite natural conditions, more precisely $\Sigma_0 = \{0\}$, each Σ_n is invariant to scaling, $\Sigma_n \subset \Sigma_{n+1}$, and the union $\bigcup_{n\in\mathbb{N}_0} \Sigma_n$ is dense in X, as well as the slightly more involved condition that $\Sigma_n + \Sigma_n \subset \Sigma_{cn}$ for some fixed $c \in \mathbb{N}$, then an abundance of results are available for the approximation classes $A^\alpha_q(X, \Sigma)$. In particular, $A^\alpha_q(X, \Sigma)$ turns out to be a proper *linear function space*, equipped with a natural (quasi)-norm. Particular highlights of the theory are various embedding and interpolation results between the different approximation spaces.

1.3. **Our Contribution.** We introduce a novel perspective on the study of expressivity of deep neural networks by introducing the associated approximation spaces and investigating their properties. This is in contrast with the usual approach of studying the approximation fidelity of neural networks on *classical* spaces. We utilize this new viewpoint for deriving novel results on, for instance, the impact of the choice of activation functions and the depth of the networks.

Given a so-called (non-linear) activation function $\varrho : \mathbb{R} \to \mathbb{R}$, a classical setting is to consider nonlinearities α_ℓ in (1.1) corresponding to a componentwise application of the activation function for each hidden layer $1 \leq \ell < L$, and α_L being the identity. We refer to networks of this form as *strict ϱ-networks*. To introduce a framework of sufficient flexibility, we also consider nonlinearities where for each component either ϱ or the identity is applied. We refer to such networks as *generalized ϱ-networks*; the realizations of such generalized networks include various function classes such as multilayer sparse linear transforms [41], networks with skip-connections [54], ResNets [32, 67] or U-nets [58].

Let us now explain how we utilize this framework of approximation spaces. Our focus will be on approximation rates in terms of growing complexity of neural networks, which we primarily measure by their *connectivity*, since this connectivity is closely linked to the number of bytes needed to describe the network, and also to the number of floating point operations needed to apply the corresponding function to a given input. This is in line with recent results [9, 55, 68] which explicitly construct neural networks that reach an optimal approximation rate for very specific function classes, and in contrast to most of the existing literature focusing on complexity measured by the number of neurons. We also consider the approximation spaces for which the complexity of the networks is measured by the number of neurons.

In addition to letting the number of connections or neurons tend to infinity while keeping the depth of the networks fixed, we also allow the depth to evolve with the number of connections or neurons. To achieve this, we link both by a non-decreasing *depth-growth function* $\mathscr{L} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}$, where we allow the possibility of not restricting the number of layers when $\mathscr{L}(n) = \infty$. We then consider the function families $\mathtt{W}_n(\Omega \to \mathbb{R}^k, \varrho, \mathscr{L})$ (resp. $\mathtt{N}_n(\Omega \to \mathbb{R}^k, \varrho, \mathscr{L})$) made of all restrictions to a given subset $\Omega \subseteq \mathbb{R}^d$ of functions which can be represented by (generalized) ϱ-networks with input/output dimensions d and k, at most n nonzero connection weights (resp. at most n hidden neurons), and at most $\mathscr{L}(n)$ layers. Finally, given a space X of functions $\Omega \to \mathbb{R}^k$, we will use the sets $\Sigma_n = \mathtt{W}_n(X, \varrho, \mathscr{L}) := \mathtt{W}_n(\Omega \to \mathbb{R}^k, \varrho, \mathscr{L}) \cap X$ (resp. $\Sigma_n = \mathtt{N}_n(X, \varrho, \mathscr{L}) := \mathtt{N}_n(\Omega \to \mathbb{R}^k, \varrho, \mathscr{L}) \cap X$) to define the associated approximation spaces. Typical choices for X are

$$X^k_p(\Omega) := L_p(\Omega; \mathbb{R}^k) \text{ for } 0 < p < \infty \quad \text{or} \quad X^k_\infty(\Omega), \tag{1.3}$$

with $X^k_\infty(\Omega)$ the space of uniformly continuous functions on Ω that vanish at infinity, equipped with the supremum norm. For ease of notation, we will sometimes also write $X_p(\Omega) := X^1_p(\Omega)$, and $X^k_p := X^k_p(\Omega)$ (resp. $X_p := X_p(\Omega)$).

Let us now give a coarse overview of our main results, which we are able to derive with our choice of approximation spaces based on $\mathtt{W}_n(X, \varrho, \mathscr{L})$ or $\mathtt{N}_n(X, \varrho, \mathscr{L})$.

1.3.1. *Core properties of the novel approximation spaces.* We first prove that each of these two families $\Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}$ satisfies the necessary requirements for the associated approximation spaces $A^\alpha_q(X, \Sigma)$—which we denote by $W^\alpha_q(X, \varrho, \mathscr{L})$ and $N^\alpha_q(X, \varrho, \mathscr{L})$, respectively—to be amenable to various results

from approximation theory. Under certain conditions on ϱ and $\mathscr{L}$, Theorem 3.27 shows that these approximation spaces are even equipped with a convenient (quasi-)Banach *spaces* structure. The spaces $W_q^\alpha(X,\varrho,\mathscr{L})$ and $N_q^\alpha(X,\varrho,\mathscr{L})$ are nested (Lemma 3.9) and do not generally coincide (Lemma 3.10).

To prepare the ground for the analysis of the impact of depth, we then prove nestedness with respect to the depth growth function. In slightly more detail, we identify a partial order $\preceq$ and an equivalence relation $\sim$ on depth growth functions such that the following holds (Lem. 3.12 and Thm. 3.13):

(1) If $\mathscr{L}_1 \preceq \mathscr{L}_2$, then $W_q^\alpha(X,\varrho,\mathscr{L}_1) \subset W_q^\alpha(X,\varrho,\mathscr{L}_2)$ for any α, q, X and ϱ; and
(2) if $\mathscr{L}_1 \sim \mathscr{L}_2$, then $W_q^\alpha(X,\varrho,\mathscr{L}_1) = W_q^\alpha(X,\varrho,\mathscr{L}_2)$ for any α, q, X and ϱ.

The same nestedness results hold for the spaces $N_q^\alpha(X,\varrho,\mathscr{L})$. Slightly surprising and already insightful might be that under mild conditions on the activation function ϱ, the approximation classes for strict and generalized ϱ-networks are in fact identical, allowing to derive the conclusion that their expressivities coincide (see Theorem 3.8).

1.3.2. *Approximation spaces associated with ReLU-networks.* The rectified linear unit (ReLU) and its powers of exponent $r \in \mathbb{N}$—in spline theory better-known under the name of *truncated powers* [21, Chapter 5, Equation (1.1)]—are defined by

$$\varrho_r : \mathbb{R} \to \mathbb{R}, x \mapsto (x_+)^r,$$

where $x_+ = \max\{0,x\} = \varrho_1(x)$, with the ReLU activation function being ϱ_1. Considering these activation functions is motivated practically by the wide use of the ReLU [42], as well as theoretically by the existence [45, Theorem 4] of pathological activation functions giving rise to trivial—too rich—approximation spaces that satisfy $W_q^\alpha(X_p^k,\varrho,\mathscr{L}) = N_q^\alpha(X_p^k,\varrho,\mathscr{L}) = X_p^k$, for all α, q. In contrast, the classes associated to ϱ_r-networks are nontrivial for $p \in (0,\infty]$ (Theorem 4.16). Moreover, strict and generalized ϱ_r-networks yield identical approximation classes for any subset $\Omega \subseteq \mathbb{R}^d$ of nonzero measure (even unbounded), for any $p \in (0,\infty]$ (Theorem 4.2). Furthermore, for any $r \in \mathbb{N}$, these approximation classes are (quasi-)Banach spaces (Theorem 4.2), as soon as

$$L := \sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq \begin{cases} 2, & \text{if } \Omega \text{ is bounded } \textit{or } d = 1, \\ 3, & \text{otherwise.} \end{cases}.$$

The expressivity of networks with more general activation functions can be related to that of ϱ_r-networks (see Theorem 4.7) in the following sense: If ϱ is continuous and piecewise polynomial of degree at most r, then its approximation spaces are contained in those of ϱ_r-networks. In particular, if Ω is bounded or if $\mathscr{L}$ satisfies a certain growth condition, then for $1 \leq s \leq r$

$$W_q^\alpha(X,\varrho_s,\mathscr{L}) \subset W_q^\alpha(X,\varrho_r,\mathscr{L}) \qquad \text{and} \qquad N_q^\alpha(X,\varrho_s,\mathscr{L}) \subset N_q^\alpha(X,\varrho_r,\mathscr{L}).$$

Also, if ϱ is a spline of degree r and not a polynomial, then its approximation spaces match those of ϱ_r on bounded Ω. In particular, on a bounded domain Ω, the spaces associated to the leaky-ReLU [44], the parametric ReLU [33], the absolute value (as, e.g, in scattering transforms [46]) and the soft-thresholding activation function [30] are all identical to the spaces associated to the ReLU.

Studying the relation of approximation spaces of ϱ_r-networks for different r, we derive the following statement as a corollary (Corollary 4.14) of Theorem 4.7: Approximation spaces of ϱ_2-networks and ϱ_r-networks are equal for $r \geq 2$ when $\mathscr{L}$ satisfies a certain growth condition, showing a saturation from degree 2 on. Given this growth condition, for any $r \geq 2$, we obtain the following diagram:

$$\begin{array}{ccccc} W_q^\alpha(X,\varrho_1,\mathscr{L}) & \subset & W_q^\alpha(X,\varrho_2,\mathscr{L}) & = & W_q^\alpha(X,\varrho_r,\mathscr{L}), \\ \cap & & \cap & & \\ N_q^\alpha(X,\varrho_1,\mathscr{L}) & \subset & N_q^\alpha(X,\varrho_2,\mathscr{L}) & = & N_q^\alpha(X,\varrho_r,\mathscr{L}). \end{array}$$

1.3.3. *Relation to classical function spaces.* Focusing still on ReLU-networks, we show that ReLU-networks of bounded depth approximate $C_c^3(\Omega)$ functions at bounded rates (Theorem 4.17) in the sense that, for open $\Omega \subset \mathbb{R}^d$ and $L := \sup_n \mathscr{L}(n) < \infty$, we prove

$$N_q^\alpha(X,\varrho_1,\mathscr{L}) \cap C_c^3(\Omega) = \{0\} \text{ if } \alpha > 2\cdot(L-1), \qquad \text{and} \qquad W_q^\alpha(X,\varrho_1,\mathscr{L}) \cap C_c^3(\Omega) = \{0\} \text{ if } \alpha > 2\cdot\lfloor L/2\rfloor.$$

As classical function spaces (e.g. Sobolev, Besov) intersect $C_c^3(\Omega)$ nontrivially, they can only embed into $W_q^\alpha(X,\varrho_1,\mathscr{L})$ or $N_q^\alpha(X,\varrho_1,\mathscr{L})$ if the networks are somewhat deep ($L \geq 1+\alpha/2$ or $\lfloor L/2\rfloor \geq \alpha/2$, respectively), giving some insight about the impact of depth on the expressivity of neural networks.

We then study relations to the classical Besov spaces $B_{\sigma,\tau}^s(\Omega) := B_\tau^s(L_\sigma(\Omega;\mathbb{R}))$. We establish both *direct estimates*—that is, embeddings of certain Besov spaces into approximation spaces of ϱ_r-networks—and *inverse estimates*—that is, embeddings of the approximation spaces into certain Besov spaces.

The main result in the regime of direct estimates is Theorem 5.5 showing that if $\Omega \subset \mathbb{R}^d$ is a bounded Lipschitz domain, if $r \geq 2$, and if $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n)$ satisfies $L \geq 2 + 2\lceil \log_2 d \rceil$, then

$$B^{d\alpha}_{p,q}(\Omega) \hookrightarrow W^\alpha_q(X_p(\Omega), \varrho_r, \mathscr{L}) \quad \forall\, p, q \in (0,\infty] \text{ and } 0 < \alpha < \frac{r + \min\{1, p^{-1}\}}{d}. \tag{1.4}$$

For large input dimensions d, however, the condition $L \geq 2 + 2\lceil \log_2 d \rceil$ is only satisfied for quite deep networks. In the case of more shallow networks with $L \geq 3$, the embedding (1.4) still holds (for any $r \in \mathbb{N}$), but is only established for $0 < \alpha < \frac{\min\{1,p^{-1}\}}{d}$. Finally, in case of $d = 1$, the embedding (1.4) is valid as soon as $L \geq 2$ and $r \geq 1$.

Regarding inverse estimates, we first establish limits on possible embeddings (Theorem 5.7). Precisely, for $\Omega = (0,1)^d$ and any $r \in \mathbb{N}$, $\alpha, s \in (0,\infty)$, and $\sigma, \tau \in (0,\infty]$ we have, with $L := \sup_n \mathscr{L}(n) \geq 2$:

- if $\alpha < \lfloor L/2 \rfloor \cdot \min\{s, 2\}$ then $W^\alpha_q(L_p, \varrho_r, \mathscr{L})$ does *not* embed into $B^s_{\sigma,\tau}(\Omega)$;
- if $\alpha < (L-1) \cdot \min\{s, 2\}$ then $N^\alpha_q(L_p, \varrho_r, \mathscr{L})$ does *not* embed into $B^s_{\sigma,\tau}(\Omega)$.

A particular consequence is that for unbounded depth $L = \infty$, none of the spaces $W^\alpha_q(X, \varrho_r, \mathscr{L})$, $N^\alpha_q(X, \varrho_r, \mathscr{L})$ can embed into *any* Besov space of strictly positive smoothness $s > 0$.

For scalar input dimension $d = 1$, an embedding into a Besov space with the relation $\alpha = \lfloor L/2 \rfloor \cdot s$ (respectively $\alpha = (L-1)\cdot s$) is indeed achieved for $X = L_p((0,1))$, $0 < p < \infty$, $r \in \mathbb{N}$, (Theorem 5.13):

$$W^\alpha_q(L_p, \varrho_r, \mathscr{L}) \subset B^s_{\sigma,\sigma}(\Omega), \quad \text{for each } 0 < s < r+1, s \quad \alpha := \lfloor L/2 \rfloor \cdot s, \qquad \sigma := (s + 1/p)^{-1},$$
$$N^\alpha_q(L_p, \varrho_r, \mathscr{L}) \subset B^s_{\sigma,\sigma}(\Omega), \quad \text{for each } 0 < s < r+1, \qquad \alpha := (L-1)\cdot s, \quad \sigma := (s + 1/p)^{-1}.$$

1.4. **Expected Impact and Future Directions.** We anticipate our results to have an impact in a number of areas that we now describe together with possible future directions:

- *Theory of Expressivity.* We introduce a general framework to study approximation properties of deep neural networks from an approximation space viewpoint. This opens the door to transfer various results from this part of approximation theory to deep neural networks. We believe that this conceptually new approach in the theory of expressivity will lead to further insight. One interesting topic for future investigation is, for instance, to derive a finer characterization of the spaces $W^\alpha_q(X_p, \varrho_r, \mathscr{L})$, $N^\alpha_q(X_p, \varrho_r, \mathscr{L})$, for $r \in \{1, 2\}$ (with some assumptions on $\mathscr{L}$).

  Our framework is amenable to various extensions; for example the restriction to convolutional weights would allow a study of approximation spaces of convolutional neural networks.
- *Statistical Analysis of Deep Learning.* Approximation spaces characterize fundamental tradeoffs between the complexity of a network architecture and its ability to approximate (with proper choices of parameter values) a given function f. In statistical learning, a related question is to characterize which generalization bounds (also known as excess risk guarantees) can be achieved when fitting network parameters using m independent training samples. Some "oracle inequalities" [60] of this type have been recently established for idealized training algorithms minimizing the empirical risk (1.2). Our framework, in combination with existing results on the VC-dimension of neural networks [7] is expected to shed new light on such generalization guarantees through a generic approach encompassing various types of constraints on the considered architecture.
- *Design of Deep Neural Networks—Architectural Guidelines.* Our results reveal how the expressive power of a network architecture may be impacted by certain choices such as the presence of certain types of skip connections or the selected activation functions. Thus, our results provide indications on how a network architecture may be adapted without hurting its expressivity, in order to get additional degrees of freedom to ease the task of optimization-based learning algorithms and improve their performance. For instance, while we show that generalized and strict networks have (under mild assumptions on the activation function) the same expressivity, we have not yet considered so-called ResNet architectures. Yet, the empirical observation that a ResNet architecture makes it easier to train deep networks [32] calls for a better understanding of the relations between the corresponding approximations classes.

1.5. **Outline.** The paper is organized as follows.

Section 2 introduces our notations regarding neural networks and provides basic lemmata concerning the "calculus" of neural networks. The classical notion of approximation spaces is reviewed in Section 3, and therein also specialized to the setting of approximation spaces of networks, with a focus on approximation in L_p spaces. This is followed by Section 4, which concentrates on ϱ-networks with ϱ the so-called ReLU or one of its powers. Finally, Section 5 studies embeddings between $W^\alpha_q(X, \varrho_r, \mathscr{L})$ (resp. $N^\alpha_q(X, \varrho_r, \mathscr{L})$) and classical Besov spaces, with $X = X_p(\Omega)$.

2. Neural networks and their elementary properties

In this section, we formally introduce the definition of neural networks used throughout this paper, and discuss the elementary properties of the corresponding sets of functions.

2.1. Neural networks and their main characteristics.

Definition 2.1 (Neural network)**.** Let $\varrho : \mathbb{R} \to \mathbb{R}$. A *(generalized) neural network with activation function* ϱ (in short: a *ϱ-network*) is a tuple $\big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$, where each $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is an affine-linear map, $\alpha_L = \mathrm{id}_{\mathbb{R}^{N_L}}$, and each function $\alpha_\ell : \mathbb{R}^{N_\ell} \to \mathbb{R}^{N_\ell}$ for $1 \leq \ell < L$ is of the form $\alpha_\ell = \bigotimes_{j=1}^{N_\ell} \varrho_j^{(\ell)}$ for certain $\varrho_j^{(\ell)} \in \{\mathrm{id}_\mathbb{R}, \varrho\}$. Here, we use the notation

$$\bigotimes_{j=1}^{n} \theta_j : X_1 \times \cdots \times X_n \to Y_1 \times \cdots \times Y_n, (x_1, \dots, x_n) \mapsto \big(\theta_1(x_1), \dots, \theta_n(x_n)\big) \text{ for } \theta_j : X_j \to Y_j\,. \qquad \blacktriangleleft$$

Definition 2.2. A ϱ-network as above is called *strict* if $\varrho_j^{(\ell)} = \varrho$ for all $1 \leq \ell < L$ and $1 \leq j \leq N_\ell$. $\blacktriangleleft$

Definition 2.3 (Realization of a network)**.** The *realization* $\mathtt{R}(\Phi)$ of a network $\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$ as above is the function

$$\mathtt{R}(\Phi) : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}, \quad \text{with} \quad \mathtt{R}(\Phi) := \alpha_L \circ T_L \circ \cdots \circ \alpha_1 \circ T_1\,. \qquad \blacktriangleleft$$

The *complexity* of a neural network is characterized by several features.

Definition 2.4 (Depth, number of hidden neurons, number of connections)**.** Consider a neural network $\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$ with $T_\ell : \mathbb{R}^{N_\ell - 1} \to \mathbb{R}^{N_\ell}$ for $1 \leq \ell \leq L$.

- The *input-dimension* of Φ is $d_{\mathrm{in}}(\Phi) := N_0 \in \mathbb{N}$, its *output-dimension* is $d_{\mathrm{out}}(\Phi) := N_L \in \mathbb{N}$.
- The *depth* of Φ is $L(\Phi) := L \in \mathbb{N}$, corresponding to the *number of (affine) layers* of Φ. We remark that with these notations, the number of *hidden* layers is $L - 1$.
- The *number of hidden neurons* of Φ is $N(\Phi) := \sum_{\ell=1}^{L-1} N_\ell \in \mathbb{N}_0$;
- The *number of connections* (or *number of weights*) of Φ is $W(\Phi) := \sum_{\ell=1}^{L} \|T_\ell\|_{\ell^0} \in \mathbb{N}_0$, with $\|T\|_{\ell^0} := \|A\|_{\ell^0}$ for an affine map $T : x \mapsto Ax + b$ with A some matrix and b some vector; here, $\|\cdot\|_{\ell^0}$ counts the number of nonzero entries in a vector or a matrix. $\blacktriangleleft$

Remark 2.5. If $W(\Phi) = 0$ then $\mathtt{R}(\Phi)$ is constant (but not necessarily zero), and if $N(\Phi) = 0$, then $\mathtt{R}(\Phi)$ is affine-linear (but not necessarily zero or constant). $\blacklozenge$

Unlike the notation used in [9, 55], which considers $W_0(\Phi) := \sum_{\ell=1}^{L} (\|A^{(\ell)}\|_{\ell^0} + \|b^{(\ell)}\|_{\ell^0})$ where $T_\ell\, x = A^{(\ell)} x + b^{(\ell)}$, Definition 2.4 only counts the nonzero entries *of the linear part* of each T_ℓ, so that $W(\Phi) \leq W_0(\Phi)$. Yet, as shown with the following lemma, both definitions are in fact equivalent up to constant factors if one is only interested in the represented functions. The proof is in Appendix A.1.

Lemma 2.6. *For any network Φ there is a "compressed" network $\widetilde{\Phi}$ with $\mathtt{R}(\widetilde{\Phi}) = \mathtt{R}(\Phi)$ such that $L(\widetilde{\Phi}) \leq L(\Phi)$, $N(\widetilde{\Phi}) \leq N(\Phi)$, and*

$$W(\widetilde{\Phi}) \leq W_0(\widetilde{\Phi}) \leq d_{\mathrm{out}}(\Phi) + 2 \cdot W(\Phi)\,.$$

The network $\widetilde{\Phi}$ can be chosen to be strict if Φ is strict. $\blacktriangleleft$

Remark 2.7. The reason for distinguishing between a neural network and its associated realization is that for a given function $f : \mathbb{R}^d \to \mathbb{R}^k$, there might be many different neural networks Φ with $f = \mathtt{R}(\Phi)$, so that talking about the number of layers, neurons, or weights of the *function* f is not well-defined, whereas these notions certainly make sense for neural networks as defined above. A possible alternative would be to define for example

$$L(f) := \min\big\{L(\Phi) \,:\, \Phi \text{ neural network with } \mathtt{R}(\Phi) = f\big\},$$

and analogously for $N(f)$ and $W(f)$; but this has the considerable drawback that it is not clear whether there is a neural network Φ that *simultaneously* satisfies e.g. $L(\Phi) = L(f)$ and $W(\Phi) = W(f)$. Because of these issues, we prefer to properly distinguish between a neural network and its realization. $\blacklozenge$

Remark 2.8. Some of the conventions in the above definitions might appear unnecessarily complicated at first sight, but they have been chosen after careful thought. In particular:

- Many neural network architectures used in practice use the same activation function for all neurons in a common layer. If this choice of activation function even stays the same across all layers—except for the last one—one obtains a *strict* neural network.

- In applications, network architectures very similar to our "generalized" neural networks are used; examples include *residual networks* (also called "ResNets", see [32, 67]), and *networks with skip connections* [54].
- As expressed in Section 2.3, the class of realizations of *generalized* neural networks admits nice closure properties under linear combinations and compositions of functions. Similar closure properties do in general *not* hold for the class of strict networks.
- The introduction of generalized networks will be justified in Section 3.3, where we show that if one is only interested in approximation theoretic properties of the respective function class, then—at least on bounded domains $\Omega \subset \mathbb{R}^d$ for "generic" ϱ, but also on unbounded domains for the ReLU activation function and its powers—generalized networks and strict networks have identical properties. ♦

2.2. **Relations between depth, number of neurons, and number of connections.** We now investigate the relationships between the quantities describing the complexity of a neural network $\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$ with $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$.

Given the number of (hidden) neurons of the network, the other quantities can be bounded. Indeed, by definition we have $N_\ell \geq 1$ for all $1 \leq \ell \leq L-1$; therefore, the number of layers satisfies

$$L(\Phi) = 1 + \sum_{\ell=1}^{L-1} 1 \leq 1 + \sum_{\ell=1}^{L-1} N_\ell = 1 + N(\Phi)\,. \tag{2.1}$$

Similarly, as $\|T_\ell\|_{\ell^0} \leq N_{\ell-1} N_\ell$ for each $1 \leq \ell < L$, we have

$$W(\Phi) = \sum_{\ell=1}^{L} \|T_\ell\|_{\ell^0} \leq \sum_{\ell=1}^{L} N_{\ell-1} N_\ell \leq \sum_{\ell'=0}^{L-1} \sum_{\ell=1}^{L} N_{\ell'} N_\ell = (d_{\mathrm{in}}(\Phi) + N(\Phi))(N(\Phi) + d_{\mathrm{out}}(\Phi))\,, \tag{2.2}$$

showing that $W(\Phi) = \mathcal{O}([N(\Phi)]^2 + dk)$ for fixed input and output dimensions d, k. When $L(\Phi) = 2$ we have in fact $W(\Phi) = \|T_1\|_{\ell^0} + \|T_2\|_{\ell^0} \leq N_0 N_1 + N_1 N_2 = (N_0 + N_2) N_1 = (d_{\mathrm{in}}(\Phi) + d_{\mathrm{out}}(\Phi)) \cdot N(\Phi)$.

In general, one *cannot* bound the number of layers or of hidden neurons by the number of nonzero weights, as one can build arbitrarily large networks with many "dead neurons". Yet, such a bound is true if one is willing to switch to a potentially different network *which has the same realization as the original network*. To show this, we begin with the case of networks with zero connections.

Lemma 2.9. *Let $\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$ be a neural network. If there exists some $\ell \in \{1, \dots, L\}$ such that $\|T_\ell\|_{\ell^0} = 0$, then $\mathtt{R}(\Phi) \equiv c$ for some $c \in \mathbb{R}^k$ where $k = d_{\mathrm{out}}(\Phi)$.* ◀

Proof. As $\|T_\ell\|_{\ell^0} = 0$, the affine map T_ℓ is a constant map $\mathbb{R}^{N_{\ell-1}} \ni y \mapsto b^{(\ell)} \in \mathbb{R}^{N_\ell}$. Therefore, $f_\ell = \alpha_\ell \circ T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is a constant map, so that also $\mathtt{R}(\Phi) = f_L \circ \cdots \circ f_\ell \circ \cdots \circ f_1$ is constant. □

Corollary 2.10. *If $W(\Phi) < L(\Phi)$ then $\mathtt{R}(\Phi) \equiv c$ for some $c \in \mathbb{R}^k$ where $k = d_{\mathrm{out}}(\Phi)$.* ◀

Proof. Let $\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$ and observe that if $\sum_{\ell=1}^{L} \|T_\ell\|_{\ell^0} = W(\Phi) < L(\Phi) = \sum_{\ell=1}^{L} 1$ then there must exist $\ell \in \{1, \dots, L\}$ such that $\|T_\ell\|_{\ell^0} = 0$, so that we can apply Lemma 2.9. □

Indeed, constant maps play a special role as they are exactly the set of realizations of neural networks with no (nonzero) connections. Before formally stating this result, we introduce notations for families of neural networks of constrained complexity, which can have a variety of shapes as illustrated on Figure 1.

Definition 2.11. Consider $L \in \mathbb{N} \cup \{\infty\}$, $W, N \in \mathbb{N}_0 \cup \{\infty\}$, and $\Omega \subseteq \mathbb{R}^d$ a non-empty set.

- $\mathcal{NN}_{W,L,N}^{\varrho,d,k}$ denotes the set of all generalized ϱ-networks Φ with input dimension d, output dimension k, and with $W(\Phi) \leq W$, $L(\Phi) \leq L$, and $N(\Phi) \leq N$.
- $\mathcal{SNN}_{W,L,N}^{\varrho,d,k}$ denotes the subset of networks $\Phi \in \mathcal{NN}_{W,L,N}^{\varrho,d,k}$ which are strict.
- The class of all functions $f : \mathbb{R}^d \to \mathbb{R}^k$ that can be represented by (generalized) ϱ-networks with at most W weights, L layers, and N neurons is

$$\mathtt{NN}_{W,L,N}^{\varrho,d,k} := \big\{\mathtt{R}(\Phi) \,:\, \Phi \in \mathcal{NN}_{W,L,N}^{\varrho,d,k}\big\}.$$

  The set of all restrictions of such functions to Ω is denoted $\mathtt{NN}_{W,L,N}^{\varrho,d,k}(\Omega)$.
- Similarly

$$\mathtt{SNN}_{W,L,N}^{\varrho,d,k} := \big\{\mathtt{R}(\Phi) \,:\, \Phi \in \mathcal{SNN}_{W,L,N}^{\varrho,d,k}\big\}.$$

  The set of all restrictions of such functions to Ω is denoted $\mathtt{SNN}_{W,L,N}^{\varrho,d,k}(\Omega)$.

Finally, we define $\mathtt{NN}_{W,L}^{\varrho,d,k} := \mathtt{NN}_{W,L,\infty}^{\varrho,d,k}$ and $\mathtt{NN}_{W}^{\varrho,d,k} := \mathtt{NN}_{W,\infty,\infty}^{\varrho,d,k}$, as well as $\mathtt{NN}^{\varrho,d,k} := \mathtt{NN}_{\infty,\infty,\infty}^{\varrho,d,k}$. We will use similar notations for $\mathtt{SNN}$, $\mathcal{NN}$, and $\mathcal{SNN}$. ◀

Remark 2.12. If the dimensions d, k and/or the activation function ϱ are implied by the context, we will sometimes omit them from the notation. ⧫

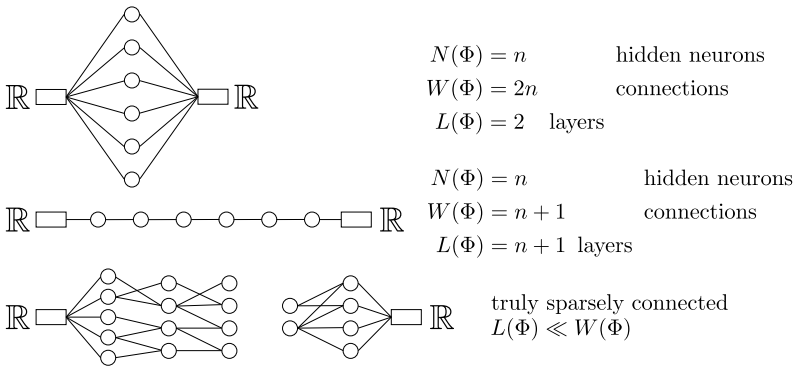

FIGURE 1. The considered network classes include a variety of networks such as: (top) shallow networks with a single hidden layer, where the number of neurons is of the same order as the number of possible connections; (middle) "narrow and deep" networks, e.g. with a single neuron per layer, where the same holds; (bottom) "truly" sparse networks that have much fewer nonzero weights than potential connections.

Lemma 2.13. *Let $\varrho : \mathbb{R} \to \mathbb{R}$, and let $d, k \in \mathbb{N}$, $N \in \mathbb{N}_0 \cup \{\infty\}$, and $L \in \mathbb{N} \cup \{\infty\}$ be arbitrary. Then*

$$\mathtt{NN}^{\varrho,d,k}_{0,L,N} = \mathtt{SNN}^{\varrho,d,k}_{0,L,N} = \mathtt{NN}^{\varrho,d,k}_{0,1,0} = \mathtt{SNN}^{\varrho,d,k}_{0,1,0} = \{f : \mathbb{R}^d \to \mathbb{R}^k \mid \exists c \in \mathbb{R}^k : f \equiv c\}\,. \qquad \blacktriangleleft$$

Proof. If $f \equiv c$ where $c \in \mathbb{R}^k$ then the affine map $T : \mathbb{R}^d \to \mathbb{R}^k, x \mapsto c$ satisfies $\|T\|_{\ell^0} = 0$ and the (strict) network $\Phi := \big((T, \mathrm{id}_{\mathbb{R}^k})\big)$ satisfies $\mathtt{R}(\Phi) \equiv c = f$, $W(\Phi) = 0$, $N(\Phi) = 0$ and $L(\Phi) = 1$. By Definition 2.11, we have $\Phi \in \mathcal{SNN}^{\varrho,d,k}_{0,1,0}$ whence $f \in \mathtt{SNN}^{\varrho,d,k}_{0,1,0}$. The inclusions $\mathtt{SNN}^{\varrho,d,k}_{0,1,0} \subset \mathtt{NN}^{\varrho,d,k}_{0,1,0} \subset \mathtt{NN}^{\varrho,d,k}_{0,L,N}$ and $\mathtt{SNN}^{\varrho,d,k}_{0,1,0} \subset \mathtt{SNN}^{\varrho,d,k}_{0,L,N} \subset \mathtt{NN}^{\varrho,d,k}_{0,L,N}$ are trivial by definition of these sets. If $f \in \mathtt{NN}^{\varrho,d,k}_{0,L,N}$ then there is $\Phi \in \mathcal{NN}^{\varrho,d,k}_{0,L,N}$ such that $f = \mathtt{R}(\Phi)$. As $W(\Phi) = 0 < 1 \leq L(\Phi)$, Corollary 2.10 yields $f = \mathtt{R}(\Phi) \equiv c$. □

Our final result in this subsection shows that any realization of a network with at most $W \geq 1$ connections can also be obtained by a network with W connections but which additionally has at most $L \leq W$ layers and at most $N \leq W$ hidden neurons. The proof is postponed to Appendix A.2.

Lemma 2.14. *Let $\varrho : \mathbb{R} \to \mathbb{R}$, $d, k \in \mathbb{N}$, $L \in \mathbb{N} \cup \{\infty\}$, and $W \in \mathbb{N}$ be arbitrary. Then we have*

$$\mathtt{NN}^{\varrho,d,k}_{W,L,\infty} = \mathtt{NN}^{\varrho,d,k}_{W,L,W} \subset \mathtt{NN}^{\varrho,d,k}_{W,W,W}\,.$$

The inclusion is an equality for $L \geq W$. In particular, $\mathtt{NN}^{\varrho,d,k}_{W} = \mathtt{NN}^{\varrho,d,k}_{W,\infty,W} = \mathtt{NN}^{\varrho,d,k}_{W,W,W}$. The same claims are valid for strict *networks, replacing the symbol* NN *by* SNN *everywhere.* ◀

To summarize, for given input and output dimensions d, k, when combining (2.2) with the above lemma, we obtain that for any network Φ there exists a network Ψ with $\mathtt{R}(\Psi) = \mathtt{R}(\Phi)$ and $L(\Psi) \leq L(\Phi)$, and such that

$$N(\Psi) \leq W(\Psi) \leq W(\Phi) \leq N^2(\Phi) + (d + k)N(\Phi) + dk. \tag{2.3}$$

When $L(\Phi) = 2$ we have in fact $N(\Psi) \leq W(\Psi) \leq W(\Phi) \leq (d + k)N(\Phi)$; see the discussion after (2.2).

Remark 2.15. (*Connectivity, flops and bits.)* A motivation for measuring a network's complexity by its connectivity is that the number of connections is directly related to several practical quantities of interest such as the number of floating point operations needed to compute the output given the input, or the number of bits needed to store a (quantized) description of the network in a computer file. This is not the case for complexity measured in terms of the number of neurons. ⧫

2.3. **Calculus with generalized neural networks.** In this section, we show as a consequence of Lemma 2.14 that the class of realizations of generalized neural networks of a given *complexity*—as measured by the number of connections $W(\Phi)$—is closed under addition and composition, as long as one is willing to increase the complexity by a constant factor. To this end, we first show that *one can increase the depth of generalized neural networks with controlled increase of the required complexity.*

Lemma 2.16. *Given $\varrho : \mathbb{R} \to \mathbb{R}$, $d, k \in \mathbb{N}$, $c := \min\{d, k\}$, $\Phi \in \mathcal{NN}^{\varrho,d,k}$, and $L_0 \in \mathbb{N}_0$, there exists $\Psi \in \mathcal{NN}^{\varrho,d,k}$ such that $\mathtt{R}(\Psi) = \mathtt{R}(\Phi)$, $L(\Psi) = L(\Phi) + L_0$, $W(\Psi) = W(\Phi) + cL_0$, $N(\Psi) = N(\Phi) + cL_0$.* ◀

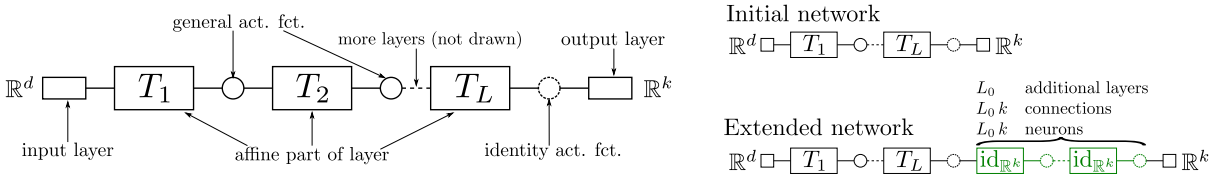

FIGURE 2. (left) Graphical convention for drawing neural networks; this convention is used everywhere except in Figure 1. (right) Depth synchronization of Lemma 2.16, identity layers are added at the output if $k < d$; in case of $d < k$ they are added at the input.

This fact appears without proof in [60, Section 5.1] under the name of *depth synchronization* for strict networks with the ReLU activation function, with $c = d$. We refine it to $c = \min\{d, k\}$ and give a proof for *generalized networks* with *arbitrary* activation function in Appendix A.3. The underlying proof idea is illustrated in Figure 2.

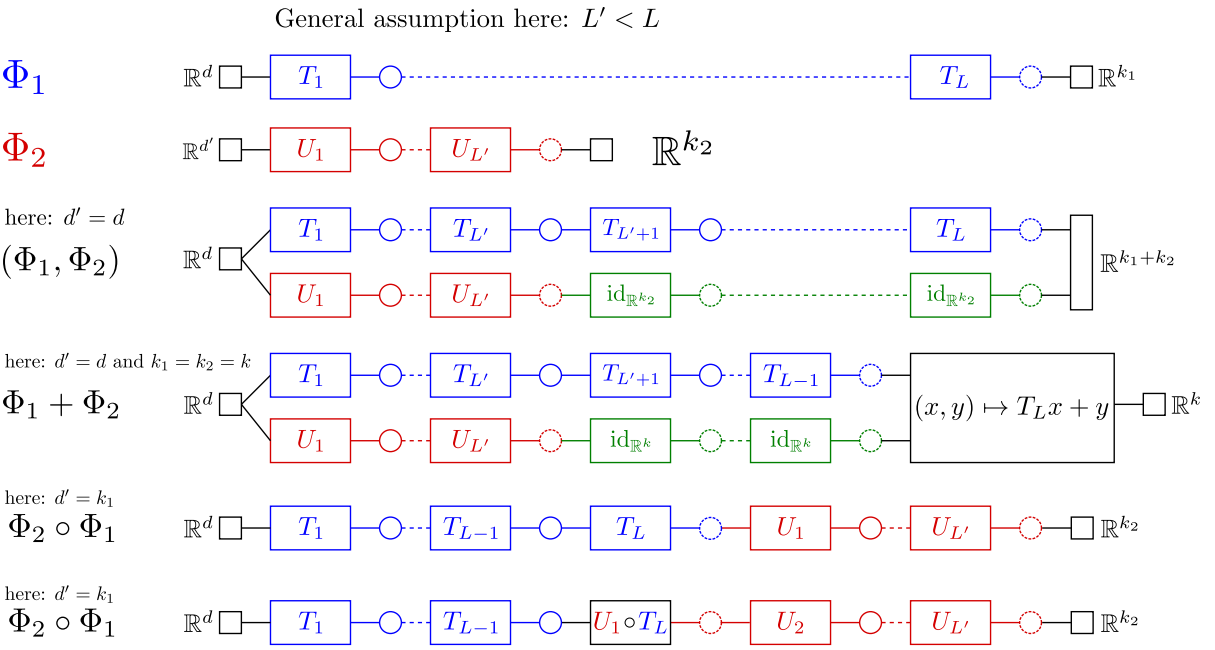

FIGURE 3. Illustration of the networks constructed in the proofs of Lemmas 2.17 and 2.18. (top) Implementation of Cartesian products; (middle) Implementation of addition; (bottom) Implementation of composition.

A consequence of the depth synchronization property is that the class of generalized networks is closed under linear combinations and Cartesian products. The proof idea behind the following lemma, whose proof is in Appendix A.4 is illustrated in Figure 3 (top and middle).

Lemma 2.17. *Consider arbitrary $d, k, n \in \mathbb{N}$, $c \in \mathbb{R}$, $\varrho : \mathbb{R} \to \mathbb{R}$, and $k_i \in \mathbb{N}$ for $i \in \{1, \dots, n\}$.*

(1) If $\Phi \in \mathcal{NN}^{\varrho,d,k}$ then $c \cdot \mathtt{R}(\Phi) = \mathtt{R}(\Psi)$ where $\Psi \in \mathcal{NN}^{\varrho,d,k}$ satisfies $W(\Psi) \leq W(\Phi)$ (with equality if $c \neq 0$), $L(\Psi) = L(\Phi)$, $N(\Psi) = N(\Phi)$. The same holds with $\mathcal{SNN}$ instead of $\mathcal{NN}$.

(2) If $\Phi_i \in \mathcal{NN}^{\varrho,d,k_i}$ for $i \in \{1, \dots, n\}$, then $(\mathtt{R}(\Phi_1), \dots, \mathtt{R}(\Phi_n)) = \mathtt{R}(\Psi)$ with $\Psi \in \mathcal{NN}^{\varrho,d,K}$, where

$$L(\Psi) = \max_{i=1,\dots,n} L(\Phi_i), \quad W(\Psi) \leq \delta + \sum_{i=1}^{n} W(\Phi_i), \quad N(\Psi) \leq \delta + \sum_{i=1}^{n} N(\Phi_i), \quad \text{and} \quad K := \sum_{i=1}^{n} k_i,$$

with $\delta := c \cdot \big(\max_{i=1,\dots,n} L(\Phi_i) - \min_i L(\Phi_i)\big)$ and $c := \min\{d, K-1\}$.

(3) If $\Phi_1, \dots, \Phi_n \in \mathcal{NN}^{\varrho,d,k}$, then $\sum_{i=1}^{n} \mathtt{R}(\Phi_i) = \mathtt{R}(\Psi)$ with $\Psi \in \mathcal{NN}^{\varrho,d,k}$, where

$$L(\Psi) = \max_i L(\Phi_i), \quad W(\Psi) \leq \delta + \sum_{i=1}^{n} W(\Phi_i), \qquad \text{and} \quad N(\Psi) \leq \delta + \sum_{i=1}^{n} N(\Phi_i),$$

with $\delta := c\,(\max_i L(\Phi_i) - \min_i L(\Phi_i))$ and $c := \min\{d, k\}$. ◀

One can also control the complexity of certain networks resulting from compositions in an intuitive way. To state and prove this, we introduce a convenient notation: For a matrix $A \in \mathbb{R}^{n\times d}$, we denote

$$\|A\|_{\ell^{0,\infty}} := \max_{i\in\{1,\dots,d\}} \|A\,e_i\|_{\ell^0} \quad \text{and} \quad \|A\|_{\ell_*^{0,\infty}} := \|A^T\|_{\ell^{0,\infty}} = \max_{i\in\{1,\dots,n\}} \|e_i^T A\|_{\ell^0}, \tag{2.4}$$

where $e_1,\dots,e_n$ is the standard basis of $\mathbb{R}^n$. Likewise, for an affine-linear map $T = A\bullet + b$, we denote $\|T\|_{\ell^{0,\infty}} := \|A\|_{\ell^{0,\infty}}$ and $\|T\|_{\ell_*^{0,\infty}} := \|A\|_{\ell_*^{0,\infty}}$.

Lemma 2.18. *Consider arbitrary $d, d_1, d_2, k, k_1 \in \mathbb{N}$ and $\varrho : \mathbb{R}\to\mathbb{R}$.*

(1) If $\Phi \in \mathcal{NN}^{\varrho,d,k}$ and $P : \mathbb{R}^{d_1}\to\mathbb{R}^d$, $Q : \mathbb{R}^k \to \mathbb{R}^{k_1}$ are two affine maps then $Q\circ \mathtt{R}(\Phi)\circ P = \mathtt{R}(\Psi)$ where $\Psi \in \mathcal{NN}^{\varrho,d_1,k_1}$ with $L(\Psi) = L(\Phi)$, $N(\Psi) = N(\Phi)$ and

$$W(\Psi) \le \|Q\|_{\ell^{0,\infty}} \cdot W(\Phi) \cdot \|P\|_{\ell_*^{0,\infty}}.$$

The same holds with $\mathcal{SNN}$ instead of $\mathcal{NN}$.

(2) If $\Phi_1 \in \mathcal{NN}^{\varrho,d,d_1}$ and $\Phi_2 \in \mathcal{NN}^{\varrho,d_1,d_2}$ then $\mathtt{R}(\Phi_2)\circ\mathtt{R}(\Phi_1) = \mathtt{R}(\Psi)$ where $\Psi \in \mathcal{NN}^{\varrho,d,d_2}$ and

$$W(\Psi) = W(\Phi_1) + W(\Phi_2), \quad L(\Psi) = L(\Phi_1) + L(\Phi_2), \quad N(\Psi) = N(\Phi_1) + N(\Phi_2) + d_1.$$

(3) Under the assumptions of Part (2), there is also $\Psi' \in \mathcal{NN}^{\varrho,d,d_2}$ such that $\mathtt{R}(\Phi_2)\circ\mathtt{R}(\Phi_1) = \mathtt{R}(\Psi')$ and

$$W(\Psi') \le W(\Phi_1) + \max\{N(\Phi_1), d\}\, W(\Phi_2), \quad L(\Psi') = L(\Phi_1)+L(\Phi_2)-1, \quad N(\Psi') = N(\Phi_1)+N(\Phi_2).$$

In this case, the same holds for $\mathcal{SNN}$ instead of $\mathcal{NN}$. ◀

The proof idea of Lemma 2.18 is illustrated in Figure 3 (bottom). The formal proof is in Appendix A.5. A direct consequence of Lemma 2.18-(1) that we will use in several places is that $x \mapsto a_2\, g(a_1 x + b_1) + b_2 \in \mathtt{NN}^{\varrho,d,k}_{W,L,N}$ whenever $g \in \mathtt{NN}^{\varrho,d,k}_{W,L,N}$, $a_1, a_2 \in \mathbb{R}$, $b_1 \in \mathbb{R}^d$, $b_2 \in \mathbb{R}^k$.

Our next result shows that if σ can be expressed as the realization of a ϱ-network then realizations of σ-networks can be re-expanded into realizations of ϱ-networks of controlled complexity.

Lemma 2.19. *Consider two activation functions ϱ, σ such that $\sigma = \mathtt{R}(\Psi_\sigma)$ for some $\Psi_\sigma \in \mathcal{NN}^{\varrho,1,1}_{w,\ell,m}$ with $L(\Psi_\sigma) = \ell \in \mathbb{N}$, $w \in \mathbb{N}_0$, $m \in \mathbb{N}$. Furthermore, assume that $\sigma \not\equiv \text{const}$.*

Then the following hold:

(1) if $\ell = 2$ then for any W, N, L, d, k we have $\mathtt{NN}^{\sigma,d,k}_{W,L,N} \subset \mathtt{NN}^{\varrho,d,k}_{Wm^2,L,Nm}$

(2) for any ℓ, W, N, L, d, k we have $\mathtt{NN}^{\sigma,d,k}_{W,L,N} \subset \mathtt{NN}^{\varrho,d,k}_{mW+wN,1+(L-1)\ell,N(1+m)}$. ◀

The proof of Lemma 2.19 is in Appendix A.6. In the case when σ is simply an s-fold composition of ϱ, we have the following improvement of Lemma 2.19.

Lemma 2.20. *Let $s \in \mathbb{N}$. Consider an activation function $\varrho : \mathbb{R}\to\mathbb{R}$, and let $\sigma := \varrho\circ\cdots\circ\varrho$, where the composition has s "factors". We have*

$$\mathtt{NN}^{\sigma,d,k}_{W,L,N} \subset \mathtt{NN}^{\varrho,d,k}_{W+(s-1)N,1+s(L-1),sN} \quad \forall\, W, N \in \mathbb{N}_0\cup\{\infty\} \text{ and } L \in \mathbb{N}\cup\{\infty\}.$$

The same holds for strict networks, replacing $\mathtt{NN}$ by $\mathtt{SNN}$ everywhere. ◀

The proof is in Appendix A.7. In our next result, we consider the case where σ cannot be exactly implemented by ϱ-networks, but only approximated arbitrarily well by such networks of uniformly bounded complexity.

Lemma 2.21. *Consider two activation functions $\varrho, \sigma : \mathbb{R}\to\mathbb{R}$. Assume that σ is continuous and that there are $w, m \in \mathbb{N}_0$, $\ell \in \mathbb{N}$ and a family $\Psi_h \in \mathcal{NN}^{\varrho,1,1}_{w,\ell,m}$ parameterized by $h \in \mathbb{R}$, with $L(\Psi_h) = \ell$, such that $\sigma_h := \mathtt{R}(\Psi_h) \xrightarrow[h\to 0]{} \sigma$ locally uniformly on $\mathbb{R}$. For any $d, k \in \mathbb{N}$, $W, N \in \mathbb{N}_0$, $L \in \mathbb{N}$ we have*

$$\mathtt{NN}^{\sigma,d,k}_{W,L,N} \subset \begin{cases} \overline{\mathtt{NN}^{\varrho,d,k}_{Wm^2,L,Nm}}, & \text{if } \ell = 2; \\ \overline{\mathtt{NN}^{\varrho,d,k}_{mW+wN,1+(L-1)\ell,N(1+m)}}, & \text{for any } \ell, \end{cases} \tag{2.5}$$

where the closure is with respect to locally uniform convergence. ◀

The proof is in Appendix A.8. In the next lemma, we establish a relation between the approximation capabilities of strict and generalized networks. The proof is given in Appendix A.9.

Lemma 2.22. *Let $\varrho : \mathbb{R}\to\mathbb{R}$ be* continuous *and assume that ϱ is differentiable at some $x_0 \in \mathbb{R}$ with $\varrho'(x_0) \neq 0$. For any $d, k \in \mathbb{N}$, $L \in \mathbb{N}\cup\{\infty\}$, $N \in \mathbb{N}_0\cup\{\infty\}$, and $W \in \mathbb{N}_0$ we have*

$$\mathtt{NN}^{\varrho,d,k}_{W,L,N} \subset \overline{\mathtt{SNN}^{\varrho,d,k}_{4W,L,2N}},$$

where the closure is with respect to locally uniform convergence. ◀

2.4. **Networks with activation functions that can represent the identity.** The convergence in Lemma 2.22 is only locally uniformly, which is not strong enough to ensure equality of the associated approximation spaces on *unbounded* domains. In this subsection we introduce a certain condition on the activation functions which ensures that strict and generalized networks yield the same approximation spaces also on unbounded domains.

Definition 2.23. We say that a function $\varrho : \mathbb{R} \to \mathbb{R}$ *can represent* $f : \mathbb{R} \to \mathbb{R}$ *with* n *terms* (where $n \in \mathbb{N}$) if $f \in \mathtt{SNN}^{\varrho,1,1}_{\infty,2,n}$; that is, if there are $a_i, b_i, c_i \in \mathbb{R}$ for $i \in \{1, \dots, n\}$, and some $c \in \mathbb{R}$ satisfying

$$f(x) = c + \sum_{i=1}^{n} a_i \cdot \varrho(b_i\, x + c_i) \qquad \forall\, x \in \mathbb{R}\,.$$

A particular case of interest is when ϱ *can represent the identity* $\mathrm{id} : \mathbb{R} \to \mathbb{R}$ *with* n *terms.* ◀

As shown in Appendix A.10, primary examples are the ReLU activation function and its powers.

Lemma 2.24. *For any $r \in \mathbb{N}$, ϱ_r can represent any polynomial of degree $\leq r$ with $2r + 2$ terms.* ◀

Lemma 2.25. *Assume that $\varrho : \mathbb{R} \to \mathbb{R}$ can represent the identity with n terms. Let $d, k \in \mathbb{N}$, $W, N \in \mathbb{N}_0$, and $L \in \mathbb{N} \cup \{\infty\}$ be arbitrary. Then $\mathtt{NN}^{\varrho,d,k}_{W,L,N} \subset \mathtt{SNN}^{\varrho,d,k}_{n^2 \cdot W, L, n \cdot N}$.* ◀

The proof of Lemma 2.25 is in Appendix A.9. The next lemma is proved in Appendix A.11.

Lemma 2.26. *If $\varrho : \mathbb{R} \to \mathbb{R}$ can represent all polynomials of degree two with n terms, then:*

(1) For $d \in \mathbb{N}_{\geq 2}$ the multiplication function $M_d : \mathbb{R}^d \to \mathbb{R}, x \mapsto \prod_{i=1}^d x_i$ satisfies

$$M_d \in \mathtt{NN}^{\varrho,d,1}_{6n(2^j-1),2j,(2n+1)(2^j-1)-1} \quad \text{with} \quad j = \lceil \log_2 d \rceil.$$

In particular, for $d = 2$ we have $M_2 \in \mathtt{SNN}^{\varrho,d,1}_{6n,2,2n}$.

(2) For $k \in \mathbb{N}$ the multiplication map $m : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}^k, (x, y) \mapsto x \cdot y$ satisfies $m \in \mathtt{NN}^{\varrho,1+k,k}_{6kn,2,2kn}$. ◀

3. Neural network approximation spaces

The overall goal of this paper is to study *approximation spaces* associated to the sequence of sets Σ_n of realizations of networks with at most n connections (resp. at most n neurons), $n \in \mathbb{N}_0$, either for fixed network depth $L \in \mathbb{N}$, or for unbounded depth $L = \infty$, or even for varying depth $L = \mathscr{L}(n)$.

In this section, we first formally introduce these approximation spaces, following the theory from [21, Chapter 7, Section 9], and then specialize these spaces to the context of neural networks. The next sections will be devoted to establishing embeddings between classical functions spaces and neural network approximation spaces, as well as nesting properties between such spaces.

3.1. **Generic tools from approximation theory.** Consider a quasi-Banach [1] space X equipped with the quasi-norm $\|\cdot\|_X$, and let $f \in X$. The *error of best approximation* of f from a nonempty set $\Gamma \subset X$ is

$$E(f, \Gamma)_X := \inf_{g \in \Gamma} \|f - g\|_X \in [0, \infty)\,. \tag{3.1}$$

In case of $X = X_p^k(\Omega)$ (as in Equation (1.3)) with $\Omega \subseteq \mathbb{R}^d$ a set of nonzero measure, the corresponding approximation error will be denoted by $E(f, \Gamma)_p$. As in [21, Chapter 7, Section 9], we consider an arbitrary family $\Sigma = (\Sigma_n)_{n \in \mathbb{N}_0}$ of subsets $\Sigma_n \subset X$ and define for $f \in X$, $\alpha \in (0, \infty)$, and $q \in (0, \infty]$ the following quantity (which will turn out to be a quasi-norm under mild assumptions on the family Σ):

$$\|f\|_{A^\alpha_q(X,\Sigma)} := \begin{cases} \left(\displaystyle\sum_{n=1}^{\infty} [n^\alpha \cdot E(f, \Sigma_{n-1})_X]^q \frac{1}{n} \right)^{1/q} \in [0, \infty], & \text{if } 0 < q < \infty, \\ \displaystyle\sup_{n \geq 1} [n^\alpha \cdot E(f, \Sigma_{n-1})_X] \in [0, \infty], & \text{if } q = \infty. \end{cases}$$

As expected, the associated *approximation class* is simply

$$A^\alpha_q(X, \Sigma) := \left\{ f \in X \,:\, \|f\|_{A^\alpha_q(X,\Sigma)} < \infty \right\}.$$

For $q = \infty$, this class is precisely the subset of elements $f \in X$ such that $E(f, \Sigma_n)_X = \mathcal{O}(n^{-\alpha})$, and the classes associated to $0 < q < \infty$ correspond to subtle variants of this subset. If we assume that

[1] See e.g. [4, Section 3] for reminders on quasi-norms and quasi-Banach spaces.

$\Sigma_n \subset \Sigma_{n+1}$ for all $n \in \mathbb{N}_0$, then the following "embeddings" can be derived directly from the definition; see [21, Chapter 7, Equation (9.2)]:

$$A_q^\alpha(X,\Sigma) \hookrightarrow A_s^\beta(X,\Sigma), \quad \text{if} \quad \alpha > \beta \quad \text{or if} \quad \alpha = \beta \quad \text{and} \quad q \leq s. \tag{3.2}$$

Note that we do not yet know that the approximation classes $A_q^\alpha(X,\Sigma)$ are (quasi)-Banach spaces. Therefore, the notation $X_1 \hookrightarrow X_2$—where for $i \in \{1,2\}$ we consider the class $X_i := \{x \in X : \|x\|_{X_i} < \infty\}$ associated to some "proto"-quasi-norm $\|\cdot\|_{X_i}$—simply means that $X_1 \subset X_2$ and $\|\cdot\|_{X_2} \leq C \cdot \|\cdot\|_{X_1}$, even though $\|\cdot\|_{X_i}$ might not be proper (quasi)-norms and X_i might not be (quasi)-Banach spaces. When the classes are indeed (quasi)-Banach spaces (see below), this corresponds to the standard notion of a continuous embedding.

As a direct consequence of the definitions, we get the following result on the relation between approximation classes using *different families* of subsets.

Lemma 3.1. *Let X be a quasi-Banach space, and let $\Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}$ and $\Sigma' = (\Sigma'_n)_{n\in\mathbb{N}_0}$ be two families of subsets $\Sigma_n, \Sigma'_n \subset X$ satisfying the following properties:*

(1) $\Sigma_0 = \{0\} = \Sigma'_0$;
(2) $\Sigma_n \subset \Sigma_{n+1}$ and $\Sigma'_n \subset \Sigma'_{n+1}$ for all $n \in \mathbb{N}_0$; and
(3) there are $c \in \mathbb{N}$ and $C > 0$ such that $E(f,\Sigma_{cm})_X \leq C \cdot E(f,\Sigma'_m)_X$ for all $f \in X, m \in \mathbb{N}$.

Then $A_q^\alpha(X,\Sigma') \hookrightarrow A_q^\alpha(X,\Sigma)$ holds for arbitrary $q \in (0,\infty]$ and $\alpha > 0$. More precisely, there is a constant $K = K(\alpha,q,c,C) > 0$ satisfying

$$\|f\|_{A_q^\alpha(X,\Sigma)} \leq K \cdot \|f\|_{A_q^\alpha(X,\Sigma')} \qquad \forall\, f \in A_q^\alpha(X,\Sigma')\,. \qquad \blacktriangleleft$$

Remark. One can alternatively assume that $E(f;\Sigma_{cm})_X \leq C \cdot E(f;\Sigma'_m)_X$ only holds for $m \geq m_0 \in \mathbb{N}$. Indeed, if this is satisfied and if we set $c' := m_0\, c$, then we see for arbitrary $m \in \mathbb{N}$ that $m_0 m \geq m_0$, so that

$$E(f;\Sigma_{c'm})_X = E(f;\Sigma_{c\cdot m_0\, m})_X \leq C \cdot E(f;\Sigma'_{m_0 m})_X \leq C \cdot E(f;\Sigma'_m)_X\,.$$

Here, the last step used that $m_0\, m \geq m$, so that $\Sigma'_m \subset \Sigma'_{m_0\, m}$. ♦

The proof of Lemma 3.1 can be found in Appendix B.1.

In [21, Chapter 7, Section 9], the authors develop a general theory regarding approximation classes of this type. To apply this theory, we merely have to verify that $\Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}$ satisfies the following list of axioms, which is identical to [21, Chapter 7, Equation (5.2)]:

(P1) $\Sigma_0 = \{0\}$;
(P2) $\Sigma_n \subset \Sigma_{n+1}$ for all $n \in \mathbb{N}_0$;
(P3) $a \cdot \Sigma_n = \Sigma_n$ for all $a \in \mathbb{R} \setminus \{0\}$ and $n \in \mathbb{N}_0$;
(P4) There is a fixed constant $c \in \mathbb{N}$ with $\Sigma_n + \Sigma_n \subset \Sigma_{cn}$ for all $n \in \mathbb{N}_0$;
(P5) $\Sigma_\infty := \bigcup_{j\in\mathbb{N}_0} \Sigma_j$ is dense in X;
(P6) for any $n \in \mathbb{N}_0$, each $f \in X$ has a best approximation from Σ_n.

As we will show in Theorem 3.27 below, Properties (P1)–(P5) hold in $X = X_p^k(\Omega)$ for an appropriately defined family Σ related to neural networks of fixed or varying network depth $L \in \mathbb{N} \cup \{\infty\}$.

Property (P6), however, can fail in this setting even for the simple case of the ReLU activation function; indeed, a combination of Lemmas 3.26 and 4.4 below shows that ReLU networks of bounded complexity can approximate the *discontinuous* function $\mathbb{1}_{[a,b]}$ arbitrarily well. Yet, since realizations of ReLU networks are always continuous, $\mathbb{1}_{[a,b]}$ is not implemented exactly by such a network; hence, no best approximation exists. Fortunately, Property (P6) is not essential for the theory from [21] to be applicable: by the arguments given in [21, Chapter 7, discussion around Equation (9.2)] (see also [4, Proposition 3.8 and Theorem 3.12]) we get the following properties of the approximation classes $A_q^\alpha(X,\Sigma)$ that turn out to be approximation *spaces*, i.e., quasi-Banach spaces.

Proposition 3.2. *If Properties (P1)–(P5) hold, then the classes $(A_q^\alpha(X,\Sigma), \|\cdot\|_{A_q^\alpha(X,\Sigma)})$ are quasi-Banach* spaces *satisfying the continuous embeddings* (3.2) *and $A_q^\alpha(X,\Sigma) \hookrightarrow X$.* ◀

Remark. Note that $\|\cdot\|_{A_q^\alpha(X,\Sigma)}$ is in general only a quasi-norm, even if X is a Banach space and $q \in [1,\infty]$. Only if one additionally knows that all the sets Σ_n are vector spaces (that is, one can choose $c = 1$ in Property (P4)), one knows for sure that $\|\cdot\|_{A_q^\alpha(X,\Sigma)}$ is a norm. ♦

Everything except for the completeness and the embedding $A_q^\alpha(X,\Sigma) \hookrightarrow X$ is shown in [21, Chapter 7, Discussion around Equation (9.2)]. For the sake of completeness, as we could not locate a reference, we provide a proof of the missing elements in Appendix B.2.

3.2. **Approximation classes of generalized networks.** We now specialize to the setting of neural networks and consider $d, k \in \mathbb{N}$, an activation function $\varrho : \mathbb{R} \to \mathbb{R}$, and a non-empty set $\Omega \subseteq \mathbb{R}^d$.

Our goal is to define a family of sets of (realizations of) ϱ-networks of "complexity" $n \in \mathbb{N}_0$. The complexity will be measured in terms of the number of connections $W \leq n$ or the number of neurons $N \leq n$, possibly with a control on how the depth L evolves with n.

Definition 3.3 (Depth growth function)**.** A *depth growth function* is a non-decreasing function

$$\mathscr{L} : \mathbb{N} \to \mathbb{N} \cup \{\infty\}, n \mapsto \mathscr{L}(n). \qquad \blacktriangleleft$$

Definition 3.4 (Approximation family, approximation spaces)**.** Given an activation function ϱ, a depth growth function $\mathscr{L}$, a subset $\Omega \subseteq \mathbb{R}^d$, and a quasi-Banach space X whose elements are (equivalence classes of) functions $f : \Omega \to \mathbb{R}^k$, we define $\mathtt{N}_0(X, \varrho, \mathscr{L}) = \mathtt{W}_0(X, \varrho, \mathscr{L}) := \{0\}$, and

$$\mathtt{W}_n(X, \varrho, \mathscr{L}) := \mathtt{NN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty}(\Omega) \cap X, \qquad (n \in \mathbb{N}), \tag{3.3}$$

$$\mathtt{N}_n(X, \varrho, \mathscr{L}) := \mathtt{NN}^{\varrho,d,k}_{\infty,\mathscr{L}(n),n}(\Omega) \cap X, \qquad (n \in \mathbb{N}). \tag{3.4}$$

To highlight the role of the activation function ϱ and the depth growth function $\mathscr{L}$ in the definition of the corresponding approximation classes, we introduce the specific notation

$$W^\alpha_q(X, \varrho, \mathscr{L}) = A^\alpha_q(X, \Sigma) \quad \text{where} \quad \Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}, \text{ with } \Sigma_n := \mathtt{W}_n(X, \varrho, \mathscr{L}), \tag{3.5}$$

$$N^\alpha_q(X, \varrho, \mathscr{L}) = A^\alpha_q(X, \Sigma) \quad \text{where} \quad \Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}, \text{ with } \Sigma_n := \mathtt{N}_n(X, \varrho, \mathscr{L}). \tag{3.6}$$

The quantities $\|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L})}$ and $\|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L})}$ are defined similarly. Notice that the input and output dimensions d, k as well as the set Ω are implicitly described by the space X. Finally, if the depth growth function is constant ($\mathscr{L} \equiv L$ for some $L \in \mathbb{N}$), we write $\mathtt{W}_n(X, \varrho, L)$, etc. $\blacktriangleleft$

Remark 3.5. By convention, $\mathtt{W}_0(X, \varrho, \mathscr{L}) = \mathtt{N}_0(X, \varrho, \mathscr{L}) = \{0\}$, while $\mathtt{NN}^{\varrho,d,k}_{0,L}$ is the set of constant functions $f \equiv c$, where $c \in \mathbb{R}^k$ is arbitrary (Lemma 2.13), and $\mathtt{NN}^{\varrho,d,k}_{\infty,L,0}$ is the set of affine functions. $\blacklozenge$

Remark 3.6. Lemma 2.14 shows that $\mathtt{NN}^{\varrho,d,k}_{W,L} = \mathtt{NN}^{\varrho,d,k}_{W,W}$ if $L \geq W \geq 1$; hence the approximation family $\mathtt{W}_n(X, \varrho, \mathscr{L})$ associated to any depth growth function $\mathscr{L}$ is also generated by the modified depth growth function $\mathscr{L}'(n) := \min\{n, \mathscr{L}(n)\}$, which satisfies $\mathscr{L}'(n) \in \{1, \dots, n\}$ for all $n \in \mathbb{N}$.

In light of Equation (2.1), a similar observation holds for $\mathtt{N}_n(X, \varrho, \mathscr{L})$ with $\mathscr{L}'(n) := \min\{n+1, \mathscr{L}(n)\}$. It will be convenient, however, to explicitly specify unbounded depth as $\mathscr{L} \equiv +\infty$ rather than the equivalent form $\mathscr{L}(n) = n$ (resp. rather than $\mathscr{L}(n) = n + 1$). $\blacklozenge$

We will further discuss the role of the depth growth function in Section 3.5. Before that, we compare approximation with generalized and strict networks.

3.3. **Approximation with generalized *vs* strict networks.** In this subsection, we show that *if one only considers the approximation theoretic properties* of the resulting function classes, then—under extremely mild assumptions on the activation function ϱ—it does not matter whether we consider strict or generalized networks, at least on *bounded* domains $\Omega \subset \mathbb{R}^d$. Here, instead of the approximating sets for *generalized* neural networks defined in (3.3)-(3.4) we wish to consider the corresponding sets for *strict* neural networks, given by $\mathtt{SW}_0(X, \varrho, \mathscr{L}) := \mathtt{SN}_0(X, \varrho, \mathscr{L}) := \{0\}$, and

$$\mathtt{SW}_n(X, \varrho, \mathscr{L}) := \mathtt{SNN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty}(\Omega) \cap X, \qquad (n \in \mathbb{N}),$$

$$\mathtt{SN}_n(X, \varrho, \mathscr{L}) := \mathtt{SNN}^{\varrho,d,k}_{\infty,\mathscr{L}(n),n}(\Omega) \cap X, \qquad (n \in \mathbb{N}),$$

and the associated approximation classes that we denote by

$$SW^\alpha_q(X, \varrho, \mathscr{L}) = A^\alpha_q(X, \Sigma) \quad \text{where} \quad \Sigma = (\Sigma_n)_{n\in\mathbb{N}_0} \quad \text{with} \quad \Sigma_n := \mathtt{SW}_n(X, \varrho, \mathscr{L})$$

$$SN^\alpha_q(X, \varrho, \mathscr{L}) = A^\alpha_q(X, \Sigma) \quad \text{where} \quad \Sigma = (\Sigma_n)_{n\in\mathbb{N}_0} \quad \text{with} \quad \Sigma_n := \mathtt{SN}_n(X, \varrho, \mathscr{L}).$$

Since generalized networks are at least as expressive as strict ones, these approximation classes embed into the corresponding classes for generalized networks, as we now formalize.

Proposition 3.7. *Consider ϱ an activation function, $\mathscr{L}$ a depth growth function, and X a quasi-Banach space of (equivalence classes of) functions from a subset $\Omega \subseteq \mathbb{R}^d$ to $\mathbb{R}^k$.For any $\alpha > 0$ and $q \in (0, \infty]$, we have $\|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L})} \leq \|\cdot\|_{SW^\alpha_q(X,\varrho,\mathscr{L})}$ and $\|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L})} \leq \|\cdot\|_{SN^\alpha_q(X,\varrho,\mathscr{L})}$; hence*

$$SW^\alpha_q(X, \varrho, \mathscr{L}) \hookrightarrow W^\alpha_q(X, \varrho, \mathscr{L}) \quad \text{and} \quad SN^\alpha_q(X, \varrho, \mathscr{L}) \hookrightarrow N^\alpha_q(X, \varrho, \mathscr{L}). \qquad \blacktriangleleft$$

Proof. We give the proof for approximation spaces associated to connection complexity; the proof is similar for the case of neuron complexity. Obviously $\mathtt{SW}_n(X,\varrho,\mathscr{L}) \subset \mathtt{W}_n(X,\varrho,\mathscr{L})$ for all $n \in \mathbb{N}_0$, so that the approximation errors satisfy $E\big(f,\mathtt{W}_n(X,\varrho,\mathscr{L})\big)_X \leq E\big(f,\mathtt{SW}_n(X,\varrho,\mathscr{L})\big)_X$ for all $n \in \mathbb{N}_0$. This implies $\|\cdot\|_{W_q^\alpha(X,\varrho,\mathscr{L})} \leq \|\cdot\|_{SW_q^\alpha(X,\varrho,\mathscr{L})}$, whence $SW_q^\alpha(X,\varrho,\mathscr{L}) \subset W_q^\alpha(X,\varrho,\mathscr{L})$. □

Under mild conditions on ϱ, the converse holds on bounded domains when approximating in L_p. This also holds on unbounded domains for activation functions *that can represent the identity.*

Theorem 3.8 (Approximation classes of strict *vs.* generalized networks)**.** Consider $d \in \mathbb{N}$, a measurable set $\Omega \subseteq \mathbb{R}^d$ with nonzero measure, and $\varrho : \mathbb{R} \to \mathbb{R}$ an activation function. Assume either that:

- Ω is bounded, ϱ is *continuous* and ϱ is differentiable at some $x_0 \in \mathbb{R}$ with $\varrho'(x_0) \neq 0$; or that
- ϱ can represent the identity $\mathrm{id} : \mathbb{R} \to \mathbb{R}, x \mapsto x$ with m terms for some $m \in \mathbb{N}$.

Then for any depth growth function $\mathscr{L}$, $k \in \mathbb{N}$, $\alpha > 0$, $p, q \in (0,\infty]$, with $X := X_p^k(\Omega)$ as in Equation (1.3), we have the identities

$$SW_q^\alpha(X,\varrho,\mathscr{L}) = W_q^\alpha(X,\varrho,\mathscr{L}) \quad \text{and} \quad SN_q^\alpha(X,\varrho,\mathscr{L}) = N_q^\alpha(X,\varrho,\mathscr{L})\,,$$

and there exists $C < \infty$ such that

$$\begin{aligned} &\|\bullet\|_{W_q^\alpha(X,\varrho,\mathscr{L})} \leq \|\bullet\|_{SW_q^\alpha(X,\varrho,\mathscr{L})} \leq C\,\|\bullet\|_{W_q^\alpha(X,\varrho,\mathscr{L})} \\ \text{and} \quad &\|\bullet\|_{N_q^\alpha(X,\varrho,\mathscr{L})} \leq \|\bullet\|_{SN_q^\alpha(X,\varrho,\mathscr{L})} \leq C\,\|\bullet\|_{N_q^\alpha(X,\varrho,\mathscr{L})}. \end{aligned}$$ ◀

Before giving the proof, let us clarify the precise choice of (quasi)-norm for the vector-valued spaces $X := X_p^k(\Omega)$ from Equation (1.3). For $f = (f_1,\dots,f_k) : \Omega \to \mathbb{R}^k$ and $0 < p < \infty$ it is defined by $\|f\|_{L_p(\Omega;\mathbb{R}^k)}^p := \sum_{\ell=1}^k \|f_\ell\|_{L_p(\Omega;\mathbb{R})}^p = \int_\Omega |f(x)|_p^p\,dx$, where $|u|_p^p := \sum_{\ell=1}^k |u_\ell|^p$ for each $u \in \mathbb{R}^k$. For $p = \infty$ we use the definition $\|f\|_\infty := \max_{\ell=1,\dots,k} \|f_\ell\|_{L_\infty(\Omega;\mathbb{R})}$.

Proof. When ϱ can represent the identity with m terms, we rely on Lemma 2.25 and on the estimate $\mathscr{L}(n) \leq \mathscr{L}(m^2 n)$ to obtain for any $n \in \mathbb{N}$ that

$$\mathtt{W}_n(X,\varrho,\mathscr{L}) = \mathtt{NN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty}(\Omega) \cap X \subset \mathtt{SNN}^{\varrho,d,k}_{m^2 n,\mathscr{L}(m^2 n),\infty} \cap X = \mathtt{SW}_{m^2 n}(X,\varrho,\mathscr{L}),$$

and similarly $\mathtt{N}_n(X,\varrho,\mathscr{L}) \subset \mathtt{SN}_{mn}(X,\varrho,\mathscr{L})$, so that

$$\begin{aligned} E\big(f,\mathtt{SW}_{m^2 n}(X,\varrho,\mathscr{L})\big)_X &\leq E\big(f,\mathtt{W}_n(X,\varrho,\mathscr{L})\big)_X \quad \forall n \in \mathbb{N}_0\,, \\ E\big(f,\mathtt{SN}_{mn}(X,\varrho,\mathscr{L})\big)_X &\leq E\big(f,\mathtt{N}_n(X,\varrho,\mathscr{L})\big)_X \quad \forall n \in \mathbb{N}_0\,. \end{aligned}$$

We now establish similar results for the case where Ω is bounded, ϱ is continuous and $\varrho'(x_0) \neq 0$ is well defined for some $x_0 \in \mathbb{R}$. We rely on Lemma 2.22. First, note by continuity of ϱ that any $f \in \mathtt{NN}^{\varrho,d,k} \supset \mathtt{SNN}^{\varrho,d,k}$ is a continuous function $f : \mathbb{R}^d \to \mathbb{R}^k$. Furthermore, since Ω is bounded, $\overline{\Omega}$ is compact, so that $f|_{\overline{\Omega}}$ is *uniformly* continuous and bounded. Clearly, this implies that $f|_\Omega$ is uniformly continuous and bounded as well. Since $X = X_p^k(\Omega)$, this implies

$$\mathtt{SW}_n(X,\varrho,\mathscr{L}) = \mathtt{SNN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty}(\Omega) \cap X = \mathtt{SNN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty}(\Omega)$$

and similarly for $\mathtt{SN}_n(X,\varrho,\mathscr{L})$. Since $\Omega \subset \mathbb{R}^d$ is bounded, locally uniform convergence on $\mathbb{R}^d$ implies convergence in $X_p^k(\Omega)$. Hence for any $n \in \mathbb{N}_0$, using that $\mathscr{L}(n) \leq \mathscr{L}(4n)$, Lemma 2.22 yields

$$\mathtt{W}_n(X,\varrho,\mathscr{L}) \subset \overline{\mathtt{SNN}^{\varrho,d,k}_{4n,\mathscr{L}(4n),\infty}(\Omega)}^{X_p^k(\Omega)} \subset \overline{\mathtt{SW}_{4n}(X,\varrho,\mathscr{L})}^{X_p^k(\Omega)}\,,$$

where the closure is taken with respect to the topology induced by $\|\cdot\|_{X_p^k(\Omega)}$. Similarly, we have

$$\mathtt{N}_n(X,\varrho,\mathscr{L}) \subset \overline{\mathtt{SNN}^{\varrho,d,k}_{\infty,\mathscr{L}(2n),2n}(\Omega)}^{X_p^k(\Omega)} \subset \overline{\mathtt{SN}_{2n}(X,\varrho,\mathscr{L})}^{X_p^k(\Omega)}\,.$$

Now for an arbitrary subset $\Gamma \subset X_p^k(\Omega)$, observe by continuity of $\|\cdot\|_{X_p^k(\Omega)}$ that

$$\inf_{\theta\in\Gamma} \|f-\theta\|_{X_p^k(\Omega)} = \inf_{\theta\in\overline{\Gamma}} \|f-\theta\|_{X_p^k(\Omega)}\,;$$

that is, if one is only interested in the distance of functions f to the set Γ, then switching from Γ to its closure $\overline{\Gamma}$ (computed in X_p^k) does not change the resulting distance. Therefore,

$$\begin{aligned} E\big(f,\mathtt{SW}_{4n}(X,\varrho,\mathscr{L})\big)_X &\leq E\big(f,\mathtt{W}_n(X,\varrho,\mathscr{L})\big)_X \quad \forall n \in \mathbb{N}_0\,, \\ E\big(f,\mathtt{SN}_{2n}(X,\varrho,\mathscr{L})\big)_X &\leq E\big(f,\mathtt{N}_n(X,\varrho,\mathscr{L})\big)_X \quad \forall n \in \mathbb{N}_0\,. \end{aligned}$$

In both settings (ϱ can represent the identity, or Ω is bounded and ϱ differentiable at x_0), Lemma 3.1 shows $\|\cdot\|_{SW_q^\alpha(X,\varrho,\mathscr{L})} \leq C\|\cdot\|_{W_q^\alpha(X,\varrho,\mathscr{L})}$ and $\|\cdot\|_{SN_q^\alpha(X,\varrho,\mathscr{L})} \leq C\|\cdot\|_{N_q^\alpha(X,\varrho,\mathscr{L})}$ for some $C \in (0,\infty)$. The conclusion follows using Proposition 3.7. □

3.4. **Connectivity *vs.* number of neurons.**

Lemma 3.9. *Consider $\varrho : \mathbb{R} \to \mathbb{R}$ an activation function, $\mathscr{L}$ a depth growth function, $d, k \in \mathbb{N}$, $p \in (0,\infty]$ and a measurable $\Omega \subseteq \mathbb{R}^d$ with nonzero measure. With $X := X_p^k(\Omega)$, we have for any $\alpha > 0$ and $q \in (0,\infty]$*

$$W_q^\alpha(X,\varrho,\mathscr{L}) \hookrightarrow N_q^\alpha(X,\varrho,\mathscr{L}) \quad \hookrightarrow W_q^{\alpha/2}(X,\varrho,\mathscr{L}),$$
$$\text{and} \quad SW_q^\alpha(X,\varrho,\mathscr{L}) \hookrightarrow SN_q^\alpha(X,\varrho,\mathscr{L}) \hookrightarrow SW_q^{\alpha/2}(X,\varrho,\mathscr{L}),$$

and there exists $c > 0$ such that

$$\|\cdot\|_{W_q^\alpha(X,\varrho,\mathscr{L})} \geq \|\cdot\|_{N_q^\alpha(X,\varrho,\mathscr{L})} \quad \geq c\|\cdot\|_{W_q^{\alpha/2}(X,\varrho,\mathscr{L})},$$
$$\text{and} \quad \|\cdot\|_{SW_q^\alpha(X,\varrho,\mathscr{L})} \geq \|\cdot\|_{SN_q^\alpha(X,\varrho,\mathscr{L})} \geq c\|\cdot\|_{SW_q^{\alpha/2}(X,\varrho,\mathscr{L})}.$$

When $L := \sup_n \mathscr{L}(n) = 2$ (i.e., for shallow networks) the exponent $\alpha/2$ can be replaced by α; that is, $W_q^\alpha(X,\varrho,\mathscr{L}) = N_q^\alpha(X,\varrho,\mathscr{L})$ with equivalent norms. ◀

Remark. We will see in Lemma 3.10 below that $W_q^\alpha(X,\varrho,\mathscr{L}) \neq N_q^\alpha(X,\varrho,\mathscr{L})$ if, for instance, $\varrho = \varrho_r$ is a power of the ReLU, if Ω is bounded, and if $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n)$ satisfies $3 \leq L < \infty$. In general, however, one cannot expect the spaces to be always distinct. For instance, if ϱ is the activation function constructed in [45, Theorem 4], if $L \geq 3$ and if Ω is bounded, then both $W_q^\alpha(X_p(\Omega),\varrho,\mathscr{L})$ and $N_q^\alpha(X_p(\Omega),\varrho,\mathscr{L})$ coincide with $X_p(\Omega)$. ♦

Proof. We give the proof for generalized networks. By Lemma 2.14 and Equation (2.3),

$$\mathtt{NN}_{n,\mathscr{L}(n),\infty}^{\varrho,d,k} \subset \mathtt{NN}_{n,\mathscr{L}(n),n}^{\varrho,d,k} \subset \mathtt{NN}_{\infty,\mathscr{L}(n),n}^{\varrho,d,k} \subset \mathtt{NN}_{n^2+(d+k)n+dk,\mathscr{L}(n),n}^{\varrho,d,k} \subset \mathtt{NN}_{n^2+(d+k)n+dk,\mathscr{L}(n),\infty}^{\varrho,d,k}$$

for any $n \in \mathbb{N}$. Hence, the approximation errors satisfy

$$E(f,\mathtt{W}_n(X,\varrho,\mathscr{L}))_X \geq E(f,\mathtt{N}_n(X,\varrho,\mathscr{L}))_X \geq E(f,\mathtt{W}_{n^2+(d+k)n+dk}(X,\varrho,\mathscr{L}))_X. \tag{3.7}$$

By the first inequality in (3.7), $\|\cdot\|_{W_q^\alpha(X,\varrho,\mathscr{L})} \geq \|\cdot\|_{N_q^\alpha(X,\varrho,\mathscr{L})}$ and $W_q^\alpha(X,\varrho,\mathscr{L}) \subset N_q^\alpha(X,\varrho,\mathscr{L})$.

When $L = 2$, by the remark below Equation (2.3) we get $\mathtt{NN}_{\infty,\mathscr{L}(n),n}^{\varrho,d,k} \subset \mathtt{NN}_{(d+k)n,\mathscr{L}(n),\infty}^{\varrho,d,k}$; hence $E(f,\mathtt{N}_n(X,\varrho,\mathscr{L}))_X \geq E(f,\mathtt{W}_{(d+k)n}(X,\varrho,\mathscr{L}))_X$ so that Lemma 3.1 shows $W_q^\alpha(X,\varrho,\mathscr{L}) \supset N_q^\alpha(X,\varrho,\mathscr{L})$, with a corresponding (quasi)-norm estimate; hence, these spaces coincide with equivalent (quasi)-norms.

For the general case, observe that $n^2+(d+k)n+dk \leq (n+\gamma)^2$ with $\gamma := \max\{d,k\}$. Let us first consider the case $q < \infty$. In this case, we note that if $(n+\gamma)^2+1 \leq m \leq (n+\gamma+1)^2$, then $n^2 \leq m \leq (2\gamma+2)^2\, n^2$, and thus $m^{\alpha q-1} \lesssim n^{2\alpha q-2}$, where the implied constant only depends on α, q, and γ. This implies

$$\sum_{m=(n+\gamma)^2+1}^{(n+\gamma+1)^2} m^{\alpha q-1} \leq C \cdot n^{2\alpha q-1} \qquad \forall\, n \in \mathbb{N}$$

where $C = C(\alpha,q,\gamma) < \infty$, since the sum has $((n+\gamma)+1)^2 - (n+\gamma)^2 = 2n+2\gamma+1 \leq 4n(2\gamma+1)$ many summands. By the second inequality in (3.7) we get for any $n \in \mathbb{N}$

$$\begin{aligned}\sum_{m=(n+\gamma)^2+1}^{(n+1+\gamma)^2} [m^\alpha E(f,\mathtt{W}_{m-1}(X,\varrho,\mathscr{L}))_X]^q \tfrac{1}{m} &\leq \left(\sum_{m=(n+\gamma)^2+1}^{(n+1+\gamma)^2} m^{\alpha q-1}\right) \cdot \big[E(f,\mathtt{W}_{(n+\gamma)^2}(X,\varrho,\mathscr{L}))_X\big]^q \\ &\leq C \cdot n^{2\alpha q-1} \cdot \big[E(f,\mathtt{W}_{n^2+(d+k)n+dk}(X,\varrho,\mathscr{L}))_X\big]^q \\ &\leq C \cdot n^{2\alpha q-1} \cdot \big[E(f,\mathtt{N}_n(X,\varrho,\mathscr{L}))_X\big]^q.\end{aligned}$$

It follows that

$$\begin{aligned}\sum_{m\geq 1+(\gamma+1)^2} [m^\alpha E(f,\mathtt{W}_{m-1}(X,\varrho,\mathscr{L}))_X]^q \tfrac{1}{m} &= \sum_{n\in\mathbb{N}} \sum_{m=(n+\gamma)^2+1}^{(n+1+\gamma)^2} [m^\alpha E(f,\mathtt{W}_{m-1}(X,\varrho,\mathscr{L}))_X]^q \tfrac{1}{m} \\ &\leq C \sum_{n\in\mathbb{N}} n^{2\alpha q-1} \cdot \big[E(f,\mathtt{N}_n(X,\varrho,\mathscr{L}))_X\big]^q \leq C\|f\|^q_{N_q^{2\alpha}(X,\varrho,\mathscr{L})}.\end{aligned}$$

To conclude we use that $\sum_{m=1}^{(\gamma+1)^2} [m^\alpha E(f,\mathtt{W}_{m-1}(X,\varrho,\mathscr{L}))_X]^q \tfrac{1}{m} \leq C'\|f\|_X^q \leq C'\|f\|^q_{N_q^{2\alpha}(X,\varrho,\mathscr{L})}$ with $C' = \sum_{m=1}^{(\gamma+1)^2} m^{\alpha q-1}$.

The proof for $q = \infty$ is similar. The proof for strict networks follows along similar lines. □

The final result in this subsection shows that the inclusions in Lemma 3.9 are quite sharp.

Lemma 3.10. *For $r \in \mathbb{N}$, define $\varrho_r : \mathbb{R} \to \mathbb{R}, x \mapsto (x_+)^r$.*

Let $\Omega \subset \mathbb{R}^d$ be bounded and measurable with nonempty interior. Let $L, L' \in \mathbb{N}_{\geq 2}$, let $r_1, r_2 \in \mathbb{N}$, let $p_1, p_2, q_1, q_2 \in (0, \infty]$, and $\alpha, \beta > 0$. Then the following hold:

(1) If $W^\alpha_{q_1}(X_{p_1}(\Omega), \varrho_{r_1}, L) \subset N^\beta_{q_2}(X_{p_2}(\Omega), \varrho_{r_2}, L')$, then $L' - 1 \geq \frac{\beta}{\alpha} \cdot \lfloor L/2 \rfloor$.

(2) If $N^\beta_{q_2}(X_{p_2}(\Omega), \varrho_{r_2}, L') \subset W^\alpha_{q_1}(X_{p_1}(\Omega), \varrho_{r_1}, L)$, then $\lfloor L/2 \rfloor \geq \frac{\alpha}{\beta} \cdot (L' - 1)$.

In particular, if $W^\alpha_{q_1}(X_{p_1}(\Omega), \varrho_{r_1}, L) = N^\alpha_{q_2}(X_{p_2}(\Omega), \varrho_{r_2}, L)$, then $L = 2$. ◀

The proof of this result is given in Appendix E.

3.5. **Role of the depth growth function.** In this subsection, we investigate the relation between approximation classes associated to different depth growth functions. First we define a comparison rule between depth growth functions.

Definition 3.11 (Comparison between depth growth functions)**.** The depth growth function $\mathscr{L}$ is *dominated* by the depth growth function $\mathscr{L}'$ (denoted $\mathscr{L} \preceq \mathscr{L}'$ or $\mathscr{L}' \succeq \mathscr{L}$) if there are $c, n_0 \in \mathbb{N}$ such that

$$\forall\, n \geq n_0 : \quad \mathscr{L}(n) \leq \mathscr{L}'(cn)\,. \tag{3.8}$$

Observe that $\mathscr{L} \leq \mathscr{L}'$ implies $\mathscr{L} \preceq \mathscr{L}'$.

The two depth growth functions are *equivalent* (denoted $\mathscr{L} \sim \mathscr{L}'$) if $\mathscr{L} \preceq \mathscr{L}'$ and $\mathscr{L} \succeq \mathscr{L}'$, that is to say if there exist $c, n_0 \in \mathbb{N}$ such that for each $n \geq n_0$, $\mathscr{L}(n) \leq \mathscr{L}'(cn)$ and $\mathscr{L}'(n) \leq \mathscr{L}(cn)$. This defines an equivalence relation on the set of depth growth functions. ◀

Lemma 3.12. *Consider two depth growth functions $\mathscr{L}$, $\mathscr{L}'$. If $\mathscr{L} \preceq \mathscr{L}'$, then for each $\alpha > 0$ and $q \in (0, \infty]$, there is a constant $C = C(\mathscr{L}, \mathscr{L}', \alpha, q) \in [1, \infty)$ such that:*

$$W^\alpha_q(X, \varrho, \mathscr{L}) \hookrightarrow W^\alpha_q(X, \varrho, \mathscr{L}') \quad \text{and} \quad \|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L}')} \leq C \cdot \|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L})}$$
$$N^\alpha_q(X, \varrho, \mathscr{L}) \hookrightarrow N^\alpha_q(X, \varrho, \mathscr{L}') \quad \text{and} \quad \|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L}')} \leq C \cdot \|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L})}$$

for each activation function $\varrho : \mathbb{R} \to \mathbb{R}$, each (bounded or unbounded) set $\Omega \subset \mathbb{R}^d$, and each quasi-Banach space X of (equivalence classes of) functions $f : \Omega \to \mathbb{R}^k$.

The same holds with $SW^\alpha_q(X, \varrho, \mathscr{L})$ (resp. $SN^\alpha_q(X, \varrho, \mathscr{L})$) instead of $W^\alpha_q(X, \varrho, \mathscr{L})$ (resp. $N^\alpha_q(X, \varrho, \mathscr{L})$).

The constant C depends only on the constants $c, n_0 \in \mathbb{N}$ involved in (3.8) and on α, q. ◀

Proof. Let $c, n_0 \in \mathbb{N}$ as in Equation (3.8). For $n \geq n_0$, we then have $\mathscr{L}(n) \leq \mathscr{L}'(cn)$, and hence

$$\mathtt{NN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty} \subset \mathtt{NN}^{\varrho,d,k}_{n,\mathscr{L}'(cn),\infty} \subset \mathtt{NN}^{\varrho,d,k}_{cn,\mathscr{L}'(cn),\infty}\,,$$
$$\mathtt{NN}^{\varrho,d,k}_{\infty,\mathscr{L}(n),n} \subset \mathtt{NN}^{\varrho,d,k}_{\infty,\mathscr{L}'(cn),n} \subset \mathtt{NN}^{\varrho,d,k}_{\infty,\mathscr{L}'(cn),cn}\,,$$

from which we easily get

$$E\big(f, \mathtt{W}_{cn}(X, \varrho, \mathscr{L}')\big)_X \leq E\big(f, \mathtt{W}_n(X, \varrho, \mathscr{L})\big)_X \qquad \forall\, n \geq n_0\,,$$
$$E\big(f, \mathtt{N}_{cn}(X, \varrho, \mathscr{L}')\big)_X \leq E\big(f, \mathtt{N}_n(X, \varrho, \mathscr{L})\big)_X \qquad \forall\, n \geq n_0\,.$$

Now, Lemma 3.1 and the associated remark complete the proof. Exactly the same proof works for strict networks; one just has to replace `NN` by `SNN` everywhere. □

As a direct consequence of Lemma 3.12, we see that equivalent depth growth functions induce the same approximation spaces.

Theorem 3.13. If $\mathscr{L}, \mathscr{L}'$ are two depth-growth functions satisfying $\mathscr{L} \sim \mathscr{L}'$, then for any $\alpha > 0$ and $q \in (0, \infty]$, there is a constant $C \in [1, \infty)$ such that

$$W^\alpha_q(X, \varrho, \mathscr{L}) = W^\alpha_q(X, \varrho, \mathscr{L}') \quad \text{and} \quad \tfrac{1}{C}\,\|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L})} \leq \|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L}')} \leq C \cdot \|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L})}$$
$$N^\alpha_q(X, \varrho, \mathscr{L}) = N^\alpha_q(X, \varrho, \mathscr{L}') \quad \text{and} \quad \tfrac{1}{C}\,\|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L})} \leq \|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L}')} \leq C \cdot \|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L})}$$

for each activation function ϱ, each $\Omega \subset \mathbb{R}^d$, and each quasi-Banach space X of (equivalence classes of) functions $f : \Omega \to \mathbb{R}^k$. The same holds with $SW^\alpha_q(X, \varrho, \mathscr{L})$ (resp. $SN^\alpha_q(X, \varrho, \mathscr{L})$) instead of $W^\alpha_q(X, \varrho, \mathscr{L})$ (resp. $N^\alpha_q(X, \varrho, \mathscr{L})$). The constant C depends only on the constants $c, n_0 \in \mathbb{N}$ in Definition 3.11 and on α, q. ◀

Theorem 3.13 shows in particular that if $L := \sup_n \mathscr{L}(n) < \infty$, then $W^\alpha_q(X, \varrho, \mathscr{L}) = W^\alpha_q(X, \varrho, L)$ with equivalent "proto-norms" (and similarly with $N^\alpha_q(X, \varrho, \mathscr{L})$ instead of $W^\alpha_q(X, \varrho, \mathscr{L})$ or with strict networks instead of generalized ones). Indeed, it is easy to see that $\mathscr{L} \sim \mathscr{L}'$ if $\sup_n \mathscr{L}(n) = \sup_n \mathscr{L}'(n) = L < \infty$.

Lemma 3.14. *Consider $\mathscr{L}$ a depth growth function and $\varepsilon > 0$.*
(1) if $\mathscr{L} + \varepsilon \preceq \mathscr{L}$ then $\mathscr{L} + b \sim \mathscr{L}$ for each $b \geq 0$;
(2) if $e^{\varepsilon}\mathscr{L} \preceq \mathscr{L}$ then $a\mathscr{L} + b \sim \mathscr{L}$ for each $a \geq 1$, $b \geq 1 - a$. ◀

Proof. For the first claim, we first show by induction on $k \in \mathbb{N}$ that $\mathscr{L} + k\varepsilon \preceq \mathscr{L}$. For $k = 1$ this holds by assumption. For the induction step, recall that $\mathscr{L}+k\varepsilon \preceq \mathscr{L}$ simply means that there are $c, n_0 \in \mathbb{N}$ such that $\mathscr{L}(n) + k\varepsilon \leq \mathscr{L}(cn)$ for all $n \in \mathbb{N}_{\geq n_0}$. Therefore, if $n \geq n_0$ then $\mathscr{L}(n) + (k+1)\varepsilon \leq \mathscr{L}(cn) + \varepsilon \leq \mathscr{L}(c^2 n)$ since $n' = cn \geq n \geq n_0$. Now, note that if $\mathscr{L} \leq \mathscr{L}'$, then also $\mathscr{L} \preceq \mathscr{L}'$. Therefore, given $b \geq 0$ we choose $k \in \mathbb{N}$ such that $b \leq k\varepsilon$ and get $\mathscr{L} \preceq \mathscr{L} + b \preceq \mathscr{L} + k\varepsilon \preceq \mathscr{L}$, so that all these depth-growth functions are equivalent.

For the second claim, a similar induction yields $e^{k\varepsilon}\mathscr{L} \preceq \mathscr{L}$ for all $k \in \mathbb{N}$. Now, given $a \geq 1$ and $b \geq 1 - a$, we choose $k \in \mathbb{N}$ such that $a + b_+ \leq e^{k\varepsilon}$, where $b_+ = \max\{0, b\}$. There are now two cases: If $b \geq 0$, then clearly $\mathscr{L} \leq a\mathscr{L} \leq a\mathscr{L} + b$. If otherwise $b < 0$, then $b\mathscr{L} \leq b$, since $\mathscr{L} \geq 1$, and hence $\mathscr{L} = a\mathscr{L} + (1-a)\mathscr{L} \leq a\mathscr{L} + b\mathscr{L} \leq a\mathscr{L} + b$. Therefore, we see in both cases that $\mathscr{L} \leq a\mathscr{L} + b_+ \leq (a + b_+)\mathscr{L} \leq e^{k\varepsilon}\,\mathscr{L} \preceq \mathscr{L}$. □

The following two examples discuss elementary properties of poly-logarithmic and polynomial growth functions, respectively.

Example 3.15. *Assume there are $q \geq 1$, $\alpha, \beta > 0$ such that $|\mathscr{L}(n) - \alpha \log^q n| \leq \beta$ for all $n \in \mathbb{N}$.*

Choosing $c \in \mathbb{N}$ such that $\varepsilon := \alpha \log^q c - 2\beta > 0$, we have

$$\mathscr{L}(n) + \varepsilon \leq \alpha \log^q n + \beta + \varepsilon = \alpha \log^q n + \alpha \log^q c - \beta \leq \alpha(\log c + \log n)^q - \beta = \alpha \log^q(cn) - \beta \leq \mathscr{L}(cn)$$

for all $n \in \mathbb{N}$; hence $\mathscr{L} + \varepsilon \preceq \mathscr{L}$. Here, we used that $x^q + y^q = \|(x,y)\|_{\ell^q}^q \leq \|(x,y)\|_{\ell^1}^q = (x+y)^q$ for $x, y \geq 0$.

By Lemma 3.14 we get $\mathscr{L} \sim \mathscr{L}+b$ for arbitrary $b \geq 0$. Moreover as $\lfloor \alpha \log^q n \rfloor \leq \alpha \log^q n \leq \mathscr{L}(n)+\beta$ we have $\max(1, \lfloor \alpha \log^q(\cdot) \rfloor) \preceq \mathscr{L}+\beta \sim \mathscr{L}$. Similarly $\mathscr{L}(n) \leq \lfloor \alpha \log^q n \rfloor + \beta + 1$ hence $\mathscr{L} \sim \max(1, \lfloor \alpha \log^q(\cdot) \rfloor)$.

Example 3.16. *Assume there are $\gamma > 0$ and $C \geq 1$ such that $1/C \leq \mathscr{L}(n)/n^{\gamma} \leq C$ for all $n \in \mathbb{N}$.*

Choosing any integer $c \geq (2C^2)^{1/\gamma}$ we have $2C^2 c^{-\gamma} \leq 1$, and hence

$$2\mathscr{L}(n) \leq 2Cn^{\gamma} \leq 2Cc^{-\gamma}(cn)^{\gamma} \leq 2Cc^{-\gamma}C\mathscr{L}(cn) = 2C^2 c^{-\gamma}\mathscr{L}(cn) \leq \mathscr{L}(cn)$$

for all $n \in \mathbb{N}$; hence $2\mathscr{L} \preceq \mathscr{L}$. By Lemma 3.14 we get $\mathscr{L} \sim a\mathscr{L} + b$ for each $a \geq 1, b \geq 1 - a$. Moreover, we have $\lceil n^{\gamma} \rceil \leq n^{\gamma} + 1 \leq C\mathscr{L}(n) + 1$ for all $n \in \mathbb{N}$ hence $\lceil (\cdot)^{\gamma} \rceil \preceq C\mathscr{L} + 1 \sim \mathscr{L}$. Similarly $\mathscr{L}(n) \leq Cn^{\gamma} \leq C\lceil n^{\gamma} \rceil$, and thus $\mathscr{L} \sim \lceil (\cdot)^{\gamma} \rceil$.

In the next sections we conduct preliminary investigations on the role of the (finite or infinite) depth L in terms of the associated approximation spaces for ϱ_r-networks. A general understanding of the role of depth growth largely remains an open question. A very surprising result in this direction was recently obtained by Yarotsky [69].

Remark 3.17. It is not difficult to show that approximation classes defined on nested sets $\Omega' \subset \Omega \subset \mathbb{R}^d$ satisfy natural restriction properties. More precisely, the map

$$W_q^{\alpha}(X_p^k(\Omega), \varrho, \mathscr{L}) \to W_q^{\alpha}(X_p^k(\Omega'), \varrho, \mathscr{L}), f \mapsto f|_{\Omega'}$$

is well-defined and bounded (meaning, $\|f|_{\Omega'}\|_{W_q^{\alpha}(X_p^k(\Omega'),\varrho,\mathscr{L})} \leq \|f\|_{W_q^{\alpha}(X_p^k(\Omega),\varrho,\mathscr{L})}$), and the same holds for the spaces N_q^{α} instead of W_q^{α}.

Furthermore, the approximation classes of vector-valued functions $f : \Omega \to \mathbb{R}^k$ are cartesian products of real-valued function classes; that is,

$$W_q^{\alpha}(X_p^k(\Omega; \mathbb{R}^k), \varrho, \mathscr{L}) \to \left(W_q^{\alpha}(X_p^k(\Omega; \mathbb{R}), \varrho, \mathscr{L})\right)^k, f \mapsto (f_1, \ldots, f_k)$$

is bijective and $\|f\|_{W_q^{\alpha}(X_p^k(\Omega;\mathbb{R}^k),\varrho,\mathscr{L})} \asymp \sum_{\ell=1}^k \|f_\ell\|_{W_q^{\alpha}(X_p^k(\Omega;\mathbb{R}),\varrho,\mathscr{L})}$. Again, the same holds for the spaces N_q^{α} instead of W_q^{α}. For the sake of brevity, we omit the easy proofs. ♦

3.6. **Approximation classes are approximation spaces.** We now verify that the main axioms needed to apply Proposition 3.2 are satisfied. Properties (P1)–(P4) hold without any further assumptions:

Lemma 3.18. *Let $\varrho : \mathbb{R} \to \mathbb{R}$ be arbitrary, and let $\mathscr{L}$ be a depth growth function. The sets Σ_n defined in* (3.5)–(3.6) *satisfy Properties (P1)–(P4) on Page 12, with $c = 2 + \min\{d, k\}$ for Property (P4).* ◀

Proof. We generically write $\Sigma_n(X, \varrho, \mathscr{L})$ to indicate either $\mathtt{W}_n(X, \varrho, \mathscr{L})$ or $\mathtt{N}_n(X, \varrho, \mathscr{L})$.

Property (P1). We have $\Sigma_0(X, \varrho, \mathscr{L}) = \{0\}$ by definition. For later use, let us also verify that $0 \in \Sigma_n(X, \varrho, \mathscr{L})$ for $n \in \mathbb{N}$. Indeed, Lemma 2.13 shows $0 \in \mathtt{NN}_{0,1,0}^{\varrho,d,k} \subset \mathtt{NN}_{n,L,m}^{\varrho,d,k}$ for all $n, m, L \in \mathbb{N} \cup \{\infty\}$, and hence $0 \in \Sigma_n(X, \varrho, \mathscr{L})$ for all $n \in \mathbb{N}$.

Property (P2). The inclusions $\mathtt{NN}^{\varrho,d,k}_{W,L,\infty} \subset \mathtt{NN}^{\varrho,d,k}_{W+1,L',\infty}$ and $\mathtt{NN}^{\varrho,d,k}_{\infty,L,N} \subset \mathtt{NN}^{\varrho,d,k}_{\infty,L',N+1}$ for $W, N \in \mathbb{N}_0$ and $L, L' \in \mathbb{N} \cup \{\infty\}$ with $L \leq L'$ hold by the very definition of these sets. As $\mathscr{L}$ is non-decreasing (that is, $\mathscr{L}(n+1) \geq \mathscr{L}(n)$), we thus get $\Sigma_n(X, \varrho, \mathscr{L}) \subset \Sigma_{n+1}(X, \varrho, \mathscr{L})$ for all $n \in \mathbb{N}$. As seen in the proof of Property (P1), this also holds for $n = 0$.

Property (P3). By Lemma 2.17-(1), if $f \in \mathtt{NN}^{\varrho,d,k}_{W,L,N}$, then $a \cdot f \in \mathtt{NN}^{\varrho,d,k}_{W,L,N}$ for any $a \in \mathbb{R}$. Therefore, $a \cdot \Sigma_n(X, \varrho, \mathscr{L}) \subset \Sigma_n(X, \varrho, \mathscr{L})$ for each $a \in \mathbb{R}$ and $n \in \mathbb{N}$. The converse is proved similarly for $a \neq 0$; hence $a \cdot \Sigma_n(X, \varrho, \mathscr{L}) = \Sigma_n(X, \varrho, \mathscr{L})$ for each $a \in \mathbb{R} \setminus \{0\}$ and $n \in \mathbb{N}$. For $n = 0$, this holds trivially.

Property (P4). The claim is trivial for $n = 0$. For $n \in \mathbb{N}$, let $f_1, f_2 \in \Sigma_n(X, \varrho, \mathscr{L})$ be arbitrary.

For the case of $\Sigma_n(X, \varrho, \mathscr{L}) = \mathtt{W}_n(X, \varrho, \mathscr{L})$, let $g_1, g_2 \in \mathtt{NN}^{\varrho,d,k}_{n,\mathscr{L}(n),\infty}$ such that $f_i = g_i|_\Omega$. Lemma 2.14 shows that $g_i \in \mathtt{NN}^{\varrho,d,k}_{n,L',\infty}$ with $L' := \min\{\mathscr{L}(n), n\}$. By Lemma 2.17-(3), setting $c_0 := \min\{d, k\}$, and $W' := 2n + c_0 \cdot (L' - 1) \leq (2 + c_0)n$, we have $g_1 + g_2 \in \mathtt{NN}^{\varrho,d,k}_{W',L'} \subset \mathtt{NN}^{\varrho,d,k}_{(2+c_0)n,\mathscr{L}((2+c_0)n)}$ where for the last inclusion we used that $L' \leq \mathscr{L}(n)$, that $\mathscr{L}$ is non-decreasing, and that $n \leq (2 + c_0)n$.

For the case of $\Sigma_n(X, \varrho, \mathscr{L}) = \mathtt{N}_n(X, \varrho, \mathscr{L})$, consider similarly $g_1, g_2 \in \mathtt{NN}^{\varrho,d,k}_{\infty,\mathscr{L}(n),n}$ such that $f_i = g_i|_\Omega$. By (2.1), $g_i \in \mathtt{NN}^{\varrho,d,k}_{\infty,L',n}$ with $L' := \min\{\mathscr{L}(n), n + 1\}$. By Lemma 2.17-(3) again, setting $c_0 := \min\{d, k\}$, and $N' := 2n + c_0 \cdot (L' - 1) \leq (2 + c_0)n$, we get $g_1 + g_2 \in \mathtt{NN}^{\varrho,d,k}_{\infty,L',N'} \subset \mathtt{NN}^{\varrho,d,k}_{\infty,\mathscr{L}((2+c_0)n),(2+c_0)n}$.

By Definitions (3.3)–(3.4), this shows in all cases that $f_1 + f_2 \in \Sigma_{(2+c_0)n}(X, \varrho, \mathscr{L})$. □

We now focus on Property (P5), in the function space $X = X^k_p(\Omega)$ with $p \in (0, \infty]$ and $\Omega \subset \mathbb{R}^d$ a measurable set with nonzero measure. First, as proved in Appendix B.3, these spaces are indeed complete, and each $f \in X^k_p(\Omega)$ can be extended to an element $\widetilde{f} \in X^k_p(\mathbb{R}^d)$.

Definition 3.19 (L_p-domain)**.** For brevity, in the rest of the paper we refer to $\Omega \subseteq \mathbb{R}^d$ as an *L_p-domain* if, and only if, it is Borel-measurable with nonzero measure. ◀

Lemma 3.20. *Consider $\Omega \subseteq \mathbb{R}^d$ an L_p-domain, $k \in \mathbb{N}$, and $C_0(\mathbb{R}^d; \mathbb{R}^k)$ the space of continuous functions $f : \mathbb{R}^d \to \mathbb{R}^k$ that vanish at infinity.*

For $0 < p < \infty$, we have $X^k_p(\Omega) = \{f|_\Omega : f \in X^k_p(\mathbb{R}^d)\}$; likewise, $X^k_\infty(\Omega) = \{f|_\Omega : f \in C_0(\mathbb{R}^d; \mathbb{R}^k)\}$. The spaces $X^k_p(\Omega)$ are quasi-Banach spaces. ◀

In light of definitions (3.3)–(3.4), we have

$$\bigcup_{n \in \mathbb{N}_0} \mathtt{W}_n(X, \varrho, \mathscr{L}) = \bigcup_{n \in \mathbb{N}_0} \mathtt{N}_n(X, \varrho, \mathscr{L}) = \mathtt{NN}^{\varrho,d,k}_{\infty,L,\infty}(\Omega) \cap X =: \Sigma_\infty(X, \varrho, \mathscr{L}),$$

with $L := \sup_n \mathscr{L}(n) \in \mathbb{N} \cup \{+\infty\}$. Properties (P3) and (P4) imply that $\Sigma_\infty(X, \varrho, \mathscr{L})$ is a linear space. We study its density in X, dealing first with a few degenerate cases.

3.6.1. *Degenerate cases.* Property (P5) can fail to hold for certain activation functions: when ϱ is a polynomial and $\mathscr{L}$ is bounded, the set $\Sigma_\infty(X, \varrho, \mathscr{L})$ only contains polynomials of bounded degree, hence for nontrivial Ω, $\Sigma_\infty(X, \varrho, \mathscr{L})$ is not dense in X. Property (P5) fails again for networks with a single hidden layer ($L = 2$) and certain domains such as $\Omega = \mathbb{R}^d$. Indeed, the realization of any network in $\mathtt{NN}^{\varrho,d,k}_{\infty,2,\infty}$ is a finite linear combination of ridge functions $x \mapsto \varrho(A_i x + b_i)$. A ridge function is in $L_p(\mathbb{R}^d)$ ($p < \infty$) only if it is zero. Moreover, one can check that if a linear combination of ridge functions belongs to $L_p(\mathbb{R}^d)$ ($1 \leq p \leq 2$), then it vanishes, hence $\Sigma_\infty(X, \varrho, \mathscr{L}) = \{0\}$.

3.6.2. *Non-degenerate cases.* We now show that Property (P5) holds under proper assumptions on the activation function ϱ, the depth growth function $\mathscr{L}$, and the domain Ω. The proof uses the celebrated universal approximation theorem for multilayer feedforward networks [43]. In light of the above observations we introduce the following definition:

Definition 3.21. An activation function $\varrho : \mathbb{R} \to \mathbb{R}$ is called *non-degenerate* if the following hold:

(1) ϱ is Borel measurable;
(2) ϱ is locally bounded, that is, ϱ is bounded on $[-R, R]$ for each $R > 0$;
(3) there is a closed null-set $A \subset \mathbb{R}$ such that ϱ is continuous at every $x_0 \in \mathbb{R} \setminus A$;
(4) there does not exist a polynomial $p : \mathbb{R} \to \mathbb{R}$ such that $\varrho(x) = p(x)$ for almost all $x \in \mathbb{R}$. ◀

Remark. A *continuous* activation function is non-degenerate if and only if it is not a polynomial. ♦

These are precisely the assumptions imposed on the activation function in [43], where the following version of the universal approximation theorem is shown:

Theorem 3.22 ([43, Theorem 1])**.** Let $\varrho : \mathbb{R} \to \mathbb{R}$ be a non-degenerate activation function, $K \subset \mathbb{R}^d$ be compact, $\varepsilon > 0$, and $f : \mathbb{R}^d \to \mathbb{R}$ be continuous. Then there is $N \in \mathbb{N}$ and suitable $b_j, c_j \in \mathbb{R}$, $w_j \in \mathbb{R}^d$, $1 \leq j \leq N$ such that $g : \mathbb{R}^d \to \mathbb{R}, x \mapsto \sum_{j=1}^N c_j \, \varrho(\langle w_j, x\rangle + b_j)$ satisfies $\|f - g\|_{L_\infty(K)} \leq \varepsilon$. ◀

We prove in Appendix B.4 that Property (P5) holds under appropriate assumptions:

Theorem 3.23 (Density)**.** Consider $\varrho : \mathbb{R} \to \mathbb{R}$ a Borel measurable, locally bounded activation function, $\mathscr{L}$ a depth growth function, and $p \in (0, \infty]$. Set $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n) \in \mathbb{N} \cup \{+\infty\}$.

(1) Let $\Omega \subset \mathbb{R}^d$ be a *bounded* L_p-domain, and assume that $L \geq 2$.
   (a) For $p \in (0, \infty)$ we have $\mathtt{NN}^{\varrho,d,k}_{\infty,\infty,\infty}(\Omega) \subset X^k_p(\Omega)$;
   (b) For $p = \infty$ the same holds if ϱ is continuous;
   (c) For $p \in (0, \infty)$, if ϱ is non-degenerate then $\Sigma_\infty(X^k_p(\Omega), \varrho, \mathscr{L})$ is dense in $X^k_p(\Omega)$;
   (d) For $p = \infty$, the same holds if ϱ is non-degenerate and continuous.

(2) Assume that the L_p-closure of $\mathtt{NN}^{\varrho,d,1}_{\infty,L,\infty} \cap X_p(\mathbb{R}^d)$ contains a function $g : \mathbb{R}^d \to \mathbb{R}$ such that:
   (a) There is a non-increasing function $\mu : [0, \infty) \to [0, \infty)$ satisfying $\int_{\mathbb{R}^d} \mu(|x|)\, dx < \infty$ and furthermore $|g(x)| \leq \mu(|x|)$ for all $x \in \mathbb{R}^d$.
   (b) $\int_{\mathbb{R}^d} g(x)\, dx \neq 0$; note that this integral is well-defined, since $\int_{\mathbb{R}^d} |g(x)|\, dx \leq \int_{\mathbb{R}^d} \mu(|x|)\, dx < \infty$.

   Then $\Sigma_\infty(X^k_p(\Omega), \varrho, \mathscr{L})$ is dense in $X^k_p(\Omega)$ for every L_p-domain $\Omega \subseteq \mathbb{R}^d$ and every $k \in \mathbb{N}$. ◀

Remark. Claim (2) applies to any L_p-domain, bounded or not. Furthermore, it should be noted that the first assumption (the existence of μ) is always satisfied if g is bounded and has compact support. ♦

Corollary 3.24. *Property (P5) holds for any* bounded L_p*-domain* $\Omega \subset \mathbb{R}^d$ *and* $p \in (0, \infty]$ *as soon as* $\sup_n \mathscr{L}(n) \geq 2$ *and* ϱ *is continuous and not a polynomial.* ◀

Corollary 3.25. *Property (P5) holds for any* (even unbounded) L_p*-domain* $\Omega \subseteq \mathbb{R}^d$ *and* $p \in (0, \infty]$ *as soon as* $L := \sup_n \mathscr{L}(n) \geq 2$ *and as long as* ϱ *is continuous and such that* $\mathtt{NN}^{\varrho,d,1}_{\infty,L,\infty}$ *contains a* compactly supported, bounded, non-negative *function* $g \neq 0$. ◀

In Section 4, we show that the assumptions of Corollary 3.25 indeed hold when ϱ is the ReLU or one of its powers, provided $L \geq 3$ (or $L \geq 2$ in input dimension $d = 1$). This is a consequence of the following lemma, whose proof we defer to Appendix B.5.

Lemma 3.26. *Consider* $\varrho : \mathbb{R} \to \mathbb{R}$ *and* $W, N, L \in \mathbb{N}$. *Assume there is* $\sigma \in \mathtt{NN}^{\varrho,1,1}_{W,L,N}$ *such that*

$$\sigma(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ 1, & \text{if } x \geq 1 \end{cases} \qquad \text{and} \qquad 0 \leq \sigma(x) \leq 1 \quad \forall\, x \in \mathbb{R}\,. \tag{3.9}$$

Then the following hold:

(1) For $d \in \mathbb{N}$ *and* $0 < \varepsilon < \frac{1}{2}$ *there is* $h \in \mathtt{NN}^{\varrho,d,1}_{2dW(N+1),2L-1,(2d+1)N}$ *with* $0 \leq h \leq 1$, $\operatorname{supp}(h) \subset [0,1]^d$, *and*

$$|h(x) - \mathbb{1}_{[0,1]^d}(x)| \leq \mathbb{1}_{[0,1]^d \setminus [\varepsilon, 1-\varepsilon]^d}(x) \quad \forall\, x \in \mathbb{R}^d\,. \tag{3.10}$$

For input dimension $d = 1$, *this holds for some* $h \in \mathtt{NN}^{\varrho,1,1}_{2W,L,2N}$.

(2) There is $L' \leq 2L - 1$ *(resp.* $L' \leq L$ *for input dimension* $d = 1$*) such that for each hyper-rectangle* $[a, b] := \prod_{i=1}^d [a_i, b_i]$ *with* $d \in \mathbb{N}$ *and* $-\infty < a_i < b_i < \infty$, *each* $p \in (0, \infty)$, *and each* $\varepsilon > 0$, *there is a compactly supported, nonnegative function* $0 \leq g \leq 1$ *such that* $\operatorname{supp}(g) \subset [a, b]$,

$$\|g - \mathbb{1}_{[a,b]}\|_{L_p(\mathbb{R}^d)} < \varepsilon,$$

and $g = \mathtt{R}(\Phi)$ *for some* $\Phi \in \mathcal{NN}^{\varrho,d,1}_{2dW(N+1),L',(2d+1)N}$ *with* $L(\Phi) = L'$. *For input dimension* $d = 1$, *this holds for some* $\Phi \in \mathcal{NN}^{\varrho,1,1}_{2W,L',2N}$ *with* $L(\Phi) = L'$. ◀

With the elements established so far, we immediately get the following theorem.

Theorem 3.27. Consider $\varrho : \mathbb{R} \to \mathbb{R}$ an activation function, $\mathscr{L}$ a depth growth function, $d \in \mathbb{N}$, $p \in (0, \infty]$ and $\Omega \subseteq \mathbb{R}^d$ an L_p-domain. Set $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n) \in \mathbb{N} \cup \{+\infty\}$. Assume that at least one of the following properties holds:

(1) ϱ is continuous and not a polynomial, $L \geq 2$, *and* Ω *is bounded*;
(2) $\mathtt{NN}^{\varrho,d,1}_{\infty,L,\infty} \cap X_p(\mathbb{R}^d)$ contains some compactly supported, bounded, non-negative $g \neq 0$.

Then for every $k \in \mathbb{N}$, $\alpha > 0$, $q \in (0, \infty]$, and with $X = X^k_p(\Omega)$ as in Equation (1.3), we have:

- Properties (P1)–(P5) are satisfied for $\Sigma_n = \mathtt{W}_n(X, \varrho, \mathscr{L})$ (resp. for $\Sigma_n = \mathtt{N}_n(X, \varrho, \mathscr{L})$);
- $\big(W^\alpha_q(X, \varrho, \mathscr{L}), \|\cdot\|_{W^\alpha_q(X,\varrho,\mathscr{L})}\big)$ and $\big(N^\alpha_q(X, \varrho, \mathscr{L}), \|\cdot\|_{N^\alpha_q(X,\varrho,\mathscr{L})}\big)$ are (quasi)-Banach spaces. ◀

In particular, if ϱ is continuous and satisfies the assumptions of Lemma 3.26 for some $L \in \mathbb{N}$ and if $\sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq 2L-1$ (or $\sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq L$ in case of $d=1$), then the conclusions of Theorem 3.27 hold on *any* L_p-domain.

3.7. **Discussion and perspectives.** One could envision defining approximation classes where the sets Σ_n incorporate additional constraints besides $L \leq \mathscr{L}(n)$. For the theory to hold, one must however ensure either that: a) the additional constraints are weak enough to ensure the approximation errors (and therefore the approximation spaces) are unchanged—cf. the discussion of strict *vs* generalized networks; or, more interestingly, that b) the constraint gets sufficiently relaxed when n grows, to ensure compatibility with the additivity property.

As an example, constraints of potential interest include a lower (resp. upper) bound on the minimum width $\min_{1\leq\ell\leq L-1} N_\ell$ (resp. maximum width $\max_{1\leq\ell\leq L-1} N_\ell$), since they impact the memory needed to compute "in place" the output of the network.

While network families with a fixed lower bound on their minimum width do satisfy the additivity Property (P4), this is no longer the case of families with a *fixed* upper bound on their *maximum* width. Consider now a complexity-dependent upper bound $f(n)$ for the maximum width. Since "adding" two networks of a given width yields one with width at most doubled, the additivity property will be preserved provided that $2f(n) \leq f(cn)$ for some $c \in \mathbb{N}$ and all $n \in \mathbb{N}$. This can, e.g., be achieved with $f(n) := \lfloor \alpha n \rfloor$, with the side effect that for $n < 1/\alpha$ the set Σ_n only contains affine functions.

4. Approximation spaces of the ReLU and its powers

The choice of activation function has a decisive influence on the approximation spaces $W_q^\alpha(X, \varrho, \mathscr{L})$ and $N_q^\alpha(X, \varrho, \mathscr{L})$. As evidence of this, consider the following result.

Theorem 4.1 ([45, Theorem 4])**.** There exists an *analytic* squashing function[2] $\varrho : \mathbb{R} \to \mathbb{R}$ such that: for any $d \in \mathbb{N}$, any continuous function from $\Omega = [0,1]^d$ to $\mathbb{R}$ can be approximated arbitrarily well in the uniform norm by a strict ϱ-network with $L = 3$ layers and $W \leq 21d^2 + 15d + 3$ connections. ◀

Consider the pathological activation function ϱ from Theorem 4.1 and a depth growth function $\mathscr{L}$ satisfying $L := \sup_n \mathscr{L}(n) \geq 2$. Since ϱ is continuous and not a polynomial, we can apply Theorem 3.27; hence $W_q^\alpha(X, \varrho, \mathscr{L})$ and $N_q^\alpha(X, \varrho, \mathscr{L})$ are well defined quasi-Banach spaces for each bounded L_p-domain Ω, $p \in (0, \infty]$ and $X = X_p^k(\Omega)$. Yet, if $L \geq 3$ there is n_0 so that $\mathscr{L}(n) \geq 3$ for $n \geq n_0$, and the set $\Sigma_n(X, \varrho, \mathscr{L})$ is dense in X for any $p \in (0, \infty]$ provided that $n \geq \max\{n_0, 21d^2 + 15d + 3\}$; hence $E(f, \Sigma_n(X, \varrho, \mathscr{L}))_X = 0$ for any $f \in X$ and any such n, showing that $W_q^\alpha(X, \varrho, \mathscr{L}) = N_q^\alpha(X, \varrho, \mathscr{L}) = X$ with equivalent (quasi)-norms.

The approximation spaces generated by pathological activation functions such as in Theorem 4.1 are so degenerate that they are uninteresting both from a practical perspective (computing a near best approximation with such an activation function is hopeless) and from a theoretical perspective (the whole scale of approximation spaces collapses to $X_p^k(\Omega)$).

Much more interesting is the study of approximation spaces generated by commonly used activation functions such as the ReLU ϱ_1 or its powers ϱ_r, $r \in \mathbb{N}$. For any L_p-domain, generalized and strict ϱ_r-networks indeed yield well-defined approximations spaces that coincide.

Theorem 4.2 (Approximation spaces of generalized and strict ϱ_r-networks)**.** Let $r \in \mathbb{N}$ and define $\varrho_r : \mathbb{R} \to \mathbb{R}, x \mapsto (x_+)^r$, where $x_+ := \max\{0, x\}$. Consider $X := X_p^k(\Omega)$ with $p \in (0, \infty]$, $d, k \in \mathbb{N}$ and $\Omega \subseteq \mathbb{R}^d$ an arbitrary L_p-domain. Let $\mathscr{L}$ be any depth growth function.

(1) For each $\alpha > 0, q \in (0, \infty]$, $r \in \mathbb{N}$ we have

$$SW_q^\alpha(X, \varrho_r, \mathscr{L}) = W_q^\alpha(X, \varrho_r, \mathscr{L}) \quad \text{and} \quad SN_q^\alpha(X, \varrho_r, \mathscr{L}) = N_q^\alpha(X, \varrho_r, \mathscr{L})$$

and there is $C < \infty$ such that

$$\|\cdot\|_{W_q^\alpha(X,\varrho_r,\mathscr{L})} \leq \|\cdot\|_{SW_q^\alpha(X,\varrho_r,\mathscr{L})} \leq C\|\cdot\|_{W_q^\alpha(X,\varrho_r,\mathscr{L})}\,,$$
$$\|\cdot\|_{N_q^\alpha(X,\varrho_r,\mathscr{L})} \leq \|\cdot\|_{SN_q^\alpha(X,\varrho_r,\mathscr{L})} \leq C\|\cdot\|_{N_q^\alpha(X,\varrho_r,\mathscr{L})}\,.$$

(2) If the depth growth function $\mathscr{L}$ satisfies

$$\sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq \begin{cases} 2, & \text{if } \Omega \text{ is bounded } \textit{or } d = 1 \\ 3, & \text{otherwise} \end{cases}$$

then, for each $\alpha > 0$, $q \in (0, \infty]$, $r \in \mathbb{N}$ and $\varrho := \varrho_r$, the following hold:

[2] A function $\sigma : \mathbb{R} \to \mathbb{R}$ is a *squashing function* if it is nondecreasing with $\lim_{x\to-\infty} \sigma(x) \to 0$ and $\lim_{x\to\infty} \sigma(x) \to 1$; see [36, Definition 2.3].

- Properties (P1)–(P5) are satisfied for $\Sigma_n = \mathtt{W}_n(X, \varrho, \mathscr{L})$ (resp. for $\Sigma_n = \mathtt{N}_n(X, \varrho, \mathscr{L})$);
- $\big(W_q^\alpha(X, \varrho, \mathscr{L}), \|\cdot\|_{W_q^\alpha(X,\varrho,\mathscr{L})}\big)$ and $\big(N_q^\alpha(X, \varrho, \mathscr{L}), \|\cdot\|_{N_q^\alpha(X,\varrho,\mathscr{L})}\big)$ are (quasi)-Banach spaces. ◀

Remark 4.3. For a bounded domain or when $d = 1$, the second claim holds for any depth growth function allowing at least one hidden layer. In the other cases, the restriction to at least two hidden layers is unavoidable (except for some exotic unbounded domains with vanishing mass at infinity) as the only realization of a ϱ_r-network of depth two that belongs to $X_p(\mathbb{R}^d)$ is the zero network. ⧫

Proof of Theorem 4.2. By Lemma 2.24, ϱ_r can represent the identity using $2r + 2$ terms. By Theorem 3.8, this establishes the first claim. The second claim follows from Theorem 3.27, once we that show we can apply the latter. For bounded Ω, this is clear, since ϱ_r is continuous and not a polynomial, and hence non-degenerate. For general Ω, we relate ϱ_r to B-splines to establish the following lemma (which we prove below).

Lemma 4.4. *For any $r \in \mathbb{N}$ there is $\sigma_r \in \mathtt{SNN}^{\varrho_r,1,1}_{2(r+1),2,r+1}$ satisfying* (3.9). ◀

Combined with Lemma 3.26, we obtain the existence of a compactly supported, continuous, non-negative function $g \neq 0$ such that $g \in \mathtt{NN}^{\varrho_r,d,1}_{\infty,3,\infty} \cap X_p(\mathbb{R}^d)$ (respectively $g \in \mathtt{NN}^{\varrho_r,d,1}_{\infty,2,\infty} \cap X_p(\mathbb{R})$ for input dimension $d = 1$). Hence, Theorem 3.27 is applicable. □

Definition 4.5 (B-splines)**.** For any function $f : \mathbb{R} \to \mathbb{R}$, define $\Delta f : x \mapsto f(x) - f(x-1)$. Let $\varrho_0 := \mathbb{1}_{[0,\infty)}$ denote the Heaviside function, and $\beta_+^{(0)} := \mathbb{1}_{[0,1)} = \Delta\varrho_0$ the B-spline of degree 0. The B-spline of degree n is obtained by convolving $\beta_+^{(0)}$ with itself $n + 1$ times:

$$\beta_+^{(n)} := \underbrace{\beta_+^{(0)} \star \ldots \star \beta_+^{(0)}}_{n+1 \text{ factors}}.$$

For $n \geq 0$, $\beta_+^{(n)}$ is non-negative and is zero except for $x \in [0, n + 1]$. We have $\beta_+^{(n)} \in C_c^{n-1}(\mathbb{R})$ for $n \geq 1$. Indeed, this follows since $\varrho_n \in C^{n-1}(\mathbb{R})$, and since it is known (see [65, Equation (10)], noting that [65] uses *centered* B-splines) that the B-spline of degree n can be decomposed as

$$\beta_+^{(n)} = \frac{\Delta^{n+1}\varrho_n}{n!} = \frac{1}{n!}\sum_{k=0}^{n+1}\binom{n+1}{k}(-1)^k\varrho_n(x-k). \tag{4.1}$$

◀

Proof of Lemma 4.4. For $n \geq 0$, $\beta_+^{(n)}$ is non-negative and is zero except for $x \in [0, n + 1]$. Its primitive

$$g_n(x) := \int_0^x \beta_+^{(n)}(t)dt$$

is thus non-decreasing, with $g_n(x) = 0$ for $x \leq 0$ and $g_n(x) = g_n(n+1)$ for $x \geq n+1$. Since $\beta_+^{(n)} \in C_c^{n-1}(\mathbb{R})$ for $n \geq 1$, we have $g_n \in C^n(\mathbb{R})$ for $n \geq 1$. Furthermore, $g_0 \in C^0(\mathbb{R})$ since $\beta_+^{(0)}$ is bounded.

For $r \geq 1$, the above facts imply that the function $\sigma_r(x) := g_{r-1}(rx)/g_{r-1}(r)$ belongs to $C^{r-1}(\mathbb{R})$ and satisfies (3.9). To conclude, we now prove that $\sigma_r \in \mathtt{SNN}^{\varrho_r,1,1}_{2(r+1),2,r+1}$. For $0 \leq k \leq n + 1$ we have

$$\int_0^x \varrho_n(t-k)dt = \begin{cases} 0, & \text{if } x \leq k \\ \int_k^x (t-k)^n dt = \int_0^{x-k} t^n dt = \frac{(x-k)^{n+1}}{n+1}, & \text{otherwise} \end{cases} = \frac{\varrho_{n+1}(x-k)}{n+1}.$$

By (4.1) it follows that

$$g_n(x) = \frac{1}{(n+1)!}\sum_{k=0}^{n+1}\binom{n+1}{k}(-1)^k\varrho_{n+1}(x-k),$$

and hence

$$\sigma_r(x) = \frac{g_{r-1}(rx)}{g_{r-1}(r)} = \frac{1}{r!\ g_{r-1}(r)}\sum_{k=0}^{r}\binom{r}{k}(-1)^k\varrho_r(rx-k).$$

Setting $\alpha_1 := \varrho_r \otimes \ldots \otimes \varrho_r : \mathbb{R}^{r+1} \to \mathbb{R}^{r+1}$ as well as $T_1 : \mathbb{R} \to \mathbb{R}^{r+1}, x \mapsto (rx - k)_{k=0}^r$ and

$$T_2 : \mathbb{R}^{r+1} \to \mathbb{R}, y = (y_k)_{k=0}^r \mapsto \tfrac{1}{r!\ g_{r-1}(r)}\sum_{k=0}^{r}\binom{r}{k}(-1)^k y_k$$

and $\Phi := \big((T_1, \alpha_1), (T_2, \mathrm{id}_{\mathbb{R}})\big)$, it is then easy to check that $\sigma_r = \mathtt{R}(\Phi)$. Obviously $L(\Phi) = 2$, $N(\Phi) = r + 1$, and $\|T_i\|_{\ell^0} = r + 1$ for $i = 1, 2$, hence as Φ is strict we have $\Phi \in \mathcal{SNN}^{\varrho_r,1,1}_{2(r+1),2,r+1}$. □

4.1. **Piecewise polynomial activation functions *vs.* ϱ_r.** In this subsection, we show that approximation spaces of ϱ_r-networks contain the approximation spaces of continuous piecewise polynomial activation functions, and match those of (free-knot) spline activation functions.

Definition 4.6. Consider an interval $I \subseteq \mathbb{R}$. A function $f : I \to \mathbb{R}$ is *piecewise polynomial* if there are *finitely* many intervals $I_i \subset I$ such that $I = \bigcup_i I_i$ and $f|_{I_i}$ is a polynomial. It is *of degree at most* $r \in \mathbb{N}$ when each $f|_{I_i}$ is of degree at most r, and *with at most* $n \in \mathbb{N}$ *pieces* (or *with at most* $n-1 \in \mathbb{N}_0$ *breakpoints*) when there are at most n such intervals. The set of piecewise polynomials of degree at most r with at most n pieces is denoted $\mathtt{PPoly}_n^r(I)$, and we set $\mathtt{PPoly}^r(I) := \cup_{n\in\mathbb{N}}\mathtt{PPoly}_n^r(I)$.

A function $f \in \mathtt{Spline}_n^r(I) := \mathtt{PPoly}_n^r(I) \cap C^{r-1}(I)$ is called a *free-knot spline* of degree at most r with at most n pieces (or at most $n-1$ breakpoints). We set $\mathtt{Spline}^r(I) := \cup_{n\in\mathbb{N}}\mathtt{Spline}_n^r(I)$. ◀

Theorem 4.7. Consider a depth growth function $\mathscr{L}$, an L_p domain $\Omega \subset \mathbb{R}^d$, and let $X = X_p^k(\Omega)$ with $d, k \in \mathbb{N}$, $p \in (0,\infty]$. Let $r \in \mathbb{N}$, set $\varrho_r : \mathbb{R} \to \mathbb{R}, x \mapsto (x_+)^r$, and let $\alpha > 0$, $q \in (0,\infty]$.

(1) If $\varrho : \mathbb{R} \to \mathbb{R}$ is continuous and piecewise polynomial of degree at most r then,

$$W_q^\alpha(X,\varrho,\mathscr{L}) \hookrightarrow \begin{cases} W_q^\alpha(X,\varrho_r,\max(\mathscr{L}+1,2)), & \text{if } d=1 \\ W_q^\alpha(X,\varrho_r,\max(\mathscr{L}+1,3)), & \text{if } d\geq 2. \end{cases} \tag{4.2}$$

Moreover if Ω is bounded, or if $r=1$, or if $\mathscr{L}+1 \preceq \mathscr{L}$, then we further have

$$W_q^\alpha(X,\varrho,\mathscr{L}) \hookrightarrow W_q^\alpha(X,\varrho_r,\mathscr{L}). \tag{4.3}$$

(2) If $\varrho \in \mathtt{Spline}^r(\mathbb{R})$ is not a polynomial and Ω is bounded, then we have (with equivalent norms)

$$W_q^\alpha(X,\varrho,\mathscr{L}) = W_q^\alpha(X,\varrho_r,\mathscr{L}). \tag{4.4}$$

(3) For any $s \in \mathbb{N}$ we have

$$W_q^\alpha(X,\varrho_{r^s},\mathscr{L}) \hookrightarrow W_q^\alpha(X,\varrho_r,1+s(\mathscr{L}-1)). \tag{4.5}$$

The same results hold with $N_q^\alpha(X,\cdot,\cdot)$ instead of $W_q^\alpha(X,\cdot,\cdot)$. ◀

Examples 3.15 and 3.16 provide important examples of depth growth functions $\mathscr{L}$ with $\mathscr{L}+1 \preceq \mathscr{L}$, so that (4.3) holds on any domain.

Remark 4.8 (Nestedness). For $1 \leq r' \leq r$, the function $\varrho := \varrho_{r'}$ is indeed a continuous piecewise polynomial with two pieces of degree at most r. Theorem 4.7 thus implies that if Ω is bounded or $\mathscr{L}+1 \preceq \mathscr{L}$, then $W_q^\alpha(X,\varrho_{r'},\mathscr{L}) \hookrightarrow W_q^\alpha(X,\varrho_r,\mathscr{L})$ and $N_q^\alpha(X,\varrho_{r'},\mathscr{L}) \hookrightarrow N_q^\alpha(X,\varrho_r,\mathscr{L})$. We will see in Corollary 4.14 below that if $2\mathscr{L} \preceq \mathscr{L}$, then these embeddings are indeed equalities if $2 \leq r' \leq r$. ♦

The main idea behind the proof of Theorem 4.7 given below is to combine Lemma 2.19 and its consequences with the following results proved in Appendices C.1–C.2.

Lemma 4.9. *Consider $\varrho : \mathbb{R} \to \mathbb{R}$ a continuous piecewise polynomial function with at most $n \in \mathbb{N}$ pieces of degree at most $r \in \mathbb{N}$. With*[3] *$w := 2\cdot(4^r-1)/3$ and $m := 2^r - 1$ we have*

$$\varrho \in \overline{\mathtt{NN}^{\varrho_r,1,1}_{4(r+1)+(n-1)w,2,2(r+1)+(n-1)m}},$$

where the closure is with respect to the topology of locally uniform convergence. For $r=1$ (that is, when ϱ is continuous and piecewise affine with at most $n \in \mathbb{N}$ pieces and $\varrho_r = \varrho_1$), we even have $\varrho \in \mathtt{SNN}^{\varrho_r,1,1}_{2(n+1),2,n+1}$. ◀

Lemma 4.10. *Consider $r \in \mathbb{N}$ and $\varrho \in \mathtt{Spline}^r(\mathbb{R})$. If ϱ is not a polynomial then $\varrho_r \in \overline{\mathtt{NN}^{\varrho,1,1}_{5^r r!,2,3^r r!}}$, where the closure is with respect to locally uniform convergence.* ◀

For bounded Ω, locally uniform convergence on $\mathbb{R}^d$ implies convergence in $X = X_p^k(\Omega)$ for all $p \in (0,\infty]$. To similarly "upgrade" locally uniform convergence to convergence in X on unbounded domains, we use the following localization lemma which is proved in Appendix C.3.

Lemma 4.11. *Consider $d, k \in \mathbb{N}$, $r \in \mathbb{N}_{\geq 2}$. There is $c = c(d,k,r) \in \mathbb{N}$ such that*[4] *for any $W, L, N \in \mathbb{N}$, $g \in \mathtt{NN}^{\varrho_r,d,k}_{W,L,N}$, $R \geq 1, \delta > 0$, there is $g_{R,\delta} \in \mathtt{NN}^{\varrho_r,d,k}_{cW,\max\{L+1,3\},cN}$, such that*

$$|g_{R,\delta}(x) - (\mathbb{1}_{[-R,R]^d}\cdot g)(x)| \leq 2\cdot|g(x)|\cdot\mathbb{1}_{[-R-\delta,R+\delta]^d\setminus[-R,R]^d}(x) \qquad \forall\, x \in \mathbb{R}^d. \tag{4.6}$$

For $d=1$ the same holds with $\max\{L+1,2\}$ layers instead of $\max\{L+1,3\}$. ◀

[3] Note that $4 = 1 \mod 3$ and hence $4^n - 1 = 0 \mod 3$, so that $w \in \mathbb{N}$.

[4] Notice the restriction to $W, N \geq 1$; in fact, the result of Lemma 4.11 as stated cannot hold for $W = 0$ or $N = 0$.

The following proposition describes how one can "upgrade" the locally uniform convergence to convergence in $X_p(\Omega)$, at the cost of slightly increasing the depth of the approximating networks.

Proposition 4.12. *Consider $\Omega \subset \mathbb{R}^d$ an L_p-domain and $X = X_p^k(\Omega)$ with $d, k \in \mathbb{N}$, $p \in (0, \infty]$. Assume $\varrho \in \overline{\mathtt{NN}_{\infty,2,m}^{\varrho_r,1,1}}$ where the closure is with respect to locally uniform convergence and $r \in \mathbb{N}_{\geq 2}$, $m \in \mathbb{N}$. For any $W, N \in \mathbb{N}_0 \cup \{\infty\}$, $L \in \mathbb{N} \cup \{\infty\}$ we have, with closure in X,*

$$\mathtt{NN}_{W,L,N}^{\varrho,d,k}(\Omega) \cap X \subset \overline{\mathtt{NN}_{cWm^2,\max\{L+1,3\},cNm}^{\varrho_r,d,k}(\Omega) \cap X}^{X},$$

where $c = c(d, k, r) \in \mathbb{N}$ is as in Lemma 4.11. If $d = 1$ the same holds with $\max\{L + 1, 2\}$ layers instead of $\max\{L + 1, 3\}$. If Ω is bounded, or if $\varrho \in \mathtt{NN}_{\infty,2,m}^{\varrho_r,1,1}$ with $r = 1$, then the same holds with $c = 1$ and L layers instead of $\max\{L + 1, 3\}$ (resp. instead of $\max\{L + 1, 2\}$ when $d = 1$). ◀

The proof is in Appendix C.4. We are now equipped to prove Theorem 4.7.

Proof of Theorem 4.7. We give the proof for $W_q^\alpha(X, \cdot, \mathscr{L})$; minor adaptations yield the results for $N_q^\alpha(X, \cdot, \mathscr{L})$.

For Claim (1), first note that Lemma 4.9 shows that there is some $m \in \mathbb{N}$ satisfying $\varrho \in \overline{\mathtt{NN}_{\infty,2,m}^{\varrho_r,1,1}}$, where the closure is with respect to locally uniform convergence. Define $\ell := 3$ if $d \geq 2$ (resp. $\ell := 2$ if $d = 1$) and $\widetilde{\mathscr{L}} := \max\{\mathscr{L} + 1, \ell\}$ (resp. $\widetilde{\mathscr{L}} := \mathscr{L}$ when Ω is bounded or $r = 1$) and consider $c \in \mathbb{N}$ as in Proposition 4.12. Thus, since $\widetilde{\mathscr{L}}$ is non-decreasing, by Proposition 4.12 and Lemma 2.14 we have for all $n \in \mathbb{N}$

$$\begin{aligned}\mathtt{W}_n(X, \varrho, \mathscr{L}) = \mathtt{NN}_{n,\mathscr{L}(n),\infty}^{\varrho,d,k}(\Omega) \cap X &\subset \overline{\mathtt{NN}_{cnm^2,\widetilde{\mathscr{L}}(n),\infty}^{\varrho_r,d,k}(\Omega) \cap X}^{X} \\ &\subset \overline{\mathtt{NN}_{cnm^2,\widetilde{\mathscr{L}}(cnm^2),\infty}^{\varrho_r,d,k}(\Omega) \cap X}^{X} = \overline{\mathtt{W}_{cm^2n}(X, \varrho_r, \widetilde{\mathscr{L}})}^{X}.\end{aligned}$$

Hence, for any $f \in X$ and $n \in \mathbb{N}$

$$E(f, \mathtt{W}_n(X, \varrho, \mathscr{L}))_X \geq E\big(f, \mathtt{W}_{cm^2n}(X, \varrho_r, \widetilde{\mathscr{L}})_X\big).$$

Thus, Lemma 3.1 yields (4.2). When Ω is bounded or $r = 1$, as $\widetilde{\mathscr{L}} = \mathscr{L}$, this yields (4.3). When $\mathscr{L} + 1 \preceq \mathscr{L}$, as $\widetilde{\mathscr{L}} \leq \max\{\mathscr{L} + 1, \ell\} \leq \mathscr{L} + \ell + 1$, we have $\widetilde{\mathscr{L}} \preceq \mathscr{L} + \ell + 1 \preceq \mathscr{L}$ by Lemma 3.14, yielding again (4.3) by Lemma 3.12.

For Claim (2), if Ω is bounded and $\varrho \in \mathtt{Spline}^r(\mathbb{R})$ is not a polynomial, combining Lemma 4.10 with Lemma 2.21, we similarly get the converse to (4.3). This establishes (4.4).

We now prove Claim (3). Since $\varrho_{r^s} = \varrho_r \circ \cdots \circ \varrho_r$ (where ϱ appears s times), Lemma 2.20 shows that $\mathtt{NN}_{W,L,N}^{\varrho_{r^s},d,k} \subset \mathtt{NN}_{W+(s-1)N,1+s(L-1),sN}^{\varrho_r,d,k}$ for all W, L, N. Combining this with Lemma 2.14, we obtain

$$\mathtt{NN}_{n,\mathscr{L}(n),\infty}^{\varrho_{r^s},d,k} \subset \mathtt{NN}_{n,\mathscr{L}(n),n}^{\varrho_{r^s},d,k} \subset \mathtt{NN}_{sn,1+s(\mathscr{L}(n)-1),sn}^{\varrho_r,d,k} \subset \mathtt{NN}_{sn,1+s(\mathscr{L}(sn)-1),\infty}^{\varrho_r,d,k} \quad \forall\, n \in \mathbb{N}.$$

Therefore, we get for any $f \in X$ and $n \in \mathbb{N}$

$$E(f, \mathtt{W}_n(X, \varrho_{r^s}, \mathscr{L}))_X \geq E\big(f, \mathtt{W}_{sn}(X, \varrho_r, 1 + s(\mathscr{L} - 1))_X\big).$$

Hence, we can finally apply Lemma 3.1 to obtain (4.5). □

Remark 4.13. Inspecting the proofs, we see that if $\varrho \in \mathtt{Spline}^r$ has exactly one breakpoint then $\varrho \in \mathtt{NN}_{w,2,m}^{\varrho_r,1,1}$ and $\varrho_r \in \mathtt{NN}_{w,2,m}^{\varrho,1,1}$ for some $w, m \in \mathbb{N}$. This is stronger than $\varrho \in \overline{\mathtt{NN}_{w,2,m}^{\varrho_r,1,1}}$ (resp. than $\varrho_r \in \overline{\mathtt{NN}_{w,2,m}^{\varrho,1,1}}$) and implies (4.4) with equivalent norms *even on unbounded domains*. Examples include the leaky ReLU [44], the parametric ReLU [33], and the absolute value which is used in scattering transforms [46].

Another spline of degree one is soft-thresholding, $\sigma(x) := x(1 - \lambda/|x|)_+$, which appears in Iterative Shrinkage Thresholding Algorithms (ISTA) for ℓ^1 sparse recovery in the context of linear inverse problems [28, Chap. 3] and has been used in the Learned ISTA (LISTA) method [30]. As $\sigma \in \mathtt{Spline}^1$, using soft-thresholding as an activation function on bounded Ω is exactly as expressive as using the ReLU. ♦

4.2. **Saturation property of approximation spaces with polynomial depth growth.** For certain depth growth functions, the approximation spaces of ϱ_r-networks are independent of the choice of $r \geq 2$.

Corollary 4.14. *With the notations of Theorem 4.7, if $2\mathscr{L} \preceq \mathscr{L}$ then for every $r \in \mathbb{N}_{\geq 2}$ we have*

$$W_q^\alpha(X, \varrho_1, \mathscr{L}) \hookrightarrow W_q^\alpha(X, \varrho_2, \mathscr{L}) = W_q^\alpha(X, \varrho_r, \mathscr{L}), \tag{4.7}$$

$$N_q^\alpha(X, \varrho_1, \mathscr{L}) \hookrightarrow N_q^\alpha(X, \varrho_2, \mathscr{L}) = N_q^\alpha(X, \varrho_r, \mathscr{L}), \tag{4.8}$$

where the equality is with equivalent quasi-norms. ◀

Example 4.15. *By Example 3.16, for polynomially growing depth we do have $2\mathscr{L} \preceq \mathscr{L}$. This includes the case $\mathscr{L}(n) = n + 1$, which gives the same approximation spaces as $\mathscr{L} \equiv \infty$; see Remark 3.6.*

In words, approximation spaces of ϱ_r-networks with appropriate depth growth have a saturation property: increasing the degree r beyond $r = 2$ does not pay off in terms of the considered function spaces. Note, however, that the constants in the norm equivalence may still play a qualitative role in practice.

Proof. We prove (4.7), the proof of (4.8) is similar. By Lemma 3.14, since $2\mathscr{L} \preceq \mathscr{L}$ we have $a\mathscr{L} + b \sim \mathscr{L}$ for all $a \geq 1$, $b \geq 1 - a$. In particular, $\mathscr{L} + 1 \preceq \mathscr{L}$ hence (4.3) holds with $\varrho = \varrho_{r'}$, $r' \in \mathbb{N}$, $1 \leq r' \leq r$. Combined with (4.5) and Lemma 3.12, since $r \leq 2^r$ for $r \in \mathbb{N}$ we see

$$W_q^\alpha(X, \varrho_r, \mathscr{L}) \hookrightarrow W_q^\alpha(X, \varrho_{2^r}, \mathscr{L}) \hookrightarrow W_q^\alpha(X, \varrho_2, 1 + r(\mathscr{L} - 1)) \hookrightarrow W_q^\alpha(X, \varrho_2, \mathscr{L}) \hookrightarrow W_q^\alpha(X, \varrho_r, \mathscr{L})$$

for all $r \in \mathbb{N}_{\geq 2}$. In the middle we used that $1 + r(\mathscr{L} - 1) \preceq 1 + r\mathscr{L} \preceq (1 + r)\mathscr{L} \preceq \mathscr{L}$. □

4.3. **Piecewise polynomial activation functions yield non-trivial approximation spaces.** In light of the pathological example of Theorem 4.1, it is important to check that the approximation spaces $W_q^\alpha(X_p^k(\Omega), \varrho, \mathscr{L})$ and $N_q^\alpha(X_p^k(\Omega), \varrho, \mathscr{L})$ with $\varrho = \varrho_r$, $r \in \mathbb{N}$, are *non-trivial*: they are proper subspaces of $X_p^k(\Omega)$. This is what we prove for any continuous and piecewise polynomial activation function ϱ.

Theorem 4.16. Let $\varrho : \mathbb{R} \to \mathbb{R}$ be continuous and piecewise polynomial (with finitely many pieces), let $\Omega \subset \mathbb{R}^d$ be measurable with nonempty interior, and let $s > 0$. Let $p, q \in (0, \infty]$, $k \in \mathbb{N}$, $\alpha \in (0, \infty)$, and $X = X_p^k(\Omega)$. Finally, let $\mathscr{L}$ be a depth-growth function satisfying $\sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq 2$. Then $W_q^\alpha(X, \varrho, \mathscr{L}) \subsetneq X$ and $N_q^\alpha(X, \varrho, \mathscr{L}) \subsetneq X$. ◀

The proof is given at the end of Appendix E.

4.4. **ReLU-networks of bounded depth have limited expressiveness.** In this subsection, we show that approximation spaces of ReLU-networks *of bounded depth* and high approximation rate α are non-trivial in a very explicit sense: they fail to contain any nonzero function in $C_c^3(\mathbb{R}^d)$. This quite general obstruction to the expressiveness of shallow ReLU-networks, and to the embedding of "classical" function spaces into the approximation spaces of shallow ReLU-networks, is obtained by translating [55, Theorem 4.5] into the language of approximation spaces.

Theorem 4.17. Let $\Omega \subseteq \mathbb{R}^d$ be an *open* L_p-domain, $p, q \in (0, \infty]$, $X = X_p(\Omega)$, $L \in \mathbb{N}$, and $\alpha > 0$.

- If $C_c^3(\Omega) \cap W_q^\alpha(X, \varrho_1, L) \neq \{0\}$ then $\lfloor L/2 \rfloor \geq \alpha/2$;
- If $C_c^3(\Omega) \cap N_q^\alpha(X, \varrho_1, L) \neq \{0\}$ then $L - 1 \geq \alpha/2$. ◀

Before we give a proof we immediately highlight a consequence.

Corollary 4.18. *Let Y be a function space such that $C_c^3(\Omega) \cap Y \neq \{0\}$ where $\Omega \subseteq \mathbb{R}^d$ is an* open *L_p-domain. For $p \in (0, \infty]$, $X = X_p(\Omega)$, $L \in \mathbb{N}$, $\alpha > 0$ and $q \in (0, \infty]$ we have*

- *If $Y \subset W_q^\alpha(X, \varrho_1, L)$ then $\lfloor L/2 \rfloor \geq \alpha/2$;*
- *If $Y \subset N_q^\alpha(X, \varrho_1, L)$ then $L - 1 \geq \alpha/2$.* ◀

Remark. All "classical" function spaces (Sobolev, Besov, or modulation spaces, ...) include $C_c^\infty(\Omega)$, hence this shows that none of these spaces embed into $W_q^\alpha(X, \varrho_1, L)$ (resp. into $N_q^\alpha(X, \varrho_1, L)$) for $\alpha > 2L$. In other words, *to achieve embeddings into approximation spaces of ReLU-networks with a good approximation rate, one needs depth!* ♦

Proof of Theorem 4.17. The claimed estimates are trivially satisfied in case of $L = 1$; hence, we will assume $L \geq 2$ in what follows.

Let $f \in C_c^3(\Omega)$ be not identically zero. We derive necessary criteria on L which have to be satisfied if $f \in W_q^\alpha(X, \varrho_1, L)$ or $f \in N_q^\alpha(X, \varrho_1, L)$. By Equation (3.2), we have $W_q^\alpha(X, \varrho_1, L) \subset W_\infty^\alpha(X, \varrho_1, L)$ and the same for $N_q^\alpha(X, \varrho_1, L)$; thus, it suffices to consider the case $q = \infty$.

Extending f by zero outside Ω, we can assume $f \in C_c^3(\mathbb{R}^d)$ with $\operatorname{supp} f \subset \Omega$. We claim that there is $x_0 \in \operatorname{supp}(f) \subset \Omega$ with $\operatorname{Hess}_f(x_0) \neq 0$, where Hess_f denotes the Hessian of f. If this was false, we would have $\operatorname{Hess}_f \equiv 0$ on all of $\mathbb{R}^d$, and hence $\nabla f \equiv v$ for some $v \in \mathbb{R}^d$. This would imply $f(x) = \langle v, x \rangle + b$ for all $x \in \mathbb{R}^d$, with $b = f(0)$. However since $f \equiv 0$ on the nonempty open set $\mathbb{R}^d \setminus \operatorname{supp}(f)$, this would entail $v = 0$, and then $f \equiv 0$, contradicting our choice of f.

Now, choose $r > 0$ such that $\Omega_0 := B_r(x_0) \subset \Omega$. Then $f|_{\Omega_0}$ is *not* an affine-linear function, so that [55, Proposition C.5] yields a constant $C_1 = C_1(f, p) > 0$ satisfying

$$\|f - g\|_{L^p(\Omega_0)} \geq C_1 \cdot P^{-2} \quad \text{for each } P\text{-piecewise slice affine function } g : \mathbb{R}^d \to \mathbb{R}. \tag{4.9}$$

Here, a function $g : \mathbb{R}^d \to \mathbb{R}$ is called *P-piecewise slice affine* if for arbitrary $x_0, v \in \mathbb{R}^d$ the function $g_{x_0,v} : \mathbb{R} \to \mathbb{R}, t \mapsto g(x_0 + tv)$ is piecewise affine-linear with at most P pieces; that is, $g_{x_0,v} \in \mathtt{PPoly}^1_P(\mathbb{R})$.

Now, Lemma 5.19 (which will be proved independently) shows that there is a constant $K = K(L) \in \mathbb{N}$ such that

$$\mathtt{NN}^{\varrho_1,1,1}_{W,L,\infty} \subset \mathtt{PPoly}^1_{K\cdot W^{\lfloor L/2\rfloor}}(\mathbb{R}) \qquad \text{and} \qquad \mathtt{NN}^{\varrho_1,1,1}_{\infty,L,N} \subset \mathtt{PPoly}^1_{K\cdot N^{L-1}}(\mathbb{R})$$

for all $N \in \mathbb{N}$. Furthermore, if $g \in \mathtt{NN}^{\varrho_1,d,1}_{W,L,N}$, then Lemma 2.18 shows $g_{x_0,v} \in \mathtt{NN}^{\varrho_1,1,1}_{W,L,N}$; here, we used that the affine map $T : \mathbb{R} \to \mathbb{R}^d, t \mapsto x_0 + tv$ satisfies $\|T\|_{\ell^{0,\infty}_*} \leq 1$. In combination, we see that each $g \in \mathtt{NN}^{\varrho_1,d,1}_{W,L,\infty}$ is P-piecewise slice affine with $P = K \cdot W^{\lfloor L/2\rfloor}$, and each $g \in \mathtt{NN}^{\varrho_1,d,1}_{\infty,L,N}$ is P-piecewise slice affine with $P = K \cdot N^{L-1}$.

Now, if $f \in W^\alpha_\infty(X, \varrho_1, L)$, then there is a constant $C_2 = C_2(f, \alpha, p) > 0$ such that for each $n \in \mathbb{N}$ there is $g_n \in \mathtt{NN}^{\varrho_1,d,1}_{n,L,\infty}$ satisfying $\|f - g_n\|_{L^p(\Omega_0)} \leq \|f - g_n\|_X \leq C_2 \cdot n^{-\alpha}$. Furthermore, since g_n is P-piecewise slice affine with $P = K \cdot n^{\lfloor L/2\rfloor}$, Equation (4.9) shows that $K^{-2}C_1 \cdot n^{-2\lfloor L/2\rfloor} \leq \|f - g_n\|_{L^p(\Omega_0)} \leq C_2 \cdot n^{-\alpha}$. Since this holds for all $n \in \mathbb{N}$, we get $\alpha - 2\lfloor L/2\rfloor \leq 0$, as claimed.

The proof in case of $f \in N^\alpha_q(X, \varrho_1, L)$ is almost identical, and hence omitted. □

Our next result shows that for networks of *fixed* depth, neural networks using the activation function ϱ_r with $r \geq 2$ are strictly more expressive than ReLU networks—at least in the regime of very high approximation rates.

Corollary 4.19. *Consider $\Omega \subseteq \mathbb{R}^d$ an open L_p-domain, $p \in (0, \infty]$, $X = X_p(\Omega)$, $L \in \mathbb{N}$. In case of $d = 1$, assume that $r \geq 4$ and $L \geq 2$, or that $r \in \{2, 3\}$ and $L \geq 3$. In case of $d > 1$, assume instead that $r \geq 4$ and $L \geq 3$, or that $r \in \{2, 3\}$ and $L \geq 5$. Then the following hold:*

$$\alpha > 2\lfloor L/2\rfloor \Longrightarrow W^\alpha_q(X, \varrho_r, L) \not\hookrightarrow W^\alpha_q(X, \varrho_1, L) \quad \text{and} \quad \alpha > 2L \Longrightarrow N^\alpha_q(X, \varrho_r, L) \not\hookrightarrow N^\alpha_q(X, \varrho_1, L). \quad \blacktriangleleft$$

Proof. We use Lemma 4.20 below to get $W^\alpha_q(X, \varrho_r, L) \cap C^3_c(\Omega) \neq \{0\}$ and $N^\alpha_q(X, \varrho_r, L) \cap C^3_c(\Omega) \neq \{0\}$, and we conclude using Corollary 4.18. □

Lemma 4.20. *Consider $d, r, L \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ an* open *L_p-domain, $p \in (0, \infty]$, $X = X_p(\Omega)$. In case of $d = 1$, assume that $r \geq 4$ and $L \geq 2$, or that $r \in \{2, 3\}$ and $L \geq 3$. In case of $d > 1$, assume instead that $r \geq 4$ and $L \geq 3$, or that $r \in \{2, 3\}$ and $L \geq 5$.*

Then for each $\alpha > 0$ and $q \in (0, \infty]$, we have $N^\alpha_q(X, \varrho_r, L) \cap C^3_c(\Omega) \neq \{0\} \neq W^\alpha_q(X, \varrho_r, L) \cap C^3_c(\Omega)$. $\blacktriangleleft$

Proof. Since Ω is an L_p-domain, it is non empty. Being open, Ω thus contains a hyper-rectangle $[a, b] := \prod_{i=1}^d [a_i, b_i] \subset \Omega$, where $a_i < b_i$.

For $r' \geq 2$, let $\sigma_{r'} \in \mathtt{SNN}^{\varrho_{r'},1,1}_{2(r'+1),2,r'+1}$ be the function constructed in Lemma 4.4. As $\sigma_{r'}$ satisfies (3.9), the function g built from $\sigma_{r'}$ in Lemma 3.26-(2) for small enough ε is nonzero and satisfies $\operatorname{supp}(g) \subset [a, b] \subset \Omega$ and $g \in \mathtt{NN}^{\varrho_{r'},d,1}_{\infty,3,\infty}$ (resp. $g \in \mathtt{NN}^{\varrho_{r'},d,1}_{\infty,2,\infty}$ when $d = 1$). Note that if $r' \geq 4$ then $\varrho_{r'} \in C^3(\mathbb{R})$, hence $g \in C^3_c(\mathbb{R}^d) \setminus \{0\}$.

When $r \geq 4$, set $r' := r$ so that $g \in \mathtt{NN}^{\varrho_r,d,1}_{\infty,3,\infty}$ ($g \in \mathtt{NN}^{\varrho_r,d,1}_{\infty,2,\infty}$ when $d = 1$). When $r \in \{2, 3\}$ set $r' := r^2 \geq 4$. As $\varrho_{r'} = \varrho_r \circ \varrho_r$, Lemma 2.20 with $s = 2$ yields $g \in \mathtt{NN}^{\varrho_r,d,1}_{\infty,5,\infty}$ ($g \in \mathtt{NN}^{\varrho_r,d,1}_{\infty,3,\infty}$ for $d = 1$).

It is not hard to see that our assumptions regarding L imply in each case for n large enough that $g|_\Omega \in \mathtt{W}_n(X, \varrho_r, L) \cap \mathtt{N}_n(X, \varrho_r, L)$, and hence $0 \neq g|_\Omega \in W^\alpha_q(X, \varrho_r, L) \cap C^3_c(\Omega) \cap N^\alpha_q(X, \varrho_r, L)$. □

5. Direct and inverse estimates with Besov spaces

In this section we characterize certain embeddings

- of Besov spaces into $W^\alpha_q(X, \varrho_r, \mathscr{L})$ and $N^\alpha_q(X, \varrho_r, \mathscr{L})$; these are called *direct estimates*;
- of $W^\alpha_q(X, \varrho_r, \mathscr{L})$ and $N^\alpha_q(X, \varrho_r, \mathscr{L})$ into Besov spaces; these are called *inverse estimates*.

Since the approximation classes for output dimension $k > 1$ are k-fold cartesian products of the classes for $k = 1$ (cf. Remark 3.17), we focus on scalar output dimension $k = 1$. We will use so-called *Jackson inequalities* and *Bernstein inequalities*, as well as the notion of real interpolation spaces. These concepts are recalled in Section 5.1, while Besov spaces and some of their properties are briefly recalled in Section 5.2 before we proceed to our main results.

5.1. Reminders on interpolation theory.

Given two quasi-normed vector spaces $(Y_J, \|\cdot\|_{Y_J})$ and $(Y_B, \|\cdot\|_{Y_B})$ with $Y_J \hookrightarrow X$ and $Y_B \hookrightarrow X$ for a given quasi-normed linear space $(X, \|\cdot\|_X)$, we say that Y_J *fulfills a Jackson inequality with exponent $\gamma > 0$* with respect to the family $\Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}$, if there is a constant $C_J > 0$ such that

$$E(f, \Sigma_n)_X \leq C_J \cdot n^{-\gamma} \cdot \|f\|_{Y_J} \qquad \forall\, f \in Y_J \text{ and } n \in \mathbb{N}. \tag{J}$$

We say that Y_B *fulfills a Bernstein inequality with exponent* $\gamma > 0$ with respect to $\Sigma = (\Sigma_n)_{n\in\mathbb{N}_0}$, if there is a constant $C_B > 0$ such that

$$\|\varphi\|_{Y_B} \leq C_B \cdot n^\gamma \cdot \|\varphi\|_X \qquad \forall\, n \in \mathbb{N} \text{ and } \varphi \in \Sigma_n. \tag{B}$$

As shown in the proof of [21, Chapter 7, Theorem 9.1], we have the following:

Proposition 5.1. *Denote by* $(X,Y)_{\theta,q}$ *the real interpolation space obtained from* X, Y*, as defined e.g. in* [21, Chapter 6, Section 7]. *Then the following hold:*

- *If* $Y_J \hookrightarrow X$ *fulfills the Jackson inequality with exponent* $\gamma > 0$*, then*
$$(X, Y_J)_{\alpha/\gamma,q} \hookrightarrow A_q^\alpha(X,\Sigma) \qquad \forall\ 0 < \alpha < \gamma \text{ and } 0 < q \leq \infty.$$
- *If* $Y_B \hookrightarrow X$ *fulfills the Bernstein inequality with exponent* $\gamma > 0$*, then*
$$A_q^\alpha(X,\Sigma) \hookrightarrow (X, Y_B)_{\alpha/\gamma,q} \qquad \forall\ 0 < \alpha < \gamma \text{ and } 0 < q \leq \infty. \qquad \blacktriangleleft$$

In particular, if the single space $Y = Y_J = Y_B$ satisfies both inequalities with the same exponent γ, then $A_q^\alpha(X,\Sigma) = (X,Y)_{\alpha/\gamma,q}$ for all $0 < \alpha < \gamma$ and $0 < q \leq \infty$.

By [21, Chapter 7, Theorem 9.3], if Σ satisfies Properties (P1)–(P5) then for $0 < \tau \leq \infty$, $0 < \alpha < \infty$ the space $Y := A_\tau^\alpha(X,\Sigma)$ satisfies matching Jackson and Bernstein inequalities with exponent $\gamma := \alpha$. The Bernstein inequality reads

$$\exists\, C = C(\alpha,\tau,X) > 0 \quad \forall\, n \in \mathbb{N} \text{ and } \varphi \in \Sigma_n : \quad \|\varphi\|_{A_\tau^\alpha(X,\Sigma)} \leq C \cdot n^\alpha \cdot \|\varphi\|_X. \tag{5.1}$$

We will also use the following well-known property of (real) interpolation spaces (see [21, Chapter 6, Theorem 7.1]): For quasi-Banach spaces X_1, X_2 and Y_1, Y_2, assume that $T : X_1 + X_2 \to Y_1 + Y_2$ is linear and such that $T|_{X_i} : X_i \to Y_i$ is bounded for $i \in \{1,2\}$. Then $T|_{(X_1,X_2)_{\theta,q}} : (X_1,X_2)_{\theta,q} \to (Y_1,Y_2)_{\theta,q}$ is well-defined and bounded for all $\theta \in (0,1)$ and $q \in (0,\infty]$.

5.2. **Reminders on Besov spaces.** We refer to [20, Section 2] for the definition of the Besov spaces $B^s_{\sigma,\tau}(\Omega) := B^s_\tau(X_\sigma(\Omega;\mathbb{R}))$ with $\sigma,\tau \in (0,\infty]$, $s \in (0,\infty)$ and with a Lipschitz domain[5] $\Omega \subset \mathbb{R}^d$ (see [1, Definition 4.9] for the precise definition of these domains).

As shown in [19, Theorem 7.1], we have for all $p, s \in (0,\infty)$ the embedding

$$B^s_{\sigma,p}((0,1)^d) \hookrightarrow L_p((0,1)^d;\mathbb{R}), \quad \text{provided} \quad \sigma = (s/d + 1/p)^{-1}.$$

Combined with the embedding $B^s_{p,q}(\Omega) \hookrightarrow B^s_{p,q'}(\Omega)$ for $q \leq q'$ (see [17, Displayed equation on Page 92]) and because of $\sigma = (s/d+1/p)^{-1} \leq p$, we see that

$$B^s_{\sigma,\sigma}((0,1)^d) \hookrightarrow B^s_{\sigma,p}((0,1)^d) \hookrightarrow L_p((0,1)^d;\mathbb{R}), \quad \text{provided} \quad \sigma = (s/d+1/p)^{-1}. \tag{5.2}$$

For the special case $\Omega = (0,1) \subset \mathbb{R}$ and each fixed $p \in (0,\infty)$, the sub-family of Besov spaces $B^s_{\sigma,\sigma}((0,1))$ with $\sigma = (s/d+1/p)^{-1}$ satisfies

$$\big(L_p((0,1);\mathbb{R}), B^s_{\sigma,\sigma}((0,1))\big)_{\theta,q} = B^{\theta s}_{q,q}((0,1)), \quad \text{for } 0 < \theta < 1, \text{ where } q = (\theta s + 1/p)^{-1}. \tag{5.3}$$

This is shown in [21, Chapter 12, Corollary 8.5].

Finally, from the definition of Besov spaces given in [20, Equation (2.2)] it is clear that

$$B^\alpha_{p,q}(\Omega) \hookrightarrow B^\beta_{p,q}(\Omega) \quad \text{if} \quad p,q \in (0,\infty] \text{ and } 0 < \beta < \alpha. \tag{5.4}$$

5.3. **Direct estimates.** In this subsection, we investigate embeddings of Besov spaces into the approximation spaces $W_q^\alpha(X,\varrho_r,\mathscr{L})$ where $\Omega \subseteq \mathbb{R}^d$ is an L_p-domain and $X := X_p(\Omega)$ with $p \in (0,\infty]$. For technical reasons, we further assume Ω to be a bounded Lipschitz domain, such as $\Omega = (0,1)^d$, see [1, Definition 4.9]. The main idea is to exploit known direct estimates for Besov spaces on such domains which give error bounds for the n-term approximations with B-spline based wavelet systems, see [19].

For $t \in \mathbb{N}_0$ and $d \in \mathbb{N}$, the tensor product B-spline is $\beta_d^{(t)}(x_1,\dots,x_d) := \beta_+^{(t)}(x_1)\,\beta_+^{(t)}(x_2)\cdots\beta_+^{(t)}(x_d)$, where $\beta_+^{(t)}$ is as introduced in Definition 4.5. Notice that $\beta_d^{(0)} = \mathbb{1}_{[0,1)^d}$.

By Lemma 4.4 there is $\sigma_1 \in \mathtt{NN}^{\varrho_r,1,1}_{2(r+1),2,r+1}$ satisfying (3.9); hence by Lemma 3.26 there is $L \leq 3$ such that for $\varepsilon > 0$, we can approximate $\beta_d^{(0)}$ with $g_\varepsilon = \mathtt{R}(\Phi_\varepsilon)$ with precision $\|\beta_d^{(0)} - g_\varepsilon\|_{L_p(\mathbb{R}^d)} < \varepsilon$, where $L(\Phi_\varepsilon) = L$ and $\Phi_\varepsilon \in \mathtt{NN}^{\varrho_r,d,1}_{w,3,m}$, for suitable $w = w(d,r), m = m(d,r) \in \mathbb{N}$. Furthermore, if $d = 1$, then Lemma 3.26 shows that the same holds for some $\Phi_\varepsilon \in \mathtt{NN}^{\varrho_r,d,1}_{w,2,m}$.

For approximating $\beta_d^{(t)}$ (with $t \in \mathbb{N}$) instead of $\beta_d^{(0)}$, we can actually do better. In fact, we prove in Appendix D.1 that one can implement $\beta_d^{(t)}$ as a ϱ_t-network, provided that $t \geq \min\{d,2\}$.

[5] Here, the term "domain" is to be understood as an open connected set.

Lemma 5.2. *Let $d, t \in \mathbb{N}$ with $t \geq \min\{d, 2\}$. Then the tensor product B-spline*

$$\beta_d^{(t)} : \mathbb{R}^d \to \mathbb{R}, \quad \beta_d^{(t)}(x) := \beta_+^{(t)}(x_1)\beta_+^{(t)}(x_2)\cdots\beta_+^{(t)}(x_d) \tag{5.5}$$

satisfies $\beta_d^{(t)} \in \mathtt{NN}_{w,L,m}^{\varrho_t,d,1}$ with $L = 2 + 2\lceil \log_2 d \rceil$ and

$$\begin{cases} w = 28d(t+1) \text{ and } m = 13d(t+1), & \text{if } d > 1, \\ w = 2(t+2) \text{ and } m = t+2, & \text{if } d = 1. \end{cases}$$ ◀

In the following, we will consider n-term approximations with respect to the continuous wavelet system generated by $\beta_d^{(t)}$. Precisely, for $a > 0$ and $b \in \mathbb{R}^d$, define $\beta_{a,b}^{(t)} := \beta_d^{(t)}(a \cdot + b)$. The continuous wavelet system generated by $\beta_d^{(t)}$ is then $\mathcal{D}_d^t := \{\beta_{a,b}^{(t)} \colon a \in (0,\infty), b \in \mathbb{R}^d\}$. For any $t \in \mathbb{N}_0$, we define $\Sigma_0(\mathcal{D}_d^t) := \{0\}$, and the reservoir of all n-term expansions from $\mathcal{D}_d^t$, $n \in \mathbb{N}$, is given by

$$\Sigma_n(\mathcal{D}_d^t) := \left\{ g = \sum_{i=1}^n c_i g_i \colon c_i \in \mathbb{R}, g_i \in \mathcal{D}_d^t \right\}.$$

The following lemma relates $\Sigma_n(\mathcal{D}_d^t)$ to $\mathtt{NN}_{cn,L,cn}^{\varrho_r,d,1}$ for a suitably chosen constant $c = c(d,r,t) \in \mathbb{N}$.

Lemma 5.3. *Consider $d \in \mathbb{N}$, $t \in \mathbb{N}_0$, $p \in (0,\infty]$, $X = X_p(\mathbb{R}^d)$.*
(1) If $t = 0$ and $p < \infty$ then, with $L := \min\{d+1, 3\}$ and $c = c(d,r) \in \mathbb{N}$, we have

$$\Sigma_n(\mathcal{D}_d^0) \subset \overline{\mathtt{NN}_{cn,L,cn}^{\varrho_r,d,1} \cap X}^X \qquad \forall\, n, r \in \mathbb{N}. \tag{5.6}$$

(2) If $t \geq \min\{d, 2\}$ then, with $L := 2 + 2\lceil \log_2 d \rceil$ we have for any $p \in (0,\infty]$ that

$$\Sigma_n(\mathcal{D}_d^t) \subset \mathtt{NN}_{cn,L,cn}^{\varrho_t,d,1} \cap X \qquad \forall\, n \in \mathbb{N}, \tag{5.7}$$

where $c = c(d,t) \in \mathbb{N}$. ◀

Proof. **Part (1):** For $t = 0$, $r \in \mathbb{N}$, $0 < p < \infty$, we have already noticed before Lemma 5.2 that there exist $w = w(d,r), m = m(d,r) \in \mathbb{N}$ such that $\beta_d^{(0)} \in \overline{\mathtt{NN}_{w,L,m}^{\varrho_r,d,1} \cap X}^X$, where $L = \min\{d+1,3\}$. Since $\beta_{a,b}^{(0)} = \beta_d^{(0)} \circ P_{a,b}$ for the affine map $P_{a,b} : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto ax + b$ and since $\|P_{a,b}\|_{\ell_*^{0,\infty}} = 1$, Lemma 2.17-(1) and Lemma 2.18-(1) yield $\beta_{a,b}^{(0)} \in \overline{\mathtt{NN}_{c,L,c}^{\varrho_r,d,1} \cap X}^X$ with $c := \max\{w,m\}$. Thus, the claim follows from Parts (1) and (3) of Lemma 2.17.

Part (2): For $t \geq \min\{d,2\}$, Lemma 5.2 shows that $\beta_{a,b}^{(t)} \in \mathtt{NN}_{c,L,c}^{\varrho_t,d,1} \cap X$ with $L = 2 + 2\lceil \log_2 d \rceil$ and $c := \max\{w,m\}$ where $w = w(d,t)$ and $m = m(d,t)$ are as in Lemma 5.2. As before, we conclude using Parts (1) and (3) of Lemma 2.17. □

Corollary 5.4. *Consider $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$ an L_p-domain, $p \in (0,\infty]$, $X = X_p(\Omega)$, $\mathscr{L}$ a depth growth function, $L := \sup_n \mathscr{L}(n) \in \mathbb{N} \cup \{\infty\}$. For $t \in \mathbb{N}_0$ define $\Sigma(\mathcal{D}_d^t) := (\Sigma_n(\mathcal{D}_d^t))_{n\in\mathbb{N}_0}$.*

(1) If $L \geq \min\{d+1, 3\}$ and $p < \infty$, then for any $r \geq 1$

$$A_q^\alpha(X, \Sigma(\mathcal{D}_d^0)) \hookrightarrow W_q^\alpha(X, \varrho_r, \mathscr{L}) \qquad \text{for each} \quad \alpha \in (0,\infty),\ q \in (0,\infty].$$

(2) If $L \geq 2 + 2\lceil \log_2 d \rceil$ then for any $r \geq \min\{d,2\}$, we have

$$A_q^\alpha(X, \Sigma(\mathcal{D}_d^r)) \hookrightarrow W_q^\alpha(X, \varrho_r, \mathscr{L}) \qquad \text{for each} \quad \alpha \in (0,\infty),\ q \in (0,\infty].$$ ◀

Proof. For the proof of Part (1) let $L_0 := \min\{d+1,3\}$, while $L_0 := 2 + 2\lceil \log_2 d \rceil$ for the proof of Part (2). Since $L \geq L_0$, there is $n_0 \in \mathbb{N}$ such that $\mathscr{L}(n) \geq L_0$ for all $n \geq n_0$.

We first start with the proof of Part (2). By Lemma 5.3-(2), with $t = r \geq \min\{d,2\}$, Equation (5.7) holds for some $c \in \mathbb{N}$. For $n \geq n_0/c$ we have $2 + 2\lceil \log_2 d \rceil = L_0 \leq \mathscr{L}(cn)$, whence

$$\Sigma_n(\mathcal{D}_d^t) \subset \overline{\mathtt{W}_{cn}(X, \varrho_r, \mathscr{L})}^X.$$

Therefore, we see that

$$E(f, \Sigma_n(\mathcal{D}_d^t))_X \geq E(f, \mathtt{W}_{cn}(X, \varrho_r, \mathscr{L}))_X \quad \forall\, f \in X \text{ and } n \geq \tfrac{n_0}{c}. \tag{5.8}$$

For the proof of Part (1), the same reasoning with (5.6) instead of (5.7) yields (5.8) with $t = 0$ and any $r \in \mathbb{N}$. For both parts, we conclude using Lemma 3.1 and the associated remark. □

Theorem 5.5. Let $\Omega \subset \mathbb{R}^d$ be a bounded Lipschitz domain of positive measure. For $p \in (0,\infty]$, define $X_p(\Omega)$ as in Equation (1.3). Let $\mathscr{L}$ be a depth growth function.

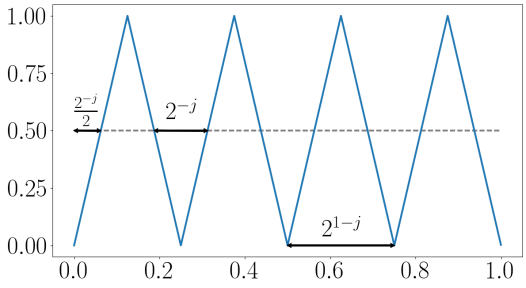

FIGURE 4. A plot of the function Δ_j (for $j = 3$).

(1) Suppose that $d = 1$ and $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq 2$. Then the following holds for each $r \in \mathbb{N}$:

$$B^s_{p,q}(\Omega) \hookrightarrow W^s_q(X_p(\Omega), \varrho_r, \mathscr{L}) \quad \forall\, p, q \in (0, \infty] \text{ and } 0 < s < r + \min\{1, p^{-1}\}. \tag{5.9}$$

(2) Suppose that $d > 1$ and $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq 3$, and let $r \in \mathbb{N}$. Define $r_0 := r$ if $r \geq 2$ and $L \geq 2 + 2\lceil \log_2 d \rceil$, and $r_0 := 0$ otherwise. Then

$$B^{sd}_{p,q}(\Omega) \hookrightarrow W^s_q(X_p(\Omega), \varrho_r, \mathscr{L}) \quad \forall\, p, q \in (0, \infty] \text{ and } 0 < s < \frac{r_0 + \min\{1, p^{-1}\}}{d}. \tag{5.10}$$

◀

Remark 5.6. If Ω is open then each Besov space $B^{sd}_{p,q}(\Omega)$ contains $C^3_c(\Omega)$. Hence, by Corollary 4.18, the embeddings (5.9) or (5.10) with $r = 1$ imply that $\lfloor L/2 \rfloor \geq s/2$. This is indeed the case, since these embeddings for $r = 1$ are only established when $L \geq 2$ and $0 < ds < 1 + \min\{p^{-1}, 1\} \leq 2$, which implies $s/2 < 1/d \leq 1 \leq \lfloor L/2 \rfloor$. ♦

Proof of Theorem 5.5. See Appendix D.2. □

5.4. **Limits on possible inverse estimates.** For networks of finite depth $\mathscr{L} \equiv L < \infty$, there are limits on possible embeddings of $W^\alpha_q(X, \varrho_1, \mathscr{L})$ (resp. of $N^\alpha_q(X, \varrho_1, \mathscr{L})$) into Besov spaces.

Theorem 5.7. Consider $\Omega = (0, 1)^d$, $p \in (0, \infty]$, $X = X_p(\Omega)$, $\mathscr{L}$ a depth growth function such that $L := \sup_n \mathscr{L}(n) \in \mathbb{N}_{\geq 2} \cup \{\infty\}$ and $r \in \mathbb{N}$. For $\sigma, \tau, q \in (0, \infty]$ and $\alpha, s \in (0, \infty)$, the following claims hold[6]:

(1) If $W^\alpha_q(X, \varrho_r, \mathscr{L}) \hookrightarrow B^s_{\sigma,\tau}(\Omega)$ then $\alpha \geq \lfloor L/2 \rfloor \cdot \min\{s, 2\}$.
(2) If $N^\alpha_q(X, \varrho_r, \mathscr{L}) \hookrightarrow B^s_{\sigma,\tau}(\Omega)$ then $\alpha \geq (L - 1) \cdot \min\{s, 2\}$. ◀

A direct consequence is that for networks of unbounded depth ($L = \infty$), none of the spaces $W^\alpha_q(X, \varrho_r, \mathscr{L})$, $N^\alpha_q(X, \varrho_r, \mathscr{L})$ embed into any Besov space of strictly positive smoothness $s > 0$.

Remark 5.8. For $L = 2$, as $\lfloor L/2 \rfloor = L - 1$ the two inequalities resulting from Theorem 5.7 match. This is natural as for $L = 2$ we know from Lemma 3.9 that $W^\alpha_q(X, \varrho_r, \mathscr{L}) = N^\alpha_q(X, \varrho_r, \mathscr{L})$. For $L \geq 3$ the inequalities no longer match. Each inequality is in fact stronger than what would be achieved by simply combining the other one with Lemma 3.9. Note also that in contrast to the direct estimate (5.10) of Theorem 5.5 where the Besov spaces are of smoothness sd, here the dimension d does not appear. ♦

The proof of Theorem 5.7 employs a particular family of oscillating functions that have a long history [31] in the analysis of neural networks and of the benefits of depth [64].

Definition 5.9 (Sawtooth functions)**.** Consider $\beta^{(1)}_+$ the B-spline of degree one, and $\Delta_1 := \beta^{(1)}_+(2\cdot)$ the "hat" function supported on $[0, 1]$. For $j \geq 1$ the univariate "sawtooth" function of order j,

$$\Delta_j = \sum_{k=0}^{2^{j-1}-1} \Delta_1(2^{j-1} \cdot - k), \tag{5.11}$$

has support in $[0, 1]$ and is made of 2^{j-1} triangular "teeth" (see Figure 4). The *multivariate* sawtooth function $\Delta_{j,d}$ is defined as $\Delta_{j,d}(x) := \Delta_j(x_1)$ for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$, $j \in \mathbb{N}$. ◀

An important property of Δ_j is that it is a realization of a ϱ_1-network of specific complexity. The proof of this lemma is in Appendix D.3.

[6] with the convention $\lfloor \infty/2 \rfloor = \infty - 1 = \infty$

Lemma 5.10. *Let $L \in \mathbb{N}_{\geq 2}$ and define $C_L := 4\,L + 2^{L-1}$. Then*

$$\Delta_j \in \mathtt{NN}^{\varrho_1,1,1}_{\infty,L,C_L\cdot 2^{j/(L-1)}} \qquad \text{and} \qquad \Delta_j \in \mathtt{NN}^{\varrho_1,1,1}_{C_L\cdot 2^{j/\lfloor L/2\rfloor},L,\infty} \qquad \forall\, j \in \mathbb{N}. \qquad \blacktriangleleft$$

Corollary 5.11. *For $L \in \mathbb{N}_{\geq 2}$, let C_L as in Lemma 5.10. Then*

$$\Delta_{j,d} \in \mathtt{NN}^{\varrho_1,d,1}_{\infty,L,C_L\cdot 2^{j/(L-1)}} \qquad \text{and} \qquad \Delta_{j,d} \in \mathtt{NN}^{\varrho_1,d,1}_{C_L\cdot 2^{j/\lfloor L/2\rfloor},L,\infty} \qquad \forall\, j \in \mathbb{N}. \qquad \blacktriangleleft$$

Proof. We have $\Delta_{j,d} = \Delta_j \circ T$ for the affine map $T : \mathbb{R}^d \to \mathbb{R}, (x_1,\dots,x_d) \mapsto x_1$, which satisfies $\|T\|_{\ell^{0,\infty}_*} = 1$. Now, the claim is a direct consequence of Lemmas 5.10 and 2.18-(1). □

We further prove in Appendix D.4 that that the Besov norm of $\Delta_{j,d}$ grows exponentially with j:

Lemma 5.12. *Let $d \in \mathbb{N}$ and $\Omega = (0,1)^d$. Let $p,q \in (0,\infty]$ and $s \in (0,\infty)$ be arbitrary. Let $\alpha \in (0,2)$ with $\alpha \leq s$. There is a constant $c = c(d,p,q,s,\alpha) > 0$ such that*

$$\|\Delta_{j,d}\|_{B^s_{p,q}(\Omega)} \geq c \cdot 2^{\alpha j} \qquad \forall\, j \in \mathbb{N}. \qquad \blacktriangleleft$$

Given this lower bound on the Besov space norm of the sawtooth function $\Delta_{j,d}$, we can now prove the limitations regarding possible inverse estimates that we announced above.

Proof of Theorem 5.7. We start with the proof for $W^\alpha_q(X,\varrho_r,\mathscr{L})$. Let us fix $\ell \in \mathbb{N}$ with $\ell \leq \lfloor L/2\rfloor$, and note that $2\ell \leq L$, so that there is some $j_0 = j_0(\ell,\mathscr{L}) \in \mathbb{N}$ such that $\mathscr{L}(2^j) \geq 2\ell$ for all $j \geq j_0$. Now, Corollary 5.11 (applied with 2ℓ instead of L) shows that $\Delta_{\ell j,d} \in \mathtt{NN}^{\varrho_1,d,1}_{C_{2\ell}2^{(\ell j)/\ell},2\ell,\infty} \subset \mathtt{NN}^{\varrho_1,d,1}_{C_{2\ell}2^j,\mathscr{L}(C_{2\ell}2^j),\infty}$ for all $j \geq j_0$ and a suitable constant $C_{2\ell} \in \mathbb{N}$. Therefore, the Bernstein inequality (5.1) yields a constant $C = C(d,\alpha,q,p) > 0$ such that

$$\|\Delta_{\ell j,d}|_\Omega\|_{W^\alpha_q(X,\varrho_1,\mathscr{L})} \leq C \cdot \left(C_{2\ell}\,2^j\right)^\alpha \cdot \|\Delta_{\ell j,d}|_\Omega\|_X \leq C^\alpha_{2\ell} C \cdot 2^{\alpha j} \quad \forall\, j \geq j_0.$$

Let $s_0 := \min\{2,s\}$, let $0 < s' < s_0$ be arbitrary, and note as a consequence of Equation (4.3) that

$$W^\alpha_q(X,\varrho_1,\mathscr{L}) \hookrightarrow W^\alpha_q(X,\varrho_r,\mathscr{L}) \hookrightarrow B^s_{\sigma,\tau}(\Omega).$$

Here we used that Ω is bounded, so that Equation (4.3) is applicable. Overall, as a consequence of this embedding and of Lemma 5.12, we obtain $c = c(d,s',s,\sigma,\tau) > 0$ and $C' = C'(\sigma,\tau,s,p,q,\alpha,\mathscr{L},\Omega) > 0$ satisfying

$$c \cdot 2^{s'\ell j} \leq \|\Delta_{\ell j,d}|_\Omega\|_{B^s_{\sigma,\tau}(\Omega)} \leq C' \cdot \|\Delta_{\ell j,d}|_\Omega\|_{W^\alpha_q(X,\varrho_1,\mathscr{L})} \leq C' C^\alpha_{2\ell} C \cdot 2^{\alpha j}$$

for all $j \geq j_0$. This implies $s' \cdot \ell \leq \alpha$. Since this holds for all $s' \in (0,s_0)$ and all $\ell \in \mathbb{N}$ with $\ell \leq \lfloor L/2\rfloor$, we get $\lfloor L/2\rfloor \cdot s_0 \leq \alpha$, as claimed.

Now, we prove the claim for $N^\alpha_q(X,\varrho_r,\mathscr{L})$. In this case, fix $\ell \in \mathbb{N}$ with $\ell + 1 \leq L$, and note that there is some $j_0 \in \mathbb{N}$ satisfying $\mathscr{L}(2^j) \geq \ell + 1$ for all $j \geq j_0$. Now, Corollary 5.11 (applied with $\ell+1$ instead of L) yields a constant $C_{\ell+1} \in \mathbb{N}$ such that $\Delta_{\ell j,d} \in \mathtt{NN}^{\varrho_1,d,1}_{\infty,\ell+1,C_{\ell+1}2^{(\ell j)/((\ell+1)-1)}} \subset \mathtt{NN}^{\varrho_1,d,1}_{\infty,\mathscr{L}(C_{\ell+1}2^j),C_{\ell+1}2^j}$ for all $j \geq j_0$. As above, the Bernstein inequality (5.1) therefore shows $\|\Delta_{\ell j,d}|_\Omega\|_{N^\alpha_q(X,\varrho_1,\mathscr{L})} \leq C^\alpha_{\ell+1} C \cdot 2^{\alpha j}$ for all $j \geq j_0$ and some constant $C = C(d,\alpha,q,p) < \infty$. Reasoning as above, we get that

$$c \cdot 2^{s'\ell j} \leq \|\Delta_{\ell j,d}|_\Omega\|_{B^s_{\sigma,\tau}(\Omega)} \leq C' \cdot \|\Delta_{\ell j,d}|_\Omega\|_{N^\alpha_q(X,\varrho_1,\mathscr{L})} \leq C' C^\alpha_{\ell+1} C \cdot 2^{\alpha j}$$

for all $j \geq j_0$ and $0 < s' < s_0 = \min\{2,s\}$. Therefore, $s' \cdot \ell \leq \alpha$. Since this holds for all $s' \in (0,s_0)$ and all $\ell \in \mathbb{N}$ with $\ell + 1 \leq L$, we get $\alpha \geq s_0 \cdot (L-1)$, as claimed. □

5.5. **Univariate inverse estimates ($d = 1$).** The "no-go theorem" (Theorem 5.7) holds for $\Omega = (0,1)^d$ in any dimension $d \geq 1$, for any $0 < p \leq \infty$. In this subsection, we show in dimension $d = 1$ that Theorem 5.7 is quite sharp. Precisely, we prove the following:

Theorem 5.13. Let $X = L_p(\Omega)$ with $\Omega = (0,1)$ and $p \in (0,\infty)$, let $r \in \mathbb{N}$, and let $\mathscr{L}$ be a depth growth function. Assume that $L := \sup_n \mathscr{L}(n) < \infty$. Setting $\nu := \lfloor L/2\rfloor$, the following statements hold:

(1) For $s \in (0,\infty)$, $\alpha \in (0,\nu s)$ and $q \in (0,\infty]$, we have

$$W^\alpha_q(X,\varrho_r,\mathscr{L}) \hookrightarrow \left(L_p(\Omega;\mathbb{R}), B^s_{\sigma,\sigma}(\Omega)\right)_{\frac{\alpha}{s\nu},q} \quad \text{where} \quad \sigma := (s + 1/p)^{-1}.$$

(2) For $\alpha \in (0,\infty)$, we have

$$W^\alpha_q(X,\varrho_r,\mathscr{L}) \hookrightarrow B^{\alpha/\nu}_{q,q}(\Omega) \quad \text{where} \quad q := (\alpha/\nu + 1/p)^{-1}.$$

The same holds for $N^\alpha_q(X,\varrho_r,\mathscr{L})$ instead of $W^\alpha_q(X,\varrho_r,\mathscr{L})$ if we set $\nu := L - 1$. $\blacktriangleleft$

The proof involves a Bernstein inequality for piecewise polynomials by Petrushev [56], and new bounds on the number of pieces of piecewise polynomials implemented by ϱ_r-networks. Petrushev considers the (nonlinear) set $\tilde{\mathtt{S}}(k,n)$ of all piecewise polynomials on $(0,1)$ of degree at most $r=k-1$ ($k\in\mathbb{N}$) with at most $n-1$ breakpoints in $[0,1]$. In the language of Definition 4.6, $\tilde{\mathtt{S}}(k,n)=\mathtt{PPoly}_n^r((0,1))$ is the set of piecewise polynomials of degree at most $r=k-1\in\mathbb{N}_0$ with at most n pieces on $(0,1)$.

By [21, Chapter 12, Theorem 8.2] (see [56, Theorem 2.2] for the original proof) the following Bernstein-type inequality holds for each family $\Sigma:=(\tilde{\mathtt{S}}(k,n))_{n\in\mathbb{N}}$, $k\in\mathbb{N}$:

Theorem 5.14 ([56, Theorem 2.2])**.** Let $\Omega=(0,1)$, $p\in(0,\infty)$, $r\in\mathbb{N}_0$, and $s\in(0,r+1)$ be arbitrary, and set $\sigma:=(s+1/p)^{-1}$. Then there is a constant $C<\infty$ such that we have

$$\|f\|_{B^s_{\sigma,\sigma}(\Omega)}\leq C\cdot n^s\cdot\|f\|_{L_p(\Omega)}\qquad\forall\, n\in\mathbb{N}\text{ and }f\in\mathtt{PPoly}_n^r(\Omega).$$ ◀

Remark 5.15. Theorem 5.14 even holds for *discontinuous* piecewise polynomial functions, see [56, Theorem 2.2]. Hence, the Besov spaces in Theorem 5.13 also contain discontinuous functions. This is natural, as ϱ_r-networks with bounded number of connections or neurons approximate indicator functions arbitrarily well (though with weight values going to infinity, see the proof of Lemma 3.26). ⧫

When f is a realization of a ϱ_r-network of depth L, it is piecewise polynomial [64]. As there are $L-1$ hidden layers, the polynomial pieces are of degree r^{L-1} at most, hence $f|_{(0,1)}\in\mathtt{PPoly}_n^{r^{L-1}}((0,1))$ for large enough n. This motivates the following definition.

Definition 5.16 (Number of pieces)**.** Define $n_r(W,L,N)$ to be the optimal bound on the number of polynomial pieces for a ϱ_r-network with $W\in\mathbb{N}_0$ connections, depth $L\in\mathbb{N}$ and $N\in\mathbb{N}_0$ neurons; that is,

$$n_r(W,L,N):=\min\left\{n\in\mathbb{N}\quad:\quad\forall\, g\in\mathtt{NN}_{W,L,N}^{\varrho_r,1,1}\,:\,g|_{(0,1)}\in\mathtt{PPoly}_n^{r^{L-1}}((0,1))\right\}.$$

Furthermore, let $n_r(W,L,\infty):=\sup_{N\in\mathbb{N}_0}n_r(W,L,N)$ and $n_r(\infty,L,N):=\sup_{W\in\mathbb{N}_0}n_r(W,L,N)$. ◀

Remark 5.17. The definition of $n_r(W,L,N)$ is independent of the non-degenerate interval $I\subset\mathbb{R}$ used for its definition. To see this, write $n_r^{(I)}(W,L,N)$ for the analogue of $n_r(W,L,N)$, but with $(0,1)$ replaced by a general non-degenerate interval $I\subset\mathbb{R}$. First, note that $n_r^{(I)}(W,L,N)\leq n_r^{(J)}(W,L,N)$ if $I\subset J$.

Next, note for $g\in\mathtt{NN}_{W,L,N}^{\varrho_r,1,1}$ and $a\in(0,\infty)$, $b\in\mathbb{R}$ that $g_{a,b}:=g(a\cdot+b)\in\mathtt{NN}_{W,L,N}^{\varrho_r,1,1}$ as well (see Lemma 2.18) and that $g|_I\in\mathtt{PPoly}_n^{r^{L-1}}(I)$ if and only $g_{a,b}|_{a^{-1}(I-b)}\in\mathtt{PPoly}_n^{r^{L-1}}(a^{-1}(I-b))$. Therefore, $n_r^{(I)}(W,L,N)=n_r^{(aI+b)}(W,L,N)$ for all $a\in(0,\infty)$ and $b\in\mathbb{R}$.

Now, if $J\subset\mathbb{R}$ is any non-degenerate interval, and if $I\subset\mathbb{R}$ is a *bounded* interval, then $aI+b\subset J$ for suitable $a>0$, $b\in\mathbb{R}$. Hence, $n_r^{(I)}=n_r^{(aI+b)}\leq n_r^{(J)}$. In particular, this shows $n_r^{(I)}=n_r^{(J)}$ for all *bounded* non-degenerate intervals $I,J\subset\mathbb{R}$.

Finally, if $g\in\mathtt{NN}_{W,L,N}^{\varrho_r,1,1}$ is arbitrary, then $g\in\mathtt{PPoly}_n^{r^{L-1}}(\mathbb{R})$ for *some* $n\in\mathbb{N}$. Thus, there are $a,b\in\mathbb{R}$, $a<b$ such that $g|_{(-\infty,a+1)}$ and $g|_{(b-1,\infty)}$ are polynomials of degree at most r^{L-1}. Let $k:=n_r^{((a,b))}(W,L,N)=n_r^{((0,1))}(W,L,N)$, so that $g|_{(a,b)}\in\mathtt{PPoly}_k^{r^{L-1}}((a,b))$. Clearly, $g\in\mathtt{PPoly}_k^{r^{L-1}}(\mathbb{R})$. Hence, $n_r^{(\mathbb{R})}(W,L,N)\leq k=n_r^{((0,1))}(W,L,N)$. ⧫

We now have the ingredients to establish the first main lemma behind the proof of Theorem 5.13.

Lemma 5.18. *Let $X=L_p(\Omega)$ with $\Omega=(0,1)$ and $p\in(0,\infty)$. Let $r\in\mathbb{N}$ and $\nu\in(0,\infty)$, and let $\mathscr{L}$ be a depth growth function such that $L:=\sup_n\mathscr{L}(n)<\infty$. Assume that*

$$\sup_{W\in\mathbb{N}}W^{-\nu}\;n_r(W,L,\infty)<\infty.\tag{5.12}$$

(1) For $s\in(0,r+1)$, $\alpha\in(0,\nu\cdot s)$, and $q\in(0,\infty]$, we have

$$W_q^\alpha(X,\varrho_r,\mathscr{L})\hookrightarrow(L_p(\Omega),B^s_{\sigma,\sigma}(\Omega))_{\frac{\alpha}{s\nu},q}\quad\text{where}\quad\sigma:=(s+1/p)^{-1}.\tag{5.13}$$

(2) For $\alpha\in(0,\nu(r+1))$, we have

$$W_q^\alpha(X,\varrho_r,\mathscr{L})\hookrightarrow B^{\alpha/\nu}_{q,q}(\Omega)\quad\text{where}\quad q:=(\alpha/\nu+1/p)^{-1}.\tag{5.14}$$

The same results hold with $N_q^\alpha(X,\varrho_r,\mathscr{L})$ instead of $W_q^\alpha(X,\varrho_r,\mathscr{L})$ if we assume instead that

$$\sup_{N\in\mathbb{N}}N^{-\nu}\;n_r(\infty,L,N)<\infty.\tag{5.15}$$

◀

Proof of Lemma 5.18. As $\mathtt{NN}_{n,\mathscr{L}(n),\infty}^{\varrho_r,1,1} \subset \mathtt{NN}_{n,L,\infty}^{\varrho_r,1,1}$ for each $n \in \mathbb{N}$, Theorem 5.14 and Equation (5.12) yield a constant $C < \infty$ such that

$$\|f\|_{B^s_{\sigma,\sigma}(\Omega)} \leq C \cdot n^{\nu s} \cdot \|f\|_{L_p(\Omega)}, \quad \text{for all } \ n \in \mathbb{N} \text{ and } f \in \mathtt{W}_n(X, \varrho_r, \mathscr{L}), \tag{5.16}$$

where $\sigma := (s+1/p)^{-1} = (s/d+1/p)^{-1}$ (recall $d = 1$). By (5.2) we further get that $Y_B := B^s_{\sigma,\sigma}(\Omega) \hookrightarrow L_p(\Omega)$, whence (5.16) is a valid Bernstein inequality for Y_B with exponent $\gamma := s \cdot \nu > \alpha$. Proposition 5.1 with $\theta := \alpha/\gamma = \alpha/(s\nu)$ and $0 < q \leq \infty$ yields (5.13).

When $0 < \alpha < \nu(r+1)$, there is $s \in (0, r+1)$ such that $0 < \alpha < \nu \cdot s$; hence, (5.13) holds for any $0 < q \leq \infty$. By (5.3), we see for $\theta := \frac{\alpha}{s\nu} \in (0,1)$ and $q := (\theta s + 1/p)^{-1} = (\alpha/\nu + 1/p)^{-1}$ that the right hand side of (5.13) is simply $B^{\theta s}_{q,q}(\Omega) = B^{\alpha/\nu}_{q,q}(\Omega)$.

The proof for $N^\alpha_q(X, \varrho_r, \mathscr{L})$ follows the same steps. □

Theorem 5.13 is a corollary of Lemma 5.18 once we establish (5.12) (resp. (5.15)). The smaller ν the better, as it yields a larger value for α/ν, hence a smoother (smaller) Besov space in (5.14).

Lemma 5.19. *Consider $L \in \mathbb{N}_{\geq 2}$, $r \in \mathbb{N}$.*

- *Property* (5.12) *holds if and only if $\nu \geq \lfloor L/2 \rfloor$;*
- *Property* (5.15) *holds if and only if $\nu \geq L - 1$.* ◀

Proof. If (5.12) holds with some exponent ν, then Lemma 5.18-(2) with $\mathscr{L} \equiv L$, arbitrary $p \in (0, \infty)$, $\alpha := \nu$ and $q := (\alpha/\nu + 1/p)^{-1}$ yields $W^\alpha_q(X, \varrho_r, \mathscr{L}) \hookrightarrow B^1_{q,q}(\Omega)$ with $\Omega := (0,1)$. If we set $s := 1$, then $\min\{s, 2\} = s = 1$. Hence, Theorem 5.7 implies $\nu = \alpha \geq \lfloor L/2 \rfloor$. The same argument shows that if (5.15) holds with some exponent ν, then $\nu \geq L - 1$. For the converse results it is clearly sufficient to establish (5.12) with $\nu = \lfloor L/2 \rfloor$ and (5.15) with $\nu = L - 1$. The proofs are in Appendix D.5. □

Proof of Theorem 5.13. We only prove Part (1) for the spaces W^α_q. The proof for the N^α_q spaces and that of Part (2) are similar.

Let $s \in (0, \infty)$ be arbitrary, and choose $r' \in \mathbb{N}$ such that $r \leq r'$ and $s \in (0, r'+1)$. Combining Lemmas 5.18 and 5.19, we get $W^\alpha_q(X, \varrho_{r'}, \mathscr{L}) \hookrightarrow \big(L_p(\Omega), B^s_{\sigma,\sigma}(\Omega)\big)_{\frac{\alpha}{s\nu},q}$. Since Ω is bounded, Theorem 4.7 shows that $W^\alpha_q(X, \varrho_r, \mathscr{L}) \hookrightarrow W^\alpha_q(X, \varrho_{r'}, \mathscr{L})$. By combining the two embeddings, we get the claim. □

References

[1] R. A. Adams and J. J. F. Fournier. *Sobolev spaces*, volume 140 of *Pure and Applied Mathematics (Amsterdam)*. Elsevier/Academic Press, Amsterdam, second edition, 2003.

[2] J. Adler and O. Öktem. Solving ill-posed inverse problems using iterative deep neural networks. *Inverse Problems*, 33:124007, 2017.

[3] C. D. Aliprantis and K. C. Border. *Infinite dimensional analysis: A hitchhiker's guide*. Springer, Berlin, third edition, 2006.

[4] J.M. Almira and U. Luther. Generalized approximation spaces and applications. *Math. Nachr.*, 263/264:3–35, 2004.

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

[6] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Mach. Learn.*, 14(1):115–133, 1994.

[7] Peter L Bartlett, Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension and pseudodimension bounds for piecewise linear neural networks. *arXiv*, March 2017.

[8] C. Beck, S. Becker, P. Grohs, N. Jaafari, and A. Jentzen. Solving stochastic differential equations and kolmogorov equations by means of deep learning. *arXiv preprint arXiv:1806.00421*, 2018.

[9] H. Bölcskei, P. Grohs, G. Kutyniok, and P. Petersen. Optimal approximation with sparsely connected deep neural networks. *SIAM J. Math. Data Sci.*, 1:8–45, 2019.

[10] T.A. Bubba, G. Kutyniok, M. Lassas, M. März, W. Samek, S. Siltanen, and V. Srinivasan. Learning the invisible: A hybrid deep learning-shearlet framework for limited angle computed tomography. *Inverse Probl.*, 2019. to appear.

[11] H.-Q. Bui and R. S. Laugesen. Affine systems that span Lebesgue spaces. *J. Fourier Anal. Appl.*, 11(5):533–556, 2005.

[12] E. J. Candès. Ridgelets: Theory and Applications, 1998. Ph.D. thesis, Stanford University.

[13] C. K. Chui, Xin Li, and H. N. Mhaskar. Neural networks for localized approximation. *Math. Comp.*, 63(208):607–623, 1994.

[14] N. Cohen, O. Sharir, and A. Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on Learning Theory*, pages 698–728, 2016.

[15] N. Cohen and A. Shashua. Convolutional rectifier networks as generalized tensor decompositions. In *International Conference on Machine Learning*, pages 955–963, 2016.

[16] G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

[17] R. A. DeVore. Nonlinear approximation. In *Acta numerica*, pages 51–150. Cambridge Univ. Press, Cambridge, 1998.

[18] R. A. DeVore, K.I. Oskolkov, and P.P. Petrushev. Approximation by feed-forward neural networks. *Ann. Numer. Math.*, 4:261–287, 1996.

[19] R. A. DeVore and V. A. Popov. Interpolation of Besov spaces. *Trans. Amer. Math. Soc.*, 305(1):397–414, January 1988.

[20] R. A. DeVore and R. C. Sharpley. Besov spaces on domains in $\mathbf{R}^d$. *Trans. Amer. Math. Soc.*, 335(2):843–864, 1993.

[21] R.A. DeVore and G.G. Lorentz. *Constructive approximation*, volume 303 of *Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]*. Springer-Verlag, Berlin, 1993.

[22] M. Elad. Deep, deep trouble. deep learning's impact on image processing, mathematics, and humanity. *SIAM News*, 2017.

[23] R. Eldan and O. Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory, COLT 2016, New York, USA, June 23-26, 2016*, pages 907–940, 2016.

[24] SW Ellacott. Aspects of the numerical analysis of neural networks. *Acta Numer.*, 3:145–202, 1994.

[25] J. Elstrodt. *Maß- und Integrationstheorie*. Springer Spektrum. Springer Spektrum, Berlin, Heidelberg, eigth edition, 2018.

[26] G.B. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and applied mathematics. Wiley, second edition, 1999.

[27] Gerald B. Folland. *A course in abstract harmonic analysis*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1995.

[28] Simon Foucart and Holger Rauhut. *A Mathematical Introduction to Compressive Sensing*. Springer, May 2012.

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

[30] Karol Gregor and Yann LeCun. Learning Fast Approximations of Sparse Coding. In *Proceedings of the 27th Annual International Conference on Machine Learning*, pages 399–406, 2010.

[31] J T Håstad. Computational Limitations for Small-Depth Circuits. ACM Doctoral Dissertation Award (1986), 1987.

[32] 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.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ICCV '15, pages 1026–1034, Washington, DC, USA, 2015. IEEE Computer Society.

[34] K. Hoffman and R. Kunze. *Linear algebra*. Second edition. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1971.

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

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

[37] H. Johnen and K. Scherer. On the equivalence of the K-functional and moduli of continuity and some applications. In *Constructive theory of functions of several variables (Proc. Conf., Math. Res. Inst., Oberwolfach, 1976)*, pages 119–140. Lecture Notes in Math., Vol. 571. Springer, Berlin, 1977.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'12, pages 1097–1105, USA, 2012. Curran Associates Inc.

[39] R. S. Laugesen. Affine synthesis onto L^p when $0 < p \leq 1$. *J. Fourier Anal. Appl.*, 14(2):235–266, 2008.

[40] P. D. Lax and M. S. Terrell. *Calculus with applications*. Undergraduate Texts in Mathematics. Springer, New York, second edition, 2014.

[41] Luc Le Magoarou and Remi Gribonval. Flexible Multi-layer Sparse Approximations of Matrices and Applications. *IEEE Journal of Selected Topics in Signal Processing*, June 2016.

[42] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[43] M. Leshno, V. Ya. Lin, A. Pinkus, and S. Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Netw.*, 6(6):861–867, 1993.

[44] A. L. Maas, A. Y. Hannun, and A. Y. Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. ICML*, volume 30, page 3, 2013.

[45] V. Maiorov and A. Pinkus. Lower bounds for approximation by MLP neural networks. *Neurocomputing*, 25(1):81–91, 1999.

[46] Stéphane Mallat. Understanding deep convolutional networks. *Phil. Trans. R. Soc. A*, 374(2065):20150203–16, March 2016.

[47] A. Mardt, L. Pasquali, H. Wu, and F. Noé. Vampnets: Deep learning of molecular kinetics. *Nature communications*, 9:5, 2018.

[48] W. S. McCulloch and W. Pitts. A logical calculus of the ideas immanent in nervous activity. *Bull. Math. Biophys.*, 5(4):115–133, 1943.

[49] H. N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Adv. Comput. Math.*, 1(1):61–80, 1993.

[50] H. N. Mhaskar. Neural networks for optimal approximation of smooth and analytic functions. *Neural Comput.*, 8(1):164–177, 1996.

[51] H. N. Mhaskar and T. Poggio. Deep vs. shallow networks: An approximation theory perspective. *Analysis and Applications*, 14(06):829–848, 2016.

[52] H.N. Mhaskar and C.A. Micchelli. Degree of approximation by neural and translation networks with a single hidden layer. *Adv. Appl. Math.*, 16(2):151–183, 1995.

[53] T. Nguyen-Thien and T. Tran-Cong. Approximation of functions and their derivatives: A neural network implementation with applications. *Appl. Math. Model.*, 23(9):687–704, 1999.

[54] A.E. Orhan and X. Pitkow. Skip Connections Eliminate Singularities. *arXiv preprint arXiv:1701.09175*, 2017.

[55] P. Petersen and F. Voigtlaender. Optimal approximation of piecewise smooth functions using deep ReLU neural networks. *Neural Netw.*, 108:296–330, 2018.

[56] P.P. Petrushev. Direct and converse theorems for spline and rational approximation and Besov spaces. In *Function spaces and applications (Lund, 1986)*, volume 1302 of *Lecture Notes in Math.*, pages 363–377. Springer, Berlin, 1988.

[57] A. Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numer.*, 8:143–195, 1999.

[58] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. Springer International Publishing, Cham, 2015.

[59] W. Rudin. *Functional analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, Inc., New York, second edition, 1991.
[60] J. Schmidt-Hieber. Nonparametric regression using deep neural networks with ReLU activation function. *arXiv preprint arXiv:1708.06633*, math.ST, 2017.
[61] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8:13890, 2017.
[62] U. Shaham, A. Cloninger, and R. R. Coifman. Provable approximation properties for deep neural networks. *Appl. Comput. Harmon. Anal.*, 44(3):537–557, May 2018.
[63] A. N. Somashekhar and J. F. Peters. *Topology with Applications*. World Scientific, 2013.
[64] M. Telgarsky. Benefits of depth in neural networks. *arXiv preprint arXiv:1602.04485*, 2016.
[65] M. A. Unser. Splines: a perfect fit for signal and image processing. *IEEE Signal Processing Magazine*, 16(6):22–38, 1999.
[66] F. Voigtlaender. *Embedding Theorems for Decomposition Spaces with Applications to Wavelet Coorbit Spaces*. PhD thesis, RWTH Aachen University, 2015. `http://publications.rwth-aachen.de/record/564979`.
[67] Z. Wu, C. Shen, and A.v.d. Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080*, 2016.
[68] D. Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Netw.*, 2017.
[69] D. Yarotsky. Optimal approximation of continuous functions by very deep relu networks. *arXiv preprint arXiv:1802.03620*, 2018.

Appendix A. Proofs for Section 2

For a matrix $A \in \mathbb{R}^{n\times d}$, we write $A^T \in \mathbb{R}^{d\times n}$ for the transpose of A. For $i \in \{1,\dots,n\}$ we write $A_{i,-} \in \mathbb{R}^{1\times d}$ for the i-th row of A, while $A_{(i)} \in \mathbb{R}^{(n-1)\times d}$ denotes the matrix obtained by deleting the i-th row of A. We use the same notation $b_{(i)}$ for vectors $b \in \mathbb{R}^n \cong \mathbb{R}^{n\times 1}$. Finally, for $j \in \{1,\dots,d\}$, $A_{[j]} \in \mathbb{R}^{n\times(d-1)}$ denotes the matrix obtained by removing the j-th column of A.

A.1. **Proof of Lemma 2.6.** Write $N_0(\Phi) := d_{\text{in}}(\Phi) + d_{\text{out}}(\Phi) + N(\Phi)$ for the total number of neurons of the network Φ, *including the "non-hidden" neurons*.

The proof is by contradiction. Assume that there is a network Φ for which the claim fails. Among all such networks, consider one with minimal value of $N_0(\Phi)$, i.e., such that the claim holds for all networks Ψ with $N_0(\Psi) < N_0(\Phi)$. Let us write $\Phi = \big((T_1,\alpha_1),\dots,(T_L,\alpha_L)\big)$ with $T_\ell\, x = A^{(\ell)}x + b^{(\ell)}$, for certain $A^{(\ell)} \in \mathbb{R}^{N_\ell\times N_{\ell-1}}$ and $b^{(\ell)} \in \mathbb{R}^{N_\ell}$.

Let us first consider the case that

$$\forall\ \ell \in \{1,\dots,L\}\ \forall\ i \in \{1,\dots,N_\ell\}\ :\ A^{(\ell)}_{i,-} \neq 0. \tag{A.1}$$

By (A.1), we get $\|A^{(\ell)}\|_{\ell^0} \geq N_\ell \geq \|b^{(\ell)}\|_{\ell^0}$, so that

$$W_0(\Phi) = \sum_{\ell=1}^{L}(\|A^{(\ell)}\|_{\ell^0} + \|b^{(\ell)}\|_{\ell^0}) \leq 2\cdot\sum_{\ell=1}^{L}\|A^{(\ell)}\|_{\ell^0} = 2W(\Phi) \leq d_{\text{out}}(\Phi) + 2W(\Phi).$$

Hence, with $\widetilde{\Phi} = \Phi$, Φ satisfies the claim of the lemma, in contradiction to our assumption.

Thus, there is some $\ell_0 \in \{1,\dots,L\}$ and some $i \in \{1,\dots,N_{\ell_0}\}$ satisfying $A^{(\ell_0)}_{i,-} = 0$. In other words, there is a neuron that is not connected to the previous layers. Intuitively, one can "remove it" without changing $\mathtt{R}(\Phi)$. This is what we now show formally.

Let us write $\alpha_\ell = \bigotimes_{j=1}^{N_\ell}\varrho_j^{(\ell)}$ for certain $\varrho_j^{(\ell)} \in \{\mathrm{id}_{\mathbb{R}},\varrho\}$, and set $\theta_\ell := \alpha_\ell \circ T_\ell$, so that $\mathtt{R}(\Phi) = \theta_L \circ \cdots \circ \theta_1$. By our choice of ℓ_0 and i, note

$$\big(\theta_{\ell_0}(x)\big)_i = \varrho_i^{(\ell_0)}\left((A^{(\ell_0)}x + b^{(\ell_0)})_i\right) = \varrho_i^{(\ell_0)}\left(\langle A^{(\ell_0)}_{i,-}, x\rangle + b_i^{(\ell_0)}\right) = \varrho_i^{(\ell_0)}(b_i^{(\ell_0)}) =: c \in \mathbb{R}, \tag{A.2}$$

for arbitrary $x \in \mathbb{R}^{N_{\ell_0-1}}$. After these initial observations, we now distinguish four cases:

Case 1 (Neuron on the output layer of size $d_{\text{out}}(\Phi) = 1$): We have $\ell_0 = L$ and $N_L = 1$, so that necessarily $i = 1$. In view of Equation (A.2), we then have $\mathtt{R}(\Phi) \equiv c$. Thus, if we choose the affine-linear map $S_1 : \mathbb{R}^{N_0} \to \mathbb{R}^1, x \mapsto c$, and set $\gamma_1 := \mathrm{id}_{\mathbb{R}}$, then the strict ϱ-network $\widetilde{\Phi} := \big((S_1,\gamma_1)\big)$ satisfies $\mathtt{R}(\widetilde{\Phi}) \equiv c \equiv \mathtt{R}(\Phi)$, and $L(\widetilde{\Phi}) = 1 \leq L(\Phi)$, as well as $W_0(\widetilde{\Phi}) = 1 = d_{\text{out}}(\Phi) \leq d_{\text{out}}(\Phi) + 2W(\Phi)$ and $N(\widetilde{\Phi}) = 0 \leq N(\Phi)$. Thus, Φ satisfies the claim of the lemma, contradicting our assumption.

Case 2 (Neuron on the output layer of size $d_{\text{out}}(\Phi) > 1$): We have $\ell_0 = L$ and $N_L > 1$. Define

$$B^{(\ell)} := A^{(\ell)},\quad c^{(\ell)} := b^{(\ell)},\quad \text{and}\quad \beta_\ell := \alpha_\ell\quad \text{for}\quad \ell \in \{1,\dots,L-1\}.$$

We then set $B^{(L)} := A^{(L)}_{(i)} \in \mathbb{R}^{(N_L-1)\times N_{L-1}}$ and $c^{(L)} := b^{(L)}_{(i)} \in \mathbb{R}^{N_L-1}$, as well as $\beta_L := \mathrm{id}_{\mathbb{R}^{N_L-1}}$.

Setting $S_\ell\, x := B^{(\ell)}x + c^{(\ell)}$ for $x \in \mathbb{R}^{N_{\ell-1}}$, the network $\Phi_0 := \big((S_1,\beta_1),\dots,(S_L,\beta_L)\big)$ then satisfies $\mathtt{R}(\Phi_0)(x) = \big(\mathtt{R}(\Phi)(x)\big)_{(i)}$ for all $x \in \mathbb{R}^{N_0}$, and $N_0(\Phi_0) = N_0(\Phi) - 1 < N_0(\Phi)$. Furthermore, if Φ is strict, then so is Φ_0.

By the "minimality" assumption on Φ, there is thus a network $\widetilde{\Phi}_0$ (which is strict if Φ is strict) with $\mathtt{R}(\widetilde{\Phi}_0) = \mathtt{R}(\Phi_0)$ and such that $L' := L(\widetilde{\Phi}_0) \le L(\Phi_0) = L(\Phi)$, as well as $N(\widetilde{\Phi}_0) \le N(\Phi_0) = N(\Phi)$, and

$$W(\widetilde{\Phi}_0) \le W_0(\widetilde{\Phi}_0) \le d_{\mathrm{out}}(\Phi_0) + 2\cdot W(\Phi_0) \le d_{\mathrm{out}}(\Phi) - 1 + 2\cdot W(\Phi).$$

Let us write $\widetilde{\Phi}_0 = \big((U_1,\gamma_1),\dots,(U_{L'},\gamma_{L'})\big)$, with affine-linear maps $U_\ell : \mathbb{R}^{M_{\ell-1}} \to \mathbb{R}^{M_\ell}$, so that $U_\ell\, x = C^{(\ell)}x + d^{(\ell)}$ for $\ell \in \{1,\dots,L'\}$ and $x \in \mathbb{R}^{M_{\ell-1}}$. Note that $M_{L'} = N_L - 1$, and define

$$\widetilde{C}^{(L')} := \begin{pmatrix} C^{(L')}_{1,-} \\ \vdots \\ C^{(L')}_{i-1,-} \\ 0 \\ C^{(L')}_{i,-} \\ \vdots \\ C^{(L')}_{M_{L'},-} \end{pmatrix} \in \mathbb{R}^{N_L \times M_{L'-1}} \quad \text{and} \quad \widetilde{d}^{(L')} := \begin{pmatrix} d^{(L')}_1 \\ \vdots \\ d^{(L')}_{i-1} \\ c \\ d^{(L')}_i \\ \vdots \\ d^{(L')}_{M_{L'}} \end{pmatrix} \in \mathbb{R}^{N_L},$$

as well as $\widetilde{\gamma}_{L'} := \mathrm{id}_{\mathbb{R}^{N_L}}$, and $\widetilde{U}_{L'} : \mathbb{R}^{M_{L'-1}} \to \mathbb{R}^{N_L}, x \mapsto \widetilde{C}^{(L')}x + \widetilde{d}^{(L')}$, and finally

$$\widetilde{\Phi} := \big((U_1,\gamma_1),\dots,(U_{L'-1},\gamma_{L'-1}),(\widetilde{U}_{L'},\widetilde{\gamma}_{L'})\big).$$

By virtue of Equation (A.2), we then have $\mathtt{R}(\widetilde{\Phi}) = \mathtt{R}(\Phi)$, and if Φ is strict, then so is Φ_0 and thus also $\widetilde{\Phi}_0$ and $\widetilde{\Phi}$. Furthermore, we have $L(\widetilde{\Phi}) = L' \le L(\Phi)$, and $N(\widetilde{\Phi}) = N(\widetilde{\Phi}_0) \le N(\Phi)$, as well as $W(\widetilde{\Phi}) \le W_0(\widetilde{\Phi}) \le 1 + W_0(\widetilde{\Phi}_0) \le d_{\mathrm{out}}(\Phi) + 2W(\Phi)$. Thus, Φ satisfies the claim of the lemma, contradicting our assumption.

Case 3 (Hidden neuron on layer ℓ_0 with $N_{\ell_0} = 1$): We have $1 \le \ell_0 < L$ and $N_{\ell_0} = 1$. In this case, Equation (A.2) implies $\theta_{\ell_0} \equiv c$, whence $\mathtt{R}(\Phi) = \theta_L \circ \cdots \circ \theta_1 \equiv \widetilde{c}$ for some $\widetilde{c} \in \mathbb{R}^{N_L}$.

Thus, if we choose the affine map $S_1 : \mathbb{R}^{N_0} \to \mathbb{R}^{N_L}, x \mapsto \widetilde{c}$, then the strict ϱ-network $\widetilde{\Phi} = \big((S_1,\gamma_1)\big)$ satisfies $\mathtt{R}(\widetilde{\Phi}) \equiv \widetilde{c} \equiv \mathtt{R}(\Phi)$ and $L(\widetilde{\Phi}) = 1 \le L(\Phi)$, as well as $W_0(\widetilde{\Phi}) \le d_{\mathrm{out}}(\Phi) \le d_{\mathrm{out}}(\Phi) + 2\,W(\Phi)$ and $N(\widetilde{\Phi}) = 0 \le N(\Phi)$. Thus, Φ satisfies the claim of the lemma, in contradiction to our choice of Φ.

Case 4 (Hidden neuron on layer ℓ_0 with $N_{\ell_0} > 1$): In this case, we have $1 \le \ell_0 < L$ and $N_{\ell_0} > 1$. Define $S_\ell := T_\ell$ and $\beta_\ell := \alpha_\ell$ for $\ell \in \{1,\dots,L\} \setminus \{\ell_0, \ell_0+1\}$, and let us choose $S_{\ell_0} : \mathbb{R}^{N_{\ell_0-1}} \to \mathbb{R}^{N_{\ell_0}-1}, x \mapsto B^{(\ell_0)}x + c^{(\ell_0)}$, where

$$B^{(\ell_0)} := A^{(\ell_0)}_{(i)}, \quad c^{(\ell_0)} := b^{(\ell_0)}_{(i)}, \quad \text{and} \quad \beta_{\ell_0} := \varrho^{(\ell_0)}_1 \otimes \cdots \otimes \varrho^{(\ell_0)}_{i-1} \otimes \varrho^{(\ell_0)}_{i+1} \otimes \cdots \otimes \varrho^{(\ell_0)}_{N_{\ell_0}}.$$

Finally, for $x \in \mathbb{R}^{N_{\ell_0}-1}$, let $\iota_c(x) := \big(x_1,\dots,x_{i-1},c,x_i,\dots,x_{N_{\ell_0}-1}\big)^T \in \mathbb{R}^{N_{\ell_0}}$, and set $\beta_{\ell_0+1} := \alpha_{\ell_0+1}$, as well as

$$S_{\ell_0+1} : \mathbb{R}^{N_{\ell_0}-1} \to \mathbb{R}^{N_{\ell_0+1}}, x \mapsto A^{(\ell_0+1)}_{[i]}\, x + c\cdot A^{(\ell_0+1)}e_i + b^{(\ell_0+1)} = A^{(\ell_0+1)}(\iota_c(x)) + b^{(\ell_0+1)},$$

where e_i is the i-th element of the standard basis of $\mathbb{R}^{N_{\ell_0}}$.

Setting $\vartheta_\ell := \beta_\ell \circ S_\ell$ and recalling that $\theta_\ell = \alpha_\ell \circ T_\ell$ for $\ell \in \{1,\dots,L\}$, we then have $\vartheta_{\ell_0}(x) = (\theta_{\ell_0}(x))_{(i)}$ for all $x \in \mathbb{R}^{N_{\ell_0-1}}$. By virtue of Equation (A.2), this implies $\theta_{\ell_0}(x) = \iota_c(\vartheta_{\ell_0}(x))$, so that

$$S_{\ell_0+1}(\vartheta_{\ell_0}(x)) = A^{(\ell_0+1)}\big(\iota_c(\vartheta_{\ell_0}(x))\big) + b^{(\ell_0+1)} = A^{(\ell_0+1)}(\theta_{\ell_0}(x)) + b^{(\ell_0+1)} = T_{\ell_0+1}(\theta_{\ell_0}(x)).$$

Recalling that $\beta_{\ell_0+1} = \alpha_{\ell_0+1}$, we thus see $\vartheta_{\ell_0+1} \circ \vartheta_{\ell_0} = \theta_{\ell_0+1} \circ \theta_{\ell_0}$, which then easily shows $\mathtt{R}(\Phi_0) = \mathtt{R}(\Phi)$ for $\Phi_0 := \big((S_1,\beta_1),\dots,(S_L,\beta_L)\big)$. Note that if Φ is strict, then so is Φ_0. Furthermore, we have $N_0(\Phi_0) = N_0(\Phi) - 1 < N_0(\Phi)$ so that by "minimality" of Φ, there is a network $\widetilde{\Phi}_0$ (which is strict if Φ is strict) satisfying $\mathtt{R}(\widetilde{\Phi}_0) = \mathtt{R}(\Phi_0) = \mathtt{R}(\Phi)$ and furthermore $L(\widetilde{\Phi}_0) \le L(\Phi_0) = L(\Phi)$, as well as $N(\widetilde{\Phi}_0) \le N(\Phi_0) \le N(\Phi)$, and finally $W(\widetilde{\Phi}_0) \le W_0(\widetilde{\Phi}_0) \le d_{\mathrm{out}}(\Phi_0) + 2W(\Phi_0) \le d_{\mathrm{out}}(\Phi) + 2W(\Phi)$. Thus, the claim holds for Φ, contradicting our assumption. □

A.2. **Proof of Lemma 2.14.** We begin by showing $\mathtt{NN}^{\varrho,d,k}_{W,L,W} \subset \mathtt{NN}^{\varrho,d,k}_{W,W,W}$. Let $f \in \mathtt{NN}^{\varrho,d,k}_{W,L,W}$. By definition there is $\Phi \in \mathcal{NN}^{\varrho,d,k}_{W,L,W}$ such that $f = \mathtt{R}(\Phi)$. Note that $W(\Phi) \leq W$, and let us distinguish two cases: If $L(\Phi) \leq W(\Phi)$ then $L(\Phi) \leq W$, whence in fact $\Phi \in \mathcal{NN}^{\varrho,d,k}_{W,W,W}$ and $f \in \mathtt{NN}^{\varrho,d,k}_{W,W,W}$ as claimed. Otherwise, $W(\Phi) < L(\Phi)$ and by Corollary 2.10 we have $f = \mathtt{R}(\Phi) \equiv c$ for some $c \in \mathbb{R}^k$. Therefore, Lemma 2.13 shows that $f \in \mathtt{NN}^{\varrho,d,k}_{0,1,0} \subset \mathtt{NN}^{\varrho,d,k}_{W,W,W}$, where the inclusion holds by definition of these sets.

The inclusion $\mathtt{NN}^{\varrho,d,k}_{W,L,W} \subset \mathtt{NN}^{\varrho,d,k}_{W,L,\infty}$ is trivial. Similarly, if $L \geq W$ then trivially $\mathtt{NN}^{\varrho,d,k}_{W,W,W} \subset \mathtt{NN}^{\varrho,d,k}_{W,L,W}$.

Thus, it remains to show $\mathtt{NN}^{\varrho,d,k}_{W,L,\infty} \subset \mathtt{NN}^{\varrho,d,k}_{W,L,W}$. To prove this, we will show that for each network $\Phi = \big((T_1,\alpha_1),\dots,(T_K,\alpha_K)\big) \in \mathcal{NN}^{\varrho,d,k}_{W,L,\infty}$ (so that necessarily $K \leq L$) with $N(\Phi) > W$, one can find a neural network $\Phi' \in \mathcal{NN}^{\varrho,d,k}_{W,L,\infty}$ with $\mathtt{R}(\Phi') = \mathtt{R}(\Phi)$, and such that $N(\Phi') < N(\Phi)$. If Φ is strict, then we show that Φ' can also be chosen to be strict. The desired inclusion can then be obtained by repeating this "compression" step until one reaches the point where $N(\Phi') \leq W$.

For each $\ell \in \{1,\dots,K\}$, let $b^{(\ell)} \in \mathbb{R}^{N_\ell}$ and $A^{(\ell)} \in \mathbb{R}^{N_\ell \times N_{\ell-1}}$ be such that $T_\ell = A^{(\ell)} \bullet + b^{(\ell)}$. Since $\Phi \in \mathcal{NN}^{\varrho,d,k}_{W,L,\infty}$, we have $W(\Phi) \leq W$. In combination with $N(\Phi) > W$, this implies

$$\sum_{\ell=1}^{K-1} N_\ell = N(\Phi) > W \geq W(\Phi) = \sum_{\ell=1}^{K} \|A^{(\ell)}\|_{\ell_0} \geq \sum_{\ell=1}^{K-1} \sum_{i=1}^{N_\ell} \|A^{(\ell)}_{i,-}\|_{\ell^0}\,.$$

Therefore, $K > 1$, and there must be some $\ell_0 \in \{1,\dots,K-1\}$ and $i \in \{1,\dots,N_{\ell_0}\}$ with $A^{(\ell_0)}_{i,-} = 0$. We now distinguish two cases:

Case 1 (Single neuron on layer ℓ_0): We have $N_{\ell_0} = 1$. In this case, $A^{(\ell_0)} = 0$ and hence $T_{\ell_0} \equiv b^{(\ell_0)}$. Therefore, $\mathtt{R}(\Phi)$ is constant; say $\mathtt{R}(\Phi) \equiv c \in \mathbb{R}^k$. Choose $S_1 : \mathbb{R}^d \to \mathbb{R}^k, x \mapsto c$, and $\beta_1 := \mathrm{id}_{\mathbb{R}^k}$. Then $\mathtt{R}(\Phi) \equiv c \equiv \mathtt{R}(\Phi')$ for the strict ϱ-network $\Phi' := \big((S_1,\beta_1)\big) \in \mathcal{NN}^{\varrho,d,k}_{0,1,0} \subset \mathcal{NN}^{\varrho,d,k}_{W,L,\infty}$, which indeed satisfies $N(\Phi') = 0 \leq W < N(\Phi)$.

Case 2 (Multiple neurons on layer ℓ_0): We have $N_{\ell_0} > 1$. Recall that $\ell_0 \in \{1,\dots,K-1\}$, so that $\ell_0 + 1 \in \{1,\dots,K\}$. Now define $S_\ell := T_\ell$ and $\beta_\ell := \alpha_\ell$ for $\ell \in \{1,\dots,K\} \setminus \{\ell_0, \ell_0+1\}$. Further, define

$$S_{\ell_0} : \mathbb{R}^{N_{\ell_0-1}} \to \mathbb{R}^{N_{\ell_0}-1}, \quad \text{with} \quad (S_{\ell_0}\,x)_j := \begin{cases} (T_{\ell_0}\,x)_j, & \text{if } j < i, \\ (T_{\ell_0}\,x)_{j+1}, & \text{if } j \geq i \end{cases} \quad \text{for } j \in \{1,\dots,N_{\ell_0}-1\}.$$

Using the notation $A_{(i)}, b_{(i)}$ from the beginning of Appendix A, this means $S_{\ell_0}\,x = A^{(\ell_0)}_{(i)} x + b^{(\ell_0)}_{(i)} = (T_{\ell_0}\,x)_{(i)}$.

Finally, writing $\alpha_\ell = \varrho^{(\ell)}_1 \otimes \cdots \otimes \varrho^{(\ell)}_{N_\ell}$ for $\ell \in \{1,\dots,K\}$, define $\beta_{\ell_0+1} := \alpha_{\ell_0+1}$, as well as

$$\beta_{\ell_0} := \varrho^{(\ell_0)}_1 \otimes \cdots \otimes \varrho^{(\ell_0)}_{i-1} \otimes \varrho^{(\ell_0)}_{i+1} \otimes \cdots \otimes \varrho^{(\ell_0)}_{N_{\ell_0}} \quad : \quad \mathbb{R}^{N_{\ell_0}-1} \to \mathbb{R}^{N_{\ell_0}-1}\,,$$

and

$$\begin{aligned} S_{\ell_0+1} : \mathbb{R}^{N_{\ell_0}-1} \to \mathbb{R}^{N_{\ell_0+1}}, y \mapsto\ & T_{\ell_0+1}\left(y_1,\dots,y_{i-1},\varrho^{(\ell_0)}_i(b^{(\ell_0)}_i),y_i,\dots,y_{N_{\ell_0}-1}\right) \\ & = A^{(\ell_0+1)}_{[i]} y + b^{(\ell_0+1)} + \varrho^{(\ell_0)}_i(b^{(\ell_0)}_i) \cdot A^{(\ell_0+1)}\,e_i, \end{aligned}$$

where $e_i \in \mathbb{R}^{N_{\ell_0}}$ denotes the i-th element of the standard basis, and where $A_{[i]}$ is the matrix obtained from a given matrix A by removing its i-th column.

Now, for arbitrary $x \in \mathbb{R}^{N_{\ell_0-1}}$, let $y := S_{\ell_0}\,x \in \mathbb{R}^{N_{\ell_0}-1}$ and $z := T_{\ell_0}\,x \in \mathbb{R}^{N_{\ell_0}}$. Because of $A^{(\ell_0)}_{i,-} = 0$, we then have $z_i = b^{(\ell_0)}_i$. Further, by definition of S_{ℓ_0}, we have $y_j = (T_{\ell_0}\,x)_j = z_j$ for $j < i$, and $y_j = (T_{\ell_0}\,x)_{j+1} = z_{j+1}$ for $j \geq i$. All in all, this shows

$$\begin{aligned} S_{\ell_0+1}\big(\beta_{\ell_0}(S_{\ell_0}x)\big) &= S_{\ell_0+1}(\beta_{\ell_0}(y)) \\ &= T_{\ell_0+1}\left(\varrho^{(\ell_0)}_1(y_1),\dots,\varrho^{(\ell_0)}_{i-1}(y_{i-1}),\varrho^{(\ell_0)}_i(b^{(\ell_0)}_i),\varrho^{(\ell_0)}_{i+1}(y_i),\dots,\varrho^{(\ell_0)}_{N_{\ell_0}}(y_{N_{\ell_0}-1})\right) \\ &= T_{\ell_0+1}\left(\varrho^{(\ell_0)}_1(z_1),\dots,\varrho^{(\ell_0)}_{i-1}(z_{i-1}),\varrho^{(\ell_0)}_i(z_i),\varrho^{(\ell_0)}_{i+1}(z_{i+1}),\dots,\varrho^{(\ell_0)}_{N_{\ell_0}}(z_{N_{\ell_0}})\right) \\ &= T_{\ell_0+1}\big(\alpha_{\ell_0}(z)\big) = T_{\ell_0+1}\big(\alpha_{\ell_0}(T_{\ell_0}x)\big)\,. \end{aligned}$$

Recall that this holds for all $x \in \mathbb{R}^{N_{\ell_0-1}}$. From this, it is not hard to see $\mathtt{R}(\Phi) = \mathtt{R}(\Phi')$ for the network $\Phi' := \big((S_1,\beta_1),\dots,(S_K,\beta_K)\big) \in \mathcal{NN}^{\varrho,d,k}_{\infty,K,\infty} \subset \mathcal{NN}^{\varrho,d,k}_{\infty,L,\infty}$. Note that Φ' is a strict network if Φ is strict. Finally, directly from the definition of Φ', we see $W(\Phi') \leq W(\Phi) \leq W$, so that $\Phi' \in \mathcal{NN}^{\varrho,d,k}_{W,L,\infty}$. Also, $N(\Phi') = N(\Phi) - 1 < N(\Phi)$, as desired. □

A.3. **Proof of Lemma 2.16.** Write $\Phi = \big((T_1,\alpha_1),\dots,(T_L,\alpha_L)\big)$ with $L = L(\Phi)$. If $L_0 = 0$, we can simply choose $\Psi = \Phi$. Thus, let us assume $L_0 > 0$, and distinguish two cases:

Case 1: If $k \leq d$, so that $c = k$, set

$$\Psi := \Big((T_1,\alpha_1),\dots,(T_L,\alpha_L),\underbrace{(\mathrm{id}_{\mathbb{R}^k},\mathrm{id}_{\mathbb{R}^k}),\dots,(\mathrm{id}_{\mathbb{R}^k},\mathrm{id}_{\mathbb{R}^k})}_{L_0 \text{ terms}}\Big),$$

and note that the affine map $T := \mathrm{id}_{\mathbb{R}^k}$ satisfies $\|T\|_{\ell^0} = k = c$, and hence $W(\Psi) = W(\Phi) + c\,L_0$. Furthermore, $\mathtt{R}(\Psi) = \mathtt{R}(\Phi)$, $L(\Psi) = L(\Phi) + L_0$, and $N(\Psi) = N(\Phi) + cL_0$. Here we used crucially that the definition of *generalized* neural networks allows us to use the identity as the activation function for some neurons.

Case 2: If $d < k$, so that $c = d$, the proof proceeds as in the previous case, but with

$$\Psi := \Big(\underbrace{(\mathrm{id}_{\mathbb{R}^d},\mathrm{id}_{\mathbb{R}^d}),\dots,(\mathrm{id}_{\mathbb{R}^d},\mathrm{id}_{\mathbb{R}^d})}_{L_0 \text{ terms}},(T_1,\alpha_1),\dots,(T_L,\alpha_L)\Big). \qquad \square$$

A.4. **Proof of Lemma 2.17.** For the proof of the first part, denoting $\Phi = \big((T_1,\alpha_1),\dots,(T_L,\alpha_L)\big)$, we set $\Psi := \big((T_1,\alpha_1),\dots,(c\cdot T_L,\alpha_L)\big)$. By Definition 2.1 we have $\alpha_L = \mathrm{id}_{\mathbb{R}^k}$, hence one easily sees $\mathtt{R}(\Psi) = c\cdot\mathtt{R}(\Phi)$. If Φ is strict, then so is Ψ. By construction Φ and Ψ have the same number of layers and neurons, and $W(\Psi) \leq W(\Phi)$ with equality if $c \neq 0$.

For the second and third part, we proceed by induction, using two auxiliary claims.

Lemma A.1. *Let $\Psi_1 \in \mathcal{NN}^{\varrho,d,k_1}$ and $\Psi_2 \in \mathcal{NN}^{\varrho,d,k_2}$. There is a network $\Psi \in \mathcal{NN}^{\varrho,d,k_1+k_2}$ with $L(\Psi) = \max\{L(\Psi_1),L(\Psi_2)\}$ such that $\mathtt{R}(\Psi) = g$, where $g : \mathbb{R}^d \to \mathbb{R}^{k_1+k_2}, x \mapsto \big(\mathtt{R}(\Psi_1)(x),\mathtt{R}(\Psi_2)(x)\big)$. Furthermore, setting $c := \min\big\{d,\max\{k_1,k_2\}\big\}$, Ψ can be chosen to satisfy*

$$\begin{aligned} W(\Psi) &\leq W(\Psi_1) + W(\Psi_2) + c\cdot|L(\Psi_2) - L(\Psi_1)| \\ N(\Psi) &\leq N(\Psi_1) + N(\Psi_2) + c\cdot|L(\Psi_2) - L(\Psi_1)|\,. \end{aligned} \qquad \blacktriangleleft$$

Lemma A.2. *Let $\Psi_1,\Psi_2 \in \mathcal{NN}^{\varrho,d,k}$. There is $\Psi \in \mathcal{NN}^{\varrho,d,k}$ with $L(\Psi) = \max\{L(\Psi_1),L(\Psi_2)\}$ such that $\mathtt{R}(\Psi) = \mathtt{R}(\Psi_1) + \mathtt{R}(\Psi_2)$ and, with $c = \min\{d,k\}$,*

$$\begin{aligned} W(\Psi) &\leq W(\Psi_1) + W(\Psi_2) + c\cdot|L(\Psi_2) - L(\Psi_1)| \\ N(\Psi) &\leq N(\Psi_1) + N(\Psi_2) + c\cdot|L(\Psi_2) - L(\Psi_1)|\,. \end{aligned} \qquad \blacktriangleleft$$

Proof of Lemmas A.1 and A.2. Set $L := \max\{L(\Psi_1),L(\Psi_2)\}$ and $L_i := L(\Psi_i)$ for $i \in \{1,2\}$. By Lemma 2.16 applied to Ψ_i and $L_0 = L - L_i \in \mathbb{N}_0$, we get for each $i \in \{1,2\}$ a network $\Psi_i' \in \mathcal{NN}^{\varrho,d,k_i}$ with $\mathtt{R}(\Psi_i') = \mathtt{R}(\Psi_i)$ and such that $L(\Psi_i') = L$, as well as $W(\Psi_i') \leq W(\Psi_i) + c(L - L_i)$ and furthermore $N(\Psi_i') \leq N(\Psi_i) + c(L - L_i)$. By choice of L, we have $(L - L_1) + (L - L_2) = |L_1 - L_2|$, whence $W(\Psi_1') + W(\Psi_2') \leq W(\Psi_1) + W(\Psi_2) + c\,|L_1 - L_2|$, and $N(\Psi_1') + N(\Psi_2') \leq N(\Psi_1) + N(\Psi_2) + c\,|L_1 - L_2|$.

First we deal with the pathological case $L = 1$. In this case, each Ψ_i' is of the form $\Psi_i' = \big((T_i,\mathrm{id}_{\mathbb{R}^k})\big)$, with $T_i : \mathbb{R}^d \to \mathbb{R}^k$ an affine-linear map. For proving Lemma A.1, we set $\Psi := \big((T,\mathrm{id}_{\mathbb{R}^{k_1+k_2}})\big)$ with the affine-linear map $T : \mathbb{R}^d \to \mathbb{R}^{k_1+k_2},\ x \mapsto \big(T_1(x),T_2(x)\big)$, so that $\mathtt{R}(\Psi) = g$. For proving Lemma A.2, we set $\Psi := \big((T,\mathrm{id}_{\mathbb{R}^k})\big)$ with $T = T_1 + T_2$, so that $\mathtt{R}(\Psi) = T_1 + T_2 = \mathtt{R}(\Psi_1') + \mathtt{R}(\Psi_2') = \mathtt{R}(\Psi_1) + \mathtt{R}(\Psi_2)$. Finally, we see for both cases that $N(\Psi) = 0 = N(\Psi_1') + N(\Psi_2')$ and

$$W(\Psi) = \|T\|_{\ell^0} \leq \|T_1\|_{\ell^0} + \|T_2\|_{\ell^0} = W(\Psi_1') + W(\Psi_2')\,.$$

This establishes the result for the case $L = 1$.

For $L > 1$, write $\Psi_1' = \big((T_1,\alpha_1),\dots,(T_L,\alpha_L)\big)$ and $\Psi_2' = \big((S_1,\beta_1),\dots,(S_L,\beta_L)\big)$ with affine-linear maps $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ and $S_\ell : \mathbb{R}^{M_{\ell-1}} \to \mathbb{R}^{M_\ell}$ for $\ell \in \{1,\dots,L\}$. Let us define $\theta_\ell := \alpha_\ell \otimes \beta_\ell$ for $\ell \in \{1,\dots,L\}$—except for $\ell = L$ when proving Lemma A.2, in which case we set $\theta_L := \mathrm{id}_{\mathbb{R}^k}$. Next, set

$$R_1 : \mathbb{R}^d \to \mathbb{R}^{N_1+M_1}, x \mapsto (T_1x, S_1x) \quad \text{and} \quad R_\ell : \mathbb{R}^{N_{\ell-1}+M_{\ell-1}} \to \mathbb{R}^{N_\ell+M_\ell}, (x,y) \mapsto (T_\ell\,x, S_\ell\,y)$$

for $2 \leq \ell \leq L$—except if $\ell = L$ when proving Lemma A.2. In this latter case, we instead define R_L as $R_L : \mathbb{R}^{N_{L-1}+M_{L-1}} \to \mathbb{R}^k, (x,y) \mapsto T_L\,x + S_L\,y$. Finally set $\Psi := \big((R_1,\theta_1),\dots,(R_L,\theta_L)\big)$.

When proving Lemma A.1, it is straightforward to verify that Ψ satisfies

$$\mathtt{R}(\Psi)(x) = \big(\mathtt{R}(\Psi_1')(x),\mathtt{R}(\Psi_2')(x)\big) = \big(\mathtt{R}(\Psi_1)(x),\mathtt{R}(\Psi_2)(x)\big) = g(x) \qquad \forall\, x \in \mathbb{R}^d\,.$$

Similarly, when proving Lemma A.2, one can easily check that $\mathtt{R}(\Psi) = \mathtt{R}(\Psi_1') + \mathtt{R}(\Psi_2') = \mathtt{R}(\Psi_1) + \mathtt{R}(\Psi_2)$.

Further, for arbitrary $\ell \in \{1, \dots, L\}$, we have $\|R_\ell\|_{\ell^0} \leq \|T_\ell\|_{\ell^0} + \|S_\ell\|_{\ell^0}$ so that

$$W(\Psi) = \sum_{\ell=1}^{L} \|R_\ell\|_{\ell^0} \leq \sum_{\ell=1}^{L} (\|T_\ell\|_{\ell^0} + \|S_\ell\|_{\ell^0}) = W(\Psi_1') + W(\Psi_2')\,.$$

Finally, $N(\Psi) = \sum_{\ell=1}^{L-1}(N_\ell + M_\ell) = N(\Psi_1') + N(\Psi_2')$. Given the estimates for $W(\Psi_1') + W(\Psi_2')$ and $N(\Psi_1') + N(\Psi_2')$ stated at the beginning of the proof, this yields the claim. □

Let us now return to the proof of Parts 2 and 3 of Lemma 2.17. Set $f_i := \mathtt{R}(\Phi_i)$ and $L_i := L(\Phi_i)$. We first show that we can without loss of generality assume $L_1 \leq \cdots \leq L_n$. To see this, note that there is a permutation $\sigma \in S_n$ such that if we set $\Gamma_j := \Phi_{\sigma(j)}$, then $L(\Gamma_1) \leq \cdots \leq L(\Gamma_n)$. Furthermore, $\sum_{j=1}^{n} \mathtt{R}(\Gamma_j) = \sum_{j=1}^{n} \mathtt{R}(\Phi_j)$. Finally, there is a permutation matrix $P \in \mathrm{GL}(\mathbb{R}^d)$ such that

$$P \circ \big(\mathtt{R}(\Gamma_1), \dots, \mathtt{R}(\Gamma_n)\big) = \big(\mathtt{R}(\Phi_1), \dots, \mathtt{R}(\Phi_n)\big) = (f_1, \dots, f_n) = g\,.$$

Since the permutation matrix P has exactly one non-zero entry per row and column, we have $\|P\|_{\ell^{0,\infty}} = 1$ in the notation of Equation (2.4). Therefore, the first part of Lemma 2.18 (which will be proven independently) shows that $g \in \mathtt{NN}_{W,L,N}^{\varrho,d,K}$, provided that $\big(\mathtt{R}(\Gamma_1), \dots, \mathtt{R}(\Gamma_n)\big) \in \mathtt{NN}_{W,L,N}^{\varrho,d,K}$. These considerations show that we can assume $L(\Phi_1) \leq \cdots \leq L(\Phi_n)$ without loss of generality.

We now prove the following claim by induction on $j \in \{1, \dots, n\}$: There is $\Theta_j \in \mathcal{NN}^{\varrho,d,K_j}$ satisfying $W(\Theta_j) \leq \sum_{i=1}^{j} W(\Phi_i) + c\,(L_j - L_1)$, and $N(\Theta_j) = \sum_{i=1}^{j} N(\Phi_i) + c\,(L_j - L_1)$, as well as $L(\Theta_j) = L_j$, and such that $\mathtt{R}(\Theta_j) = g_j := \sum_{i=1}^{j} f_i$ and $K_j := k$ for the summation, respectively such that $\mathtt{R}(\Theta_j) = g_j := (f_1, \dots, f_j)$ and $K_j := \sum_{i=1}^{j} k_i$ for the cartesian product. Here, c is as in the corresponding claim of Lemma 2.17.

Specializing to $j = n$ then yields the conclusion of Lemma 2.17.

We now proceed to the induction. The claim trivially holds for $j = 1$—just take $\Theta_1 = \Phi_1$. Assuming that the claim holds for some $j \in \{1, \dots, n-1\}$, we define $\Psi_1 := \Theta_j$ and $\Psi_2 := \Phi_{j+1}$. Note that $L(\Psi_1) = L(\Theta_j) = L_j \leq L_{j+1} = L(\Psi_2)$. For the summation, by Lemma A.2 there is a network $\Psi \in \mathcal{NN}^{\varrho,d,k}$ with $L(\Psi) = L_{j+1}$ and $\mathtt{R}(\Psi) = \mathtt{R}(\Psi_1) + \mathtt{R}(\Psi_2) = \mathtt{R}(\Theta_j) + \mathtt{R}(\Phi_{j+1}) = g_j + f_{j+1} = g_{j+1}$, and such that

$$W(\Psi) \leq W(\Psi_1) + W(\Psi_2) + c' \cdot |L(\Psi_2) - L(\Psi_1)| \leq W(\Theta_j) + W(\Phi_{j+1}) + c' \cdot (L_{j+1} - L_j)$$

and likewise $N(\Psi) \leq N(\Theta_j) + N(\Phi_{j+1}) + c' \cdot (L_{j+1} - L_j)$, where $c' = \min\{d, k\} = c$. For the cartesian product, Lemma A.1 yields a network $\Psi \in \mathcal{NN}^{\varrho,d,K_j+k_{j+1}} = \mathcal{NN}^{\varrho,d,K_{j+1}}$ satisfying

$$\mathtt{R}(\Psi) = \big(\mathtt{R}(\Psi_1), \mathtt{R}(\Psi_2)\big) = \big(\mathtt{R}(\Theta_j), \mathtt{R}(\Phi_{j+1})\big) = g_{j+1}$$

and such that, setting $c' := \min\big\{d, \max\{K_j, k_{j+1}\}\big\} \leq \min\{d, K-1\} = c$, we have

$$W(\Psi) \leq W(\Psi_1) + W(\Psi_2) + c' \cdot |L(\Psi_2) - L(\Psi_1)| = W(\Theta_j) + W(\Phi_{j+1}) + c' \cdot (L_{j+1} - L_j)$$

and $N(\Psi) \leq N(\Theta_j) + N(\Phi_{j+1}) + c' \cdot (L_{j+1} - L_j)$.

With $\Theta_{j+1} := \Psi$ we get $\mathtt{R}(\Theta_{j+1}) = g_{j+1}$, $L(\Theta_{j+1}) = L_{j+1}$ and, by the induction hypothesis,

$$W(\Theta_{j+1}) \leq \sum_{i=1}^{j} W(\Phi_i) + c\,(L_j - L_1) + W(\Phi_{j+1}) + c\,(L_{j+1} - L_j) = \sum_{i=1}^{j+1} W(\Phi_i) + c\,(L_{j+1} - L_1)\,.$$

Similarly, $N(\Theta_{j+1}) \leq \sum_{i=1}^{j+1} N(\Phi_i) + c \cdot (L_{j+1} - L_1)$. This completes the induction and the proof. □

A.5. **Proof of Lemma 2.18.** We prove each part of the lemma individually.

Part (2)**:** Let $\Phi_1 = \big((T_1, \alpha_1), \dots, (T_{L_1}, \alpha_{L_1})\big) \in \mathcal{NN}^{\varrho,d,d_1}$ and $\Phi_2 = \big((S_1, \beta_1), \dots, (S_{L_2}, \beta_{L_2})\big) \in \mathcal{NN}^{\varrho,d_1,d_2}$ Define

$$\Psi := \big((T_1, \alpha_1), \dots, (T_{L_1}, \alpha_{L_1}), (S_1, \beta_1), \dots, (S_{L_2}, \beta_{L_2})\big)\,.$$

We emphasize that Ψ is indeed a *generalized* ϱ-network, since all T_ℓ and all S_ℓ are affine-linear (with "fitting" dimensions), and since all α_ℓ and all β_ℓ are $\otimes$-products of ϱ and $\mathrm{id}_{\mathbb{R}}$, with $\beta_{L_2} = \mathrm{id}_{\mathbb{R}^{d_2}}$. Furthermore, we clearly have $L(\Psi) = L_1 + L_2 = L(\Phi_1) + L(\Phi_2)$, and

$$W(\Psi) = \sum_{\ell=1}^{L_1} \|T_\ell\|_{\ell^0} + \sum_{\ell'=1}^{L_2} \|S_{\ell'}\|_{\ell^0} = W(\Phi_1) + W(\Phi_2).$$

Clearly, $N(\Psi) = N(\Phi_1) + d_1 + N(\Phi_2)$. Finally, the property $\mathtt{R}(\Psi) = \mathtt{R}(\Phi_2) \circ \mathtt{R}(\Phi_1)$ is a direct consequence of the definition of the realization of neural networks.

Part (1)**:** Let $\Phi = \big((T_1,\alpha_1),\dots,(T_L,\alpha_L)\big) \in \mathcal{NN}^{\varrho,d,k}$. We give the proof for $Q \circ \mathtt{R}(\Phi)$, since the proof for $\mathtt{R}(\Phi) \circ P$ is similar but simpler; the general statement in the lemma then follows from the identity $Q \circ \mathtt{R}(\Phi) \circ P = (Q \circ \mathtt{R}(\Phi)) \circ P = \mathtt{R}(\Psi_1) \circ P$.

We first treat the special case $\|Q\|_{\ell^{0,\infty}} = 0$ which implies $\|Q\|_{\ell^0} = 0$, and hence $Q \circ \mathtt{R}(\Phi) \equiv c$ for some $c \in \mathbb{R}^{k_1}$. Choose $N_0,\dots,N_L$ such that $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ for $\ell \in \{1,\dots,L\}$, and define $S_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}, x \mapsto 0$ for $\ell \in \{1,\dots,L-1\}$ and $S_L : \mathbb{R}^{N_{L-1}} \to \mathbb{R}^{k_1}, x \mapsto c$. It is then not hard to see that the network $\Psi := \big((S_1,\alpha_1),\dots,(S_L,\alpha_L)\big)$ satisfies $L(\Psi) = L(\Phi)$ and $N(\Psi) = N(\Phi)$, as well as $W(\Psi) = 0$ and $\mathtt{R}(\Psi) \equiv c = Q \circ \mathtt{R}(\Phi)$.

We now consider the case $\|Q\|_{\ell^{0,\infty}} \geq 1$. Define $U_\ell := T_\ell$ for $\ell \in \{1,\dots,L-1\}$ and $U_L := Q \circ T_L$. By Definition 2.1 we have $\alpha_L = \mathrm{id}_{\mathbb{R}^k}$, whence $\Psi := \big((U_1,\alpha_1),\dots,(U_{L-1},\alpha_{L-1}),(U_L,\mathrm{id}_{\mathbb{R}^{k_1}})\big) \in \mathcal{NN}^{\varrho,d,k_1}_{\infty,L,N(\Phi)}$ satisfies $\mathtt{R}(\Psi) = Q \circ \mathtt{R}(\Phi)$. To control $W(\Psi)$, we use the following lemma. The proof is slightly deferred.

Lemma A.3. *Let $p,q,r \in \mathbb{N}$ be arbitrary.*

(1) For arbitrary affine-linear maps $T : \mathbb{R}^p \to \mathbb{R}^q$ and $S : \mathbb{R}^q \to \mathbb{R}^r$, we have

$$\|S \circ T\|_{\ell^0} \leq \|S\|_{\ell^{0,\infty}} \cdot \|T\|_{\ell^0} \quad \text{and} \quad \|S \circ T\|_{\ell^0} \leq \|S\|_{\ell^0} \cdot \|T\|_{\ell^{0,\infty}_*} .$$

(2) For affine-linear maps $T_1,\dots,T_n$, we have $\|T_1 \otimes \cdots \otimes T_n\|_{\ell^0} \leq \sum_{i=1}^n \|T_i\|_{\ell^0}$, as well as

$$\|T_1 \otimes \cdots \otimes T_n\|_{\ell^{0,\infty}} \leq \max_{i\in\{1,\dots,n\}} \|T_i\|_{\ell^{0,\infty}} \quad \text{and} \quad \|T_1 \otimes \cdots \otimes T_n\|_{\ell^{0,\infty}_*} \leq \max_{i\in\{1,\dots,n\}} \|T_i\|_{\ell^{0,\infty}_*} . \qquad \blacktriangleleft$$

Let us continue with the proof from above. By definition, $\|U_\ell\|_{\ell^0} = \|T_\ell\|_{\ell^0} \leq \|Q\|_{\ell^{0,\infty}} \cdot \|T_\ell\|_{\ell^0}$ for $\ell \in \{1,\dots,L-1\}$. By Lemma A.3 we also have $\|U_L\|_{\ell^0} \leq \|Q\|_{\ell^{0,\infty}} \cdot \|T_L\|_{\ell^0}$, and hence

$$W(\Psi) = \sum_{\ell=1}^{L} \|U_\ell\|_{\ell^0} \leq \|Q\|_{\ell^{0,\infty}} \sum_{\ell=1}^{L} \|T_\ell\|_{\ell^0} = \|Q\|_{\ell^{0,\infty}} \cdot W(\Phi).$$

Finally, if Φ is strict, then Ψ is strict as well; thus, the claim also holds with $\mathtt{SNN}$ instead of $\mathtt{NN}$.

Part (3): Let $\Phi_1 = \big((T_1,\alpha_1),\dots,(T_L,\alpha_L)\big) \in \mathcal{NN}^{\varrho,d,d_1}$ and $\Phi_2 = \big((S_1,\beta_1),\dots,(S_K,\beta_K)\big) \in \mathcal{NN}^{\varrho,d_1,d_2}$.

We distinguish two cases: First, if $L = 1$, then $\mathtt{R}(\Phi_1) = T_1$. Since $T_1 : \mathbb{R}^d \to \mathbb{R}^{d_1}$, this implies $\|T_1\|_{\ell^{0,\infty}_*} \leq d$. Thus, Part (1) shows that

$$\mathtt{R}(\Phi_2) \circ \mathtt{R}(\Phi_1) = \mathtt{R}(\Phi_2) \circ T_1 \in \mathtt{NN}^{\varrho,d,d_2}_{d\cdot W(\Phi_2),K,N(\Phi_2)} \subset \mathtt{NN}^{\varrho,d,d_2}_{W(\Phi_1)+N\cdot W(\Phi_2),L+K-1,N(\Phi_1)+N(\Phi_2)},$$

where $N := \max\{N(\Phi_1),d\}$.

Let us now assume that $L > 1$. In this case, define

$$\Psi := \big((T_1,\alpha_1),\dots,(T_{L-1},\alpha_{L-1}),(S_1 \circ T_L,\beta_1),(S_2,\beta_2)\dots,(S_K,\beta_K)\big) .$$

It is not hard to see that $N(\Psi) \leq N(\Phi_1) + N(\Phi_2)$ and—because of $\alpha_L = \mathrm{id}_{\mathbb{R}^{d_1}}$—that

$$\mathtt{R}(\Psi) = (\beta_K \circ S_K) \circ \cdots \circ (\beta_1 \circ S_1) \circ (\alpha_L \circ T_L) \circ \cdots \circ (\alpha_1 \circ T_1) = \mathtt{R}(\Phi_2) \circ \mathtt{R}(\Phi_1).$$

Note $T_\ell : \mathbb{R}^{M_{\ell-1}} \to \mathbb{R}^{M_\ell}$ for certain $M_0,\dots,M_L \in \mathbb{N}$. Since $L > 1$, we have $M_{L-1} \leq N(\Phi_1) \leq N$. Furthermore, since $T_L : \mathbb{R}^{M_{L-1}} \to \mathbb{R}^{M_L}$, we get $\|T_L\|_{\ell^{0,\infty}_*} \leq M_{L-1} \leq N$ directly from the definition. Thus, Lemma A.3 shows $\|S_1 \circ T_L\|_{\ell^0} \leq \|S_1\|_{\ell^0} \cdot \|T_L\|_{\ell^{0,\infty}_*} \leq N \cdot \|S_1\|_{\ell^0}$. Therefore, and since $N \geq 1$, we see that

$$W(\Psi) = \sum_{\ell=1}^{L-1} \|T_\ell\|_{\ell^0} + \|S_1 \circ T_L\|_{\ell^0} + \sum_{\ell=2}^{K} \|S_\ell\|_{\ell^0} \leq W(\Phi_1) + N\cdot\|S_1\|_{\ell^0} + N\cdot\sum_{\ell=2}^{K} \|S_\ell\|_{\ell^0} = W(\Phi_1) + N\cdot W(\Phi_2).$$

Finally, note that if Φ_1,Φ_2 are strict networks, then so is Ψ. □

Proof of Lemma A.3. The stated estimates follow directly from the definitions by direct computations and are thus left to the reader. For instance, the main observation for proving that $\|BA\|_{\ell^0} \leq \|B\|_{\ell^{0,\infty}} \cdot \|A\|_{\ell^0}$ is that

$$\|Ax\|_{\ell^0} = \Big\|\sum_{i=1}^{p} x_i \cdot Ae_i\Big\|_{\ell^0} \leq \sum_{i\,:\,x_i\neq 0} \|Ae_i\|_{\ell^0} \leq \|x\|_{\ell^0} \cdot \|A\|_{\ell^{0,\infty}} \quad \text{for} \quad A \in \mathbb{R}^{q\times p} \text{ and } x \in \mathbb{R}^p. \qquad \square$$

A.6. **Proof of Lemma 2.19.** We start with an auxiliary lemma.

Lemma A.4. *Consider two activation functions ϱ, σ such that $\sigma = \mathtt{R}(\Psi_\sigma)$ for some $\Psi_\sigma \in \mathcal{NN}_{w,\ell,m}^{\varrho,1,1}$ with $L(\Psi_\sigma) = \ell \in \mathbb{N}$, $w \in \mathbb{N}_0$, $m \in \mathbb{N}$. Furthermore, assume that $\sigma \not\equiv \text{const}$.*

Then, for any $d \in \mathbb{N}$ and $\alpha_i \in \{\mathrm{id}_\mathbb{R}, \sigma\}$, $1 \le i \le d$ we have $\alpha_1 \otimes \cdots \otimes \alpha_d = \mathtt{R}(\Phi)$ for some network

$$\Phi = \big((U_1, \gamma_1), \dots, (U_\ell, \gamma_\ell)\big) \in \mathcal{NN}_{dw,\ell,dm}^{\varrho,d,d}$$

satisfying $\|U_1\|_{\ell^{0,\infty}} \le m$, $\|U_1\|_{\ell^{0,\infty}_*} \le 1$, $\|U_\ell\|_{\ell^{0,\infty}} \le 1$, and $\|U_\ell\|_{\ell^{0,\infty}_*} \le m$.

If Ψ_σ is a strict network and $\alpha_i = \sigma$ for all i, then Φ can be chosen to be a strict network. ◀

Proof of Lemma A.4. First we show that any $\alpha \in \{\mathrm{id}_\mathbb{R}, \sigma\}$ satisfies $\alpha = \mathtt{R}(\Psi_\alpha)$ for some network

$$\Psi_\alpha = \big((U_1^\alpha, \gamma_1^\alpha), \dots, (U_\ell^\alpha, \gamma_\ell^\alpha)\big) \in \mathcal{NN}_{w,\ell,m}^{\varrho,1,1}$$

with $\|U_1^\alpha\|_{\ell^{0,\infty}} \le m$, $\|U_1^\alpha\|_{\ell^{0,\infty}_*} \le 1$, $\|U_\ell^\alpha\|_{\ell^{0,\infty}} \le 1$ and $\|U_\ell^\alpha\|_{\ell^{0,\infty}_*} \le m$.

For $\alpha = \sigma$ we have $\alpha = \mathtt{R}(\Psi_\sigma)$ where Ψ_σ is of the form $\Psi_\sigma = \big((T_1, \beta_1), \dots, (T_\ell, \beta_\ell)\big) \in \mathcal{NN}_{w,\ell,m}^{\varrho,1,1}$. For $\alpha = \mathrm{id}_\mathbb{R}$, observe that $\alpha = \mathtt{R}(\Psi_{\mathrm{id}_\mathbb{R}})$ with

$$\Psi_{\mathrm{id}_\mathbb{R}} := \big((T_1', \mathrm{id}_\mathbb{R}), \dots, (T_\ell', \mathrm{id}_\mathbb{R})\big) := \big((\mathrm{id}_\mathbb{R}, \mathrm{id}_\mathbb{R}), \dots, (\mathrm{id}_\mathbb{R}, \mathrm{id}_\mathbb{R})\big),$$

where it is easy to see that $N(\Psi_{\mathrm{id}_\mathbb{R}}) = \ell - 1 \le m$ and $W(\Psi_{\mathrm{id}_\mathbb{R}}) = \ell \le w$. Indeed, Equation (2.1) shows that $\ell = L(\Psi_\sigma) \le 1 + N(\Psi_\sigma) \le 1 + m$. On the other hand, since $\sigma \not\equiv \text{const}$, Corollary 2.10 shows that $\ell = L(\Psi_\sigma) \le W(\Psi_\sigma) \le w$.

Denoting by N_i the number of neurons in the i-th layer of Ψ_σ (where layer 0 is the input layer, and layer ℓ the output layer), we get because of $\Psi_\sigma \in \mathcal{NN}_{w,\ell,m}^{\varrho,1,1}$ that $N_i \le m$ for $1 \le i \le L-1$. Furthermore, since $T_1 : \mathbb{R} \to \mathbb{R}^{N_1}$, we have $\|T_1\|_{\ell^{0,\infty}} \le N_1 \le m$ and $\|T_1\|_{\ell^{0,\infty}_*} \le 1$. Similarly, as $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}$ we have $\|T_\ell\|_{\ell^{0,\infty}} \le 1$ and $\|T_\ell\|_{\ell^{0,\infty}_*} \le m$. The same bounds trivially hold for T_1' and T_ℓ'.

We now prove the claim of the lemma by induction on d. The result is trivial for $d = 1$ using $\Phi = \Psi_{\alpha_1}$. Assuming it is true for $d \in \mathbb{N}$, we prove it for $d+1$.

Define $\alpha = \alpha_1 \otimes \cdots \otimes \alpha_d$ and $\overline{\alpha} = \alpha_1 \otimes \cdots \otimes \alpha_{d+1} = \alpha \otimes \alpha_{d+1}$. By induction, there are networks $\Psi_1 = \big((V_1, \lambda_1), \dots, (V_\ell, \lambda_\ell)\big) \in \mathcal{NN}_{dw,\ell,dm}^{\varrho,d,d}$ and $\Psi_2 = \big((W_1, \mu_1), \dots, (W_\ell, \mu_\ell)\big) \in \mathcal{NN}_{w,\ell,m}^{\varrho,1,1}$ such that $\mathtt{R}(\Psi_1) = \alpha$ and $\mathtt{R}(\Psi_2) = \alpha_{d+1}$ and such that $\|V_1\|_{\ell^{0,\infty}} \le m$, $\|V_1\|_{\ell^{0,\infty}_*} \le 1$, $\|V_\ell\|_{\ell^{0,\infty}} \le 1$, and $\|V_\ell\|_{\ell^{0,\infty}_*} \le m$, and likewise for W_1 instead of V_1 and W_ℓ instead of V_ℓ.

Define $U_i := V_i \otimes W_i$ and $\gamma_i := \lambda_i \otimes \mu_i$ for $1 \le i \le \ell$, and $\Phi := \big((U_1, \gamma_1), \dots, (U_\ell, \gamma_\ell)\big)$. One can check that $\mathtt{R}(\Phi) = \overline{\alpha}$. Moreover, Lemma A.3 shows that $\|U_i\|_{\ell^0} = \|V_i\|_{\ell^0} + \|W_i\|_{\ell^0}$ for $1 \le i \le \ell$, whence $W(\Phi) = W(\Psi_1) + W(\Psi_2) \le dw + d = (d+1)w$ and similarly $N(\Phi) = N(\Psi_1) + N(\Psi_2) \le (d+1)m$. Finally, Lemma A.3 shows that

$$\|U_1\|_{\ell^{0,\infty}} \le \max\big\{\|V_1\|_{\ell^{0,\infty}}, \|W_1\|_{\ell^{0,\infty}}\big\} \le m, \qquad \|U_1\|_{\ell^{0,\infty}_*} \le \max\big\{\|V_1\|_{\ell^{0,\infty}_*}, \|W_1\|_{\ell^{0,\infty}_*}\big\} \le 1,$$
$$\|U_\ell\|_{\ell^{0,\infty}} \le \max\big\{\|V_\ell\|_{\ell^{0,\infty}}, \|W_\ell\|_{\ell^{0,\infty}}\big\} \le 1, \qquad \|U_\ell\|_{\ell^{0,\infty}_*} \le \max\big\{\|V_\ell\|_{\ell^{0,\infty}_*}, \|W_\ell\|_{\ell^{0,\infty}_*}\big\} \le m\,.$$

Clearly, if Ψ_σ is strict, and if $\alpha_i = \sigma$ for all i, then the same induction shows that Φ can be chosen to be a strict network. □

Proof of Lemma 2.19. For the first statement with $\ell = 2$ consider $f = \mathtt{R}(\Psi)$ for some

$$\Psi = \big((S_1, \alpha_1), \dots, (S_{K-1}, \alpha_{K-1}), (S_K, \mathrm{id}_{\mathbb{R}^k})\big) \in \mathcal{NN}_{W,L,N}^{\sigma,d,k}.$$

In case of $K = 1$, we trivially have $\Psi \in \mathcal{NN}_{W,L,N}^{\varrho,d,k}$, so that we can assume $K \ge 2$ in the following.

Denoting by N_i the number of neurons at the i-th layer of Ψ, Lemma A.4 yields for each $i \in \{1, \dots, K-1\}$ a network $\Phi_i = \big((U_1^i, \gamma_i), (U_2^i, \mathrm{id}_{\mathbb{R}^{N_i}})\big) \in \mathcal{NN}_{N_i w, 2, N_i m}^{\varrho, N_i, N_i}$ satisfying $\alpha_i = \mathtt{R}(\Phi_i)$ and $\gamma_i : \mathbb{R}^{N(\Phi_i)} \to \mathbb{R}^{N(\Phi_i)}$ with $N(\Phi_i) \le N_i m$ and finally $\|U_1^i\|_{\ell^{0,\infty}} \le m$ and $\|U_2^i\|_{\ell^{0,\infty}_*} \le m$. With $T_1 := U_1^1 \circ S_1$, $T_K := S_K \circ U_2^{K-1}$, $T_i := U_1^i \circ S_i \circ U_2^{i-1}$ for $2 \le i \le K-1$ and

$$\Phi := \big((T_1, \gamma_1), \dots, (T_{K-1}, \gamma_{K-1}), (T_K, \mathrm{id}_{\mathbb{R}^k})\big)\,,$$

one can check that $f = \mathtt{R}(\Phi)$.

By Lemma A.3, $\|T_i\|_{\ell^0} \le \|U_1^i\|_{\ell^{0,\infty}} \|S_i\|_{\ell^0} \|U_2^{i-1}\|_{\ell^{0,\infty}_*} \le m^2 \|S_i\|_{\ell^0}$ for $2 \le i \le K-1$, and the same overall bound also holds for $i \in \{1, K\}$. As a result we get $L(\Phi) = K \le L$ as well as

$$\frac{W(\Phi)}{m^2} = \sum_{i=1}^K \frac{\|T_i\|_{\ell^0}}{m^2} \le \sum_{i=1}^K \|S_i\|_{\ell^0} = W(\Psi) \le W \quad \text{and} \quad \frac{N(\Phi)}{m} = \sum_{i=1}^{K-1} \frac{N(\Phi_i)}{m} \le \sum_{i=1}^{K-1} N_i = N(\Psi) \le N.$$

For the second statement, we prove by induction on $L \in \mathbb{N}$ that $\mathtt{NN}_{W,L,N}^{\sigma,d,k} \subset \mathtt{NN}_{mW+Nw, 1+(L-1)\ell, N(1+m)}^{\varrho,d,k}$.

For $L = 1$, it is easy to see $\mathtt{NN}^{\sigma,d,k}_{W,1,N} = \mathtt{NN}^{\varrho,d,k}_{W,1,N}$, simply because on the last (and for $L = 1$ only) layer, the activation function is always given by $\mathrm{id}_{\mathbb{R}^k}$. Thus, the claim follows from the trivial inclusion $\mathtt{NN}^{\varrho,d,k}_{W,1,N} \subset \mathtt{NN}^{\varrho,d,k}_{mW+Nw,1,N(1+m)}$, since $m \geq 1$.

Now, assuming the claim holds true for L, we prove it for $L+1$. Consider $f \in \mathtt{NN}^{\sigma,d,k}_{W,L+1,N}$. In case of $f \in \mathtt{NN}^{\sigma,d,k}_{W,L,N}$, we get $f \in \mathtt{NN}^{\varrho,d,k}_{mW+Nw,1+(L-1)\ell,N(1+m)} \subset \mathtt{NN}^{\varrho,d,k}_{mW+Nw,1+((L+1)-1)\ell,N(1+m)}$ by the induction hypothesis. In the remaining case where $f \notin \mathtt{NN}^{\sigma,d,k}_{W,L,N}$, there is a network $\Psi \in \mathcal{NN}^{\sigma,d,k}_{W,L+1,N}$ of the form $\Psi = \big((S_1,\alpha_1),\dots,(S_L,\alpha_L),(S_{L+1},\mathrm{id}_{\mathbb{R}^k})\big)$ such that $f = \mathtt{R}(\Psi)$. Observe that $S_{L+1} : \mathbb{R}^{\overline{k}} \to \mathbb{R}^k$ with $\overline{k} := N_L$ the number of neurons of the last hidden layer. Defining $\Psi_1 := \big((S_1,\alpha_1),\dots,(S_{L-1},\alpha_{L-1}),(S_L,\mathrm{id}_{\mathbb{R}^{\overline{k}}})\big)$, we have $\Psi_1 \in \mathcal{NN}^{\sigma,d,\overline{k}}_{\overline{W},L,\overline{N}}$ where $\overline{W} := W(\Psi_1)$ and $\overline{N} := N(\Psi_1)$ satisfy

$$\overline{W} + \|S_{L+1}\|_{\ell^0} \leq W(\Psi) \leq W \quad \text{and} \quad \overline{N} + \overline{k} \leq N(\Psi) \leq N.$$

Define $g := \mathtt{R}(\Psi_1)$, so that $f = S_{L+1} \circ \alpha_L \circ g$. We now exhibit a ϱ-network Φ (instead of the σ-network Ψ) of controlled complexity such that $f = \mathtt{R}(\Phi)$. As $g := \mathtt{R}(\Psi_1) \in \mathtt{NN}^{\sigma,d,\overline{k}}_{\overline{W},L,\overline{N}}$, the induction hypothesis shows that $g = \mathtt{R}(\Phi_1)$ for some network

$$\Phi_1 = \big((T_1,\beta_1),\dots,(T_{K-1},\beta_{K-1})(T_K,\mathrm{id}_{\mathbb{R}^{\overline{k}}})\big) \in \mathcal{NN}^{\varrho,d,\overline{k}}_{m\overline{W}+\overline{N}w,1+(L-1)\ell,\overline{N}(1+m)}\,.$$

Moreover, Lemma A.4 shows that $\alpha_L = \mathtt{R}(\Phi_2)$ for a network

$$\Phi_2 = \big((U_1,\gamma_1),\dots,(U_{\ell-1},\gamma_{\ell-1}),(U_\ell,\mathrm{id}_{\mathbb{R}^{\overline{k}}})\big) \in \mathcal{NN}^{\varrho,\overline{k},\overline{k}}_{\overline{k}w,\ell,\overline{k}m}$$

with $\|U_\ell\|_{\ell^{0,\infty}_*} \leq m$. By construction, we have $f = S_{L+1} \circ \alpha_L \circ g = \mathtt{R}(\Phi)$ for the network

$$\Phi := \big((T_1,\beta_1),\dots,(T_{K-1},\beta_{K-1}),(T_K,\mathrm{id}_{\mathbb{R}^{\overline{k}}}),(U_1,\gamma_1),\dots,(U_{\ell-1},\gamma_{\ell-1}),(S_{L+1}\circ U_\ell,\mathrm{id}_{\mathbb{R}^k})\big).$$

To conclude, we observe that $L(\Phi) = K + \ell \leq 1 + (L-1)\ell + \ell = 1 + \big((L+1)-1\big)\ell$, as well as

$$\begin{aligned} W(\Phi) &= W(\Phi_1) + \big(W(\Phi_2) - \|U_\ell\|_{\ell^0}\big) + \|S_{L+1}\circ U_\ell\|_{\ell^0} \\ (\text{Lemma A.3}) &\leq m\overline{W} + \overline{N}w + W(\Phi_2) + \|S_{L+1}\|_{\ell^0}\cdot\|U_\ell\|_{\ell^{0,\infty}_*} \\ &\leq m\overline{W} + \overline{N}w + \overline{k}w + m\cdot\|S_{L+1}\|_{\ell^0} \leq mW + Nw. \end{aligned}$$

Finally, we also have $N(\Phi) = N(\Phi_1) + \overline{k} + N(\Phi_2) \leq \overline{N}(1+m) + \overline{k} + \overline{k}\cdot m = (\overline{N}+\overline{k})(1+m) \leq N(1+m)$. □

A.7. **Proof of Lemma 2.20.** Let $\Psi = \big((S_1,\alpha_1),\dots,(S_{K-1},\alpha_{K-1}),(S_K,\mathrm{id}_{\mathbb{R}^k})\big) \in \mathcal{NN}^{\sigma,d,k}_{W,L,N}$ be arbitrary and $g = \mathtt{R}(\Psi)$. We prove that there is some $\Phi \in \mathcal{NN}^{\varrho,d,k}_{W+(s-1)N,1+s(L-1),sN}$ such that $g = \mathtt{R}(\Phi)$. This is easy to see if $s = 1$ or $K = 1$; hence we now assume $K \geq 2$ and $s \geq 2$. Denoting by N_ℓ the number of neurons at the ℓ-th layer of Ψ, for $1 \leq \ell \leq K-1$ we have $\alpha_\ell = \alpha_\ell^{(1)} \otimes \ldots \otimes \alpha_\ell^{(N_\ell)}$ where $\alpha_\ell^{(i)} \in \{\mathrm{id}_{\mathbb{R}}, \sigma\}$. For $1 \leq \ell \leq L-1$, $1 \leq j \leq K_\ell$, $1 \leq i \leq s$, define

$$\beta^{(j)}_{s(\ell-1)+i} := \begin{cases} \varrho, & \text{if } \alpha_\ell^{(j)} = \sigma, \\ \mathrm{id}_{\mathbb{R}}, & \text{otherwise} \end{cases}$$

and let $\beta_{s(\ell-1)+i} := \beta^{(1)}_{s(\ell-1)+i} \otimes \ldots \otimes \beta^{(N_\ell)}_{s(\ell-1)+i}$. Define also $T_{s(\ell-1)+1} := S_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ and $T_{s(\ell-1)+i} := \mathrm{id}_{\mathbb{R}^{N_\ell}}$ for $2 \leq i \leq s$. It is painless to check that

$$\begin{aligned} \alpha_\ell \circ S_\ell &= \beta_{s(\ell-1)+s} \circ T_{s(\ell-1)+s} \circ \cdots \circ \beta_{s(\ell-1)+2} \circ T_{s(\ell-1)+2} \circ \beta_{s(\ell-1)+1} \circ T_{s(\ell-1)+1} \\ &= \beta_{s\ell} \circ T_{s\ell} \circ \cdots \circ \beta_{s(\ell-1)+1} \circ T_{s(\ell-1)+1}\,, \end{aligned}$$

and hence

$$g = S_K \circ \alpha_{K-1} \circ S_{K-1} \circ \cdots \circ \alpha_1 \circ S_1 = S_K \circ \beta_{s(K-1)} \circ T_{s(K-1)} \circ \cdots \circ \beta_1 \circ T_1\,.$$

That is to say, $g = \mathtt{R}(\Phi)$ with

$$\Phi := \big((T_1,\beta_1),\dots,(T_{s(K-1)},\beta_{s(K-1)}),(S_K,\mathrm{id}_{\mathbb{R}^k})\big) \in \mathcal{NN}^{\varrho,d,k}_{W',1+s(K-1),sN} \subset \mathcal{NN}^{\varrho,d,k}_{W',1+s(L-1),sN},$$

where we compute

$$\begin{aligned} W' &:= \|S_K\|_{\ell^0} + \sum_{j=1}^{s(K-1)} \|T_j\|_{\ell^0} = \|S_K\|_{\ell^0} + \sum_{\ell=1}^{K-1} \sum_{i=1}^{s} \|T_{s(\ell-1)+i}\|_{\ell^0} \\ &= \|S_K\|_{\ell^0} + \sum_{\ell=1}^{K-1} \Big(\|T_{s(\ell-1)+1}\|_{\ell^0} + \sum_{i=2}^{s} \|T_{s(\ell-1)+i}\|_{\ell^0} \Big) \\ &= \|S_K\|_{\ell^0} + \sum_{\ell=1}^{K-1} (\|S_\ell\|_{\ell^0} + (s-1)N_\ell) = \sum_{\ell=1}^{K} \|S_\ell\|_{\ell^0} + (s-1) \sum_{\ell=1}^{K-1} N_\ell \\ &= W(\Psi) + (s-1)N(\Psi) \leq W + (s-1)N\,. \end{aligned}$$

We conclude as claimed that $\Phi \in \mathcal{NN}^{\varrho,d,k}_{W+(s-1)N,1+s(L-1),sN}$. Finally, if Ψ is strict, then so is Φ. □

A.8. **Proof of Lemma 2.21.** For $f \in \mathtt{NN}^{\sigma,d,k}_{W,L,N}$ there is $\Phi = \big((S_1,\alpha_1),\dots,(S_{L'},\alpha_{L'})\big) \in \mathcal{NN}^{\sigma,d,k}_{W,L',N}$ with $L(\Phi) = L' \leq L$ and such that $f = \mathtt{R}(\Phi)$. Replace each occurrence of the activation function σ by σ_h in the nonlinearities α_j to define a σ_h-network $\Phi_h := \big((S_1,\alpha_1^{(h)}),\dots,(S_{L'},\alpha_{L'}^{(h)})\big) \in \mathcal{NN}^{\sigma_h,d,k}_{W,L',N}$ and its realization $f_h := \mathtt{R}(\Phi_h) \in \mathtt{NN}^{\sigma_h,d,k}_{W,L',N}$. Since σ is continuous and $\sigma_h \to \sigma$ locally uniformly on $\mathbb{R}$ as $h \to 0$, we get by Lemma A.7 (which is proved independently below) that $f_h \to f$ locally uniformly on $\mathbb{R}^d$. To conclude for $\ell = 2$ observe that $\sigma_h = \mathtt{R}(\Psi_h)$ with $\Psi_h \in \mathcal{NN}^{\varrho,1,1}_{w,\ell,m}$ and $L(\Psi_h) = \ell$, whence Lemma 2.19 yields

$$f_h \in \mathtt{NN}^{\sigma_h,d,k}_{W,L',N} \subset \mathtt{NN}^{\varrho,d,k}_{Wm^2,L',Nm} \subset \mathtt{NN}^{\varrho,d,k}_{Wm^2,L,Nm}\,.$$

For arbitrary ℓ we similarly conclude that

$$f_h \in \mathtt{NN}^{\sigma_h,d,k}_{W,L',N} \subset \mathtt{NN}^{\varrho,d,k}_{W+Nw,1+(L'-1)(\ell+1),N(2+m)} \subset \mathtt{NN}^{\varrho,d,k}_{W+Nw,1+(L-1)(\ell+1),N(2+m)}\,. \qquad \square$$

A.9. **Proof of Lemmas 2.22 and 2.25.** In this section, we provide a unified proof for Lemmas 2.22 and 2.25. To be able to handle both claims simultaneously, the following concept will be important.

Definition A.5. For each $d,k \in \mathbb{N}$, let us fix a subset $\mathcal{G}_{d,k} \subset \{f : \mathbb{R}^d \to \mathbb{R}^k\}$ and a topology $\mathcal{T}_{d,k}$ on the space of all functions $f : \mathbb{R}^d \to \mathbb{R}^k$. Let $\mathcal{G} := (\mathcal{G}_{d,k})_{d,k\in\mathbb{N}}$ and $\mathcal{T} := (\mathcal{T}_{d,k})_{d,k\in\mathbb{N}}$. The tuple $(\mathcal{G},\mathcal{T})$ is called a *network compatible topology family* if it satisfies the following:

(1) We have $\{T : \mathbb{R}^d \to \mathbb{R}^k \mid T \text{ affine-linear}\} \subset \mathcal{G}_{d,k}$ for all $d,k \in \mathbb{N}$.

(2) If $p \in \mathbb{N}$ and for each $i \in \{1,\dots,p\}$, we are given a sequence $(f_i^{(n)})_{n\in\mathbb{N}_0}$ of functions $f_i^{(n)} : \mathbb{R} \to \mathbb{R}$ satisfying $f_i^{(0)} \in \mathcal{G}_{1,1}$ and $f_i^{(n)} \xrightarrow[n\to\infty]{\mathcal{T}_{1,1}} f_i^{(0)}$, then $f_1^{(n)} \otimes \cdots \otimes f_p^{(n)} \xrightarrow[n\to\infty]{\mathcal{T}_{p,p}} f_1^{(0)} \otimes \cdots \otimes f_p^{(0)}$ and $f_1^{(0)} \otimes \cdots \otimes f_p^{(0)} \in \mathcal{G}_{p,p}$.

(3) If $f_n : \mathbb{R}^d \to \mathbb{R}^k$ and $g_n : \mathbb{R}^k \to \mathbb{R}^\ell$ for all $n \in \mathbb{N}_0$ and if $f_0 \in \mathcal{G}_{d,k}$ and $g_0 \in \mathcal{G}_{k,\ell}$ as well as $f_n \xrightarrow[n\to\infty]{\mathcal{T}_{d,k}} f_0$ and $g_n \xrightarrow[n\to\infty]{\mathcal{T}_{k,\ell}} g_0$, then $g_0 \circ f_0 \in \mathcal{G}_{d,\ell}$ and $g_n \circ f_n \xrightarrow[n\to\infty]{\mathcal{T}_{d,\ell}} g_0 \circ f_0$. ◀

Remark. Roughly speaking, the above definition introduces certain topologies $\mathcal{T}_{d,k}$ and certain sets of "good functions" $\mathcal{G}_{d,k}$ such that—for limit functions that are "good"—convergence in the topology is compatible with taking $\otimes$-products and with composition.

By induction, it is easy to see that if $p \in \mathbb{N}$ and if for each $i \in \{1,\dots,p\}$ we are given a sequence $(f_i^{(n)})_{n\in\mathbb{N}}$ with $f_i^{(n)} : \mathbb{R}^{d_{i-1}} \to \mathbb{R}^{d_i}$ and $f_i^{(0)} \in \mathcal{G}_{d_{i-1},d_i}$ as well as $f_i^{(n)} \xrightarrow[n\to\infty]{\mathcal{T}_{d_{i-1},d_i}} f_i^{(0)}$, then also $f_p^{(0)} \circ \cdots \circ f_1^{(0)} \in \mathcal{G}_{d_0,d_p}$, as well as $f_p^{(n)} \circ \cdots \circ f_1^{(0)} \xrightarrow[n\to\infty]{\mathcal{T}_{d_0,d_p}} f_p^{(0)} \circ \cdots \circ f_1^{(0)}$. Indeed, the base case of the induction is contained in Definition A.5. Now, assuming that the claim holds for $p \in \mathbb{N}$, we prove it for $p+1$. To this end, let $F_1^{(n)} := f_p^{(n)} \circ \cdots \circ f_1^{(n)}$ and $F_2^{(n)} := f_{p+1}^{(n)}$. By induction, we know $F_1^{(0)} \in \mathcal{G}_{d_0,d_p}$ and $F_1^{(n)} \xrightarrow[n\to\infty]{\mathcal{T}_{d_0,d_p}} F_1^{(0)}$. Since also $F_2^{(0)} = f_{p+1}^{(0)} \in \mathcal{G}_{d_p,d_{p+1}}$, Definition A.5 implies $F_2^{(0)} \circ F_1^{(0)} \in \mathcal{G}_{d_0,d_{p+1}}$ and $F_2^{(n)} \circ F_1^{(n)} \xrightarrow[n\to\infty]{\mathcal{T}_{d_0,d_{p+1}}} F_2^{(0)} \circ F_1^{(0)}$, which is precisely the claim for $p+1$ instead of p. ♦

We now have the following important result:

Proposition A.6. *Let $\varrho : \mathbb{R} \to \mathbb{R}$, and let $(\mathcal{G},\mathcal{T})$ be a network compatible topology family satisfying the following*

- $\varrho \in \mathcal{G}_{1,1}$;

- *There is some $n \in \mathbb{N}$ such that for each $m \in \mathbb{N}$ there are affine-linear maps $E_m : \mathbb{R} \to \mathbb{R}^n$ and $D_m : \mathbb{R}^n \to \mathbb{R}$ such that $F_m := D_m \circ (\varrho \otimes \cdots \otimes \varrho) \circ E_m : \mathbb{R} \to \mathbb{R}$ satisfies $F_m \xrightarrow[m\to\infty]{\mathcal{T}_{1,1}} \mathrm{id}_{\mathbb{R}}$.*

Then we have for arbitrary $d, k \in \mathbb{N}$, $W, N \in \mathbb{N}_0 \cup \{\infty\}$ and $L \in \mathbb{N} \cup \{\infty\}$ the inclusion

$$\mathtt{NN}^{\varrho,d,k}_{W,L,N} \subset \overline{\mathtt{SNN}^{\varrho,d,k}_{n^2W,L,nN}},$$

where the closure is a sequential closure which is taken with respect to the topology $\mathcal{T}_{d,k}$. ◀

Remark. Before we give the proof of Proposition A.6, we explain a convention that will be used in the proof. Precisely, in the definition of $W(\Phi)$, we always assume that the affine-linear maps T_ℓ are of the form $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$. Clearly, the expressivity of networks will not change if instead of the spaces $\mathbb{R}^{N_1}, \dots, \mathbb{R}^{N_{L-1}}$, one uses finite-dimensional vector spaces $V_1, \dots, V_{L-1}$ with $\dim V_i = N_i$. The only nontrivial question is the interpretation of $\|T_\ell\|_{\ell^0}$ for an affine-linear map $T_\ell : V_{\ell-1} \to V_\ell$, since for the case of $\mathbb{R}^{N_\ell}$, we chose the standard basis for obtaining the matrix representation of T_ℓ, while for general vector spaces V_ℓ, there is no such canonical choice of basis. Yet, in the proof below, we will consider the case $V_\ell = \mathbb{R}^{n_1} \times \cdots \times \mathbb{R}^{n_m}$. In this case, there is a canonical way of identifying V_ℓ with $\mathbb{R}^{N_\ell}$ for $N_\ell = \sum_{j=1}^m n_j$, and there is also a canonical choice of "standard basis" in the space V_ℓ. We will use this convention in the proof below to simplify the notation. ♦

Proof of Proposition A.6. Let $\Phi \in \mathcal{NN}^{\varrho,d,k}_{W,L,N}$. We will construct a sequence $(\Phi_m)_{m\in\mathbb{N}} \subset \mathcal{SNN}^{\varrho,d,k}_{n^2W,L,nN}$ satisfying $\mathtt{R}(\Phi_m) \xrightarrow[m\to\infty]{\mathcal{T}_{d,k}} \mathtt{R}(\Phi)$. To this end, note that $\Phi = \big((T_1, \alpha_1), \dots, (T_K, \alpha_K)\big)$ for some $K \leq L$ and that there are $N_0, \dots, N_K \in \mathbb{N}$ (with $N_0 = d$ and $N_K = k$) such that $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is affine-linear for each $\ell \in \{1, \dots, K\}$.

Let us first consider the special case $K = 1$. By definition of a neural network, we have $\alpha_K = \mathrm{id}_{\mathbb{R}^k}$, so that Φ is already a *strict* ϱ-network. Therefore, we can choose $\Phi_m := \Phi \in \mathcal{SNN}^{\varrho,d,k}_{W,L,N} \subset \mathcal{SNN}^{\varrho,d,k}_{n^2W,L,nN}$ for all $m \in \mathbb{N}$.

From now on we assume $K \geq 2$. For brevity, set $\varrho_1 := \varrho$ and $\varrho_2 := \mathrm{id}_{\mathbb{R}}$, as well as $D(1) := 1$ and $D(2) := n$, and furthermore

$$\begin{aligned} E_1^{(m)} &:= \mathrm{id}_{\mathbb{R}} : \mathbb{R} \to \mathbb{R}^{D(1)} \quad \text{and} \quad E_2^{(m)} := E_m : \mathbb{R} \to \mathbb{R}^{D(2)}, \\ \text{as well as} \quad D_1^{(m)} &:= \mathrm{id}_{\mathbb{R}} : \mathbb{R}^{D(1)} \to \mathbb{R} \quad \text{and} \quad D_2^{(m)} := D_m : \mathbb{R}^{D(2)} \to \mathbb{R}. \end{aligned}$$

By definition of a generalized ϱ-network, for each $\ell \in \{1, \dots, K\}$ there are $\iota_1^{(\ell)}, \dots, \iota_{N_\ell}^{(\ell)} \in \{1, 2\}$ with $\alpha_\ell = \varrho_{\iota_1^{(\ell)}} \otimes \cdots \otimes \varrho_{\iota_{N_\ell}^{(\ell)}}$, and with $\iota_j^{(K)} = 2$ for all $j \in \{1, \dots, N_K\}$. Now, define $V_0 := \mathbb{R}^d = \mathbb{R}^{N_0}$, $V_K := \mathbb{R}^k = \mathbb{R}^{N_K}$, and

$$V_\ell := \mathbb{R}^{D(\iota_1^{(\ell)})} \times \cdots \times \mathbb{R}^{D(\iota_{N_\ell}^{(\ell)})} \cong \mathbb{R}^{\sum_{i=1}^{N_\ell} D(\iota_i^{(\ell)})} \quad \text{for} \quad 1 \leq \ell \leq K-1.$$

Since we eventually want to obtain strict networks Φ_m, furthermore set

$$\beta^{(1)} := \varrho : \mathbb{R}^{D(1)} \to \mathbb{R}^{D(1)} \qquad \text{and} \qquad \beta^{(2)} := \varrho \otimes \cdots \otimes \varrho : \mathbb{R}^{D(2)} \to \mathbb{R}^{D(2)}.$$

Using these maps, finally define $\beta_K := \mathrm{id}_{\mathbb{R}^k}$, as well as

$$\beta_\ell := \beta^{(\iota_1^{(\ell)})} \otimes \cdots \otimes \beta^{(\iota_{N_\ell}^{(\ell)})} : V_\ell \to V_\ell \quad \text{for} \quad 1 \leq \ell \leq K-1.$$

Finally, for $\ell \in \{1, \dots, K\}$ and $m \in \mathbb{N}$, define affine-linear maps

$$P_\ell^{(m)} := E^{(m)}_{\iota_1^{(\ell)}} \otimes \cdots \otimes E^{(m)}_{\iota_{N_\ell}^{(\ell)}} : \mathbb{R}^{N_\ell} \to V_\ell \qquad \text{and} \qquad Q_\ell^{(m)} := D^{(m)}_{\iota_1^{(\ell)}} \otimes \cdots \otimes D^{(m)}_{\iota_{N_\ell}^{(\ell)}} : V_\ell \to \mathbb{R}^{N_\ell}.$$

The crucial observation is that by assumption regarding the maps D_m, E_m, we have

$$\begin{aligned} D_2^{(m)} \circ \beta^{(2)} \circ E_2^{(m)} &= F_m \xrightarrow[m\to\infty]{\mathcal{T}_{1,1}} \mathrm{id}_{\mathbb{R}} = \varrho_2, \\ \text{and} \quad D_1^{(m)} \circ \beta^{(1)} \circ E_1^{(m)} &= \mathrm{id}_{\mathbb{R}} \circ \varrho \circ \mathrm{id}_{\mathbb{R}} = \varrho = \varrho_1. \end{aligned} \tag{A.3}$$

Finally, for the construction of the strict networks Φ_m, we define for $m \in \mathbb{N}$

$$\begin{aligned} S_1^{(m)} &:= P_1^{(m)} \circ T_1 && : \mathbb{R}^d = \mathbb{R}^{N_0} = V_0 \to V_1, \\ S_K^{(m)} &:= T_K \circ Q_{K-1}^{(m)} && : V_{K-1} \to \mathbb{R}^{N_K} = \mathbb{R}^k = V_K, \\ \text{and} \quad S_\ell^{(m)} &:= P_\ell^{(m)} \circ T_\ell \circ Q_{\ell-1}^{(m)} && : V_{\ell-1} \to V_\ell \qquad \text{for } 2 \leq \ell \leq K-1, \end{aligned}$$

and then set $\Phi_m := \big((S_1^{(m)}, \beta_1), \dots, (S_K^{(m)}, \beta_K)\big)$. Because of $D(\iota_{i^{(\ell)}}) \in \{1, n\}$, we obtain

$$N(\Phi_m) = \sum_{\ell=1}^{K-1} \dim V_\ell = \sum_{\ell=1}^{K-1} \sum_{i=1}^{N_\ell} D(\iota_i^{(\ell)}) \leq \sum_{\ell=1}^{K-1} nN_\ell = nN(\Phi) \leq nN\,.$$

Furthermore, by the second part of Lemma A.3 and in view of the product structure of $P_\ell^{(m)}$, we have

$$\|P_\ell^{(m)}\|_{\ell^{0,\infty}} \leq \max\big\{\|E_1^{(m)}\|_{\ell^{0,\infty}}, \|E_2^{(m)}\|_{\ell^{0,\infty}}\big\} \leq \max\{D(1), D(2)\} \leq n,$$

for arbitrary $\ell \in \{1, \dots, K\}$, simply because $E_j^{(m)} : \mathbb{R} \to \mathbb{R}^{D(j)}$ for $j \in \{1, 2\}$. Likewise,

$$\|Q_\ell^{(m)}\|_{\ell_*^{0,\infty}} \leq \max\big\{\|D_1^{(m)}\|_{\ell_*^{0,\infty}}, \|D_2^{(m)}\|_{\ell_*^{0,\infty}}\big\} \leq \max\{D(1), D(2)\} \leq n,$$

because $D_j^{(m)} : \mathbb{R}^{D(j)} \to \mathbb{R}$ for $j \in \{1, 2\}$. By the first part of Lemma A.3, we thus see for $2 \leq \ell \leq K-1$ that

$$\|S_\ell^{(m)}\|_{\ell^0} \leq \|P_\ell^{(m)}\|_{\ell^{0,\infty}} \cdot \|T_\ell\|_{\ell^0} \cdot \|Q_{\ell-1}^{(m)}\|_{\ell_*^{0,\infty}} \leq n^2 \cdot \|T_\ell\|_{\ell^0}\,.$$

Similar arguments yield $\|S_1^{(m)}\|_{\ell^0} \leq n \cdot \|T_1\|_{\ell^0} \leq n^2 \cdot \|T_1\|_{\ell^0}$ and $\|S_K^{(m)}\|_{\ell^0} \leq n \cdot \|T_K\|_{\ell^0} \leq n^2 \cdot \|T_K\|_{\ell^0}$. All in all, this implies $W(\Phi_m) \leq n^2 \cdot W(\Phi) \leq n^2 W$, as desired.

Now, since $\varrho_1 = \varrho \in \mathcal{G}_{1,1}$ by the assumptions of the current proposition, since $\varrho_2 = \mathrm{id}_{\mathbb{R}} \in \mathcal{G}_{1,1}$ as an affine-linear map, and since $(\mathcal{G}, \mathcal{T})$ is a network compatible topology family, we see for all $1 \leq \ell \leq K-1$ that $\alpha_\ell = \varrho_{\iota_1^{(\ell)}} \otimes \cdots \otimes \varrho_{\iota_{N_\ell}^{(\ell)}} \in \mathcal{G}_{N_\ell, N_\ell}$ and furthermore that

$$\begin{aligned} Q_\ell^{(m)} \circ \beta_\ell \circ P_\ell^{(m)} &= \left(D_{\iota_1^{(\ell)}}^{(m)} \circ \beta^{(\iota_1^{(\ell)})} \circ E_{\iota_1^{(\ell)}}^{(m)}\right) \otimes \cdots \otimes \left(D_{\iota_{N_\ell}^{(\ell)}}^{(m)} \circ \beta^{(\iota_{N_\ell}^{(\ell)})} \circ E_{\iota_{N_\ell}^{(\ell)}}^{(m)}\right) \\ \big(\text{Eq. (A.3) and compatibility of } (\mathcal{G},\mathcal{T}) \text{ with } \otimes\big) &\xrightarrow[m\to\infty]{\mathcal{T}_{N_\ell, N_\ell}} \varrho_{\iota_1^{(\ell)}} \otimes \cdots \otimes \varrho_{\iota_{N_\ell}^{(\ell)}} = \alpha_\ell\,. \end{aligned} \tag{A.4}$$

Finally, since $\beta_K = \mathrm{id}_{\mathbb{R}^k} = \alpha_K \in \mathcal{G}_{k,k}$, and since $(\mathcal{G}, \mathcal{T})$ is a network compatible topology family and thus compatible with compositions (as long as the "factors" of the limit are "good", which is satisfied here, since $\alpha_\ell \in \mathcal{G}_{N_\ell, N_\ell}$ as we just saw and since $T_\ell \in \mathcal{G}_{N_{\ell-1}, N_\ell}$ as an affine-linear map), we see that

$$\begin{aligned} \mathtt{R}(\Phi_m) &= \beta_K \circ S_K^{(m)} \circ \cdots \circ \beta_1 \circ S_1^{(m)} \\ &= \alpha_K \circ T_K \circ (Q_{K-1}^{(m)} \circ \beta_{K-1} \circ P_{K-1}^{(m)}) \circ T_{K-1} \circ \cdots \circ (Q_1^{(m)} \circ \beta_1 \circ P_1^{(m)}) \circ T_1 \\ \big(\text{Eq. (A.4)}\big) &\xrightarrow[m\to\infty]{\mathcal{T}_{d,k}} \alpha_K \circ T_K \circ \alpha_{K-1} \circ T_{K-1} \circ \cdots \circ \alpha_1 \circ T_1 = \mathtt{R}(\Phi)\,, \end{aligned}$$

and hence $\mathtt{R}(\Phi) \in \overline{\mathtt{SNN}^{\varrho,d,k}_{n^2W,L,nN}}$. □

Now, we use Proposition A.6 to prove Lemma 2.25.

Proof of Lemma 2.25. For $d, k \in \mathbb{N}$, let $\mathcal{G}_{d,k} := \{f : \mathbb{R}^d \to \mathbb{R}^k\}$, and let $\mathcal{T}_{d,k} = 2^{\mathcal{G}_{d,k}}$ be the discrete topology on the set $\{f : \mathbb{R}^d \to \mathbb{R}^k\}$. This means that every set is open, so that the only convergent sequences are those that are eventually constant. It is easy to see that $(\mathcal{G}, \mathcal{T})$ is a network compatible topology family and $\varrho \in \mathcal{G}_{1,1}$.

Finally, by assumption of Lemma 2.25, there are $a_i, b_i, c_i \in \mathbb{R}$ for $i \in \{1, \dots, n\}$ and some $c \in \mathbb{R}$ such that $x = c + \sum_{i=1}^n a_i\, \varrho(b_i\, x + c_i)$ for all $x \in \mathbb{R}$. If we define $E_m : \mathbb{R} \to \mathbb{R}^n, x \mapsto (b_1\, x + c_1, \dots, b_n\, x + c_n)$ and $D_m : \mathbb{R}^n \to \mathbb{R}, y \mapsto c + \sum_{i=1}^n a_i\, y_i$, then E_m, D_m are affine-linear, and $\mathrm{id}_{\mathbb{R}} = D_m \circ (\varrho \otimes \cdots \otimes \varrho) \circ E_m$ for all $m \in \mathbb{N}$. Thus, all assumptions of Proposition A.6 are satisfied, so that this proposition implies $\mathtt{NN}^{\varrho,d,k}_{W,L,N} \subset \overline{\mathtt{SNN}^{\varrho,d,k}_{n^2W,L,nN}} = \mathtt{SNN}^{\varrho,d,k}_{n^2W,L,nN}$ for all $d, k \in \mathbb{N}$, $W, N \in \mathbb{N}_0 \cup \{\infty\}$ and $L \in \mathbb{N} \cup \{\infty\}$. Here, we used that the (sequential) closure of a set M with respect to the discrete topology is simply the set M itself. □

Finally, we will use Proposition A.6 to provide a proof of Lemma 2.22. To this end, the following lemma is essential.

Lemma A.7. *Let $(f_n)_{n\in\mathbb{N}_0}$ and $(g_n)_{n\in\mathbb{N}_0}$ be sequences of functions $f_n : \mathbb{R}^d \to \mathbb{R}^k$ and $g_n : \mathbb{R}^k \to \mathbb{R}^\ell$. Assume that f_0, g_0 are continuous and that $f_n \xrightarrow[n\to\infty]{} f_0$ and $g_n \xrightarrow[n\to\infty]{} g_0$ with locally uniform convergence. Then $g_0 \circ f_0$ is continuous, and $g_n \circ f_n \xrightarrow[n\to\infty]{} g_0 \circ f_0$ with locally uniform convergence.* ◀

Proof. Locally uniform convergence on $\mathbb{R}^d$ is equivalent to uniform convergence on bounded sets. Furthermore, the continuous function f_0 is bounded on each bounded set $K \subset \mathbb{R}^d$; by uniform convergence, this implies that $K' := \{f(x)\colon x \in K\} \cup \{f_n(x)\colon n \in \mathbb{N} \text{ and } x \in K\} \subset \mathbb{R}^k$ is bounded as well. Hence, the continuous function g_0 is *uniformly* continuous on K'. From these observations, the claim follows easily; the details are left to the reader. □

Given this auxiliary result, we can now prove Lemma 2.22.

Proof of Lemma 2.22. For $d, k \in \mathbb{N}$, define $\mathcal{G}_{d,k} := \{f : \mathbb{R}^d \to \mathbb{R}^k \mid f \text{ continuous}\}$, and let $\mathcal{T}_{d,k}$ denote the topology of locally uniform convergence on $\{f : \mathbb{R}^d \to \mathbb{R}^k\}$. We claim that $(\mathcal{G}, \mathcal{T})$ is a network compatible topology family. Indeed, the first condition in Definition A.5 is trivial, and the third condition holds thanks to Lemma A.7. Finally, it is not hard to see that if $f_i^{(n)} : \mathbb{R} \to \mathbb{R}$ satisfy $f_i^{(n)} \to f_i^{(0)}$ locally uniformly for all $i \in \{1, \dots, p\}$, then $f_1^{(n)} \otimes \cdots \otimes f_p^{(n)} \xrightarrow[n\to\infty]{} f_1^{(0)} \otimes \cdots \otimes f_p^{(0)}$ locally uniformly. This proves the second condition in Definition A.5.

We want to apply Proposition A.6 with $n = 2$. We have $\varrho \in \mathcal{G}_{1,1}$, since ϱ is continuous by the assumptions of Lemma 2.22. Thus, it remains to construct sequences $(E_m)_{m\in\mathbb{N}}, (D_m)_{m\in\mathbb{N}}$ of affine-linear maps $E_m : \mathbb{R} \to \mathbb{R}^2$ and $D_m : \mathbb{R}^2 \to \mathbb{R}$ such that $D_m \circ (\varrho \otimes \varrho) \circ E_m \to \mathrm{id}_{\mathbb{R}}$ with locally uniform convergence. Once these are constructed, Proposition A.6 shows that $\mathtt{NN}^{\varrho,d,k}_{W,L,N} \subset \overline{\mathtt{SNN}^{\varrho,d,k}_{4W,L,2N}}$, where the closure is with respect to locally uniform convergence. This is precisely what is claimed in Lemma 2.22.

To construct E_m, D_m, let us set $a := \varrho'(x_0) \neq 0$. By definition of the derivative, for arbitrary $m \in \mathbb{N}$ and $\varepsilon_m := |a|/m$, there is some $\delta_m > 0$ satisfying

$$\big|\big(\varrho(x_0 + h) - \varrho(x_0)\big)/h - a\big| \leq \varepsilon_m = |a|/m \qquad \forall\, h \in \mathbb{R} \text{ with } 0 < |h| \leq \delta_m\,. \tag{A.5}$$

Now, define affine-linear maps

$$E_m : \mathbb{R} \to \mathbb{R}^2, x \mapsto \Big(x_0 + m^{-1/2} \cdot \delta_m \cdot x\,,\, x_0\Big)^T \quad \text{and} \quad D_m : \mathbb{R}^2 \to \mathbb{R}, (y_1, y_2) \mapsto \sqrt{m} \cdot (y_1 - y_2)/(a \cdot \delta_m)\,,$$

and set $F_m := D_m \circ (\varrho \otimes \varrho) \circ E_m$.

Finally, let $x \in \mathbb{R}$ be arbitrary with $0 < |x| \leq \sqrt{m}$, and set $h := \delta_m \cdot x/\sqrt{m}$, so that $0 < |h| \leq \delta_m$. By multiplying Equation (A.5) with $|h|/|a|$, we then get

$$\begin{aligned} &\big|a^{-1} \cdot \big(\varrho(x_0 + h) - \varrho(x_0)\big) - h\big| \leq \frac{|h|}{m} \\ (\text{multiply by } \sqrt{m}/\delta_m) \Longrightarrow &\left|\frac{\sqrt{m}}{a \cdot \delta_m}\left(\varrho\left(x_0 + \frac{\delta_m \cdot x}{\sqrt{m}}\right) - \varrho(x_0)\right) - x\right| \leq \frac{|h|}{\delta_m \cdot \sqrt{m}} = \frac{|x|}{m} \leq \frac{1}{\sqrt{m}}\,, \end{aligned}$$

where the last step used that $|x| \leq \sqrt{m}$. This estimate is trivially valid for $x = 0$. Put differently, we have thus shown $|F_m(x) - x| \leq 1/\sqrt{m}$ for all $x \in \mathbb{R}$ with $|x| \leq \sqrt{m}$. That is, $F_m \xrightarrow[m\to\infty]{} \mathrm{id}_{\mathbb{R}}$ with locally uniform convergence. □

A.10. **Proof of Lemma 2.24.** We will need the following lemma that will also be used elsewhere.

Lemma A.8. *For $f : \mathbb{R} \to \mathbb{R}$ and $a \in \mathbb{R}$, let $T_a f : \mathbb{R} \to \mathbb{R}, x \mapsto T_a f(x) = f(x - a)$. Furthermore, for $n \in \mathbb{N}_0$, let $X^n : \mathbb{R} \to \mathbb{R}, x \mapsto x^n$ and $V_n := \operatorname{span}\{T_a X^n : a \in \mathbb{R}\}$, with the convention $X^0 \equiv 1$.*

We have $V_n = \mathbb{R}_{\deg\leq n}[x]$, that is, V_n is the space of all polynomials of degree at most n. ◀

Proof. Clearly, $V_n \subset \mathbb{R}_{\deg\leq n}[x] =: V$, where $\dim V = n+1$. Therefore, it suffices to show that V_n contains $n+1$ linearly independent elements. In fact, we show that whenever $a_1, \dots, a_{n+1} \in \mathbb{R}$ are pairwise distinct, then the family $(T_{a_i} X^n)_{i=1,\dots,n+1} \subset V_n$ is linearly independent.

To see this, suppose that $\theta_1, \dots, \theta_{n+1} \in \mathbb{R}$ are such that $0 \equiv \sum_{i=1}^{n+1} \theta_i\, T_{a_i} X^n$. A direct computation using the binomial theorem shows that this implies $0 \equiv \sum_{\ell=0}^{n} \big[\binom{n}{\ell}(-1)^\ell X^{n-\ell} \sum_{i=1}^{n+1} \theta_i a_i^\ell\big]$. By comparing the coefficients of X^t, this leads to $0 = \big(\sum_{i=1}^{n+1} a_i^\ell\, \theta_i\big)_{\ell=0,\dots,n} = A^T\theta$, where $\theta = (\theta_1, \dots, \theta_{n+1}) \in \mathbb{R}^n$, and where the *Vandermonde matrix* $A := (a_i^j)_{i=1,\dots,n+1, j=0,\dots,n} \in \mathbb{R}^{(n+1)\times(n+1)}$ is invertible; see [34, Equation (4-15)]. Hence, $\theta = 0$, showing that $(T_{a_i} X^n)_{i=1,\dots,n+1}$ is a linearly independent family. □

Proof of Lemma 2.24. First, note

$$\varrho_r(x) + (-1)^r\, \varrho_r(-x) = \begin{cases} \varrho_r(x) = (x_+)^r = x^r, & \text{if } x \geq 0 \\ (-1)^r \varrho_r(-x) = (-1)^r[(-x)_+]^r = (-1)^r(-x)^r = x^r, & \text{if } x < 0\,. \end{cases} \tag{A.6}$$

Next, Lemma A.8 shows that $V_r = \mathbb{R}_{\deg\leq r}[x]$ has dimension $r+1$. Thus, given any polynomial $f \in \mathbb{R}_{\deg\leq r}[x]$, there are $a_1,\dots,a_{r+1} \in \mathbb{R}$ and $b_1,\dots,b_{r+1} \in \mathbb{R}$ such that for all $x \in \mathbb{R}$

$$f(x) = \sum_{\ell=1}^{r+1} a_\ell \cdot (T_{b_\ell} X^r)(x) \overset{(A.6)}{=} \sum_{\ell=1}^{r+1} a_\ell \cdot [\varrho_r(x-b_\ell) + (-1)^r \varrho_r\big(-(x+b_\ell)\big)]. \qquad \square$$

A.11. **Proof of Lemma 2.26.** For Part (1), define $w_j := 6n(2^j-1)$ and $m_j := (2n+1)(2^j-1)-1$. We will prove below by induction on $j \in \mathbb{N}$ that $M_{2^j} \in \mathtt{NN}^{\varrho,2^j,1}_{w_j,2j,m_j}$. Let us see first that this implies the result. For arbitrary $d \in \mathbb{N}_{\geq 2}$ and $j = \lceil \log_2 d \rceil$ it is not hard to see that

$$P : \mathbb{R}^d \to \mathbb{R}^{2^j}, x \mapsto (x, 1_{2^j-d}) = (x, 0_{2^j-d}) + (0_d, 1_{2^j-d})$$

is affine-linear with $\|P\|_{\ell^{0,\infty}_*} = 1$ (cf. Equation (2.4)) and that $M_d = M_{2^j} \circ P$. Using Lemma 2.18-(1) we get $M_d \in \mathtt{NN}^{\varrho,2^j,1}_{w_j,2j,m_j}$ as claimed.

We now proceed to the induction. As a preliminary, note that by assumption there are $a \in \mathbb{R}$, $\alpha_1,\dots,\alpha_n \in \mathbb{R}$ and $\beta_1,\dots,\beta_n \in \mathbb{R}$ such that for all $x \in \mathbb{R}$

$$x^2 = a + \sum_{\ell=1}^{n} \beta_\ell \, \varrho(x-\alpha_\ell).$$

Put differently, the affine-linear maps $T_1 : \mathbb{R} \to \mathbb{R}^n, x \mapsto (x-\alpha_\ell)_{\ell=1}^n$ and $T_2 : \mathbb{R}^n \to \mathbb{R}, y \mapsto a + \sum_{\ell=1}^n \beta_\ell \, y_\ell$ satisfy $x^2 = T_2 \circ (\varrho \otimes \cdots \otimes \varrho) \circ T_1(x)$ for all $x \in \mathbb{R}$, where the $\otimes$-product has n factors. Since $x \cdot y = \frac{1}{4}\big((x+y)^2 - (x-y)^2\big)$ for all $x, y \in \mathbb{R}$, if we define the maps $T_0 : \mathbb{R}^2 \to \mathbb{R}^2, (x,y) \mapsto (x+y, x-y)$ and $T_3 : \mathbb{R}^2 \to \mathbb{R}, (u,v) \mapsto \frac{1}{4}(u-v)$, then for all $x, y \in \mathbb{R}$

$$x \cdot y = \tfrac{1}{4} \cdot \big((x+y)^2 - (x-y)^2\big) = \big(S_2 \circ \overbrace{(\varrho \otimes \cdots \otimes \varrho)}^{2n \text{ factors}} \circ S_1\big)(x,y).$$

where $S_1 := (T_1 \otimes T_1) \circ T_0 : \mathbb{R}^2 \to \mathbb{R}^{2n}$ and $S_2 := T_3 \circ (T_2 \otimes T_2) : \mathbb{R}^{2n} \to \mathbb{R}$. As $\|S_1\|_{\ell^0} \leq 4n$ and $\|S_2\|_{\ell^0} \leq 2n$ we obtain $M_2 = \mathtt{R}(\Phi_1)$ where $\Phi_1 = \big((S_1, \varrho \otimes \cdots \otimes \varrho), (S_2, \mathrm{id})\big) \in \mathcal{NN}^{\varrho,2,1}_{6n,2,2n}$. This establishes our induction hypothesis for $j=1$: $M_2 \in \mathtt{SNN}^{\varrho,2,1}_{6n,2,2n} \subset \mathtt{NN}^{\varrho,2,1}_{w_j,2^j,m_j}$ for $j=1$.

We proceed to the actual induction step. Define the affine maps $U_1, U_2 : \mathbb{R}^{2^{j+1}} \to \mathbb{R}^{2^j}$ by

$$U_1(x) := (x_1,\dots,x_{2^j}) =: \overline{x} \quad \text{and} \quad U_2(x) := (x_{2^j+1},\dots,x_{2^{j+1}}) =: x' \quad \text{for} \quad x \in \mathbb{R}^{2^{j+1}}.$$

With these definitions, observe that $M_{2^{j+1}}(x) = M_{2^j}(\overline{x}) M_{2^j}(x') = M_2\big(M_{2^j}(U_1(x)), M_{2^j}(U_2(x))\big)$.

By the induction hypothesis there is a network $\Phi_j = \big((V_1,\alpha_1),\dots,(V_L,\mathrm{id})\big) \in \mathcal{NN}^{\varrho,2^j,1}_{w_j,2j,m_j}$ with $L(\Phi_j) = L \leq 2j$ such that $M_{2^j} = \mathtt{R}(\Phi_j)$. Since $\|U_i\|_{\ell^{0,\infty}_*} = 1$, the second part of Lemma A.3 shows $\|V_1 \circ U_i\|_{\ell^0} \leq \|V_1\|_{\ell^0}$, whence $M_{2^j} \circ U_i = \mathtt{R}(\Psi_i)$, where $\Psi_i = \big((V_1 \circ U_i, \alpha_1), (V_2,\alpha_2), \dots, (V_L, \mathrm{id})\big)$ satisfies $W(\Psi_i) \leq W(\Phi_j)$, $N(\Psi_i) \leq N(\Phi_j)$, $L(\Psi_i) = L$, and $\Psi_i \in \mathcal{NN}^{\varrho,2^j,1}_{w_j,2j,m_j}$. Thus, Lemma A.1 shows that $f := (M_{2^j} \circ U_1, M_{2^j} \circ U_2) \in \mathtt{NN}^{\varrho,2^{j+1},2}_{2w_j,2j,2m_j}$. Since $M_2 \in \mathtt{NN}^{\varrho,2,1}_{6n,2,2n}$, Lemma 2.18-(2) shows that $M_{2^{j+1}} = M_2 \circ f \in \mathtt{NN}^{\varrho,2^{j+1},1}_{2w_j+6n,2j+2,2m_j+2n+2}$.

To conclude the proof of Part (1), note that $2w_j + 6n = 12n(2^j-1) + 6n = 6n(2^{j+1}-1) = w_{j+1}$ and $2m_j + 2n + 2 = 2(2n+1)(2^j-1) + 2n = (2n+1)(2^{j+1}-2) + 2n + 1 - 1 = m_{j+1}$.

To prove Part (2), we recall from Part (1) that $M_2 : \mathbb{R}^2 \to \mathbb{R}, (x,y) \mapsto x \cdot y$ satisfies $M_2 = \mathtt{R}(\Psi)$ with $\Psi \in \mathcal{SNN}^{\varrho,2,1}_{6n,2,2n}$ and $L(\Psi) = 2$. Next, let $P^{(i)} : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R} \times \mathbb{R}, (x,y) \mapsto (x, y_i)$ for each $i \in \{1,\dots,k\}$, and note that $P^{(i)}$ is linear with $\|P^{(i)}\|_{\ell^{0,\infty}} = 1 = \|P^{(i)}\|_{\ell^{0,\infty}_*}$. Lemma 2.18-(1) shows that $M_2 \circ P^{(i)} = \mathtt{R}(\Psi_i)$ where $\Psi_i \in \mathcal{SNN}^{\varrho,1+k,1}_{6n,2,2n}$ and $L(\Psi_i) = L(\Psi) = 2$. To conclude, observe $(M_2 \circ P^{(i)})(x,y) = x \cdot y_i = [m(x,y)]_i$ for $m : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}^k, (x,y) \mapsto x \cdot y$. Therefore, Lemma 2.17-(2) shows that $m = (M_2 \circ P^{(1)}, \dots, M_2 \circ P^{(k)}) \in \mathtt{NN}^{\varrho,1+k,k}_{6kn,2,2kn}$, as desired. $\square$

Appendix B. Proofs for Section 3

B.1. **Proof of Lemma 3.1.** Let $f \in A^\alpha_q(X,\Sigma')$. For the sake of brevity, set $\varepsilon_n := E(f,\Sigma_n)_X$ and $\delta_n := E(f,\Sigma'_n)_X$ for $n \in \mathbb{N}_0$. First, observe that $\varepsilon_n \leq \|f\|_X = \delta_0$ for all $n \in \mathbb{N}_0$. Furthermore, we have by assumption that $\varepsilon_{cm} \leq \delta_m$ for all $m \in \mathbb{N}$. Now, setting $m_n := \lfloor \frac{n-1}{c} \rfloor \in \mathbb{N}$ for $n \in \mathbb{N}_{\geq c+1}$, note that $n - 1 \geq c\, m_n$, and hence $\varepsilon_{n-1} \leq \varepsilon_{c\, m_n} \leq C \cdot \delta_{m_n}$. Therefore, we see

$$\varepsilon_{n-1} \leq \delta_0 \text{ if } 1 \leq n \leq c \qquad \text{and} \qquad \varepsilon_{n-1} \leq C \cdot \delta_{m_n} \text{ if } n \geq c+1.$$

Next, note for $n \in \mathbb{N}_{\geq c+1}$ that $m_n \geq 1$ and $m_n \geq \frac{n-1}{c} - 1$, whence $n \leq c\, m_n + c + 1 \leq (2c+1) m_n$. Therefore, $n^\alpha \leq (2c+1)^\alpha m_n^\alpha$. Likewise, since $m_n \leq n$, we have $n^{-1} \leq m_n^{-1}$ for all $n \in \mathbb{N}_{\geq c+1}$.

There are now two cases. First, if $q < \infty$, and if we set $K := K(\alpha, q, c) := \sum_{n=1}^{c} n^{\alpha q - 1}$, then

$$\begin{aligned} \|f\|_{A^\alpha_q(X,\Sigma)}^q &= \sum_{n=1}^{\infty} [n^\alpha\, \varepsilon_{n-1}]^q \, \frac{1}{n} \leq \delta_0^q \cdot \sum_{n=1}^{c} n^{\alpha q - 1} + C^q \sum_{n=c+1}^{\infty} (n^\alpha\, \delta_{m_n})^q \frac{1}{n} \\ &\leq K\, \delta_0^q + C^q\, (2c+1)^{\alpha q} \sum_{n=c+1}^{\infty} (m_n^\alpha\, \delta_{m_n})^q \frac{1}{m_n}\,. \end{aligned}$$

Further, for $n \in \mathbb{N}_{\geq c+1}$ satisfying $m_n = m$ for some $m \in \mathbb{N}$, we have $m \leq \frac{n-1}{c} < m+1$, which easily implies $|\{n \in \mathbb{N}_{\geq c+1} \colon m_n = m\}| \leq |\{n \in \mathbb{N} \colon cm + 1 \leq n < cm + c + 1\}| = c$. Thus,

$$\begin{aligned} \sum_{n=c+1}^{\infty} (m_n^\alpha\, \delta_{m_n})^q \frac{1}{m_n} &= \sum_{m=1}^{\infty} (m^\alpha\, \delta_m)^q \cdot \frac{1}{m} \cdot |\{n \in \mathbb{N}_{\geq c+1} \colon m_n = m\}| \\ &\leq c \sum_{m=1}^{\infty} (m^\alpha\, \delta_m)^q \frac{1}{m} \leq c \sum_{m=1}^{\infty} (m^\alpha\, \delta_{m-1})^q \frac{1}{m} = c\, \|f\|^q_{A^\alpha_q(X,\Sigma')}\,. \end{aligned}$$

Overall, we thus see for $q < \infty$ that

$$\|f\|^q_{A^\alpha_q(X,\Sigma)} \leq (K + C^q (2c+1)^{\alpha q} c) \cdot \|f\|^q_{A^\alpha_q(X,\Sigma')} < \infty\,,$$

where the constant $K + C^q(2c+1)^{\alpha q} c$ only depends on α, q, c, C.

The adaptations for the (easier) case $q = \infty$ are left to the reader. □

B.2. **Proof of Proposition 3.2.** In [21, Chapter 7, Discussion around Equation (9.2)] it was shown that the embedding (3.2) holds. All other properties claimed in Proposition 3.2 follow by combining Remark 3.5, Proposition 3.8, and Theorem 3.12 in [4]. □

B.3. **Proof of Lemma 3.20.** For $p \in (0, \infty)$, the claim is clear, since it is well-known that $L_p(\Omega; \mathbb{R}^k)$ is complete, and since one can extend each $g \in X_p^k(\Omega) = L_p(\Omega; \mathbb{R}^k)$ by zero to a function $f \in L^p(\Omega; \mathbb{R}^k)$ satisfying $g = f|_\Omega$.

Now, we consider the case $p = \infty$. We first prove completeness of $X_\infty^k(\Omega)$. Let $(f_n)_{n\in\mathbb{N}} \subset X_\infty^k(\Omega)$ be a Cauchy sequence. It is well-known that there is a continuous function $f : \Omega \to \mathbb{R}^k$ such that $f_n \to f$ uniformly. In fact (see for instance [63, Theorem 12.8]), f is uniformly continuous. It remains to show that f vanishes at infinity. Let $\varepsilon > 0$ be arbitrary, and choose $n \in \mathbb{N}$ such that $\|f - f_n\|_{\sup} \leq \frac{\varepsilon}{2}$. Since f_n vanishes at ∞, there is $R > 0$ such that $|f_n(x)| \leq \frac{\varepsilon}{2}$ for $x \in \Omega$ with $|x| \geq R$. Therefore, $|f(x)| \leq \varepsilon$ for such x, proving that $f \in X_\infty^k(\Omega)$, while $\|f - f_n\|_{X_\infty^k(\Omega)} \to 0$ follows from the uniform convergence $f_n \to f$.

Finally, we prove that $X_\infty^k(\Omega) = \{f|_\Omega : f \in C_0(\mathbb{R}^d; \mathbb{R}^k)\}$. By considering components it is enough to prove that $\{f|_\Omega : f \in C_0(\mathbb{R}^d)\} = X_\infty(\Omega)$. To see that $\{f|_\Omega : f \in C_0(\mathbb{R}^d)\} \subset X_\infty(\Omega)$, simply note that[7] if $f \in C_0(\mathbb{R}^d)$, then f is not only continuous, but in fact *uniformly* continuous. Therefore, $f|_\Omega$ is also uniformly continuous (and vanishes at infinity), whence $f|_\Omega \in X_\infty(\Omega)$.

For proving $X_\infty(\Omega) \subset \{f|_\Omega : f \in C_0(\mathbb{R}^d)\}$, we will use the notion of the *one-point compactification* $Z_\infty := \{\infty\} \cup Z$ of a locally compact Hausdorff space Z (where we assume that $\infty \notin Z$); see [26, Proposition 4.36]. The topology on Z_∞ is given by $\mathcal{T}_Z := \{U \colon U \subset Z \text{ open}\} \cup \{Z_\infty \setminus K \colon K \subset Z \text{ compact}\}$. Then, $(Z_\infty, \mathcal{T}_Z)$ is a compact Hausdorff space and the topology induced on Z as a subspace of Z_∞ coincides with the original topolog on Z; see [26, Proposition 4.36]. Furthermore, if $A \subset Z$ is *closed*, then a direct verification shows that the relative topology on A_∞ as a subset of Z_∞ coincides with the topology $\mathcal{T}_A$.

Now, let $g \in X_\infty(\Omega)$. Since g is uniformly continuous, it follows (see [3, Lemma 3.11]) that there is a uniformly continuous function $\widetilde{g} : A \to \mathbb{R}$ satisfying $g = \widetilde{g}|_\Omega$, with $A := \overline{\Omega} \subset \mathbb{R}^d$ the closure of Ω in $\mathbb{R}^d$.

Since $g \in C_0(\Omega)$, it is not hard to see that $\widetilde{g} \in C_0(A)$. Hence, [26, Proposition 4.36] shows that the function $G : A_\infty \to \mathbb{R}$ defined by $G(x) = \widetilde{g}(x)$ for $x \in A$ and $G(\infty) = 0$ is continuous. Since $A_\infty \subset (\mathbb{R}^d)_\infty$ is compact, the Tietze extension theorem (see [26, Theorem 4.34]) shows that there is a continuous extension $H : (\mathbb{R}^d)_\infty \to \mathbb{R}$ of G. Again by [26, Proposition 4.36], this implies that $f := H|_{\mathbb{R}^d} \in C_0(\mathbb{R}^d)$. By construction, we have $g = f|_\Omega$. □

B.4. **Proof of Theorem 3.23.**

[7] For instance, [26, Proposition 4.35] shows that each function in $C_0(\mathbb{R}^d)$ is a uniform limit of continuous, compactly supported functions, [27, Proposition (2.6)] shows that such functions are uniformly continuous, while [63, Theorem 12.8] shows that the uniform continuity is preserved by the uniform limit.

B.4.1. *Proof of Claims 1a-1b.* We use the following lemma.

Lemma B.1. *Let $\mathcal{C}$ be one of the following classes of functions:*

- *locally bounded functions;*
- *Borel-measurable functions;*
- *continuous functions;*
- *Lipschitz continuous functions;*
- *locally Lipschitz continuous functions.*

If the activation function ϱ belongs to $\mathcal{C}$, then any $f \in \mathtt{NN}^{\varrho,d,k}$ also belongs to $\mathcal{C}$. ◀

Proof. First, note that each affine-linear map $T : \mathbb{R}^d \to \mathbb{R}^k$ belongs to *all* of the mentioned classes. Furthermore, note that since $\mathbb{R}^d$ is locally compact, a function $f : \mathbb{R}^d \to \mathbb{R}^k$ is locally bounded [locally Lipschitz] if and only if f is bounded [Lipschitz continuous] on each *bounded* set. From this, it easily follows that each class $\mathcal{C}$ is closed under composition. Finally, it is not hard to see that if $f_1, \dots, f_n : \mathbb{R} \to \mathbb{R}$ all belong to the class $\mathcal{C}$, then so does $f_1 \otimes \cdots \otimes f_n : \mathbb{R}^n \to \mathbb{R}^n$.

Combining these facts with the definition of the realization of a neural network, we get the claim. □

As ϱ is locally bounded and Borel measurable, by Lemma B.1 each $g \in \mathtt{NN}^{\varrho,d,k}$ is locally bounded and measurable. As Ω is bounded, we get $g|_\Omega \in L_p(\Omega;\mathbb{R}^k)$ for all $p \in (0,\infty]$, and hence $g \in X_p^k(\Omega)$ if $p < \infty$. This establishes claim 1a. Finally, if $p = \infty$, then by our additional assumption that ϱ is continuous, g is continuous by Lemma B.1. On the compact set $\overline{\Omega}$, g is thus uniformly continuous and bounded, so that $g|_\Omega$ is uniformly continuous and bounded as well, that is, $g|_\Omega \in X_\infty^k(\Omega)$. This establishes claim 1b. □

B.4.2. *Proof of claims 1c-1d.* We first consider the case $p < \infty$. Let $f \in X_p^k(\Omega) = L_p(\Omega;\mathbb{R}^k)$ and $\varepsilon > 0$. For each $i \in \{1,\dots,k\}$, extend the i-th component function f_i by zero to a function $g_i \in L_p(\mathbb{R}^d)$. As is well-known (see for instance [25, Chapter VI, Theorem 2.31]), $C_c^\infty(\mathbb{R}^d)$ is dense in $L_p(\mathbb{R}^d)$, so that we find $h_i \in C_c^\infty(\mathbb{R}^d)$ satisfying $\|g_i - h_i\|_{L_p} < \varepsilon$. Choose $R > 0$ satisfying $\operatorname{supp}(h_i) \subset [-R,R]^d$ and $\Omega \subset [-R,R]^d$. By the universal approximation theorem (Theorem 3.22), we can find $\gamma_i \in \mathtt{NN}^{\varrho,d,1}_{\infty,2,\infty} \subset \mathtt{NN}^{\varrho,d,1}_{\infty,L,\infty}$ satisfying $\|h_i - \gamma_i\|_{L_\infty([-R,R]^d)} \leq \varepsilon/(4R)^{d/p}$. Note that the inclusion $\mathtt{NN}^{\varrho,d,1}_{\infty,2,\infty} \subset \mathtt{NN}^{\varrho,d,1}_{\infty,L,\infty}$ used above is (only) true since we are considering *generalized* neural networks, and since $L \geq 2$.

Using the elementary estimate $(a+b)^p \leq (2\max\{a,b\})^p \leq 2^p(a^p + b^p)$, we see

$$|\gamma_i(x) - g_i(x)|^p \leq \big(|\gamma_i(x) - h_i(x)| + |h_i(x) - g_i(x)|\big)^p \leq 2^p\Big(\frac{\varepsilon^p}{(4R)^d} + |h_i(x) - g_i(x)|^p\Big) \quad \forall\, x \in [-R,R]^d\,,$$

which easily implies $\|\gamma_i - g_i\|^p_{L_p([-R,R]^d)} \leq 2^p(\varepsilon^p + \|h_i - g_i\|^p_{L_p([-R,R]^d)}) \leq 2^{1+p}\varepsilon^p$.

Lemma 2.17 shows that $\gamma := (\gamma_1,\dots,\gamma_k) \in \mathtt{NN}^{\varrho,d,k}_{\infty,L,\infty}$, whence $\gamma|_\Omega \in \Sigma_\infty(X_p^k(\Omega),\varrho,\mathscr{L})$ by claims 1a-1b of Theorem 3.23. Finally, since $g_i|_\Omega = f_i$, we have

$$\|f - \gamma|_\Omega\|^p_{L_p(\Omega)} \leq \sum_{i=1}^k \|g_i - \gamma_i\|^p_{L_p([-R,R]^d)} \leq 2^{1+p} k \cdot \varepsilon^p\,.$$

Since $\varepsilon > 0$ was arbitrary, this proves the desired density.

Now, we consider the case $p = \infty$. Let $f \in X_\infty^k(\Omega)$. Lemma 3.20 shows that there is a continuous function $g : \mathbb{R}^d \to \mathbb{R}^k$ such that $f = g|_\Omega$. Since $L \geq 2$, we can apply the universal approximation theorem (Theorem 3.22) to each of the component functions g_i of $g = (g_1,\dots,g_k)$ to obtain functions $\gamma_i \in \mathtt{NN}^{\varrho,d,1}_{\infty,2,\infty} \subset \mathtt{NN}^{\varrho,d,1}_{\infty,L,\infty}$ satisfying $\|g_i - \gamma_i\|_{L_\infty([-R,R]^d)} \leq \varepsilon$, where we chose $R > 0$ so large that $\Omega \subset [-R,R]^d$. Lemma 2.17 shows that $\gamma := (\gamma_1,\dots,\gamma_k) \in \mathtt{NN}^{\varrho,d,k}_{\infty,L,\infty}$, whence $\gamma|_\Omega \in \Sigma_\infty(X_p^k(\Omega),\varrho,\mathscr{L})$ by claims 1a-1b of Theorem 3.23, since ϱ is continuous. Finally, since $g_i|_\Omega = f_i$, we have

$$\sup_{x\in\Omega} \|f(x) - \gamma(x)\|_{\ell^\infty} \leq \sup_{x\in[-R,R]^d} \max_{i\in\{1,\dots,k\}} |g_i(x) - \gamma_i(x)| \leq \varepsilon\,.$$

Since $\varepsilon > 0$ was arbitrary, this proves the desired density. □

B.4.3. *Proof of Claim (2).* Set $\mathcal{V} := \mathtt{NN}^{\varrho,d,1}_{\infty,L} \cap X_p(\mathbb{R}^d)$. Lemma 2.17 easily shows that $\mathcal{V}$ is a vector space. Furthermore, Lemma 2.18 shows that if $f \in \mathcal{V}$, $A \in \mathrm{GL}(\mathbb{R}^d)$, and $b \in \mathbb{R}^d$, then $f(A \bullet + b) \in \mathcal{V}$ as well. Clearly, all these properties also hold for $\overline{\mathcal{V}}$ instead of $\mathcal{V}$, where the closure is taken in $X_p(\mathbb{R}^d)$.

It suffices to show that $\mathcal{V}$ is dense in $X_p(\mathbb{R}^d)$. Indeed, suppose for the moment that this is true. Let $f \in X_p^k(\Omega)$ be arbitrary. By applying Lemma 3.20 to each of the component functions f_i of f, we see for each $i \in \{1,\dots,k\}$ that there is a function $F_i \in X_p(\mathbb{R}^d)$ such that $f_i = F_i|_\Omega$. Now, let $\varepsilon > 0$ be arbitrary, and set $p_0 := \min\{1,p\}$. Since $\mathcal{V}$ is dense in $X_p(\mathbb{R}^d)$, there is for each $i \in \{1,\dots,k\}$ a function $G_i \in \mathcal{V}$ such that

$\|G_i - F_i\|_{L_p}^{p_0} \leq \varepsilon^{p_0}/k$. Lemma 2.17 shows $g := (G_1|_\Omega, \dots, G_k|_\Omega) \in \mathtt{NN}_{\infty,L}^{\varrho,d,k}(\Omega) \cap X_p^k(\Omega) = \Sigma_\infty(X_p^k(\Omega), \varrho, \mathscr{L})$, and it is not hard to see that $\|f - g\|_{X_p^k(\Omega)}^{p_0} \leq \sum_{i=1}^k \|F_i - G_i\|_{X_p^k(\Omega)}^{p_0} \leq \varepsilon^{p_0}$, and hence $\|f - g\|_{X_p^k(\Omega)} \leq \varepsilon$. As $\varepsilon > 0$ and $g \in X_p^k(\Omega)$ were arbitrary, this proves that $\Sigma_\infty(X_p^k(\Omega), \varrho, \mathscr{L})$ is dense in $X_p^k(\Omega)$, as desired.

It remains to show that $\mathcal{V} \subset X_p(\mathbb{R}^d)$ is dense. To prove this, we distinguish three cases:

Case 1 ($p \in [1, \infty)$): First, the existence of the "radially decreasing L_1-majorant" μ for g, [11, Lemma A.2] shows that $P|g| \in L_\infty(\mathbb{R}^d) \subset L_p^{\mathrm{loc}}(\mathbb{R}^d)$, where $P|g|$ is a certain *periodization* of $|g|$ whose precise definition is immaterial for us. Since $g \in L_p(\mathbb{R}^d)$ and $P|g| \in L_p^{\mathrm{loc}}(\mathbb{R}^d)$, and $\int_{\mathbb{R}^d} g(x)\,dx \neq 0$, [11, Corollary 1] implies that $\mathcal{V}_0 := \operatorname{span}\{g_{j,k} \colon j \in \mathbb{N}, k \in \mathbb{Z}^d\}$ is dense in $L_p(\mathbb{R}^d)$, where $g_{j,k}(x) = 2^{jd/p} \cdot g(2^j x - k)$. As a consequence of the properties of the space $\mathcal{V}$ that we mentioned above, and since $g \in \overline{\mathcal{V}}$, we have $\mathcal{V}_0 \subset \overline{\mathcal{V}}$. Hence, $\mathcal{V} \subset L_p(\mathbb{R}^d)$ is dense, and we have $L_p(\mathbb{R}^d) = X_p(\mathbb{R}^d)$ since $p < \infty$.

Case 2 ($p \in (0, 1)$): Since $g \in L_1(\mathbb{R}^d) \cap L_p(\mathbb{R}^d)$ with $\int_{\mathbb{R}^d} g(x)\,dx \neq 0$, [39, Theorem 4 and Proposition 5(a)] show that $\mathcal{V}_0 \subset L_p(\mathbb{R}^d)$ is dense, where the space $\mathcal{V}_0$ is defined precisely as for $p \in [1, \infty)$. The rest of the proof is as for $p \in [1, \infty)$.

Case 3 ($p = \infty$): Note $X_p(\mathbb{R}^d) = C_0(\mathbb{R}^d)$. Let us assume towards a contradiction that $\mathcal{V}$ is not dense in $C_0(\mathbb{R}^d)$. By the Hahn-Banach theorem (see for instance [26, Theorem 5.8]), there is a bounded linear functional $\varphi \in (C_0(\mathbb{R}^d))^*$ such that $\varphi \not\equiv 0$, but $\varphi \equiv 0$ on $\overline{\mathcal{V}}$.

By the Riesz representation theorem for C_0 (see [26, Theorem 7.17]), there is a finite real-valued Borel-measure μ on $\mathbb{R}^d$ such that $\varphi(f) = \int_{\mathbb{R}^d} f(x)\,d\mu(x)$ for all $f \in C_0(\mathbb{R}^d)$. Thanks to the Jordan decomposition theorem (see [26, Theorem 3.4]), there are finite positive Borel measures μ_+ and μ_- such that $\mu = \mu_+ - \mu_-$.

Let $f \in C_0(\mathbb{R}^d)$ be arbitrary. For $a > 0$, define $g_a : \mathbb{R}^d \to \mathbb{R}, x \mapsto a^d\, g(ax)$, and note $T_x g_a \in \overline{\mathcal{V}}$ (and hence $\varphi(T_x g_a) = 0$) for all $x \in \mathbb{R}^d$, where $T_x g_a(y) = g_a(y - x)$. By Fubini's theorem and the change of variables $y = -z$, we get

$$\begin{aligned} \int_{\mathbb{R}^d} (f * g_a)(x)\,d\mu(x) &= \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} f(z)\, g_a(x - z)\,dz\,d\mu(x) \\ &= \int_{\mathbb{R}^d} f(-y) \int_{\mathbb{R}^d} g_a(y + x)\,d\mu(x)\,dy = \int_{\mathbb{R}^d} f(-y)\,\varphi(T_{-y}\, g_a)\,dy = 0 \end{aligned} \tag{B.1}$$

for all $a \geq 1$. Here, Fubini's theorem was applied to each of the integrals $\int (f * g_a)(x)\,d\mu_\pm(x)$, which is justified since

$$\int\int |f(z)\, g_a(x - z)|\,dz\,d\mu_\pm(x) \leq \mu_\pm(\mathbb{R}^d)\,\|f\|_{L_\infty}\,\|T_z g_a\|_{L_1} = \mu_\pm(\mathbb{R}^d)\,\|f\|_{L_\infty}\,\|g_a\|_{L_1} < \infty\,.$$

Now, since $f \in C_0(\mathbb{R}^d)$ is bounded and uniformly continuous, [26, Theorem 8.14] shows $f * g_a \to f$ uniformly as $a \to \infty$. Therefore, (B.1) implies $\varphi(f) = \int_{\mathbb{R}^d} f(x)\,d\mu(x) = \lim_{a\to\infty} \int_{\mathbb{R}^d} (f * g_a)(x)\,d\mu(x) = 0$, since μ is a finite measure. This implies $\varphi \equiv 0$ on $C_0(\mathbb{R}^d)$, which is the desired contradiction. □

B.5. **Proof of Lemma 3.26. Part (1):** Define

$$t : \mathbb{R} \to \mathbb{R}, x \mapsto \sigma\big(x/\varepsilon\big) - \sigma\big(1 + (x - 1)/\varepsilon\big)\,.$$

A straightforward calculation using the properties of σ shows that

$$t(x) = \begin{cases} 0, & \text{if } x \in \mathbb{R} \setminus [0, 1], \\ 1, & \text{if } x \in [\varepsilon, 1 - \varepsilon]. \end{cases} \tag{B.2}$$

We claim that $0 \leq t \leq 1$. To see this, first note that if $r \geq 1$, then $\sigma(x - r) \leq \sigma(x)$ for all $x \in \mathbb{R}$. Indeed, if $x \leq r$, then $\sigma(x - r) = 0 \leq \sigma(x)$; otherwise, if $x > r$, then $x \geq 1$, and hence $\sigma(x - r) \leq 1 = \sigma(x)$. Since $r := \frac{1}{\varepsilon} - 1 \geq 1$, we thus see that $t(x) = \sigma(\frac{x}{\varepsilon}) - \sigma(\frac{x}{\varepsilon} - r) \geq 0$ for all $x \in \mathbb{R}$. Finally, we trivially have $t(x) \leq \sigma(\frac{x}{\varepsilon}) \leq 1$ for all $x \in \mathbb{R}$.

Now, if we define

$$g_0 : \mathbb{R}^d \to \mathbb{R}, x \mapsto \sigma\left(1 + \sum_{i=1}^d t(x_i) - d\right),$$

we see $0 \leq g_0 \leq 1$. Furthermore, for $x \in [\varepsilon, 1 - \varepsilon]^d$, we have $t(x_i) = 1$ for all $i \in \{1, \dots, d\}$, whence $g_0(x) = 1$. Likewise, if $x \notin [0, 1]^d$, then $t(x_i) = 0$ for at least one $i \in \{1, \dots, d\}$. Since $0 \leq t(x_i) \leq 1$ for

all i, this implies $\sum_{i=1}^d t(x_i) - d \leq -1$, and thus $g_0(x) = 0$. All in all, and because of $0 \leq g_0 \leq 1$, these considerations imply that $\operatorname{supp}(g_0) \subset [0,1]^d$ and

$$|g_0(x) - \mathbb{1}_{[0,1]^d}(x)| \leq \mathbb{1}_{[0,1]^d \setminus [\varepsilon, 1-\varepsilon]^d}(x) \quad \forall\, x \in \mathbb{R}^d\,. \tag{B.3}$$

Now, for proving the general case of Part (1), let $h := g_0$, while $h := t$ in case of $d = 1$. As a consequence of Equations (B.3) and (B.2) and of $0 \leq t \leq 1$, we then see that Condition (3.10) is satisfied in both cases. Thus, all that needs to be shown is that $h = g_0 \in \mathtt{NN}^{\varrho,d,1}_{2dW(N+1),2L-1,(2d+1)N}$ or that $h = t \in \mathtt{NN}^{\varrho,1,1}_{2W,L,2N}$ in case of $d = 1$. We will verify both of these properties in the proof of Part (2) of the lemma.

Part (2): We first establish the claim for the special case $[a,b] = [0,1]^d$. With λ denoting the d-dimensional Lebesgue measure, and with h as constructed in Part (1), we deduce from (3.10) that

$$\|h - \mathbb{1}_{[0,1]^d}\|_{L_p}^p \leq \lambda([0,1]^d \setminus [\varepsilon, 1-\varepsilon]^d) = \left[1 - (1-2\varepsilon)^d\right].$$

Since the right-hand side vanishes as $\varepsilon \to 0$, this proves the claim for the special case $[a,b] = [0,1]^d$, once we show $h = \mathtt{R}(\Phi)$ for Φ with appropriately many layers, neurons, and nonzero weights.

By assumption on σ, there is $L_0 \leq L$ such that $\sigma = \mathtt{R}(\Phi_\sigma)$ for some $\Phi_\sigma \in \mathcal{NN}^{\varrho,1,1}_{W,L_0,N}$ with $L(\Phi_\sigma) = L_0$. For $i \in \{1,\dots,d\}$ set $f_{i,1} : \mathbb{R}^d \to \mathbb{R}, x \mapsto \sigma(\frac{x_i}{\varepsilon})$ and $f_{i,2} : \mathbb{R}^d \to \mathbb{R}, x \mapsto -\sigma(1 + \frac{x_i - 1}{\varepsilon})$. By Lemma 2.18-(1) there exist $\Psi_{i,1}, \Psi_{i,2} \in \mathcal{NN}^{\varrho,d,1}_{W,L_0,N}$ with $L(\Psi_{i,1}) = L(\Psi_{i,2}) = L_0$ for any $i \in \{1,\dots,d\}$ such that $f_{i,1} = \mathtt{R}(\Psi_{i,1})$ and $f_{i,2} = \mathtt{R}(\Psi_{i,2})$. Lemma 2.17-(3) then shows that

$$F : \mathbb{R}^d \to \mathbb{R}, x \mapsto \sum_{i=1}^d t(x_i) = \sum_{i=1}^d f_{i,1}(x) + \sum_{i=1}^d f_{i,2}(x)$$

satisfies $F = \mathtt{R}(\Phi_F)$ for some $\Phi_F \in \mathcal{NN}^{\varrho,d,1}_{2dW,L_0,2dN}$ with $L(\Phi_F) = L_0$. Hence, Lemma 2.18-(1) shows that $G : \mathbb{R}^d \to \mathbb{R}, x \mapsto 1 + \sum_{i=1}^d t(x_i) - d$ satisfies $G = \mathtt{R}(\Phi_G)$ for some $\Phi_G \in \mathcal{NN}^{\varrho,d,1}_{2dW,L_0,2dN}$ with $L(\Phi_G) = L_0$.

In case of $d = 1$, set $L' := L_0$ and recall that $h = t = F$, where we saw above that $F = \mathtt{R}(\Phi_F)$ and $\Phi_F \in \mathcal{NN}^{\varrho,1,1}_{2W,L_0,2N}$ with $L(\Phi_F) = L_0$. For general $d \in \mathbb{N}$ set $L' := 2L_0 - 1$ and recall that $h = g_0 = \sigma \circ G$. Hence, Lemma 2.18-(3) shows $h = \mathtt{R}(\Phi_h)$ for some $\Phi_h \in \mathcal{NN}^{\varrho,d,1}$ with $L(\Phi_h) = L'$, $N(\Phi_h) \leq (2d+1)N$ and $W(\Phi_h) \leq 2dW + \max\{2dN, d\}W \leq 2dW(N+1)$.

It remains to transfer the result from $[0,1]^d$ to the general case $[a,b]$. To this end, define the invertible affine-linear map

$$T_0 : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto \big((b_i - a_i)^{-1} \cdot (x_i - a_i)\big)_{i \in \{1,\dots,d\}}\,.$$

A direct calculation shows $\mathbb{1}_{[0,1]^d} \circ T_0 = \mathbb{1}_{T_0^{-1}[0,1]^d} = \mathbb{1}_{[a,b]}$. Since $\|T_0\|_{\ell_*^{0,\infty}} = 1$, the first part of Lemma 2.18 shows that $g := h \circ T_0 = \mathtt{R}(\Phi)$ for some $\Phi \in \mathcal{NN}^{\varrho,d,1}_{2dW(N+1),2L_0-1,(2d+1)N}$ with $L(\Phi) = 2L_0 - 1 = L'$ (resp. $g := h \circ T_0 = \mathtt{R}(\Phi)$ for some $\Phi \in \mathcal{NN}^{\varrho,1,1}_{2W,L_0,2N}$ with $L(\Phi) = L_0 = L'$ in case of $d = 1$) with h as above. Moreover, by an application of the change-of-variables-formula, we get

$$\begin{aligned}\|g - \mathbb{1}_{[a,b]}\|_{L_p} &= \|h \circ T_0 - \mathbb{1}_{[0,1]^d} \circ T_0\|_{L_p} \\ &= \left|\det \operatorname{diag}\big((b_i - a_i)^{-1}\big)_{i \in \{1,\dots,d\}}\right|^{-1/p} \cdot \|g - \mathbb{1}_{[0,1]^d}\|_{L_p} = \|g - \mathbb{1}_{[0,1]^d}\|_{L_p} \cdot \prod_{i=1}^d (b_i - a_i)^{1/p}\,.\end{aligned}$$

As seen above, the first factor can be made arbitrarily small by choosing $\varepsilon \in (0, \frac{1}{2})$ suitably. Since the second factor is constant, this proves the claim. □

Appendix C. Proofs for Section 4

C.1. **Proof of Lemma 4.9.** We begin with three auxiliary results.

Lemma C.1. *Let $f : \mathbb{R} \to \mathbb{R}$ be* continuously *differentiable. Define $f_h : \mathbb{R} \to \mathbb{R}, x \mapsto h^{-1} \cdot (f(x+h) - f(x))$ for $h \in \mathbb{R} \setminus \{0\}$. Then $f_h \to f$ as $h \to 0$ with locally uniform convergence on $\mathbb{R}$.* ◀

Proof. This is an easy consequence of the mean-value theorem, using that f' is locally uniformly continuous. For more details, we refer to [40, Theorem 4.14]. □

Since ϱ_{r+1} is continuously differentiable with $\varrho'_{r+1} = \varrho_r$, the preceding lemma immediately implies the following result.

Corollary C.2. *For $r \in \mathbb{N}$, $h > 0$, $\sigma_h : \mathbb{R} \to \mathbb{R}, x \mapsto (r+1)^{-1} \cdot h^{-1} \cdot \big(\varrho_{r+1}(x+h) - \varrho_{r+1}(x)\big)$ we have $\sigma_h = \mathtt{R}(\Psi_h)$ where $\Psi_h \in \mathcal{SNN}^{\varrho_{r+1},1,1}_{4,2,2}$, $L(\Psi_h) = 2$, and $\lim_{h \to 0} \sigma_h = \varrho_r$ with locally uniform convergence on $\mathbb{R}$.* ◀

We need one more auxiliary result for the proof of Lemma 4.9.

Corollary C.3. *For any $d, k, r \in \mathbb{N}$, $j \in \mathbb{N}_0$, $W, N \in \mathbb{N}_0$, $L \in \mathbb{N}$ we have*

$$\mathtt{NN}^{\varrho_r,d,k}_{W,L,N} \subset \overline{\mathtt{NN}^{\varrho_{r+j},d,k}_{4^j W,L,2^j N}} \tag{C.1}$$

where closure is with respect to locally uniform convergence on $\mathbb{R}^d$. ◀

Proof. We prove by induction on δ that the result holds for any $0 \le j \le \delta$. This is trivial for $\delta = 0$. By Corollary C.2 we can apply Lemma 2.21 to $\varrho := \varrho_{r+1}$ and $\sigma := \varrho_r$ (which is continuous) with $w = 4$, $\ell = 2$, $m = 2$. This yields for any $W, N \in \mathbb{N}_0$, $L \in \mathbb{N}$ that $\mathtt{NN}^{\varrho_r,d,k}_{W,L,N} \subset \overline{\mathtt{NN}^{\varrho_{r+1},d,k}_{4W,L,2N}}$, which shows that our induction hypothesis is valid for $\delta = 1$. Assume now that the hypothesis holds for some $\delta \in \mathbb{N}$, and consider $W, N \in \mathbb{N}_0$, $r, L \in \mathbb{N}$, $0 \le j \le \delta + 1$. If $j \le \delta$ then the induction hypothesis yields (C.1), so there only remains to check the case $j = \delta + 1$. By the induction hypothesis, for $r' = r + \delta$, $W' = 4^\delta W$, $N' = 2^\delta N$, $j = 1$ we have $\mathtt{NN}^{\varrho_{r+\delta},d,k}_{4^\delta W,L,2^\delta N} \subset \overline{\mathtt{NN}^{\varrho_{r+\delta+1},d,k}_{4^{\delta+1} W,L,2^{\delta+1} N}}$. Finally, $\mathtt{NN}^{\varrho_r,d,k}_{W,L,N} \subset \overline{\mathtt{NN}^{\varrho_{r+\delta},d,k}_{4^\delta W,L,2^\delta N}} \subset \overline{\mathtt{NN}^{\varrho_{r+\delta+1},d,k}_{4^{\delta+1} W,L,2^{\delta+} N}}$ by the induction hypothesis for $j = \delta$. □

Proof of Lemma 4.9. The proof is by induction on n. For $n = 1$, ϱ is a polynomial of degree at most r. By Lemma 2.24, ϱ_r can represent any such polynomial with $2(r+1)$ terms, whence $\varrho \in \mathtt{NN}^{\varrho_r,1,1}_{4(r+1),2,2(r+1)}$. When $r = 1$, ϱ is an affine function; hence there are $a, b \in \mathbb{R}$ such that $\varrho(x) = b + ax = b + a\varrho_1(x) - a\varrho_1(-x)$ for all x, showing that $\varrho \in \mathtt{SNN}^{\varrho_1,1,1}_{4,2,2} = \mathtt{SNN}^{\varrho_1,1,1}_{2(n+1),2,n+1}$.

Assuming the result true for $n \in \mathbb{N}$, we prove it for $n + 1$. Consider ϱ made of $n + 1$ polynomial pieces: $\mathbb{R}$ is the disjoint union of $n + 1$ intervals I_i, $0 \le i \le n$ and there are polynomials p_i such that $\varrho(x) = p_i(x)$ on the interval I_i for $0 \le i \le n$. Without loss of generality order the intervals by increasing "position" and define $\bar\varrho(x) = \varrho(x)$ for $x \in \cup_{i=0}^{n-1} I_i = \mathbb{R} \setminus I_n$, and $\bar\varrho(x) = p_{n-1}(x)$ on I_n. It is not hard to see that $\bar\varrho$ is continuous and made of n polynomial pieces, the last one being $p_{n-1}(x)$ on $I_{n-1} \cup I_n$. Observe that $\varrho(x) = \bar\varrho(x) + f(x - t_n)$ where $\{t_n\} = \overline{I_{n-1}} \cap \overline{I_n}$ is the breakpoint between the intervals I_{n-1} and I_n, and

$$f(x) := \varrho(x + t_n) - \bar\varrho(x + t_n) = \begin{cases} 0 & \text{for } x < 0 \\ p_n(x + t_n) - p_{n-1}(x + t_n) & \text{for } x \ge 0. \end{cases}$$

Note that $q(x) := p_n(x + t_n) - p_{n-1}(x + t_n)$ satisfies $q(0) = f(0) = 0$, since ϱ is continuous. Because q is a polynomial of degree at most r, there are $a_i \in \mathbb{R}$ such that $q(x) = \sum_{i=1}^r a_i\, x^i$. This shows that $f = \sum_{i=1}^r a_i \varrho_i$. In case of $r = 1$, this shows that $f \in \mathtt{SNN}^{\varrho_1,1,1}_{2,2,1}$. For $r \ge 2$, since $\varrho_i \in \mathtt{NN}^{\varrho_i,1,1}_{2,2,1}$, Corollary C.3 yields $\varrho_i \in \overline{\mathtt{NN}^{\varrho_r,1,1}_{2\cdot 4^{r-i},2,2^{r-i}}}$, where the closure is with respect to the topology of locally uniform convergence. Observing that $2\sum_{i=1}^r 4^{r-i} = 2 \cdot (4^r - 1)/3 = w$ and $\sum_{i=1}^r 2^{r-i} = 2^r - 1 = m$, Lemma 2.17-(3) implies that[8] $f \in \overline{\mathtt{NN}^{\varrho_r,1,1}_{w,2,m}}$. Since $P : \mathbb{R} \to \mathbb{R}, x \mapsto x + t_n$ is affine with $\|P\|_{\ell^{0,\infty}} = \|P\|_{\ell^{0,\infty}_*} = 1$, by the induction hypothesis, Lemma 2.18-(1) and Lemma 2.17-(3) again, we get

$$\varrho(\bullet) = \bar\varrho(\bullet) + f(\cdot - t_n) \in \overline{\mathtt{NN}^{\varrho_r,1,1}_{4(r+1)+(n-1)w+w,2,2(r+1)+(n-1)m+m}} = \overline{\mathtt{NN}^{\varrho_r,1,1}_{4(r+1)+(n+1-1)w,2,2(r+1)+(n+1-1)m}}.$$

For $r = 1$, it is not hard to see $\varrho \in \mathtt{SNN}^{\varrho_1,1,1}_{2(n+1)+2,2,n+1+1} = \mathtt{SNN}^{\varrho_1,1,1}_{2((n+1)+1),2,(n+1)+1}$. □

C.2. **Proof of Lemma 4.10.** First we show that if $s \in \mathbb{N}$ and if $\varrho \in \mathtt{Spline}^s$ is not a polynomial, then there are $\alpha, \beta, t_0 \in \mathbb{R}$, $\varepsilon > 0$ and p a polynomial of degree at most $s - 1$ such that

$$\varrho_s(z) = \alpha\varrho(t_0 + z) + \beta\varrho(t_0 - z) - p(z) \quad \forall z \in [-\varepsilon, +\varepsilon]. \tag{C.2}$$

Consider any $t_0 \in \mathbb{R}$. Since $\varrho \in \mathtt{Spline}^s$, there are $\varepsilon > 0$ and two polynomials p_-, p_+ of degree at most s, with matching $s - 1$ first derivatives at t_0, such that

$$\varrho(x) = \begin{cases} p_+(x) & \text{for } x \in [t_0, t_0 + \varepsilon] \\ p_-(x) & \text{for } x \in [t_0 - \varepsilon, t_0]. \end{cases}$$

Since ϱ is not a polynomial, there is t_0 such that the s-th derivatives of $p_\pm$ at t_0 do not match, i.e. $a_- := p_-^{(s)}(t_0)/s! \ne p_+^{(s)}(t_0)/s! =: a_+$. A Taylor expansion yields

$$\varrho(t_0 + z) = \begin{cases} q(z) + a_+ z^s & \text{for } z \in [0, \varepsilon] \\ q(z) + a_- z^s & \text{for } z \in [-\varepsilon, 0], \end{cases}$$

[8] This implicitly uses that ϱ_i is *not* affine-linear, so that $\varrho_i \in \overline{\mathtt{NN}^{\varrho_r,1,1}_{2\cdot 4^{r-i},2,2^{r-i}} \setminus \mathtt{NN}^{\varrho_r,1,1}_{\infty,1,\infty}}$.

where $q(z) := \sum_{n=0}^{s-1} p_{\pm}^{(n)}(t_0) z^n/n!$ is a polynomial of degree at most $s-1$. As a result, for $|z| \le \varepsilon$

$$a_+ \cdot [\varrho(t_0+z) - q(z)] - (-1)^s a_- \cdot [\varrho(t_0-z) - q(-z)] = \begin{cases} (a_+^2 - a_-^2) \cdot z^s & \text{for } z \in [0,\varepsilon] \\ 0 & \text{for } z \in [-\varepsilon, 0] \end{cases} = (a_+^2 - a_-^2) \cdot \varrho_s(z).$$

Since $a_+ \neq a_-$, setting $\alpha := a_+/(a_+^2 - a_-^2)$ and $\beta := (-1)^{s+1} a_-/(a_+^2 - a_-^2)$, as well as $p(x) := \alpha q(z) + \beta q(-z)$ we get as claimed $\varrho_s(z) = \alpha\varrho(z+t_0) + \beta\varrho(-z+t_0) - p(z)$ for every $|z| \le \varepsilon$.

Now consider $r \in \mathbb{N}$. Given $R > 0$ we now set

$$f_R(x) := (\tfrac{R}{\varepsilon})^r \left[\alpha\varrho(\varepsilon x/R + t_0) + \beta\varrho(-\varepsilon x/R + t_0) - p(\varepsilon x/R)\right]$$

with $\alpha, \beta, t_0, \varepsilon, p$ from (C.2). Observe that $\varrho_r(x) = (R/\varepsilon)^r \varrho_r(\varepsilon x/R) = f_R(x)$ for all $x \in [-R, R]$, so that f_R converges locally uniformly to ϱ_r on $\mathbb{R}$.

We show by induction on $r \in \mathbb{N}$ that $f_R \in \mathtt{NN}_{w,2,m}^{\varrho,1,1}$ where $w = w(r), m = m(r) \in \mathbb{N}$ only depend on r. For $r = 1$, this trivially holds as the polynomial p in (C.2) is a constant; hence $f_R \in \mathtt{NN}_{4,2,2}^{\varrho,1,1}$.

Assuming the result true for some $r \in \mathbb{N}$ we now prove it for $r+1$. Consider $\varrho \in \mathtt{Spline}^{r+1}$ that is not a polynomial. The polynomial p in (C.2) with $s = r+1$ is of degree at most r; hence by Lemma 2.24 there are $c, a_i, b_i, c_i \in \mathbb{R}$ such that $p(x) = c + \sum_{i=1}^{r+1} a_i\, \varrho_r(b_i x + c_i)$ for all $x \in \mathbb{R}$. Now, observe that since $\varrho \in \mathtt{Spline}^{r+1}$ is not a polynomial, its derivative satisfies $\varrho' \in \mathtt{Spline}^r$ and is not a polynomial either. The induction hypothesis yields $\varrho_r \in \overline{\mathtt{NN}_{w,2,m}^{\varrho',1,1}}$ for $w = w(r), m = m(r) \in \mathbb{N}$. It is not hard to check that this implies $p \in \overline{\mathtt{NN}_{2(r+1)w,2,(r+1)m}^{\varrho',1,1}}$. Finally, as $\varrho'(x)$ is the locally uniform limit of $(\varrho(x+h) - \varrho(x))/h$ as $h \to 0$ (see Lemma C.1), we obtain $p \in \overline{\mathtt{NN}_{4(r+1)w,2,2(r+1)m}^{\varrho,1,1}}$ thanks to Lemma 2.21. Combined with the definition of f_R we obtain $f_R \in \overline{\mathtt{NN}_{4(r+1)w+4,2,2(r+1)m+2}^{\varrho,1,1}}$.

Finally we quantify w, m: First of all, note that $w(1) = 4 \le 5$ and $m(1) = 2 \le 3$; furthermore, $w(r+1) \le 4(r+1)w(r) + 4 \le 5(r+1)w(r)$ and $m(r+1) \le 2(r+1)m + 2 \le 3(r+1)m$. An induction therefore yields $w(r) \le 5^r r!$ and $m(r) \le 3^r r!$. □

C.3. **Proof of Lemma 4.11. Step 1:** In this step, we construct $\theta_{R,\delta} \in \mathtt{NN}_{w,\ell,m}^{\varrho_r,d,1}$ satisfying

$$|\theta_{R,\delta}(x) - \mathbb{1}_{[-R,R]^d}(x)| \le 2 \cdot \mathbb{1}_{[-R-\delta,R+\delta]^d \setminus [-R,R]^d} \qquad \forall\, x \in \mathbb{R}^d\,, \tag{C.3}$$

with $\ell = 3$ (resp. $\ell = 2$ if $d = 1$) and with w, m only depending on d and r.

The affine map $P : \mathbb{R}^d \to \mathbb{R}^d, x = (x_i)_{i=1}^d \mapsto \left(\frac{x_i}{2(R+\delta)} + \frac{1}{2}\right)_{i=1}^d$ satisfies $\|P\|_{\ell^{0,\infty}} = \|P\|_{\ell_*^{0,\infty}} = 1$. For $x \in \mathbb{R}^d$, we have $x \in [-R-\delta, R+\delta]^d$ if and only if $P(x) \in [0,1]^d$, and $x \in [-R,R]^d$ if and only if $P(x) \in [\varepsilon, 1-\varepsilon]^d$, where $\varepsilon := \frac{2\delta}{2(R+\delta)}$; thus, $\mathbb{1}_{[-R,R]^d}(P^{-1}x) = \mathbb{1}_{[\varepsilon,1-\varepsilon]^d}(x)$ for all $x \in \mathbb{R}^d$.

Next, by combining Lemmas 4.4 and 3.26 (see in particular Equation (3.10)), we obtain $f \in \mathtt{NN}_{w,\ell,m}^{\varrho_r,d,1}$ (with the above mentioned properties for w, ℓ, m and $m \ge d$) such that $|f(x) - \mathbb{1}_{[0,1]^d}(x)| \le \mathbb{1}_{[0,1]^d \setminus [\varepsilon,1-\varepsilon]^d}$ for all $x \in \mathbb{R}^d$. Therefore, the function $\theta_{R,\delta} := f \circ P$ satisfies

$$\begin{aligned} |\theta_{R,\delta}(x) - \mathbb{1}_{[-R,R]^d}(x)| &= |f(Px) - \mathbb{1}_{[-R,R]^d}(P^{-1}Px)| \\ &\le |f(Px) - \mathbb{1}_{[0,1]^d}(Px)| + |\mathbb{1}_{[0,1]^d}(Px) - \mathbb{1}_{[\varepsilon,1-\varepsilon]^d}(Px)| \\ &\le 2 \cdot \mathbb{1}_{[-R-\delta,R+\delta]^d \setminus [-R,R]^d}(x) \end{aligned}$$

for all $x \in \mathbb{R}^d$. Finally, by Lemma 2.18-(1), we have $\theta_{R,\delta} \in \mathtt{NN}_{w,\ell,m}^{\varrho_r,d,1}$.

Step 2: Consider $g \in \mathtt{NN}_{W,L,N}^{\varrho_r,d,k}$ and define $g_{R,\delta}(x) := \theta_{R,\delta}(x) \cdot g(x)$ for all $x \in \mathbb{R}^d$. The desired estimate (4.6) is an easy consequence of (C.3). It only remains to show that one can implement $g_{R,\delta}$ with a ϱ_r-network of controlled complexity.

Since we assume $W \ge 1$ we can use Lemma 2.14; combining it with Equation (2.1) we get $g \in \mathtt{NN}_{W,L',N'}^{\varrho_r,d,k}$ with $L' = \min\{L, W, N+1\}$ and $N' = \min\{N, W\}$. Lemma 2.17-(2) yields $(\theta_{R,\delta}, g) \in \mathtt{NN}_{w',L'',m'}^{\varrho_r,d,k+1}$ with $L'' = \max\{L', \ell\}$ as well as $w' = W + w + \min\{d,k\} \cdot (L''-1)$ and $m' = N' + m + \min\{d,k\} \cdot (L''-1)$. Since $L''-1 = \max\{L'-1, \ell-1\} \le \max\{W-1, \ell-1\} \le W + \ell - 2$ and $N' \le W$, we get

$$\begin{aligned} w' &\le W + w + \min\{d,k\} \cdot (W + \ell - 2) = W \cdot (1 + \min\{d,k\}) + c_1 \\ m' &\le W + m + \min\{d,k\} \cdot (W + \ell - 2) = W \cdot (1 + \min\{d,k\}) + c_2. \end{aligned}$$

where c_1, c_2 only depend on d, k, r.

As $r \ge 2$, Lemma 2.24 shows that ϱ_r can represent any polynomial of degree two with $n = 2(r+1)$ terms. Thus, Lemma 2.26 shows that the multiplication map $m : \mathbb{R} \times \mathbb{R}^k \to \mathbb{R}^k, (x,y) \mapsto x \cdot y$ satisfies

$m \in \mathtt{NN}^{\varrho_r,1+k,k}_{12k(r+1),2,4k(r+1)}$. Finally, Lemma 2.18-(3) proves that $g_{R,\delta} = m \circ (\theta_{R,\delta}, g) \in \mathtt{NN}^{\varrho_r,d,k}_{w'',L''',m''}$, where $L''' = L'' + 1$ and $m'' = m' + 4k(r+1) = N' + m + \min\{d,k\} \cdot (L''-1) + 4k(r+1)$ as well as $w'' = w' + \max\{m', d\} \cdot 12k(r+1)$.

As $L'' = \max\{L', \ell\} \leq \max\{L, \ell\}$ we have $L''' \leq \max\{L+1, 4\}$ (respectively $L''' \leq \max\{L+1, 3\}$ if $d = 1$). Furthermore, since $m' \geq m \geq d$ we have $\max\{m', d\} = m'$. Because of $W \geq 1$, we thus see that

$$w'' = w' + m' \cdot 12k(r+1) \leq W \cdot (1 + \min\{d,k\}) \cdot (1 + 12k(r+1)) + c_3 \leq c_4 W$$

where c_3, c_4 only depend on d, k, r. Finally, $L'' - 1 = \max\{L'-1, \ell-1\} \leq \max\{N, \ell-1\} \leq N + \ell - 1$. Since $N' \leq N$, we get $m'' \leq N \cdot (1 + \min\{d,k\}) + c_5 \leq c_6 N$ where again c_5, c_6 only depend on d, k, r. To conclude, we set $c := \max\{c_4, c_6\}$. □

C.4. **Proof of Proposition 4.12.** When $r = 1$ and $\varrho \in \mathtt{NN}^{\varrho_r,1,1}_{\infty,2,m}$ the result follows from Lemma 2.19.

Now, consider $f \in \mathtt{NN}^{\varrho,d,k}_{W,L,N}$ such that $f|_\Omega \in X$. Since $\varrho \in \overline{\mathtt{NN}^{\varrho_r,1,1}_{\infty,2,m}}$, Lemma 2.21 shows that

$$\mathtt{NN}^{\varrho,d,k}_{W,L,N} \subset \overline{\mathtt{NN}^{\varrho_r,d,k}_{Wm^2,L,Nm}}, \quad \text{with closure in the topology of locally uniform convergence on } \mathbb{R}^d. \tag{C.4}$$

For bounded Ω, locally uniform convergence implies convergence in $X = X^k_p(\Omega)$ for all $p \in (0,\infty]$ hence the result.

For unbounded Ω we need to work a bit harder. First we deal with the degenerate case where $W = 0$ or $N = 0$. If $W = 0$ then by Lemma 2.13 f is a constant map; hence $f \in \mathtt{NN}^{\varrho_r,d,k}_{0,1,0}$. If $N = 0$ then f is affine-linear with $\|f\|_{\ell^0} \leq W$; hence $f \in \mathtt{NN}^{\varrho_r,d,k}_{W,1,0}$. In both cases the result trivially holds.

From now on we assume that $W, N \geq 1$. Consider $\varepsilon > 0$. By the dominated convergence theorem (in case of $p < \infty$) or our special choice of $X^k_\infty(\Omega)$ (cf. Equation (1.3)) (in case of $p = \infty$) we see that there is some $R \geq 1$ such that

$$\|f - f \cdot \mathbb{1}_{[-R,R]^d}\|_{L_p(\Omega;\mathbb{R}^k)} \leq \varepsilon' := \frac{\varepsilon}{8^{1/\min\{1,p\}}}.$$

Denoting by $\lambda(\cdot)$ the Lebesgue measure, (C.4) implies that there is $g \in \mathtt{NN}^{\varrho_r,d,k}_{Wm^2,L,Nm}$ such that

$$\|f - g\|_{L_\infty([-R-1,R+1]^d;\mathbb{R}^k)} \leq \varepsilon' \big/ \left[\lambda([-R-1,R+1]^d)\right]^{1/p}.$$

Consider $c = c(d,k,r)$, $\ell = \min\{d+1, 3\}$, $L' = \max\{L+1, \ell\}$ and the function $g_{R,1} \in \mathtt{NN}^{\varrho_r,d,k}_{cWm^2,L',cNm}$ from Lemma 4.11. By (4.6) and the fact that $\|\cdot\|^{\min\{1,p\}}_{L_p}$ is subadditive, we see

$$\begin{aligned}
\|f - g_{R,1}\|^{\min\{1,p\}}_{L_p(\Omega;\mathbb{R}^k)} &\leq \|f - f\cdot\mathbb{1}_{[-R,R]^d}\|^{\min\{1,p\}}_{L_p(\Omega;\mathbb{R}^k)} + \|(f-g)\mathbb{1}_{[-R,R]^d}\|^{\min\{1,p\}}_{L_p(\Omega;\mathbb{R}^k)} \\
&\quad + \|g\cdot\mathbb{1}_{[-R,R]^d}(x) - g_{R,1}\|^{\min\{1,p\}}_{L_p(\Omega;\mathbb{R}^k)} \\
&\leq \frac{\varepsilon^{\min\{1,p\}}}{8} + \Big(\|f-g\|_{L_\infty([-R-1,R+1]^d;\mathbb{R}^k)} \cdot [\lambda([-R,R]^d)]^{1/p}\Big)^{\min\{1,p\}} \\
&\quad + \big(\|\,2\cdot|g|\cdot\mathbb{1}_{[-R-1,R+1]^d\setminus[-R,R]^d}\|_{L_p(\Omega)}\big)^{\min\{1,p\}}, \\
&\leq \frac{\varepsilon^{\min\{1,p\}}}{2} + \big(\|\,2\cdot|g|\cdot\mathbb{1}_{[-R-1,R+1]^d\setminus[-R,R]^d}\|_{L_p(\Omega)}\big)^{\min\{1,p\}}.
\end{aligned}$$

To estimate the final term, note that

$$\begin{aligned}
&\Big(\|\,|g|\cdot\mathbb{1}_{[-R-1,R+1]^d\setminus[-R,R]^d}\|_{L_p(\Omega)}\Big)^{\min\{1,p\}} \\
&\leq \Big(\|\,|g-f|\cdot\mathbb{1}_{[-R-1,R+1]^d\setminus[-R,R]^d}\|_{L_p(\Omega)}\Big)^{\min\{1,p\}} + \Big(\|\,|f|\cdot\mathbb{1}_{[-R-1,R+1]^d\setminus[-R,R]^d}\|_{L_p(\Omega)}\Big)^{\min\{1,p\}} \\
&\leq \Big(\|f-g\|_{L_\infty([-R-1,R+1]^d;\mathbb{R}^k)} \cdot [\lambda([-R-1,R+1]^d)]^{1/p}\Big)^{\min\{1,p\}} + \Big(\|f - f\cdot\mathbb{1}_{[-R,R]^d}\|_{L_p(\Omega;\mathbb{R}^k)}\Big)^{\min\{1,p\}} \\
&\leq \frac{\varepsilon^{\min\{1,p\}}}{8} + \frac{\varepsilon^{\min\{1,p\}}}{8}.
\end{aligned}$$

Because of $2^{\min\{1,p\}} \leq 2$, this implies $\Big(\|\,2\cdot|g|\cdot\mathbb{1}_{[-R-1,R+1]^d\setminus[-R,R]^d}\|_{L_p(\Omega)}\Big)^{\min\{1,p\}} \leq \frac{\varepsilon^{\min\{1,p\}}}{2}$. Overall, we thus see that $\|f - g_{R,1}\|_{L_p(\Omega;\mathbb{R}^k)} \leq \varepsilon < \infty$. Because of $f|_\Omega \in X$, this implies in particular that $g_{R,1}|_\Omega \in X$. Since $\varepsilon > 0$ was arbitrary, we get as desired that $f|_\Omega \in \overline{\mathtt{NN}^{\varrho_r,d,k}_{cWm^2,L',cNm} \cap X}^X$, where the closure is taken in X. □

Appendix D. Proofs for Section 5

D.1. **Proof of Lemma 5.2.** In light of (4.1) we have $\beta_+^{(t)} \in \mathtt{NN}^{\varrho_t,1,1}_{2(t+2),2,t+2}$. This yields the result for $d=1$, including when $t=1$.

For $d \geq 2$ and $t \geq \min\{d,2\} = 2$, define $f_j : \mathbb{R}^d \to \mathbb{R}$ by $f_j := \beta_+^{(t)} \circ \pi_j$ with $\pi_j : \mathbb{R}^d \to \mathbb{R}, x \mapsto x_j$, $j = 1,\dots,d$. By Lemma 2.18–(1) together with the fact that $\|\pi_j\|_{\ell_*^{0,\infty}} = 1$ we get $f_j \in \mathtt{NN}^{\varrho_t,d,1}_{2(t+2),2,t+2}$. Form the vector function $f := (f_1, f_2, \dots, f_d)$. Using Lemma 2.17-(2), we deduce $f \in \mathtt{NN}^{\varrho_t,d,d}_{2d(t+2),2,d(t+2)}$.

As $t \geq 2$, by Lemma 2.24, ϱ_t can represent any polynomial of degree two with $n = 2(t+1)$ terms. Hence, for $d \geq 2$, by Lemma 2.26 the multiplication function $M_d : \mathbb{R}^d \to \mathbb{R}, (x_1,\dots,x_d) \mapsto x_1 \cdots x_d$ satisfies $M_d \in \mathtt{NN}^{\varrho_t,d,1}_{4n(2^j-1),2j,(2n+1)(2^j-1)-1}$ with $j := \lceil \log_2 d \rceil$. By definition, $2^{j-1} < d \leq 2^j$, hence $2^j - 1 \leq 2(d-1)$ and $6n(2^j-1) \leq 12n(d-1) = 24(t+1)(d-1)$, as well as

$$(2n+1)(2^j-1) - 1 \leq (4n+2)(d-1) - 1 = (8t+10)(d-1) - 1,$$

so that $M_d \in \mathtt{NN}^{\varrho_t,d,1}_{24(t+1)(d-1),2j,(8t+10)(d-1)-1}$. As $\beta_d^{(t)} = M_d \circ f$, by Lemma 2.18–(2) we get

$$\beta_d^{(t)} \in \mathtt{NN}^{\varrho_t,d,1}_{2d(t+2)+24(t+1)(d-1),2j+2,d(t+2)+(8t+10)(d-1)-1+d}.$$

To conclude, we observe that

$$\begin{aligned} 2d(t+2) + 24(t+1)(d-1) &\leq d(2t+4+24t+24) = d(26t+28) \leq 28d(t+1) \\ d(t+2) + (8t+10)(d-1) - 1 + d &\leq d(t+2+8t+10+1) = d(9t+13) \leq 13d(t+1). \end{aligned}$$

□

D.2. **Proof of Theorem 5.5.** We divide the proof into three steps.

Step 1 (Recalling results from [19]): Using the tensor B-splines $\beta_d^{(t)}$ introduced in Equation (5.5), define $N := N^{(\tau)} := \beta_d^{(\tau-1)}$ for $\tau \in \mathbb{N}$, and note that this coincides with the definition of N in [19, Equation (4.1)]. Next, as in [19, Equations (4.2) and (4.3)], for $k \in \mathbb{N}_0$ and $j \in \mathbb{Z}^d$, define $N_k^{(\tau)}(x) := N^{(\tau)}(2^k x)$ and $N_{j,k}^{(\tau)}(x) := N^{(\tau)}(2^k x - j)$. Furthermore, let $\Omega_0 := (-\frac{1}{2}, \frac{1}{2})^d$ denote the unit cube, and set

$$\Lambda^{(\tau)}(k) := \big\{ j \in \mathbb{Z}^d \colon N_{j,k}^{(\tau)}|_{\Omega_0} \not\equiv 0 \big\} \quad \text{and} \quad \Sigma_k^{(\tau)} := \operatorname{span}\big\{ N_{j,k}^{(\tau)} \colon j \in \Lambda(k) \big\},$$

and finally $s_k^{(\tau)}(f)_p := \inf_{g \in \Sigma_k^{(\tau)}} \|f-g\|_{L_p}$ for $f \in X_p(\Omega_0)$ and $k \in \mathbb{N}_0$. Setting $\lambda^{(\tau,p)} := \tau - 1 + \min\{1, p^{-1}\}$, [19, Theorem 5.1] shows

$$\|f\|_{B^\alpha_{p,q}(\Omega_0)} \asymp \|f\|_{L_p} + \big\| \big(s_k^{(\tau)}(f)_p\big)_{k\in\mathbb{N}_0} \big\|_{\ell_q^\alpha} \quad \forall\, p,q \in (0,\infty], \alpha \in (0, \lambda^{(\tau,p)}), \text{ and } f \in B^\alpha_{p,q}(\Omega_0). \tag{D.1}$$

Here, $\|(c_k)_{k\in\mathbb{N}_0}\|_{\ell_q^\alpha} = \|(2^{\alpha k} c_k)_{k\in\mathbb{N}_0}\|_{\ell^q}$; see [19, Equation (5.1)].

Step 2 (Proving the embedding $B^{d\alpha}_{p,q}(\Omega_0) \hookrightarrow A_q^\alpha(X_p(\Omega_0), \Sigma(\mathcal{D}_d^{\tau-1}))$): Define $\Sigma(\mathcal{D}_d^t) := (\Sigma_n(\mathcal{D}_d^t))_{n\in\mathbb{N}_0}$. In this step, we show that $B^{d\alpha}_{p,q}(\Omega_0) \hookrightarrow A_q^\alpha(X_p(\Omega_0), \Sigma(\mathcal{D}_d^{\tau-1}))$ for any $\tau \in \mathbb{N}$ and all $p,q \in (0,\infty]$ and $\alpha > 0$ with $0 < d\alpha < \lambda^{(\tau,p)}$.

To this end, we first show that if we choose $X = X_p(\Omega_0)$, then the family $\Sigma(\mathcal{D}_d^{\tau-1})$ satisfies the properties (P1)–(P5). To see this, we first have to show $\Sigma_n(\mathcal{D}_d^{\tau-1}) \subset X_p(\Omega_0)$. For $p < \infty$, this is trivial, since $N^{(\tau)} = \beta_d^{(\tau-1)}$ is bounded and measurable. For $p = \infty$ this holds as well, since if $\tau \geq 2$, then $N^{(\tau)} = \beta_d^{(\tau-1)}$ is continuous; finally, the case $\tau = 1$ cannot occur for $p = \infty$, since this would imply

$$0 < d\alpha < \lambda^{(\tau,p)} = \tau - 1 + \min\{1, p^{-1}\} = 0.$$

Next, Properties (P1)–(P4) are trivially satisfied. Finally, the density of $\bigcup_{n=0}^\infty \Sigma_n(\mathcal{D}_d^{\tau-1})$ in $X_p(\Omega_0)$ is well-known for $\tau = 1$, since then $\beta_0^{(\tau-1)} = \mathbb{1}_{[0,1)^d}$ and $p < \infty$. For $\tau \geq 2$, the density follows with the same arguments that were used for the case $p = \infty$ in Section B.4.3.

Next, note that $\operatorname{supp} N^{(\tau)} \subset [0,\tau]^d$ and thus $\operatorname{supp} N_{j,k}^{(\tau)} \subset 2^{-k}(j + [0,\tau]^d)$. Therefore, if $j \in \Lambda^{(\tau)}(k)$, then $\varnothing \neq \Omega_0 \cap \operatorname{supp} N_{j,k}$, so that there is some $x \in \Omega_0 \cap 2^{-k}(j + [0,\tau]^d)$. This implies $j \in \mathbb{Z}^d \cap [-2^{k-1} - \tau, 2^{k-1}]^d$, and thus $|\Lambda(k)| \leq (2^k + \tau + 1)^d$. Directly by definition of $\Sigma_n(\mathcal{D}_d^t)$ and $\Sigma_k^{(\tau)}$, this implies

$$\Sigma_k^{(\tau)} \subset \Sigma_{(2^k+\tau+1)^d}(\mathcal{D}_d^{\tau-1}) \qquad \forall\, k \in \mathbb{N}_0. \tag{D.2}$$

Next, since we are assuming $0 < \alpha d < \lambda^{(\tau,p)}$, Equation (D.1) yields a constant $C_1 = C_1(p,q,\alpha,\tau,d) > 0$ such that $\|f\|_{L_p} + \big\| \big(s_k^{(\tau)}(f)_p\big)_{k\in\mathbb{N}_0} \big\|_{\ell_q^{d\alpha}} \leq C_1 \cdot \|f\|_{B^{d\alpha}_{p,q}(\Omega_0)}$ for all $f \in B^{d\alpha}_{p,q}(\Omega_0)$. Therefore, we see for

$f \in B^{d\alpha}_{p,q}(\Omega_0)$ and $q < \infty$ that

$$\begin{aligned}
\|f\|^q_{A^\alpha_q(X_p(\Omega_0),\Sigma(\mathcal{D}^{\tau-1}_d))} &= \sum_{n=1}^{\infty} n^{-1} \cdot [n^\alpha \cdot E(f, \Sigma_{n-1}(\mathcal{D}^{\tau-1}_d))_{L_p(\Omega_0)}]^q \\
&\leq \|f\|^q_{L_p} \sum_{n=1}^{(\tau+2)^d} n^{\alpha q-1} + \sum_{k=0}^{\infty} \sum_{n=(2^k+\tau+1)^d+1}^{(2^{k+1}+\tau+1)^d} n^{\alpha q-1} [E(f, \Sigma_{n-1}(\mathcal{D}^{\tau-1}_d))_{L_p(\Omega_0)}]^q \\
&\overset{(*)}{\leq} C_2 \cdot \|f\|^q_{L_p} + C_4 \sum_{k=0}^{\infty} 2^{kd} 2^{dk(\alpha q-1)} [s^{(\tau)}_k(f)_p]^q \\
&\leq (C_2 + C_4) \cdot \big(\|f\|_{L_p} + \big\|\big(s^{(\tau)}_k(f)_p\big)_{k\in\mathbb{N}_0}\big\|_{\ell^{d\alpha}_q}\big)^q \leq C_1^q \cdot (C_2 + C_4) \cdot \|f\|^q_{B^{d\alpha}_{p,q}(\Omega_0)}.
\end{aligned}$$

At the step marked with $(*)$, we used that Equation (D.2) yields $\Sigma_{n-1}(\mathcal{D}^{\tau-1}_d) \supset \Sigma_{(2^k+\tau+1)^d}(\mathcal{D}^{\tau-1}_d) \supset \Sigma^{(\tau)}_k$ for all $n \geq 1 + (2^k + \tau + 1)^d$, and furthermore that if $1 + (2^k + \tau + 1)^d \leq n \leq (2^{k+1} + \tau + 1)^d$, then $2^{dk} \leq n \leq (\tau + 3)^d \cdot 2^{dk}$, so that $n^{\alpha q-1} \leq C_3 2^{dk(\alpha q-1)}$ for some constant $C_3 = C_3(d, \tau, \alpha, q)$, and finally that $\sum_{n=(2^k+\tau+1)^d+1}^{(2^{k+1}+\tau+1)^d} 1 \leq (2^{k+1} + \tau + 1)^d \leq (\tau + 3)^d \cdot 2^{dk}$.

For $q = \infty$, the proof is similar. Setting $\ell_k := (2^k + \tau + 1)^d$ for brevity, we see with similar estimates as above that

$$\begin{aligned}
&\|f\|_{A^\alpha_\infty(X_p(\Omega_0),\Sigma(\mathcal{D}^{\tau-1}_d))} \\
&= \max\Big\{ \max_{0\leq n\leq(\tau+2)^d} n^\alpha \, E(f, \Sigma_{n-1}(\mathcal{D}^{\tau-1}_d))_{L_p(\Omega_0)}, \quad \sup_{k\in\mathbb{N}_0} \max_{\ell_k+1\leq n\leq\ell_{k+1}} n^\alpha \, E(f, \Sigma_{n-1}(\mathcal{D}^{\tau-1}_d))_{L_p(\Omega_0)} \Big\} \\
&\leq \max\Big\{ (\tau+2)^{\alpha d} \, \|f\|_{L_p(\Omega_0)}, \quad \sup_{k\in\mathbb{N}_0} (\tau+3)^{\alpha d} \, 2^{\alpha dk} s^{(\tau)}_k(f)_p \Big\} \\
&\leq (\tau+3)^{\alpha d} \big(\|f\|_{L_p(\Omega_0)} + \|(s^{(\tau)}_k(f)_p)_{k\in\mathbb{N}_0}\|_{\ell^{d\alpha}_q}\big) \leq C_1 \, (\tau+3)^{\alpha d} \, \|f\|_{B^{d\alpha}_{p,\infty}(\Omega_0)}.
\end{aligned}$$

Overall, we have shown $B^{d\alpha}_{p,q}(\Omega_0) \hookrightarrow A^\alpha_q(X_p(\Omega_0), \Sigma(\mathcal{D}^{\tau-1}_d))$ for $\tau \in \mathbb{N}$, $p, q \in (0, \infty]$ and $0 < \alpha d < \lambda^{(\tau,p)}$.

Step 3 (Proving the embeddings (5.9) **and** (5.10)**):** In case of $d = 1$, let us set $r_0 := r$, while r_0 is as in the statement of the theorem for $d > 1$. Since Ω is bounded and $\Omega_0 = (-\frac{1}{2}, \frac{1}{2})^d$, there is some $R > 0$ such that $\Omega \subset R \cdot \Omega_0$. Let us fix $p, q \in (0, \infty]$ and $s > 0$ such that $ds < r_0 + \min\{1, p^{-1}\}$.

Since Ω and $R \cdot \Omega_0$ are bounded Lipschitz domains, there exists a (not necessarily linear) *extension operator* $\mathcal{E} : B^{ds}_{p,q}(\Omega) \to B^{ds}_{p,q}(R\Omega_0)$ with the properties $(\mathcal{E}f)|_\Omega = f$ and $\|\mathcal{E}f\|_{B^{ds}_{p,q}(R\Omega_0)} \leq C \cdot \|f\|_{B^{ds}_{p,q}(\Omega)}$ for all $f \in B^{ds}_{p,q}(\Omega)$. Indeed, for $p \in [1, \infty]$ this follows from [37, Section 4, Corollary 1], since this corollary yields an extension operator $\mathcal{E} : X_p(\Omega) \to X_p(R\Omega_0)$ with the additional property that the j-th modulus of continuity ω_j satisfies $\omega_j(t, \mathcal{E}f)_{R\Omega_0} \leq M_j \cdot \omega_j(t, f)_\Omega$ for all $j \in \mathbb{N}$, all $f \in X_p(\Omega)$, and all $t \in [0, 1]$. In view of the definition of the Besov spaces (see in particular [21, Chapter 2, Theorem 10.1]), this easily implies the result. Finally, in case of $p \in (0, 1)$, the existence of the extension operator follows from [20, Theorem 6.1]. In addition to the existence of the extension operator, we will also need that the dilation operator $D_1 : B^{ds}_{p,q}(R\Omega_0) \to B^{ds}_{p,q}(\Omega_0), f \mapsto f(R\bullet)$ is well-defined and bounded, say $\|D_1\| \leq C_1$; this follows directly from the definition of the Besov spaces.

We first prove Equation (5.9), that is, we consider the case $d = 1$. To this end, define $\tau := r + 1 \in \mathbb{N}$, let $f \in B^s_{p,q}(\Omega)$ be arbitrary, and set $f_1 := D_1(\mathcal{E}f) \in B^s_{p,q}(\Omega_0)$. By applying Step 2 with $\alpha = s$ (and noting that $0 < d\alpha = s < r + \min\{1, p^{-1}\} = \lambda^{(\tau,p)}$), we get $f_1 \in A^s_q(X_p(\Omega_0), \Sigma(\mathcal{D}^r_d))$, with $\|f_1\|_{A^s_q(X_p(\Omega_0),\Sigma(\mathcal{D}^r_d))} \leq CC_1C_2 \cdot \|f\|_{B^{ds}_{p,q}(\Omega)}$, where the constant C_2 is provided by Step 2.

Next, we note that $L := \sup_{n\in\mathbb{N}} \mathscr{L}(n) \geq 2 = 2 + 2\lceil \log_2 d \rceil$ and $r \geq 1 = \min\{d, 2\}$, so that Corollary 5.4-(2) shows $f_1 \in A^s_q(X_p(\Omega_0), \Sigma(\mathcal{D}^r_d)) \hookrightarrow W^s_q(X_p(\Omega_0), \varrho_r, \mathscr{L})$. But it is an easy consequence of Lemma 2.18-(1) that the dilation operator $D_2 : W^s_q(X_p(\Omega_0), \varrho_r, \mathscr{L}) \to W^s_q(X_p(R\Omega_0), \varrho_r, \mathscr{L}), g \mapsto g(\bullet/R)$ is well-defined and bounded. Hence, we see that $D_2 f_1 \in W^s_q(X_p(R\Omega_0), \varrho_r, \mathscr{L})$ with $\|D_2 f_1\|_{W^s_q(X_p(R\Omega_0),\varrho_r,\mathscr{L})} \lesssim \|f\|_{B^{ds}_{p,q}(\Omega)}$. Now, note $D_2 f_1(x) = f_1(x/R) = \mathcal{E}f(x) = f(x)$ for all $x \in \Omega \subset R\Omega_0$, and hence $f = (D_2 f_1)|_\Omega$. Thus, Remark 3.17 implies that $f \in W^s_q(X_p(\Omega), \varrho_r, \mathscr{L})$ with $\|f\|_{W^s_q(X_p(\Omega),\varrho_r,\mathscr{L})} \lesssim \|f\|_{B^{ds}_{p,q}(\Omega)}$, as claimed.

Now, we prove Equation (5.10). To this end, define $\tau := r_0 + 1 \in \mathbb{N}$, let $f \in B^{sd}_{p,q}(\Omega)$ be arbitrary, and set $f_1 := D_1(\mathcal{E}f) \in B^{ds}_{p,q}(\Omega_0)$. Applying Step 2 with $\alpha = s$ (noting $0 < d\alpha = ds < r_0 + \min\{1, p^{-1}\} = \lambda^{(\tau,p)}$), we get $f_1 \in A^s_q(X_p(\Omega_0), \Sigma(\mathcal{D}^{r_0}_d))$, with $\|f_1\|_{A^s_q(X_p(\Omega_0),\Sigma(\mathcal{D}^{r_0}_d))} \leq CC_1C_2 \cdot \|f\|_{B^{ds}_{p,q}(\Omega)}$, where the constant C_2 is provided by Step 2.

Next, we claim that $A^s_q(X_p(\Omega_0), \Sigma(\mathcal{D}^{r_0}_d)) \hookrightarrow W^s_q(X_p(\Omega_0), \varrho_r, \mathscr{L})$. Indeed, if $r \geq 2$ and $L \geq 2 + 2\lceil \log_2 d \rceil$, then this follows from Corollary 5.4–(2). Otherwise, we have $r_0 = 0$ and $L \geq 3 \geq \min\{d+1, 3\}$, so that the claim follows from Corollary 5.4–(1); here, we note that $p < \infty$, since we would otherwise get the contradiction $0 < \alpha d < r_0 + \min\{1, p^{-1}\} = 0$. Therefore, $f_1 \in W^s_q(X_p(\Omega_0), \varrho_r, \mathscr{L})$ with $\|f_1\|_{W^s_q(X_p(\Omega_0), \varrho_r, \mathscr{L})} \lesssim \|f\|_{B^{ds}_{p,q}(\Omega)}$. The rest of the argument is exactly as in the case $d = 1$. □

D.3. **Proof of Lemma 5.10.** Lemma 5.10 shows that deeper networks can implement the sawtooth function Δ_j using less connections/neurons than more shallow networks. The reason for this is indicated by the following lemma.

Lemma D.1. *For arbitrary $j \in \mathbb{N}$, we have $\Delta_j \circ \Delta_1 = \Delta_{j+1}$.* ◀

Proof. It suffices to verify the identity on $[0,1]$, since if $x \in \mathbb{R} \setminus [0,1]$, then $\Delta_1(x) = 0 = \Delta_{j+1}(x)$, so that $\Delta_j(\Delta_1(x)) = \Delta_j(0) = 0 = \Delta_{j+1}(x)$. We now distinguish two cases for $x \in [0,1]$.

Case 1: $x \in [0, \frac{1}{2}]$. This implies $\Delta_1(x) = 2x$, and hence (recall the definition of Δ_j in Equation (5.11))

$$\Delta_j(\Delta_1(x)) = \sum_{k=0}^{2^{j-1}-1} \Delta_1\big(2^{j-1}2x - k\big) = \sum_{k=0}^{2^{j-1}-1} \Delta_1(2^{(j+1)-1}x - k) = \Delta_{j+1}(x).$$

In the last equality we used that $2^j x - k \leq 2^{j-1} - k \leq 0$ for $k \geq 2^{j-1}$, so that $\Delta_1(2^j x - k) = 0$ for those k.

Case 2: $x \in (\frac{1}{2}, 1]$. Observe that $\Delta_j(x) = \Delta_j(1-x)$ for all $x \in \mathbb{R}$ and $j \in \mathbb{N}$. Since $x' := 1 - x \in [0, 1/2]$, this identity and Case 1 yield $\Delta_j \circ \Delta_1(x) = \Delta_j \circ \Delta_1(1-x) = \Delta_{j+1}(1-x) = \Delta_{j+1}(x)$. □

Using Lemma D.1, we can now provide the proof of Lemma 5.10.

Proof of Lemma 5.10. Part (1): Write $j = k(L-1) + s$ for suitable $k \in \mathbb{N}_0$ and $0 \leq s \leq L-2$. Note that this implies $k \leq j/(L-1)$. Thanks to Lemma D.1, we have $\Delta_j = \Delta_{k+s} \circ \Delta_k \circ \cdots \circ \Delta_k$, where Δ_k occurs $L-2$ times. Furthermore, since $\Delta_k : \mathbb{R} \to \mathbb{R}$ is affine with $2 + 2^k$ pieces (see Figure 4, and note that we consider Δ_k as a function on all of $\mathbb{R}$, not just on $[0,1]$), Lemma 4.9 shows that $\Delta_k \in \mathtt{NN}^{\varrho_1,1,1}_{\infty,2,3+2^k}$. By the same reasoning, we get $\Delta_{k+s} \in \mathtt{NN}^{\varrho_1,1,1}_{\infty,2,3+2^{k+s}}$. Now, a repeated application of Lemma 2.18-(3) shows that

$$\Delta_j = \Delta_{k+s} \circ \Delta_k \circ \cdots \circ \Delta_k \in \mathtt{NN}^{\varrho_1,1,1}_{\infty,L,(L-2)(3+2^k)+3+2^{k+s}}.$$

Finally, $\Delta_j \in \mathtt{NN}^{\varrho_1,1,1}_{\infty,L,C_L \cdot 2^{j/(L-1)}}$ with $C_L := 4\,L + 2^{L-1}$ since

$$(L-2)(3+2^k)+3+2^{k+s} = 3(L-1)+(L-2+2^s)2^k \leq (4L-5+2^{L-2})2^k \leq (4L+2^{L-1})2^{j/(L-1)} = C_L \cdot 2^{j/(L-1)}.$$

Part (2): Set $\kappa := \lfloor L/2 \rfloor$ and write $j = k\kappa + s$ for $k \in \mathbb{N}_0$ and $0 \leq s \leq \kappa - 1$. Note that $k \leq j/\kappa = j/\lfloor L/2 \rfloor$. As above, $\Delta_j = \Delta_{k+s} \circ \Delta_k \circ \cdots \circ \Delta_k$, where Δ_k occurs $\kappa - 1$ times, and since $\Delta_k : \mathbb{R} \to \mathbb{R}$ is affine with $2 + 2^k$ pieces, using Lemma 4.9 again shows that $\Delta_k \in \mathtt{NN}^{\varrho_1,1,1}_{6+2^{k+1},2,\infty}$, and $\Delta_{k+s} \in \mathtt{NN}^{\varrho_1,1,1}_{6+2^{k+s+1},2,\infty}$. Now, a repeated application of Lemma 2.18-(2) shows that

$$\Delta_j = \Delta_{k+s} \circ \Delta_k \circ \cdots \circ \Delta_k \in \mathtt{NN}^{\varrho_1,1,1}_{6+2^{k+s+1}+(\kappa-1)(6+2^{k+1}),2+2\cdot(\kappa-1),\infty}.$$

Finally, $\Delta_k \in \mathtt{NN}^{\varrho_1,1,1}_{C_L 2^{j/\lfloor L/2 \rfloor},\lfloor L/2 \rfloor,\infty}$, as $2 + 2(\kappa-1) = 2\kappa \leq L$, $s + 1 \leq \kappa \leq L/2 \leq L - 1$ (since $L \geq 2$) and

$$6 + 2^{k+s+1} + (\kappa-1)(6+2^{k+1}) = 6\kappa + (2^{s+1}+2)2^k \leq (3L + 2^{L-1} + 2)2^{j/\lfloor L/2 \rfloor} \leq C_L \cdot 2^{j/\lfloor L/2 \rfloor}. \quad \square$$

D.4. **Proof of Lemma 5.12.** For $h \in \mathbb{R}^d$, we define the translation operator T_h by $(T_h f)(x) = f(x-h)$ for $f : \mathbb{R}^d \to \mathbb{R}$. With this, the h-difference operator of order k is given by $D^k_h = (D_h)^k$, where $D_h := (T_{-h} - \mathrm{id})$. For later use, we note for $a > 0$ that $D_h[f(a\bullet)](x) = (D_{ah} f)(ax)$, as can be verified by a direct calculation. By induction, this implies $D^k_h[f(a\bullet)] = (D^k_{ah} f)(a\bullet)$ for all $k \in \mathbb{N}$. Furthermore, $T_x D^k_h = D^k_h T_x$ for all $x, h \in \mathbb{R}^d$ and $k \in \mathbb{N}$.

A direct computation shows

$$\Delta_1 = \widetilde{\Delta_1} + 2\widetilde{\Delta_1}(\bullet - \tfrac{1}{4}) + \widetilde{\Delta_1}(\bullet - \tfrac{1}{2}) = (T_{1/4} + \mathrm{id})^2 \widetilde{\Delta_1} \qquad \text{where} \qquad \widetilde{\Delta_1} := \frac{1}{2}\Delta_1(2\bullet).$$

Next, note that $(T_{-1/4} - \mathrm{id})(T_{1/4} + \mathrm{id}) = T_{-1/4} - T_{1/4}$ and hence, since $T_{-1/4}$ and $T_{1/4}$ commute,

$$D^2_{1/4}\Delta_1 = (T_{-1/4} - \mathrm{id})^2 (T_{1/4} + \mathrm{id})^2 \widetilde{\Delta_1} = (T_{-1/4} - T_{1/4})^2 \widetilde{\Delta_1} = (T_{-1/2} - 2\,\mathrm{id} + T_{1/2})\widetilde{\Delta_1}. \tag{D.3}$$

Moreover by induction on $\ell \in \mathbb{N}_0$, we see that

$$\sum_{k=0}^{\ell} T_k (T_{-\frac{1}{2}} - 2\,\mathrm{id} + T_{\frac{1}{2}}) = T_{-\frac{1}{2}} + T_{\frac{2\ell+1}{2}} + 2\sum_{i=0}^{2\ell} (-1)^{i-1}\, T_{\frac{i}{2}}. \tag{D.4}$$

Define $h_j := 2^{-(j+1)}$, so that $2^{j-1}h_j = 1/4$. Since $\Delta_j = \sum_{k=0}^{2^{j-1}-1}(T_k\Delta_1)(2^{j-1}\bullet)$ (cf. Equation (5.11)), Equations (D.3) and (D.4) and the properties from the beginning of the proof yield for $x \in \mathbb{R}$ that

$$\begin{aligned}(D^2_{h_j}\Delta_j)(x) &= \sum_{k=0}^{2^{j-1}-1}[D^2_{2^{j-1}h_j}(T_k\Delta_1)](2^{j-1}x) = \Big[\sum_{k=0}^{2^{j-1}-1} T_k(D^2_{1/4}\Delta_1)\Big](2^{j-1}x)\\ &= (T_{-\frac{1}{2}}\widetilde{\Delta_1})(2^{j-1}x) + (T_{\frac{2^j-1}{2}}\widetilde{\Delta_1})(2^{j-1}x) + 2\sum_{i=0}^{2^j-2}(-1)^{i-1}(T_{\frac{i}{2}}\widetilde{\Delta_1})(2^{j-1}x).\end{aligned} \tag{D.5}$$

Recall for $g \in X_p(\Omega)$ that the r-th modulus of continuity of g is given by

$$\omega_r(g)_p(t) := \sup_{h\in\mathbb{R}^d,|h|\le t}\|D^r_h g\|_{X_p(\Omega_{r,h})} \quad \text{where} \quad \Omega_{r,h} := \{x\in\Omega\colon x+uh\in\Omega \text{ for each } u\in[0,r]\}.$$

Let $e_1 = (1,0,\dots,0) \in \mathbb{R}^d$. For $h = h_j\,e_1$, we have $\Omega_{2,h} \supset (0,\frac{1}{2})\times(0,1)^{d-1}$ since $\Omega = (0,1)^d$. Next, because of $\operatorname{supp}\widetilde{\Delta_1} = [0,\frac{1}{2}]$, the family $(T_{i/2}\widetilde{\Delta_1})_{i\in\mathbb{Z}}$ has pairwise disjoint supports (up to null-sets), and

$$\operatorname{supp}\big[(T_{\frac{i}{2}}\widetilde{\Delta_1})(2^{j-1}\bullet)\big] = 2^{-j}(i+[0,1]) \subset \big[0,\tfrac{1}{2}\big] \qquad \text{for} \quad 0\le i\le 2^{j-1}-1.$$

Combining these observations with the fact that $(T_{\frac{i}{2}}\widetilde{\Delta_1})(2^{j-1}\bullet) = \widetilde{\Delta_1}(2^{j-1}\bullet - i/2) = \Delta_1(2^j\bullet - i)/2$, Equation (D.5) yields for $p<\infty$ that

$$\begin{aligned}\|D^2_{h_j\,e_1}\Delta_{j,d}\|^p_{L_p(\Omega_{2,h_j e_1})} &\ge \sum_{i=0}^{2^{j-1}-1} 2^p\|(T_{\frac{i}{2}}\widetilde{\Delta_1})(2^{j-1}\bullet)\|^p_{L_p(2^{-j}(i+[0,1]))} = \sum_{i=0}^{2^{j-1}-1}\|\Delta_1(2^j\bullet - i)\|^p_{L_p(2^{-j}(i+[0,1]))}\\ &= \sum_{i=0}^{2^{j-1}-1} 2^{-j}\|\Delta_1\|^p_{L_p([0,1])} = \frac{\|\Delta_1\|^p_{L_p}}{2},\end{aligned}$$

and hence $\|D^2_{h_j e_1}\Delta_{j,d}\|_{L_p(\Omega_{2,h_j e_1})} \ge C_p$, where $C_p := 2^{-1/p}\,\|\Delta_1\|_{L_p}$ for $p<\infty$. Since $\Omega_{2,h_j e_1}\subset\Omega=(0,1)^d$ has at most measure 1, we have $\|\cdot\|_{L_1(\Omega_{2,h_j e_1})} \le \|\cdot\|_{L_\infty(\Omega_{2,h_j e_1})}$, hence the same holds for $p=\infty$ with $C_\infty := C_1$. By definition, this implies $\omega_2(\Delta_{j,d})_p(t) \ge C_p$ for $t \ge |h_j e_1| = 2^{-(j+1)}$.

Overall, we get by definition of the Besov quasi-norms in case of $q<\infty$ that

$$\|\Delta_{j,d}\|^q_{B^\alpha_{p,q}(\Omega)} \ge \int_0^\infty [t^{-\alpha}\omega_2(\Delta_{j,d})_p(t)]^q\frac{dt}{t} \ge C_p^q\cdot\int_{2^{-(j+1)}}^\infty t^{-\alpha q-1}\,dt = \frac{C_p^q}{\alpha q}\cdot 2^{\alpha q(j+1)},$$

and hence $\|\Delta_{j,d}\|_{B^\alpha_{p,q}(\Omega)} \ge \frac{C_p}{(\alpha q)^{1/q}}\,2^{\alpha(j+1)}$ for all $j\in\mathbb{N}$. In case of $q=\infty$, we see similarly that

$$\|\Delta_{j,d}\|_{B^\alpha_{p,q}(\Omega)} \ge \sup_{t\in(0,\infty)} t^{-\alpha}\,\omega_2(\Delta_{j,d})_p(t) \ge C_p\cdot(2^{-(j+1)})^{-\alpha} = C_p\cdot 2^{\alpha(j+1)}$$

for all $j\in\mathbb{N}$. In both cases, we used that $\alpha<2$ to ensure that we can use the modulus of continuity of order 2 to compute the Besov quasi-norm. Finally, note because of $\alpha\le s$ that $B^s_{p,q}(\Omega)\hookrightarrow B^\alpha_{p,q}(\Omega)$; see Equation (5.4). This easily implies the claim. □

D.5. **Proof of Lemma 5.19.** In this section, we prove Lemma 5.19, based on results of Telgarsky [64].

Telgarsky makes extensive use of two special classes of functions: First, a function $\sigma:\mathbb{R}\to\mathbb{R}$ is called (t,β)**-poly** (where $t\in\mathbb{N}$ and $\beta\in\mathbb{N}_0$) if there is a partition of $\mathbb{R}$ into t intervals $I_1,\dots,I_t$ such that $\sigma|_{I_j}$ is a polynomial of degree at most β for each $j\in\{1,\dots,t\}$. In the language of Definition 4.6, these are precisely those functions which belong to $\mathtt{PPoly}^\beta_t(\mathbb{R})$. The second class of functions which is important are the (t,α,β)**-semi-algebraic functions** $f:\mathbb{R}^k\to\mathbb{R}$ (where $t\in\mathbb{N}$ and $\alpha,\beta\in\mathbb{N}_0$). The definition of this class (see [64, Definition 2.1]) is somewhat technical. Luckily, we don't need the definition, all we need to know is the following result:

Lemma D.2. *(see* [64, Lemma 2.3-(1)]*) If $\sigma:\mathbb{R}\to\mathbb{R}$ is (t,β)-poly and $q:\mathbb{R}^d\to\mathbb{R}$ is a (multivariate) polynomial of degree at most $\alpha\in\mathbb{N}_0$, then $\sigma\circ q$ is $(t,\alpha,\alpha\beta)$-semi-algebraic.* ◀

In most of our proofs, we will mainly be interested in knowing that a function $\sigma:\mathbb{R}\to\mathbb{R}$ is (t,α)-poly for certain t,α. The following lemma gives a sufficient condition for this to be the case.

Lemma D.3. *(see* [64, Lemma 3.6]*) If $f:\mathbb{R}^k\to\mathbb{R}$ is (s,α,β)-semi-algebraic and if $g_1,\dots,g_k:\mathbb{R}\to\mathbb{R}$ are (t,γ)-poly, then the function $f\circ(g_1,\dots,g_k):\mathbb{R}\to\mathbb{R}$ is $\big(st(1+\alpha\gamma)\cdot k,\beta\gamma\big)$-poly.* ◀

For proving Lemma 5.19, we begin with the easier case where we count neurons instead of weights.

Proof of the second part of Lemma 5.19. We want to show that for any depth $L \in \mathbb{N}_{\geq 2}$ and degree $r \in \mathbb{N}$ there is a constant $\Lambda_{L,r} \in \mathbb{N}$ such that each function $f \in \mathtt{NN}_{\infty,L,N}^{\varrho_r,1,1}$ is $(\Lambda_{L,r} N^{L-1}, r^{L-1})$-poly. To show this, let $\Phi \in \mathcal{NN}_{\infty,L,N}^{\varrho_r,1,1}$ with $f = \mathtt{R}(\Phi)$, say $\Phi = \big((T_1,\alpha_1),\dots,(T_K,\alpha_K)\big)$, where necessarily $K \leq L$, and where each $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is affine-linear.

For $\ell \in \{1,\dots,K\}$ and $j \in \{1,\dots,N_\ell\}$, we let $f_j^{(\ell)} : \mathbb{R} \to \mathbb{R}$ denote the output of neuron j in the ℓ-th layer. Formally, let $f_j^{(1)} : \mathbb{R} \to \mathbb{R}, x \mapsto \big(\alpha_1(T_1\,x)\big)_j$, and inductively

$$f_j^{(\ell+1)} : \mathbb{R} \to \mathbb{R}, x \mapsto \Big[\alpha_{\ell+1}\Big(T_{\ell+1}\big(f_k^{(\ell)}(x)\big)_{k\in\{1,\dots,N_\ell\}}\Big)\Big]_j \quad \text{for} \quad 1 \leq \ell \leq L-1 \text{ and } 1 \leq j \leq N_{\ell+1}. \tag{D.6}$$

We prove below by induction on $\ell \in \{1,\dots,K\}$ that there is a constant $C_{\ell,r} \in \mathbb{N}$ which only depends on ℓ, r and such that $f_j^{(\ell)}$ is $\big(C_{\ell,r}\prod_{t=0}^{\ell-1} N_t, r^{\gamma(\ell)}\big)$-poly, where $\gamma(\ell) := \min\{\ell, L-1\}$. Once this is shown, we see that $f = \mathtt{R}(\Phi) = f_1^{(K)}$ is $\big(C_{K,r}\prod_{t=0}^{K-1} N_t, r^{L-1}\big)$-poly. Then, because of $N_0 = 1$, we see that

$$C_{K,r}\prod_{t=0}^{K-1} N_t \leq \Lambda_{L,r}\prod_{t=1}^{K-1} N_t \leq \Lambda_{L,r}\prod_{t=1}^{K-1} N(\Phi) \leq \Lambda_{L,r}\cdot[N(\Phi)]^{K-1} \leq \Lambda_{L,r}\cdot N^{L-1},$$

where $\Lambda_{L,r} := \max_{1\leq K\leq L} C_{K,r}$. Therefore, f is indeed $(\Lambda_{L,r}\,N^{L-1}, r^{L-1})$-poly.

Start of induction ($\ell = 1$): Note that $L \geq 2$, so that $\gamma(\ell) = \ell = 1$. We have $T_1 x = ax + b$ for certain $a, b \in \mathbb{R}^{N_1}$ and $\alpha_1 = \varrho^{(1)} \otimes \cdots \otimes \varrho^{(N_1)}$ for certain $\varrho^{(j)} \in \{\mathrm{id}_\mathbb{R}, \varrho_r\}$. Thus, $\varrho^{(j)}$ is $(2,r)$-poly, and thus $(2,1,r)$-semi-algebraic according to Lemma D.2. Therefore, Lemma D.3 shows because of $f_j^{(1)}(x) = \varrho^{(j)}(b_j + a_j x)$ that $f_j^{(1)}$ is $(2(1+1), r)$-poly, for any $j \in \{1,\dots,N_1\}$. Because of $N_0 = 1$, the claim holds for $C_{1,r} := 4$.

Induction step ($\ell \to \ell+1$): Suppose that $\ell \in \{1,\dots,K-1\}$ is such that the claim holds. Note that $\ell \leq K-1 \leq L-1$, so that $\gamma(\ell) = \ell$.

We have $T_{\ell+1}\,y = A\,y + b$ for certain $A \in \mathbb{R}^{N_{\ell+1}\times N_\ell}$ and $b \in \mathbb{R}^{N_{\ell+1}}$, and $\alpha_{\ell+1} = \varrho^{(1)} \otimes \cdots \otimes \varrho^{(N_{\ell+1})}$ for certain $\varrho^{(j)} \in \{\mathrm{id}_\mathbb{R}, \varrho_r\}$, where $\varrho^{(j)} = \mathrm{id}_\mathbb{R}$ for all $j \in \{1,\dots,N_{\ell+1}\}$ in case of $\ell = K-1$. Hence, $\varrho^{(j)}$ is $(2,r)$-poly, and even $(2,1)$-poly in case of $\ell = K-1$. Moreover, each of the polynomials $p_{j,\ell} : \mathbb{R}^{N_\ell} \to \mathbb{R}, y \mapsto (A\,y+b)_j = b_j + \sum_{t=1}^{N_\ell} A_{j,t}\,y_t$ is of degree at most 1, hence by Lemma D.2, $\varrho^{(j)} \circ p_{j,\ell}$ is $(2,1,r)$-semi-algebraic, and even $(2,1,1)$-semi-algebraic in case of $\ell = K-1$.

Each function $f_t^{(\ell)}$ is $(C_{\ell,r}\prod_{t=0}^{\ell-1} N_t, r^\ell)$-poly by the induction hypothesis. By Lemma D.3, since

$$f_j^{(\ell+1)}(x) = \varrho^{(j)}\Big(\big[A\,\big(f_t^{(\ell)}(x)\big)_{t\in\{1,\dots,N_\ell\}} + b\big]_j\Big) = (\varrho^{(j)} \circ p_{j,\ell})\big(f_1^{(\ell)}(x),\dots,f_{N_\ell}^{(\ell)}(x)\big),$$

it follows that $f_j^{(\ell+1)}$ is $(P, r^{\ell+1})$-poly [respectively, (P, r^ℓ)-poly if $\ell = K-1$], where

$$P \leq 2C_{\ell,r}(1+r^\ell)\cdot N_\ell \cdot \prod_{t=0}^{\ell-1} N_t =: C_{\ell+1,r}\cdot \prod_{t=0}^{(\ell+1)-1} N_t.$$

Finally, note in case of $\ell < K-1$ that $\ell+1 \leq K-1 \leq L-1$, and hence $\gamma(\ell+1) = \ell+1$, while in case of $\ell = K-1$ we have $\ell \leq \min\{\ell+1, L-1\} = \gamma(\ell+1)$. Therefore, each $f_j^{(\ell+1)}$ is $(C_{\ell+1,r}\cdot\prod_{t=0}^{(\ell+1)-1} N_t, r^{\gamma(\ell+1)})$-poly. This completes the induction, and thus the proof. □

The proof of the first part of Lemma 5.19 uses the same basic arguments as in the preceding proof, but in a more careful way. In particular, we will also need the following elementary lemma.

Lemma D.4. *Let $k \in \mathbb{N}$, and for each $i \in \{1,\dots,k\}$ let $f_i : \mathbb{R} \to \mathbb{R}$ be (t_i,α)-poly and continuous. Then the function $\sum_{i=1}^k f_i$ is (t,α)-poly, where $t = 1 - k + \sum_{i=1}^k t_i$.* ◀

Proof. For each $i \in \{1,\dots,k\}$, there are "breakpoints" $b_0^{(i)} := -\infty < b_1^{(i)} < \cdots < b_{t_i-1}^{(i)} < \infty =: b_{t_i}^{(i)}$ such that $f_i|_{\mathbb{R}\cap[b_j^{(i)}, b_{j+1}^{(i)}]}$ is a polynomial of degree at most α for each $0 \leq j \leq t_i - 1$. Here, we used the continuity of f_i to ensure that we can use closed intervals.

Now, let $M := \bigcup_{i=1}^k \{b_1^{(i)},\dots,b_{t_i-1}^{(i)}\}$. We have $|M| \leq \sum_{i=1}^k (t_i - 1) = t-1$, with t as in the statement of the lemma. Thus, $M = \{b_1,\dots,b_s\}$ for some $0 \leq s \leq t-1$, where $b_0 := -\infty < b_1 < \cdots < b_s < \infty =: b_{s+1}$. It is easy to see that $F := \sum_{i=1}^k f_i$ is such that $F|_{\mathbb{R}\cap[b_j,b_{j+1}]}$ is a polynomial of degree at most α for each $0 \leq j \leq s$. Thus, F is $(s+1,\alpha)$-poly and therefore also (t,α)-poly. □

Proof of the first part of Lemma 5.19. Let us first consider an arbitrary network $\Phi \in \mathcal{NN}_{W,L,\infty}^{\varrho_r,1,1}$ satisfying $L(\Phi) = L$. Let $L_0 := \lfloor L/2 \rfloor \in \mathbb{N}_0$. We claim that

$$\mathtt{R}(\Phi) \text{ is } \big(\max\{1, \Lambda_{L,r}\,W^{L_0}\}, r^{L-1}\big)\text{-poly where } \Lambda_{L,r} \in \mathbb{N} \text{ only depends on } L, r. \tag{D.7}$$

In case of $L = 1$, this is trivial, since then $\mathtt{R}(\Phi) : \mathbb{R} \to \mathbb{R}$ is affine-linear. Thus, we will assume $L \geq 2$ in what follows. Note that this entails $L_0 \geq 1$.

Let $\Phi = \big((T_1, \alpha_1), \dots, (T_L, \alpha_L)\big)$, where $T_\ell : \mathbb{R}^{N_{\ell-1}} \to \mathbb{R}^{N_\ell}$ is affine-linear. We first consider the special case that $\|T_\ell\|_{\ell^0} = 0$ for some $\ell \in \{1, \dots, L\}$. In this case, Lemma 2.9 shows that $\mathtt{R}(\Phi) \equiv c$ for some $c \in \mathbb{R}$. This trivially implies that $\mathtt{R}(\Phi)$ is $(\max\{1, \Lambda_{L,r}\, W^{L_0}\}, r^{L-1})$-poly. Thus, we can assume in the following that $\|T_\ell\|_{\ell^0} \neq 0$ for all $\ell \in \{1, \dots, L\}$. As in the proof of the first part of Lemma 5.19, we define $f_j^{(\ell)} : \mathbb{R} \to \mathbb{R}$ to be the function computed by neuron $j \in \{1, \dots, N_\ell\}$ in layer $\ell \in \{1, \dots, L\}$, cf. Equation (D.6).

Step 1. We let $L_1 := \lfloor \frac{L-1}{2} \rfloor \in \mathbb{N}_0$, and we show by induction on $t \in \{0, 1, \dots, L_1\}$ that

$$f_j^{(2t+1)} \text{ is } \Big(C_{t,r} \prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0}, r^{\gamma(t)}\Big)\text{-poly} \quad \forall\, t \in \{0, 1, \dots, L_1\} \text{ and } j \in \{1, \dots, N_{2t+1}\}, \tag{D.8}$$

where $\gamma(t) := \min\{L - 1, 2t + 1\}$ and where the constant $C_{t,r} \in \mathbb{N}$ only depends on t, r. Here, we use the convention that the empty product satisfies $\prod_{\ell=1}^{0} \|T_{2\ell}\|_{\ell^0} = 1$.

Induction start ($t = 0$): We have $T_1 x = ax + b$ for certain $a, b \in \mathbb{R}^{N_1}$ and $\alpha_1 = \varrho^{(1)} \otimes \cdots \otimes \varrho^{(N_1)}$ for certain $\varrho^{(j)} \in \{\mathrm{id}_\mathbb{R}, \varrho_r\}$. In any case, $\varrho^{(j)}$ is $(2, r)$-poly, and hence $(2, 1, r)$-semi-algebraic by Lemma D.2. Now, note $f_j^{(2t+1)}(x) = f_j^{(1)}(x) = \varrho^{(j)}\big((T_1 x)_j\big) = \varrho^{(j)}(a_j x + b_j)$, so that Lemma D.3 shows that $f_j^{(2t+1)}$ is $(2(1 + 1), r)$-poly. Thus, Equation (D.8) holds for $t = 0$ if we choose $C_{0,r} := 4$. Here, we used that $L \geq 2$ and $t = 0$, so that $L - 1 \geq 2t + 1$ and hence $\gamma(t) = 2t + 1 = 1$.

Induction step ($t \to t + 1$): Let $t \in \mathbb{N}_0$ such that $t + 1 \leq \frac{L-1}{2}$ and such that Equation (D.8) holds for t. We have $T_{2t+2} \bullet = A \bullet + b$ for certain $A \in \mathbb{R}^{N_{2t+2} \times N_{2t+1}}$ and $b \in \mathbb{R}^{N_{2t+2}}$, and furthermore $\alpha_{2t+2} = \varrho^{(1)} \otimes \cdots \otimes \varrho^{(N_{2t+2})}$ for certain $\varrho^{(j)} \in \{\mathrm{id}_\mathbb{R}, \varrho_r\}$. Recall from Appendix A that $A_{j,-} \in \mathbb{R}^{1 \times N_{2t+1}}$ denotes the j-th row of A. For $j \in \{1, \dots, N_{2t+2}\}$, we claim that

$$\begin{cases} f_j^{(2t+2)} \equiv \varrho^{(j)}(b_j), & \text{if } A_{j,-} = 0, \\ f_j^{(2t+2)} \text{ is } (C'_{t,r} \cdot M_j \cdot \prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0}, r^{2t+2})\text{-poly}, & \text{if } A_{j,-} \neq 0, \end{cases} \tag{D.9}$$

where $M_j := \|A_{j,-}\|_{\ell^0}$, and where the constant $C'_{t,r} \in \mathbb{N}$ only depends on t, r.

The first case where $A_{j,-} = 0$ is trivial. For proving the second case where $A_{j,-} \neq 0$, let us define $\Omega_j := \{i \in \{1, \dots, N_{2t+1}\} \colon A_{j,i} \neq 0\}$, say $\Omega_j = \{i_1, \dots, i_{M_j}\}$ with (necessarily) pairwise distinct $i_1, \dots, i_{M_j}$. By introducing the polynomial $p_{j,t} : \mathbb{R}^{M_j} \to \mathbb{R}, y \mapsto b_j + \sum_{m=1}^{M_j} A_{j,i_m} y_m$, we can then write

$$f_j^{(2t+2)}(x) = \varrho^{(j)}\Big(b_j + A_{j,-}\big(f_k^{(2t+1)}(x)\big)_{k \in \{1, \dots, N_{2t+1}\}}\Big) = (\varrho^{(j)} \circ p_{j,t})\big(f_{i_1}^{(2t+1)}(x), \dots, f_{i_{M_j}}^{(2t+1)}(x)\big).$$

Since $\varrho^{(j)}$ is $(2, r)$-poly and $p_{j,t}$ is a polynomial of degree at most 1, Lemma D.2 shows that $\varrho^{(j)} \circ p_{j,t}$ is $(2, 1, r)$-semi-algebraic. Furthermore, by the induction hypothesis we know that each function $f_{i_m}^{(2t+1)}$ is $(C_{t,r} \prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0}, r^{2t+1})$-poly, where we used that $\gamma(t) = 2t + 1$ since $t + 1 \leq (L - 1)/2$. Therefore—in view of the preceding displayed equation—Lemma D.3 shows that the function $f_j^{(2t+2)}$ is indeed $(C'_{t,r} \cdot M_j \cdot \prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0}, r^{2t+2})$-poly, where $C'_{t,r} := 2C_{t,r} \cdot (1 + r^{2t+1})$.

We now estimate the number of polynomial pieces of the function $f_i^{(2t+3)}$ for $i \in \{1, \dots, N_{2t+3}\}$. To this end, let $B \in \mathbb{R}^{N_{2t+3} \times N_{2t+2}}$ and $c \in \mathbb{R}^{N_{2t+3}}$ such that $T_{2t+3} = B \bullet + c$, and choose $\sigma^{(i)} \in \{\mathrm{id}_\mathbb{R}, \varrho_r\}$ such that $\alpha_{2t+3} = \sigma^{(1)} \otimes \cdots \otimes \sigma^{(N_{2t+3})}$. For $i \in \{1, \dots, N_{2t+3}\}$, let us define

$$G_{i,t} : \mathbb{R} \to \mathbb{R}, x \mapsto \sum_{j \in \{1, \dots, N_{2t+2}\} \text{ such that } A_{j,-} \neq 0} B_{i,j}\, f_j^{(2t+2)}(x).$$

In view of Equation (D.9), Lemma D.4 shows that $G_{i,t}$ is (P, r^{2t+2})-poly, where

$$\begin{aligned} P &\leq 1 - |\{j \in \{1, \dots, N_{2t+2}\} \colon A_{j,-} \neq 0\}| + C'_{t,r} \cdot \Big(\prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0}\Big) \sum_{j \in \{1, \dots, N_{2t+2}\} \text{ such that } A_{j,-} \neq 0} M_j \\ &\leq C'_{t,r} \cdot \Big(\prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0}\Big) \cdot \|A\|_{\ell^0} = C'_{t,r} \cdot \prod_{\ell=1}^{t+1} \|T_{2\ell}\|_{\ell^0} \end{aligned}$$

Here, we used that $\|T_{2t+2}\|_{\ell^0} \neq 0$ and hence $A \neq 0$, so that $|\{j \in \{1, \dots, N_{2t+2}\} \colon A_{j,-} \neq 0\}| \geq 1$.

Next, note because of Equation (D.9) and by definition of $G_{i,t}$ that there is some $\theta_{i,t} \in \mathbb{R}$ satisfying

$$f_i^{(2t+3)}(x) = \sigma^{(i)}\Big(c_i + \sum_{j=1}^{N_{2t+2}} B_{i,j}\, f_j^{(2t+2)}(x)\Big) = \sigma^{(i)}\big(\theta_{i,t} + G_{i,t}(x)\big) \quad \forall\, x \in \mathbb{R}.$$

Now there are two cases: If $2t+3 > L-1$, then $2t+3 = L$, since $t+1 \leq \frac{L-1}{2}$. Therefore, $\sigma^{(i)} = \mathrm{id}_{\mathbb{R}}$, so that we see that $f_i^{(2t+3)} = \theta_{i,t} + G_{i,t}$ is $(C'_{t,r} \cdot \prod_{\ell=1}^{t+1} \|T_{2\ell}\|_{\ell^0}, r^{2t+2})$-poly, where $2t+2 = L-1 = \gamma(t+1)$.

If $2t+3 \leq L-1$, then $\gamma(t+1) = 2t+3$. Furthermore, each $\sigma^{(i)}$ is $(2,r)$-poly and hence $(2,1,r)$-semi-algebraic by Lemma D.2. In view of the preceding displayed equation, and since $G_{i,t}$ is $(C'_{t,r} \cdot \prod_{\ell=1}^{t+1} \|T_{2\ell}\|_{\ell^0}, r^{2t+2})$-poly, Lemma D.3 shows that $f_i^{(2t+3)}$ is $\big(2(1+r^{2t+2})C'_{t,r} \cdot \prod_{\ell=1}^{t+1} \|T_{2\ell}\|_{\ell^0}, r^{2t+3}\big)$-poly.

In each case, with $C_{t+1,r} := 2(1+r^{2t+2})C'_{t,r}$, we see that Equation (D.8) holds for $t+1$ instead of t.

Step 2. We now complete the proof of Equation (D.7), by distinguishing whether L is odd or even.

If L is odd: In this case $L_1 = \lfloor \frac{L-1}{2} \rfloor = \frac{L-1}{2}$, so that we can use Equation (D.8) for the choice $t = \frac{L-1}{2}$ to see that $\mathtt{R}(\Phi) = f_1^{(L)} = f_1^{(2t+1)}$ is (P, r^{L-1})-poly, where

$$P \leq C_{t,r} \prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0} \leq C_{t,r} \prod_{\ell=1}^{(L-1)/2} W(\Phi) \leq C_{t,r} \cdot [W(\Phi)]^{(L-1)/2} \leq C_{t,r} \cdot W^{\lfloor L/2 \rfloor}.$$

If L is even: In this case, set $t := \frac{L}{2} - 1 \in \{0, 1, \dots, L_1\}$, and note $2t+1 = L-1 = \gamma(t)$. Hence, with $A \in \mathbb{R}^{1 \times N_{L-1}}$ and $b \in \mathbb{R}$ such that $T_L = A \bullet + b$, we have

$$\mathtt{R}(\Phi) = T_L\big(f_k^{(2t+1)}(x)\big)_{k \in \{1,\dots,N_{L-1}\}} = b + \sum_{k \in \{1,\dots,N_{L-1}\} \text{ such that } A_{1,k} \neq 0} A_{1,k}\, f_k^{(2t+1)}(x).$$

Therefore, thanks to Equation (D.8), Lemma D.4 shows that $\mathtt{R}(\Phi)$ is (P, r^{2t+1})-poly, where

$$\begin{aligned} P &\leq 1 - |\{k \in \{1,\dots,N_{L-1}\} \colon A_{1,k} \neq 0\}| + C_{t,r} \sum_{\substack{k \in \{1,\dots,N_{L-1}\} \\ \text{such that } A_{1,k} \neq 0}} \prod_{\ell=1}^{t} \|T_{2\ell}\|_{\ell^0} \\ &\leq C_{t,r} \cdot \|A\|_{\ell^0} \cdot \prod_{\ell=1}^{\frac{L}{2}-1} \|T_{2\ell}\|_{\ell^0} = C_{t,r} \prod_{\ell=1}^{L/2} \|T_{2\ell}\|_{\ell^0} \\ &\leq C_{t,r} \cdot [W(\Phi)]^{L/2} = C_{t,r} \cdot [W(\Phi)]^{\lfloor L/2 \rfloor} \leq C_{t,r} \cdot W^{\lfloor L/2 \rfloor}. \end{aligned}$$

In the second inequality we used $|\{k \in \{1,\dots,N_{L-1}\} \colon A_{1,k} \neq 0\}| = \|A\|_{\ell^0} = \|T_L\|_{\ell^0} \geq 1$. We have thus established Equation (D.7) in all cases.

Step 3. It remains to prove the actual claim. Let $f \in \mathtt{NN}_{W,L,\infty}^{\varrho_r,1,1}$ be arbitrary, whence $f = \mathtt{R}(\Phi)$ for some $\Phi \in \mathtt{NN}_{W,K,\infty}^{\varrho_r,1,1}$ with $L(\Phi) = K$ for some $K \in \mathbb{N}_{\leq L}$. In view of Equation (D.7), this implies that $f = \mathtt{R}(\Phi)$ is $(\max\{1, \Lambda_{K,r}\, W^{\lfloor K/2 \rfloor}\}, r^{K-1})$-poly. If we set $\Theta_{L,r} := \max_{1 \leq K \leq L} \Lambda_{K,r}$, then this easily implies that f is $(\max\{1, \Theta_{L,r}\, W^{\lfloor L/2 \rfloor}\}, r^{L-1})$-poly, as desired. □

Appendix E. The spaces $W_q^\alpha(X_p, \varrho_r, L)$ and $N_q^\alpha(X_p, \varrho_r, L)$ are distinct

In this section, we show that for a *fixed* depth $L \geq 3$ and $\Omega = (0,1)^d$ the approximation spaces defined in terms of the number of weights and in terms of the number of neurons are distinct; that is, we show

$$W_q^\alpha(X_p(\Omega), \varrho_r, L) \neq N_q^\alpha(X_p(\Omega), \varrho_r, L). \tag{E.1}$$

The proof is based on several results by Telgarsky [64], which we first collect. The first essential concept is the notion of the *crossing number* of a function.

Definition E.1. For any piecewise polynomial function $f : \mathbb{R} \to \mathbb{R}$ with finitely many pieces, define $\widetilde{f} : \mathbb{R} \to \{0,1\}, x \mapsto \mathbb{1}_{f(x) \geq 1/2}$. Thanks to our assumption on f, the sets $\widetilde{f}^{-1}(\{0\}) \subset \mathbb{R}$ and $\widetilde{f}^{-1}(\{1\}) \subset \mathbb{R}$ are finite unions of (possibly degenerate) intervals. For $i \in \{0,1\}$, denote by $I_f^{(i)} \subset 2^{\mathbb{R}}$ the set of connected components of $\widetilde{f}^{-1}(\{i\})$. Finally, set $I_f := I_f^{(0)} \cup I_f^{(1)}$ and define the *crossing number* $\mathrm{Cr}(f)$ of f as $\mathrm{Cr}(f) := |I_f| \in \mathbb{N}$. ◀

The following result gives a bound on the crossing number of f, based on bounds on the complexity of f. Here, we again use the notion of (t,β)–poly functions as introduced at the beginning of Appendix D.5.

Lemma E.2. *([64, Lemma 3.3]) If $f : \mathbb{R} \to \mathbb{R}$ is (t, α)–poly, then $\mathrm{Cr}(f) \leq t(1+\alpha)$.* ◀

Finally, we will need the following result which tells us that if $\mathrm{Cr}(f) \gg \mathrm{Cr}(g)$, then the functions $\widetilde{f}, \widetilde{g}$ introduced in Definition E.1 differ on a large number of intervals $I \in I_f$.

Lemma E.3. *([64, Lemma 3.1]) Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be piecewise polynomial with finitely many pieces. Then*

$$\frac{1}{\mathrm{Cr}(f)} \cdot \Big|\big\{I \in I_f \colon \forall x \in I : \widetilde{f}(x) \neq \widetilde{g}(x)\big\}\Big| \geq \frac{1}{2}\Big(1 - 2\frac{\mathrm{Cr}(g)}{\mathrm{Cr}(f)}\Big).$$ ◀

The first step to proving Equation (E.1) will be the following estimate:

Lemma E.4. *Let $p \in (0, \infty]$. There is a constant $C_p > 0$ such that the error of best approximation (cf. Equation (3.1)) of the "sawtooth function" Δ_j (cf. Equation (5.11)) by piecewise polynomials satisfies*

$$E(\Delta_j, \mathtt{PPoly}_N^\alpha)_{L_p((0,1))} \geq C_p \qquad \forall\, j, \alpha \in \mathbb{N}, \quad \forall\, 1 \leq N \leq \frac{2^j+1}{4(1+\alpha)}.$$ ◀

For proving this lower bound, we first need to determine the crossing number of Δ_j.

Lemma E.5. *Let $j \in \mathbb{N}$ and $\Delta_j : \mathbb{R} \to \mathbb{R}$ as in Equation (5.11). We have $\mathrm{Cr}(\Delta_j) = 1 + 2^j$ and*

$$\int_{I \cap [0,1]} \big|\Delta_j(x) - \frac{1}{2}\big|\, dx \geq 2^{-j-3} \qquad \forall\, I \in I_{\Delta_j}.$$ ◀

Proof. The formal proof is omitted as it involves tedious but straightforward computations; graphically, the claimed properties are straightforward consequences of Figure 4. □

Proof of Lemma E.4. Let $j, \alpha \in \mathbb{N}$ and let $N \in \mathbb{N}$ with $N \leq \frac{2^j+1}{4(1+\alpha)}$ and $f \in \mathtt{PPoly}_N^\alpha$ be arbitrary. Lemma E.2 shows $\mathrm{Cr}(f) \leq N(1+\alpha) \leq \frac{2^j+1}{4}$, so that Lemma E.5 implies $\theta := 1 - 2\frac{\mathrm{Cr}(f)}{\mathrm{Cr}(\Delta_j)} = 1 - 2\frac{\mathrm{Cr}(f)}{1+2^j} \geq \frac{1}{2}$. Now, recall the notation of Definition E.1, and set

$$G := \big\{I \in I_{\Delta_j} \,\big|\, \forall\, x \in I : \widetilde{\Delta_j}(x) \neq \widetilde{f}(x)\big\}.$$

By Lemma E.3, $\frac{1}{\mathrm{Cr}(\Delta_j)}|G| \geq \frac{\theta}{2} \geq \frac{1}{4}$, which means $|G| \geq \frac{1+2^j}{4} \geq 2^{j-2}$, since we have $\mathrm{Cr}(\Delta_j) = 1 + 2^j$.

For arbitrary $I \in G$, we have $\widetilde{\Delta_j}(x) \neq \widetilde{f}(x)$ for all $x \in I$, so that either $f(x) < \frac{1}{2} \leq \Delta_j(x)$ or $\Delta_j(x) < \frac{1}{2} \leq f(x)$. In both cases, we get $|\Delta_j(x) - f(x)| \geq |\Delta_j(x) - \frac{1}{2}|$. Furthermore, recall that $0 \leq \Delta_j \leq 1$, so that $|\Delta_j(x) - \frac{1}{2}| \leq \frac{1}{2} \leq 1$. Because of $\|\Delta_j - f\|_{L_p((0,1))} \geq \|\Delta_j - f\|_{L_1((0,1))}$ for $p \geq 1$, it is sufficient to prove the result for $0 < p \leq 1$. For this range of p, we see that

$$\big|\Delta_j(x) - \tfrac{1}{2}\big| = \big|\Delta_j(x) - \tfrac{1}{2}\big|^{1-p} \cdot \big|\Delta_j(x) - \tfrac{1}{2}\big|^p \leq \big|\Delta_j(x) - \tfrac{1}{2}\big|^p.$$

Overall, we get $|\Delta_j(x) - f(x)|^p \geq |\Delta_j(x) - \frac{1}{2}|^p \geq |\Delta_j(x) - \frac{1}{2}|$ for all $x \in I$ and $I \in G$. Thus,

$$\begin{aligned}\int_{[0,1]} |\Delta_j(x) - f(x)|^p\, dx &\geq \sum_{I \in G} \int_{I \cap [0,1]} |\Delta_j(x) - f(x)|^p\, dx \geq \sum_{I \in G} \int_{I \cap [0,1]} \Big|\Delta_j(x) - \frac{1}{2}\Big|\, dx \\ &\overset{(\text{Lemma } E.5)}{\geq} \sum_{I \in G} 2^{-j-3} = |G| \cdot 2^{-j-3} \geq 2^{j-2} \cdot 2^{-j-3} = 2^{-5}.\end{aligned}$$

This implies $\|\Delta_j - f\|_{L_p((0,1))} \geq 2^{-5/p} =: C_p$. □

As a consequence of the lower bound in Lemma E.4, we can now prove lower bounds for the neural network approximation space norms of the multivariate sawtooth function $\Delta_{j,d}$ (cf. Definition 5.9)

Proposition E.6. *Consider $\Omega = [0,1]^d$, $r \in \mathbb{N}$, $L \in \mathbb{N}_{\geq 2}$, $\alpha \in (0, \infty)$, $p, q \in (0, \infty]$. There is a constant $C = C(d, r, L, \alpha, p, q) > 0$ such that*

$$\|\Delta_{j,d}\|_{W_q^\alpha(X_p(\Omega), \varrho_r, L)} \geq C \cdot 2^{\alpha j/\lfloor L/2 \rfloor} \quad \text{and} \quad \|\Delta_{j,d}\|_{N_q^\alpha(X_p(\Omega), \varrho_r, L)} \geq C \cdot 2^{\alpha j/(L-1)}, \qquad \forall j \in \mathbb{N}.$$ ◀

Proof. According to Lemma 5.19, there is a constant $C_1 = C_1(r, L) \in \mathbb{N}$ such that

$$\mathtt{NN}_{W,L,\infty}^{\varrho_r,1,1} \subset \mathtt{PPoly}_{C_1 \cdot W^{\lfloor L/2 \rfloor}}^\beta \quad \text{and} \quad \mathtt{NN}_{\infty,L,N}^{\varrho_r,1,1} \subset \mathtt{PPoly}_{C_1 \cdot N^{L-1}}^\beta \quad \text{where} \quad \beta := r^{L-1}. \tag{E.2}$$

We first prove the estimate regarding $\|\Delta_{j,d}\|_{W_q^\alpha(X_p(\Omega), \varrho_r, L)}$. To this end, note that there is a constant $C_2 = C_2(L, \beta, C_1) = C_2(L, r) > 0$ such that $\big(\frac{2^{j+1}}{4C_1(1+\beta)}\big)^{1/\lfloor L/2 \rfloor} = C_2 \cdot 2^{(j+1)/\lfloor L/2 \rfloor}$. Now, let $W \in \mathbb{N}_0$ with $W \leq C_2 \cdot 2^{(j+1)/\lfloor L/2 \rfloor}$ and $F \in \mathtt{NN}_{W,L,\infty}^{\varrho_r,d,1}$ be arbitrary. Define $F_{x'} : \mathbb{R} \to \mathbb{R}, t \mapsto F((t, x'))$ for $x' \in [0,1]^{d-1}$. According to Lemma 2.18-(1) and Equation (E.2), we have $F_{x'} \in \mathtt{NN}_{W,L,\infty}^{\varrho_r,1,1} \subset \mathtt{PPoly}_{C_1 \cdot W^{\lfloor L/2 \rfloor}}^\beta$.

Since $C_1 \cdot W^{\lfloor L/2 \rfloor} \leq C_1 \cdot \frac{2^{j+1}}{4C_1(1+\beta)} = \frac{2^{j+1}}{4(1+\beta)}$, Lemma E.4 yields a constant $C_3 = C_3(p) > 0$ such that $C_3 \leq \|\Delta_j - F_{x'}\|_{L_p((0,1))}$. For $p < \infty$, Fubini's theorem shows that

$$\begin{aligned}\|\Delta_{j,d} - F\|_{L_p(\Omega)}^p &\geq \int_{[0,1]^{d-1}} \int_{[0,1]} \left|\Delta_j(x_1) - F((x_1, x'))\right|^p dx_1\, dx' \\ &= \int_{[0,1]^{d-1}} \|\Delta_j - F_{x'}\|_{L_p((0,1))}^p \, dx' \geq C_3^p \cdot \int_{[0,1]^{d-1}} dx' = C_3^p.\end{aligned}$$

Therefore,

$$E(\Delta_{j,d}, \mathtt{NN}_{W,L,\infty}^{\varrho_r,d,1})_{L_p(\Omega)} \geq C_3 > 0 \quad \forall\, W \in \mathbb{N}_0 \text{ satisfying } W \leq C_2 \cdot 2^{(j+1)/\lfloor L/2 \rfloor}. \tag{E.3}$$

Since $\| \bullet \|_{L^\infty(\Omega)} \geq \| \bullet \|_{L^1(\Omega)}$, this also holds for $p = \infty$. In light of the embedding (3.2), it is sufficient to lower bound $\|\Delta_{j,d}\|_{W_q^\alpha(X_p(\Omega),\varrho_r,L)}$ when $q = \infty$. In this case, we have

$$\begin{aligned}\|\Delta_{j,d}\|_{W_\infty^\alpha(X_p,\varrho_r,L)} &= \max\left\{\|\Delta_{j,d}\|_{L_p}\,, \quad \sup_{W\in\mathbb{N}} \left[(1+W)^\alpha \cdot E(\Delta_{j,d}, \mathtt{NN}_{W,L,\infty}^{\varrho_r,d,1})\right]\right\} \\ (\text{Equation (E.3)}) &\geq \begin{cases} \|\Delta_{j,d}\|_{L_p} \geq C_3, & \text{if } C_2 \cdot 2^{(j+1)/\lfloor L/2 \rfloor} < 1 \\ C_3 \cdot (1 + \lfloor C_2 \cdot 2^{(j+1)/\lfloor L/2 \rfloor} \rfloor)^\alpha, & \text{if } C_2 \cdot 2^{(j+1)/\lfloor L/2 \rfloor} \geq 1 \end{cases} \\ &\geq C_3\, C_2^\alpha \cdot 2^{j\alpha/\lfloor L/2 \rfloor},\end{aligned}$$

as desired. This completes the proof of the lower bound of $\|\Delta_{j,d}\|_{W_q^\alpha(X_p,\varrho_r,L)}$.

The lower bound for $\|\Delta_{j,d}\|_{N_q^\alpha(X_p,\varrho_r,L)}$ can be derived similarly. First, in the same way that we proved Equation (E.3), one can show that

$$E(\Delta_{j,d}, \mathtt{NN}_{\infty,L,N}^{\varrho_r,d,1})_{L_p(\Omega)} \geq C_3 > 0 \quad \forall\, N \in \mathbb{N}_0 \text{ satisfying } N \leq C_2' \cdot 2^{(j+1)/(L-1)},$$

for a suitable constant $C_2' = C_2'(L, r) > 0$. The remainder of the argument is then almost identical to that for estimating $\|\Delta_{j,d}\|_{W_q^\alpha(X_p(\Omega),\varrho_r,L)}$, and is thus omitted. □

As our final preparation for showing that the spaces $W_q^\alpha(X_p(\Omega), \varrho_r, L)$ and $N_q^\alpha(X_p(\Omega), \varrho_r, L)$ are distinct for $L \geq 3$ (Lemma 3.10), we will show that the lower bound derived in Proposition E.6 is sharp and extends to arbitrary measurable Ω with nonempty interior.

Theorem E.7. Let $p, q \in (0, \infty]$, $\alpha > 0$, $r \in \mathbb{N}$, $L \in \mathbb{N}_{\geq 2}$, and let $\Omega \subset \mathbb{R}^d$ be a bounded L_p-domain with non-empty interior. Consider $y \in \mathbb{R}^d$ and $s > 0$ satisfying $y + [0, s]^d \subset \Omega$ and define

$$\Delta_j^{(y,s)} : \mathbb{R}^d \to [0,1], x \mapsto \Delta_{j,d}\Big(\frac{x-y}{s}\Big) \quad \text{for } j \in \mathbb{N}.$$

Then there are $C_1, C_2 > 0$ such that for all $j \in \mathbb{N}$ the function $\Delta_j^{(y,s)}$ satisfies

$$\begin{aligned} C_1 \cdot 2^{j\alpha/\lfloor L/2 \rfloor} &\leq \|\Delta_j^{(y,s)}\|_{W_q^\alpha(X_p(\Omega),\varrho_r,L)} \leq C_2 \cdot 2^{j\alpha/\lfloor L/2 \rfloor} \\ \text{and} \quad C_1 \cdot 2^{j\alpha/(L-1)} &\leq \|\Delta_j^{(y,s)}\|_{N_q^\alpha(X_p(\Omega),\varrho_r,L)} \leq C_2 \cdot 2^{j\alpha/(L-1)}. \quad \blacktriangleleft \end{aligned}$$

Proof. For the upper bound, since Ω is bounded, Theorem 4.7 (Equation (4.3), which also holds for N_q^α instead of W_q^α) shows that it suffices to prove the claim for $r = 1$. Since $T_{y,s} : \mathbb{R}^d \to \mathbb{R}^d, x \mapsto s^{-1}(x - y)$ satisfies $\|T_{y,s}\|_{\ell_*^{0,\infty}} = 1$, a combination of Lemmas 5.10 and 2.18-(1) shows that there is a constant $C_L > 0$ satisfying

$$\Delta_j^{(y,s)} \in \mathtt{NN}_{\infty,L,\lfloor C_L \cdot 2^{j/(L-1)} \rfloor}^{\varrho_1,d,1} \qquad \text{and} \qquad \Delta_j^{(y,s)} \in \mathtt{NN}_{\lfloor C_L \cdot 2^{j/\lfloor L/2 \rfloor} \rfloor,L,\infty}^{\varrho_1,d,1} \qquad \forall\, j \in \mathbb{N}.$$

Furthermore, $\Delta_j^{(y,s)} \in X_p(\Omega)$ since Ω is bounded and $\Delta_j^{(y,s)}$ is bounded and continuous. Thus, the Bernstein inequality (5.1) yields a constant $K_1 > 0$ such that

$$\|\Delta_j^{(y,s)}\|_{N_q^\alpha(X_p(\Omega),\varrho_1,L)} \leq K_1 \cdot \lfloor C_L \cdot 2^{j/(L-1)} \rfloor^\alpha \leq K_1 C_L^\alpha \cdot 2^{j\alpha/(L-1)}$$

for all $j \in \mathbb{N}$; similarly, we get a constant $K_2 > 0$ such that

$$\|\Delta_j^{(y,s)}\|_{W_q^\alpha(X_p(\Omega),\varrho_1,L)} \leq K_2 \cdot \lfloor C_L \cdot 2^{j/\lfloor L/2 \rfloor} \rfloor^\alpha \leq K_2 C_L^\alpha \cdot 2^{j\alpha/\lfloor L/2 \rfloor}$$

for all $j \in \mathbb{N}$. Considering $C_2 := \max\{K_1, K_2\} \cdot C_L^\alpha$ establishes the desired upper bound.

For the lower bound, consider arbitrary $W, N \in \mathbb{N}_0$, $F \in \mathtt{NN}_{W,L,N}^{\varrho_r,d,1}$, and observe that by Lemma 2.18-(1) we have $F' := F \circ T_{y,s}^{-1} \in \mathtt{NN}_{W,L,N}^{\varrho_r,d,1}$. In view of Proposition E.6, the lower bound follows from the inequality

$$\|\Delta_j^{(y,s)} - F\|_{L_p(\Omega)} \geq \|\Delta_j^{(y,s)} - F\|_{L_p(y+[0,s]^d)} = \|\Delta_{j,d} \circ T_{y,s} - F' \circ T_{y,s}\|_{L_p(y+[0,s]^d)} = s^{d/p}\, \|\Delta_{j,d} - F'\|_{L_p([0,1]^d)}.$$ □

We can now prove Lemma 3.10.

Proof of Lemma 3.10. **Ad (1)** If $W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)\subset N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L')$, then the linear map

$$\iota : W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)\to N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L'), f\mapsto f$$

is well-defined. Furthermore, this map has a closed graph. Indeed, if $f_n\to f$ in $W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)$ and $f_n=\iota f_n\to g$ in $N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L')$, then the embeddings $W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)\hookrightarrow X_{p_1}(\Omega)\hookrightarrow L_{p_1}(\Omega)$ and $N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L')\hookrightarrow X_{p_2}(\Omega)\hookrightarrow L_{p_2}(\Omega)$ (see Proposition 3.2 and Theorem 4.7) imply that $f_n\to f$ in L_{p_1} and $f_n\to g$ in L_{p_2}. But L_p-convergence implies convergence in measure, so that we get $f=g$.

Now, the closed graph theorem (which applies to F-spaces (see [59, Theorem 2.15]), hence to quasi-Banach spaces, since these are F-spaces (see [66, Remark after Lemma 2.1.5])) shows that ι is continuous. Here, we used that the approximation classes $W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)$ and $N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L')$ are quasi-Banach spaces; this is proved independently in Theorem 3.27.

Since Ω has nonempty interior, there are $y\in\mathbb{R}^d$ and $s>0$ such that $y+[0,s]^d\subset\Omega$. The continuity of ι, combined with Theorem E.7, implies for the functions $\Delta_j^{(y,s)}$ from Theorem E.7 for all $j\in\mathbb{N}$ that

$$2^{j\beta/(L'-1)}\lesssim \|\Delta_j^{(y,s)}\|_{N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L')}\lesssim \|\Delta_j^{(y,s)}\|_{W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)}\lesssim 2^{j\alpha/\lfloor L/2\rfloor},$$

where the implicit constants are independent of j. Hence, $\beta/(L'-1)\leq\alpha/\lfloor L/2\rfloor$; that is, $L'-1\geq\frac{\beta}{\alpha}\cdot\lfloor L/2\rfloor$.

Ad (2) Exactly as in the argument above, we get for all $j\in\mathbb{N}$ that

$$2^{j\alpha/\lfloor L/2\rfloor}\lesssim \|\Delta_j^{(y,s)}\|_{W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)}\lesssim \|\Delta_j^{(y,s)}\|_{N^{\beta}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L')}\lesssim 2^{j\beta/(L'-1)}$$

with implied constants independent of j. Hence, $\alpha/\lfloor L/2\rfloor\leq\beta/(L'-1)$; that is, $\lfloor L/2\rfloor\geq\frac{\alpha}{\beta}\cdot(L'-1)$.

Proof of the "in particular" part: If $W^{\alpha}_{q_1}(X_{p_1}(\Omega),\varrho_{r_1},L)=N^{\alpha}_{q_2}(X_{p_2}(\Omega),\varrho_{r_2},L)$, then Parts (1) and (2) show (because of $\alpha=\beta$) that $L-1=\lfloor L/2\rfloor$. Since $L\in\mathbb{N}_{\geq 2}$, this is only possible for $L=2$. □

As a further consequence of Lemma E.4, we can now prove the non-triviality of the neural network approximation spaces, as formalized in Theorem 4.16.

Proof of Theorem 4.16. In view of the embedding $W^{\alpha}_q(X^k_p(\Omega),\varrho,\mathscr{L})\hookrightarrow N^{\alpha}_q(X^k_p(\Omega),\varrho,\mathscr{L})$ (see Lemma 3.9), it suffices to prove the claim for $N^{\alpha}_q(X^k_p(\Omega),\varrho,\mathscr{L})$. Furthermore, it is enough to consider the case $q=\infty$, since Equation (3.2) shows that $N^{\alpha}_q(X^k_p(\Omega),\varrho,\mathscr{L})\hookrightarrow N^{\alpha}_{\infty}(X^k_p(\Omega),\varrho,\mathscr{L})$. Next, in view of Remark 3.17, it suffices to consider the case $k=1$. Finally, thanks to Theorem 4.7, it is enough to prove the claim for the special case $\varrho=\varrho_r$ (for fixed but arbitrary $r\in\mathbb{N}$).

Since Ω has nonempty interior, there are $y\in\mathbb{R}^d$ and $s>0$ such that $y+[0,s]^d\subset\Omega$. Let us fix $\varphi\in C_c(\mathbb{R}^d)$ satisfying $0\leq\varphi\leq 1$ and $\varphi|_{y+[0,s]^d}\equiv 1$. With $\Delta_j^{(y,s)}$ as in Theorem E.7, define for $j\in\mathbb{N}$

$$g_j:\mathbb{R}^d\to\mathbb{R}, x\mapsto\Delta_j^{(y,s)}(x)\cdot\varphi(x).$$

Note that $g_j\in C_c(\mathbb{R}^d)$, and hence $g_j|_{\Omega}\in X$. Furthermore, since $0\leq\Delta_j^{(y,s)}\leq 1$, it is easy to see $\|g_j|_{\Omega}\|_X\leq\|g_j\|_{L_p(\mathbb{R}^d)}\leq\|\varphi\|_{L_p(\mathbb{R}^d)}=:C$ for all $j\in\mathbb{N}$.

By Theorem 4.2 and Proposition 3.2, we know that $N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})$ is a well-defined quasi-Banach space satisfying $N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})\hookrightarrow X_p(\Omega)$. Let us assume towards a contradiction that the claim of Theorem 4.16 fails; this means $N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})=X_p(\Omega)$. Using the same "closed graph theorem arguments" as in the proof of Lemma 3.10, we see that this implies $\|f\|_{N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})}\leq C'\cdot\|f\|_{X_p(\Omega)}$ for all $f\in X_p(\Omega)$ and a fixed constant $C'>0$. In particular, this implies $\|g_j|_{\Omega}\|_{N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})}\leq C'C$ for all $j\in\mathbb{N}$. In the remainder of the proof, we will show that $\|g_j|_{\Omega}\|_{N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})}\to\infty$ as $j\to\infty$, which then provides the desired contradiction.

To prove $\|g_j|_{\Omega}\|_{N^{\alpha}_{\infty}(X_p(\Omega),\varrho_r,\mathscr{L})}\to\infty$, choose $N_0\in\mathbb{N}$ satisfying $\mathscr{L}(N_0)\geq 2$, and let $N\in\mathbb{N}_{\geq N_0}$ and $f\in\mathtt{NN}^{\varrho_r,d,1}_{\infty,\mathscr{L}(N),N}$ be arbitrary. Reasoning as in the proof of Theorem E.7, since $\varphi\equiv 1$ on $y+[0,s]^d$, we see that if we set $T_{y,s}:\mathbb{R}^d\to\mathbb{R}^d, x\mapsto s^{-1}(y-x)$, then

$$\|g_j-f\|_{L_p(\Omega)}\geq\|g_j-f\|_{L_p(y+[0,s]^d)}=\|\Delta_j^{(y,s)}-f\|_{L_p(y+[0,s]^d)}=s^{d/p}\cdot\|\Delta_{j,d}-f\circ T^{-1}_{y,s}\|_{L_p([0,1]^d)}.$$

Now, given any $x'\in\mathbb{R}^{d-1}$, let us set $f_{x'}:\mathbb{R}\to\mathbb{R}, t\mapsto(f\circ T^{-1}_{y,s})((t,x'))$. As a consequence of Lemma 2.18-(1), we see $f_{x'}\in\mathtt{NN}^{\varrho_r,1,1}_{\infty,\mathscr{L}(N),N}$. According to Part 2 of Lemma 5.19, there is a constant $K_N\in\mathbb{N}$ such that $f_{x'}\in\mathtt{PPoly}^{r^{\mathscr{L}(N)-1}}_{K_N}$ Hence, Lemma E.4 yields a constant $C_2=C_2(p)>0$ such that $\|\Delta_j-f_{x'}\|_{L_p((0,1))}\geq C_2$

as soon as $2^j + 1 \geq 4\,K_N \cdot (1 + r^{\mathscr{L}(N)-1}) =: K'_N$. Because of $2^j + 1 \geq j$, this is satisfied if $j \geq K'_N$. In case of $p < \infty$, Fubini's theorem shows

$$\|\Delta_{j,d} - f \circ T_{y,s}^{-1}\|_{L_p([0,1]^d)}^p \geq \int_{[0,1]^{d-1}} \int_{[0,1]} \left|\Delta_j(t) - f_{x'}(t)\right|^p dt\, dx' = \int_{[0,1]^{d-1}} \|\Delta_j - f_{x'}\|_{L_p((0,1))}^p \, dx' \geq C_2^p,$$

whence $\|g_j - f\|_{L_p(\Omega)} \geq s^{d/p} \|\Delta_{j,d} - f \circ T_{y,s}^{-1}\|_{L_p([0,1]^d)} \geq C_2 \cdot s^{d/p}$. For $p = \infty$, the same estimate remains true because $\|\bullet\|_{L_p([0,1]^d)} \leq \|\bullet\|_{L_\infty([0,1]^d)}$. Since $f \in \mathtt{NN}_{\infty,\mathscr{L}(N),N}^{\varrho_r,d,1}$ was arbitrary, we have shown

$$E(g_j, \mathtt{NN}_{\infty,\mathscr{L}(N),N}^{\varrho_r,d,1})_{L_p(\Omega)} \geq C_2 \cdot s^{d/p} =: C_3 \qquad \forall\, N \in \mathbb{N}_{\geq N_0} \text{ and } j \geq K'_N.$$

Directly from the definition of the norm $\|g_j|_\Omega\|_{N_\infty^\alpha(X_p(\Omega),\varrho_r,\mathscr{L})}$, this implies that for arbitrary $N \in \mathbb{N}_{\geq N_0}$

$$\|g_j|_\Omega\|_{N_\infty^\alpha(X_p(\Omega),\varrho_r,\mathscr{L})} \geq (1+N)^\alpha \cdot E(g_j, \mathtt{NN}_{\infty,\mathscr{L}(N),N}^{\varrho_r,d,1})_{L_p(\Omega)} \geq C_3 \cdot (1+N)^\alpha \quad \forall\, j \geq K'_N.$$

This proves $\|g_j|_\Omega\|_{N_\infty^\alpha(X_p(\Omega),\varrho_r,\mathscr{L})} \to \infty$ as $j \to \infty$, and thus completes the proof. □

Univ Rennes, Inria, CNRS, IRISA;, F-35042 Rennes France
E-mail address: remi.gribonval@inria.fr

Institut für Mathematik, Technische Universität Berlin, Germany
E-mail address: kutyniok@math.tu-berlin.de

Department of Mathematical Sciences, Aalborg University, Denmark
E-mail address: mnielsen@math.aau.dk

Department of Scientific Computing, Katholische Universität Eichstätt-Ingolstadt
E-mail address: felix@voigtlaender.xyz